

# Analysing Twitter for Ubisoft

Ryan Greenup

May 13, 2020

## Contents

<b>Submission Statement</b>	<b>1</b>
<b>8.1 Analysing the Relationship Between Friends and Followers for Twitter Users</b>	<b>1</b>
8.1.1 Retrieve the posts from Twitter . . . . .	2
8.2.2 Count of Followers and Friends . . . . .	2
8.1.3 Summary Statistics . . . . .	4
8.1.4 Above Average Followers . . . . .	4
8.1.5 Bootstrap confidence intervals . . . . .	5
.1 a/b.) Generate a bootstrap distribution . . . . .	5
.2 c.) Estimate a Confidence Interval for the population mean Follower Counts . . .	5
.3 d.) Estimate a Confidence Interval for the population mean Friend Counts . . . .	7
8.1.6 Estimate a 97% Confidence Interval for the High Friend Count Proportion . . . . .	9
8.1.7 Is the Number of Friends Independent to the Number of Followers . . . . .	9
.1 Bin the Follower and Friend Categories . . . . .	9
.2 Find the Group frequency . . . . .	9
.3 Find the Expected Counts under each group and test for independence . . . . .	9
.4 Conclusion . . . . .	13
<b>8.2 Finding Themes in tweets</b>	<b>13</b>
8.2.8 Find Users with Above Average Friend Counts . . . . .	13
8.2.9 Find Users with Below Average Friend Counts . . . . .	15
8.2.10 Find the <i>Tweets</i> corresponding to users with high or low friend counts . . . . .	16
8.2.11 Clean the tweets . . . . .	16
.1 Create a Corpus Object . . . . .	16
.2 Process the tweets . . . . .	17
8.2.12 Display the first two tweets before/after processing . . . . .	18
8.2.13 Create a Term Document Matrix . . . . .	19
.1 Apply Weighting Manually . . . . .	19
.2 Apply Weighting with Built-in Function . . . . .	20
.3 Remove Empty Documents . . . . .	20
8.2.14 How many Clusters are there . . . . .	21
.1 Use Cosine Distance . . . . .	21
.2 Project into Euclidean Space . . . . .	22
.3 Measure Within Cluster Variance . . . . .	22

8.2.15 Find the Number of tweets in each cluster . . . . .	24
8.2.16 Visualise the Clusters . . . . .	25
8.2.17 Comment on the Visualisation . . . . .	25
8.2.18 Cluster with Highest Number Friends . . . . .	26
8.2.18 Sample Tweets from the cluster . . . . .	27
8.2.19 Identify themes in the Tweets . . . . .	27
.1 Highest Friends Count . . . . .	27
.2 Lowest Friends Count . . . . .	28
8.2.20 Create WordClouds . . . . .	28
8.2.21 Use a Dendrogram to Display the Themes . . . . .	28
.1 Highest Friend Count . . . . .	32
.2 Lowest Friend Count . . . . .	32
8.2.23 Conclusion . . . . .	34
<b>8.3 Building Networks</b>	<b>35</b>
8.3.24 Find the 10 Most Popular Friends of the Twitter Handle . . . . .	35
8.3.25 Obtain a <b>2-degree</b> egocentric graph centred at <i>Ubisoft</i> . . . . .	36
.1 Download Friends of Friends . . . . .	36
.2 Limit the number of Nodes for Want of Visualisation . . . . .	36
.3 Build the Edge List . . . . .	37
.4 Build the Graph . . . . .	37
.5 Use GGPLOT2 . . . . .	37
8.3.26 Compute the <b>closeness</b> centrality score for each user . . . . .	37
8.3.27 List the Most Central People in the graph according to the closeness centrality . .	37
8.3.28 Comment on the Results . . . . .	40
<b>Appendix</b>	<b>40</b>

## Submission Statement

<i><b>Name</b></i>	<i><b>Student Number</b></i>
Ryan Greenup	17805315
Ben Britz	19579320
Kieran Keith	19723017
Navid Niknezhad	19636740
Hassan Rajab	19413192

Contribution Statistics are reflected by the git commit log [Available here](#), all work was implemented either through code or *GitHub* issues and so the repository is a good resource for contribution metrics.

By including this statement, we the authors of this work, verify that:

- We hold a copy of this assignment that we can produce if the original is lost or damaged.
- We hereby certify that no part of this assignment/product has been copied from any other student's work or from any other source except where due acknowledgement is made in the assignment.
- No part of this assignment/product has been written/produced for us by another person except where such collaboration has been authorised by the subject lecturer/tutor concerned.

- We are aware that this work may be reproduced and submitted to plagiarism detection software programs for the purpose of detecting possible plagiarism (which may retain a copy on its database for future plagiarism checking).
- We hereby certify that we have read and understand what the School of Computing and Mathematics defines as minor and substantial breaches of misconduct as outlined in the learning guide for this unit.

## 8.1 Analysing the Relationship Between Friends and Followers for Twitter Users

### 8.1.1 Retrieve the posts from Twitter

To analyze tweets on/Ubisoft/ it will be necessary to download them in a structure that may be interpreted by a given programming language, this analysis will be done using **R**, and the by utilizing the `rtweet` package will be used to interface with the *Twitter* API. The `rtweet` package can be automatically installed and loaded in **R** using the `pacman` package, this is demonstrated in listing 1.

```

1  # Load Packages
   ↪ -----
2  setwd("~/Dropbox/Notes/DataSci/Social_Web_Analytics/SWA-Project/scripts_1")
   ↪ "/"
3
4  if (require("pacman")) {
5    library(pacman)
6  } else{
7    install.packages("pacman")
8    library(pacman)
9  }
10
11  pacman::p_load(xts, sp, gstat, ggplot2, rmarkdown, reshape2,
12                ggmap, parallel, dplyr, plotly, tidyverse,
13                reticulate, UsingR, Rmpfr, swirl, corrplot,
14                gridExtra, mise, latex2exp, tree, rpart,
15                lattice, coin, primes, epitools, maps, clipr,
16                ggmap, twitterR, ROAuth, tm, rtweet, base64enc,
17                httpuv, SnowballC, RColorBrewer, wordcloud,
18                ggwordcloud, tidyverse, boot)

```

Listing 1: Load the Packages for **R**

In order to access data through the *Twitter* API it is necessary to use tokens provided through a *Twitter* developer account, these can be loaded into `rtweet` using the `create_token()` function as shown in listing 2.

```

1  # Set up Tokens
   ↪ =====
2
3  options(RCurlOptions = list(
4    verbose = FALSE,
5    capath = system.file("CurlSSL", "cacert.pem", package = "RCurl"),
6    ssl.verifypeer = FALSE
7  ))
8
9  setup_twitter_oauth(
10   consumer_key = "*****",
11   consumer_secret =
12     ↪ "*****",
13   access_token = "*****",
14   access_secret = "*****"
15 )
16 # rtweet
   ↪ =====
17 tk <- rtweet::create_token(
18   app = "SWA",
19   consumer_key = "*****",
20   consumer_secret =
21     ↪ "*****",
22   access_token =
23     ↪ "*****",
24   access_secret = "*****",
25   set_renv = FALSE

```

Listing 2: Import the twitter tokens (redacted)

Using the `rtweet` library 1000 tweets containing a mention of *Ubisoft*, in reverse-chronological order, can be returned and then saved to disk using the `save()` function as shown in listing 3.<sup>1</sup>

```
1  n <- 1000
2  tweets.company <- search_tweets(q = 'ubisoft', n = n, token = tk,
3                                include_rts = FALSE)
4  save(tweets.company[,], file = "resources/Download_1.Rdata")
```

Listing 3: Save the Tweets to the HDD as an rdata file

## 8.2.2 Count of Followers and Friends

To identify the number of users that are contained in the *tweets* it is necessary to distinguish which of the usernames are unique, this can be achieved by using either the `unique()` or `duplicated()` functions. The `unique` function can be used to return a vector of unique elements which can be passed to the `length()` function in order to return the number of unique users, this is demonstrated in the listing 4 and provides that 81.7% of the tweets were made by unique users.

The `duplicated` function can be used to identify and omit duplicated observations, this is also shown in the listing 4.

```
1  (users <- unique(tweets.company$name)) %>% length()
2  x <- tweets.company$followers_count[!duplicated(tweets.company$name)]
3  y <- tweets.company$friends_count[!duplicated(tweets.company$name)]
4
5  ## > [1] 817
```

Listing 4: Return follower count of twitter posts

## 8.1.3 Summary Statistics

The average number of friends and followers from users who posted tweets mentioning *Ubisoft* can be returned using the `mean()` as shown in listing 5 this provides that on average each user has 586 friends and 63,620 followers.

---

<sup>1</sup>The `rtweet` package will search for tweets that contain all the words of a query regardless of uppercase or lowercase usage [5].

```

1 x<- rnorm(090)
2 y<- rnorm(090)
3 (xbar <- mean(x))
4 (ybar <- mean(y))
5
6 ## > [1] 4295.195
7 ## > [1] 435.9449

```

Listing 5: Determine the average number of friends and followers

### 8.1.4 Above Average Followers

Each user can be compared to the average number of followers, by using a logical operator on the vector (e.g.  $y > ybar$ ), this will return an output of logical values. **R** will coerce logical into 1/0 values meaning that the mean value will return the proportion of TRUE responses as shown in listing 6. This provides that:

- 2.4% of the have identified have an above-average **number of followers**.
- 20.6% of the users identified have an above-average **number of friends**.

```

1 (px_hat <- mean(x>xbar))
2 (py_hat <- mean(y>ybar))
3
4 ## > [1] 0.0244798
5 ## > [1] 0.2729498

```

Listing 6: Calculate the proportion of users with above average follower counts

### 8.1.5 Bootstrap confidence intervals

#### a/b.) Generate a bootstrap distribution

A bootstrap assumes that the population is an infinitely large repetition of the sample and maybe produced concerning follower counts by resampling with replacement/repetition, this can then be plotted using the ggplot2 library as demonstrated in listings 7 and 8 and shown in figure 1.

This shows that the population follower counts is a non-normal skew-right distribution, which is expected because the number of friends is an integer value bound by zero [7].

#### c.) Estimate a Confidence Interval for the population mean Follower Counts

To perform a bootstrap for the population mean value of follower counts it is necessary to:

```

1  ## Resample the Data
2  (bt_pop <- sample(x, size = 10^6, replace = TRUE)) %>% head()
3
4  ## > [1]    7 515 262 309 186 166

```

Listing 7: Bootstrapping a population from the sample.

```

1  ## Make the Population
2  bt_pop_data <- tibble("Followers" = bt_pop)
3  ggplot(data = bt_pop_data, aes(x = Followers)) +
4    geom_histogram(aes(y = ..density..), fill = "lightblue", bins = 35,
5    ↪   col = "pink") +
6    geom_density(col = "violetred2") +
7    scale_x_continuous(limits = c(1, 800)) +
8    theme_bw() +
9    labs(x = "Number of Followers", y = "Density",
10         title = "Bootstrapped population of Follower Numbers")

```

Listing 8: Use ggplot2 to Produce Histogram of Follower Counts Bootstrapped from the population

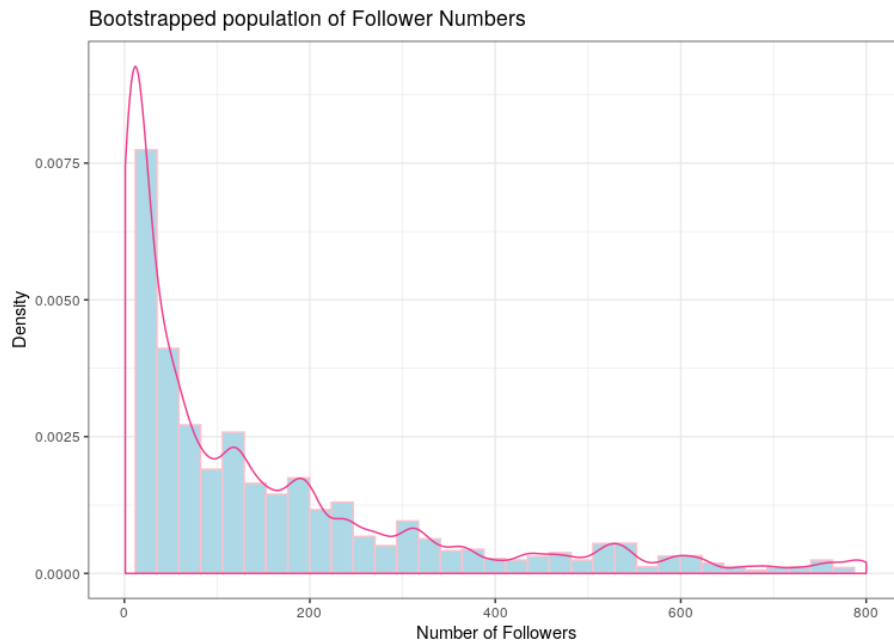


Figure 1: Histogram of the bootstrapped population of follower counts

1. Resample the data with replacement
  - i.e. randomly select values from the sample allowing for repetition
2. Measure the statistic of concern
3. Replicate this a sufficient number of times
  - i.e. Greater than or equal to 1000 times [2, Ch. 5]

This is equivalent to drawing a sample from a population that is infinitely large and constructed of repetitions of the sample. This can be performed in **R** as shown in listing 9.

```

1 xbar_boot_loop <- replicate(10^3, {
2   s <- sample(x, replace = TRUE)
3   mean(s)
4 })
5 quantile(xbar_boot_loop, c((1-0.97)/2, (1+0.97)/2))
6
7 ##          1.5%          98.5%
8 ##    588.4189 10228.7352

```

Listing 9: Confidence Interval of Mean Follower Count in Population

A 97% probability interval is such that a sample drawn from a population will contain the population mean in that interval 97% of the time, this means that it may be concluded with a high degree of certainty that the true population mean lies between 588 and 10228.

1. Alternative Approaches If this data was normally distributed it may have been appropriate to consider bootstrapping the standard error and using a  $t$  distribution, however it is more appropriate to use a percentile interval for skewed data such as this, in saying that however this method is not considered to be very accurate in the literature and is often too narrow. [3, Section 4.1]
  - It's worth noting that the normal  $t$  value bootstrap offers no advantage over using a  $t$  distribution (other than being illustrative of bootstrapping generally) [3, Section 4.1]

The boot package is a bootstrapping library common among authors in the data science sphere [4, p. 295] [10, p. 237] that implements confidence intervals consistent with work by Davison and Hinkley [8] in their textbook *Bootstrap Methods and their Application*. In this work it is provided that the  $BC_a$  method of constructing confidence intervals is superior to mere percentile methods in terms of accuracy [2, Ch. 5], a sentiment echoed in the literature. [1, 2, Ch. 5]

Such methods can be implemented in **R** by passing a function to the boot call as shown in listing 10. This provides a broader interval, suggesting that the true confidence interval could lie between 1079 and 16227 followers.



```

1  xbar_boot <- boot(data = x, statistic = mean_val, R = 10^3)
2  boot.ci(xbar_boot, conf = 0.97, type = "bca", index = 1)
3
4  ## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
5  ## Based on 1000 bootstrap replicates
6  ##
7  ## CALL :
8  ## boot.ci(boot.out = xbar_boot, conf = 0.97, type = "bca", index = 1)
9  ##
10 ## Intervals :
11 ## Level      BCa
12 ## 97%      ( 1079, 16227 )
13 ## Calculations and Intervals on Original Scale
14 ## Warning : BCa Intervals used Extreme Quantiles
15 ## Some BCa intervals may be unstable
16 ## Warning message:
17 ## In norm.inter(t, adj.alpha) : extreme order statistics used as
   ↪ endpoints

```

Listing 10: Bootstrap of population mean follower count implementing the  $BC_a$  method

#### d.) Estimate a Confidence Interval for the population mean Friend Counts

A Confidence interval for the population mean friend counts may be constructed in a likewise fashion as shown in listings 11. This provides that the 97% confidence interval for the *population mean friend count* is between 384 and 502 (or 387 and 496 if the  $BC_a$  method used, they're quite close and so the more conservative percentile method will be accepted).

### 8.1.6 Estimate a 97% Confidence Interval for the High Friend Count Proportion

In order to bootstrap a confidence interval for the proportion of users with above average follower counts, repeatedly draw random samples from an infinitely large population composed entirely of the sample, and record the sampled proportion. This can be achieved by resampling the observations of above and below as shown in the listing 12.

This provides that:

- The 97% confidence interval for the population proportion of users that have an above-average number of friends is between 0.24 and 0.31.
  - i.e. The probability, of any given sample taken from the population, containing the population mean within this interval, would be 97%; that does not, however, imply that there is a 97% probability that this interval contains the value.

```

1  # d.) Estimate a Confidence Interval for the populattion mean Friend
   ↪ Count ===
2  # Using a Percentile Method
   ↪ #####
3  ybar_boot_loop <- replicate(10^3, {
4    s <- sample(y, replace = TRUE)
5    mean(s)
6  })
7  quantile(ybar_boot_loop, c(0.015, 0.985))
8
9  # Using BCA Method
   ↪ #####
10 mean_val <- function(data, index) {
11   X = data[index]
12   return(mean(X))
13 }
14
15 xbar_boot <- boot(data = y, statistic = mean_val, R = 10^3)
16 boot.ci(xbar_boot, conf = 0.97, type = "bca", index = 1)
17
18
19 ##      1.5%      98.5%
20 ## 383.7619 501.5903
21 ##
22 ## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
23 ## Based on 1000 bootstrap replicates
24 ##
25 ## CALL :
26 ## boot.ci(boot.out = xbar_boot, conf = 0.97, type = "bca", index = 1)
27 ##
28 ## Intervals :
29 ## Level      BCa
30 ## 97%      (386.8, 496.7 )
31 ## Calculations and Intervals on Original Scale
32 ## Some BCa intervals may be unstable

```

Listing 11: Bootstrap of population mean follower count

```

1  # 8.1.6 High Friend Count Proportion
   ↪ -----
2  prop <- factor(c("Below", "Above"))
3  ## 1 is above average, 2 is below
4  py_hat_bt <- replicate(10^3, {
5      rs      <- sample(c("Below", "Above"),
6                        size = length(y),
7                        prob = c(py_hat, 1-py_hat),
8                        replace = TRUE)
9      isabove <- rs == "Above"
10     mean(isabove)
11 })
12 quantile(py_hat_bt, c(0.015, 0.985))
13
14
15 ##      1.5%      98.5%
16 ## 0.2399021 0.3072215
17 ## > > > . + > > >
18 ## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
19 ## Based on 1000 bootstrap replicates
20 ##
21 ## CALL :
22 ## boot.ci(boot.out = py_hat_boot, conf = 0.97, type = "bca")
23 ##
24 ## Intervals :
25 ## Level      BCa
26 ## 97%      ( 0.2399,  0.3072 )
27 ## Calculations and Intervals on Original Scale

```

Listing 12: Bootstrap of Proportion of Friends above average

## 8.1.7 Is the Number of Friends Independent to the Number of Followers

One method to determine whether or not the number of followers is independent of the number of friends is to bin the counts and determine whether or not the distribution of users across those counts is consistent with the hypothesis of independence.

### Bin the Follower and Friend Categories

The counts may be binned by performing a logical interval test as shown in listing 13.

```
1  ## Assign Categories
2  x_df <- data.frame(x)
3  x_df$cat[0]      <= x_df$x & x_df$x < 100] <- "Tens"
4  x_df$cat[100]    <= x_df$x & x_df$x < 1000] <- "Hundreds"
5  x_df$cat[1000]   <= x_df$x & x_df$x < 2000] <- "1Thousands"
6  x_df$cat[2000]   <= x_df$x & x_df$x < 3000] <- "2Thousands"
7  x_df$cat[3000]   <= x_df$x & x_df$x < 4000] <- "3Thousands"
8  x_df$cat[4000]   <= x_df$x & x_df$x < 5000] <- "4Thousands"
9  x_df$cat[5000]   <= x_df$x & x_df$x < Inf] <- "5ThousandOrMore"
10
11  ### Make a factor
12  x_df$cat <- factor(x_df$cat, levels = var_levels, ordered = TRUE)
13
14  ### Determine Frequencies
15  (x_freq <- table(x_df$cat) %>% as.matrix())
16
17  ## ** b) Find the Friend Count Frequency
18  ↪ =====
19  ## Assign Categories
20  y_df <- data.frame(y)
21  y_df$cat[0]      <= y_df$y & y_df$y < 100] <- "Tens"
22  y_df$cat[100]    <= y_df$y & y_df$y < 1000] <- "Hundreds"
23  y_df$cat[1000]   <= y_df$y & y_df$y < 2000] <- "1Thousands"
24  y_df$cat[2000]   <= y_df$y & y_df$y < 3000] <- "2Thousands"
25  y_df$cat[3000]   <= y_df$y & y_df$y < 4000] <- "3Thousands"
26  y_df$cat[4000]   <= y_df$y & y_df$y < 5000] <- "4Thousands"
27  y_df$cat[5000]   <= y_df$y & y_df$y < Inf] <- "5ThousandOrMore"
28
29  ### Make a factor
30  y_df$cat <- factor(y_df$cat, levels = var_levels, ordered = TRUE)
31
32  ### Determine Frequencies
33  (y_freq <- table(y_df$cat) %>% as.matrix())
```

Listing 13: Use Logical Test to Assign observations into bins

## Find the Group frequency

These values may be tabulated to count the occurrence of users among these categories as shown in listing 14 and table 1.

```

1  vals <- t(cbind(x_freq, y_freq))
2  rownames(vals) <- c("Followers.x", "followers.y")
3  vals
4
5  ##           Tens Hundreds 1Thousands 2Thousands 3Thousands 4Thousands
6  ## Followers.x 421       317       39       11       9       2
7  ## followers.y 262       476       47       15       6       9
8  ##           5ThousandOrMore
9  ## Followers.x           18
10 ## followers.y           2

```

Listing 14: Tabulate the binned counts for the distribution of users among amount and status.

Table 1: Binned Friend and Follower counts, transposed relative to code.

	<b>Followers</b>	<b>Friends</b>
<i>Tens</i>	421	262
<i>Hundreds</i>	317	476
<i>1 - Thousands</i>	39	47
<i>2 - Thousands</i>	11	15
<i>3 - Thousands</i>	9	6
<i>4 - Thousands</i>	2	9
<i>5 Thousand or More</i>	18	2

## Find the Expected Counts under each group and test for independence

The expected count of each cell, under the assumption that the two metrics are independent will be the proportion of users per bracket multiplied by the number of users in that status group. This implies that any cell will be:

- the product of the row sum, multiplied by the column sum divided by the number of counts.

This can be equivalently expressed as an outer product as shown in equation (1), in **R** this operation is denoted by the %%% operator, which is shorthand for the outer() function, this and other summary statistics may be evaluated as shown in the listing 15.

The outer product is such that:

$$\mathbf{u} \otimes \mathbf{v} = \mathbf{u}\mathbf{v}^T = \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{bmatrix} \begin{bmatrix} v_1 & v_2 & v_3 \end{bmatrix} = \begin{bmatrix} u_1v_1 & u_1v_2 & u_1v_3 \\ u_2v_1 & u_2v_2 & u_2v_3 \\ u_3v_1 & u_3v_2 & u_3v_3 \\ u_4v_1 & u_4v_2 & u_4v_3 \end{bmatrix}.$$

This means the matrix of expected frequencies can be expressed as an outer product thusly:

$$\tilde{\mathbf{e}} = \frac{1}{n} \times \begin{bmatrix} \sum_{j=1}^n [o_{1j}] \\ \sum_{j=1}^n [o_{2j}] \\ \sum_{j=1}^n [o_{3j}] \\ \sum_{j=1}^n [o_{4j}] \\ \vdots \\ \sum_{j=1}^n [o_{nj}] \end{bmatrix} \begin{bmatrix} \sum_{j=1}^n [o_{i1}] \\ \sum_{j=1}^n [o_{i2}] \\ \sum_{j=1}^n [o_{i3}] \\ \dots \\ \sum_{j=1}^n [o_{in}] \end{bmatrix}^T \quad (1)$$

```

1  ## ***** Calculate Summary Stats
2  n <- sum(vals)
3  bracket_prop <- colSums(vals) / n
4  metric_prop  <- rowSums(vals) / n
5  o <- vals
6  e <- rowSums(vals) %o% colSums(vals) / n
7  chi_obs <- sum((e-o)^2/e)

```

Listing 15: Calculate Expected frequency of values under the assumption of independence.

## 1. Testing Independence

To test whether or not the distribution of users among brackets is independent of being a follower or friend a  $\chi^2$  test may be used, this can be evaluated from a model or simulated, in **R**, the simulated test is shown in the listing 16, this provides a  $p$ -value  $< 0.0005$ , which means that the hypothesis of independence may be rejected with a high degree of certainty.

```

1  chisq.test(vals, simulate.p.value = TRUE)
2
3
4  ## Pearson's Chi-squared test with simulated p-value (based on 2000
5  ## replicates)
6  ##
7  ## data:  vals
8  ## X-squared = 88.109, df = NA, p-value = 0.0004998

```

Listing 16: Chi-Square testing for independence between friend and follower bin categories.

### (a) From First Principles

The  $\chi^2$  statistic may be performed from first principles by randomly sampling the values at the rate at which they occurred, tabulating those counts, measuring the  $\chi^2$ -value, and then repeating this many times.

Because the samples are random they must be independent and average number of positives is hence an estimate for the  $FPR$ , which is, in turn, an estimate for the  $p$ -value. This

technique is demonstrated in the listing 17, the p-value being returned as 0.0004, this value is consistent with the value produced by *R*'s built-in `chisq.test` function and so is accepted.

## Conclusion

The  $p$ -value measures the probability of rejecting the null hypothesis when it is true, i.e. the probability of a detecting a *false positive*, a very small  $p$ -value is hence good evidence that the null hypothesis should be rejected.

In saying that however the  $p$ -value is distinct from the *power* statistic, which is a measure of *the probability of accepting the alternative hypothesis* when it is true, a low  $p$ -value is not a measurement of the probability of being correct.

Nonetheless, may we conclude, with a high degree of certainty, that the follower and friend counts are not independent of one another.

## 8.2 Finding Themes in tweets

### 8.2.8 Find Users with Above Average Friend Counts

Users with Above average Friend Counts can be identified by filtering the tweets data frame for two conditions:

1. non-duplicated user-id
2. friend\_count greater than average

This can be achieved easily using the `dplyr` package as shown in the listing 18, the top 20 of these users are shown in table 4 of the appendix

### 8.2.9 Find Users with Below Average Friend Counts

Users with high friends may be determined by a similar method (or by taking the complement of the high friends) as shown in the listing 19, the lowest 20 of these users are shown in table 5 of the appendix.

### 8.2.10 Find the *Tweets* corresponding to users with high or low friend counts

The tweets corresponding to users with high and low friend counts can be identified by filtering the data frame based on the friend count and using that to index the tweets from the data frame <sup>2</sup>, alternatively, it is possible to test whether or not the ID of a user appears in the high or low vector set using the `%in%` operator as shown in the listing 20.

---

<sup>2</sup>This works because the `tm` package preserves the order of the data, this can be confirmed by using a data frame source as opposed to a vector source (e.g. in listing 21) and comparing the ID's before/after transformation.

```

1  ## ***** Create Vectors of factor levels
2  brackets <- unique(x_df$cat)
3  metrics <- c("follower", "friend")
4
5  ## ***** Simulate the data Assuming H_0
6  ## I.e. assuming that the null hypothesis is true in that
7  ## the brackets assigned to followers are independent of the friends
8  ## (this is a symmetric relation)
9
10 s <- replicate(10^4,{
11   ## Sample the set of Metrics
12   m <- sample(metrics, size = n, replace = TRUE, prob = metric_prop)
13
14   ## Sample the set of Brackets (i.e. which performance bracket the
15   ↪ user falls in)
16   b <- sample(brackets, size = n, replace = TRUE, prob = bracket_prop)
17
18   ## Make a table of results
19   o <- table(m, b)
20
21   ## Find What the expected value would be
22   e_sim <- t(colSums(o) %o% rowSums(o) / n)
23
24   ## Calculate the Chi Stat
25   chi_sim <- sum((e_sim-o)^2/e_sim)
26   chi_sim
27
28   ## Is this more extreme, i.e. would we reject the null hypothesis?
29   chi_sim > chi_obs
30
31 })
32
33 mean(s)
34
35 ## [1] 4e-04

```

Listing 17: Performing a  $\chi^2$  statistic from first principles



```

1 select <- dplyr::select
2 filter <- dplyr::filter
3 interested_vars <- c("user_id", "friends_count")
4 (friend_counts <- tweets.company %>%
5   select(interested_vars) %>%
6   filter(!duplicated(user_id)))
7
8 (high_friends <- friend_counts %>%
9   filter(friends_count > mean(friends_count, na.rm = TRUE)))
10
11 ## Export Friends List
12 write.csv(high_friends[order(
13   high_friends$friends_count,
14   decreasing = TRUE),], file = "/tmp/highfriend.csv")

```

Listing 18: Use dplyr to Filter for Users with a high Friend Count

```

1 (low_friends <- friend_counts %>%
2   filter(friends_count <= mean(friends_count, na.rm = TRUE)))
3
4 low_friends <- low_friends[order(
5   low_friends$friends_count,
6   decreasing = TRUE),]
7
8 ## Export Users
9 write.csv(low_friends[order(
10   low_friends$friends_count,
11   decreasing = FALSE),], file = "/tmp/lowfriend.csv")

```

Listing 19: Use dplyr to Filter for Users with a low Friend Count

```

1  ## Method 1
2  friend_test <- tweets.company$friends_count >
   ↪ mean(tweets.company$friends_count)
3  tweets_high <- tweets.company$text[friend_test]
4  tweets_low <- tweets.company$text[!friend_test]
5
6  ## Method 2
   ↪ :15b5a74:
7  tweets_high <- tweets.company$text[tweets.company$user_id %in%
   ↪ high_friends$user_id]
8  tweets_low <- tweets.company$text[tweets.company$user_id %in%
   ↪ low_friends$user_id]
9  tweets <- c(tweets_high, tweets_low)
10
11 ## Mark as High or Low
12 tweets_low <- cbind(tweets_low, rep("Low_Friend", length(tweets_low)))
13 tweets_high <- cbind(tweets_high, rep("High_Friend",
   ↪ length(tweets_high)))
14 tweets <- as.data.frame(rbind(tweets_high, tweets_low))
15 tweets$Friend_Status <- factor(tweets$Friend_Status)

```

Listing 20: Identify tweets corresponding to users with high and low friend counts

## 8.2.11 Clean the tweets

### Create a Corpus Object

To clean the tweets it is necessary to create a corpus object as shown in the listing 21, it is possible to pass a dataframe source to include the user ID, this isn't strictly necessary however because the `tm` package preserves order when performing transformations.

```

1  tweet_source <- tm::VectorSource(tweets$text)
2  tweet_corpus <- tm::Corpus(x = tweet_source)

```

Listing 21: Create a Corpus from the tweets

Next, it is necessary to choose an encoding, a primary consideration of this is whether or not the use of *emoji* characters will influence the model performance. There is research to suggest that emoji can be used as predictive features [6] and that they can improve *sentiment analysis* models [9] that implement a *bag of words* approach. For these reasons *emoji* characters will be preserved and **UTF-8** implemented.

In order to encode the data as *UTF-8*, the `iconv` function can be used as shown in listing 22.

```

1  encode <- function(x) {
2      iconv(x, to = "UTF-8")
3      # iconv(x, to = "latin1")
4      # iconv(x, to = "ASCII")
5  }
6
7  tweet_corpus <- tm_map(x = tweet_corpus, FUN = encode)
8  tweet_corpus_raw <- tweet_corpus

```

Listing 22: Encode the Data as UTF-8

### Process the tweets

Before analysis the tweets should be modified to remove characters that may interfere with categorizing words, this is referred to as cleaning, in particular, the following should be implemented:

1. Remove URL's
2. Remove Usernames
3. Remove numbers
4. Remove punctuation
5. Remove whitespace
6. Case fold all characters to lower case
7. Remove a set of stop words
8. Reduce each word to its stem

In particular, it is important to reduce words to lower case before removing stop words otherwise an unorthodox use of capitalization may prevent the word from being removed throughout.

The stop word `ubisoft` will also be used, this was the query term so it's expected to turn up at a very high frequency, the words `can` and `'s` also occurred quite frequently and so were removed.

The cleaning can be implemented by mapping functions over the corpus, which is fundamentally a list, this can be performed via the `tm_map` function as shown in the listing 23.

### 8.2.12 Display the first two tweets before/after processing

The tweets can be viewed from inside the corpus by selecting with the `[` function<sup>3</sup> as demonstrated in listing 24, the first *tweet* was rendered empty by the processing and the following two tweets were:

<sup>3</sup>The `[` function is actually shorthand for `Extract()`, most things in *R* are functions, this is similar to *LISP* and has to do with the origins of the language, e.g. `sum(1:10) == (sum (1:10))`, also relevant see the relevant `xkcd` in figure 11.

```

1  mystop <- c(stopwords(), "s", "can", "ubisoft", "@ubisoft",
  ↪ "#ubisoft")# <<stphere>>
2
3  clean_corp <- function(corpus) {
4    ## Remove URL's
5    corpus <- tm_map(corpus, content_transformer(function(x)
  ↪ gsub("(f|ht)tp(s?)://\\S+", "", x)))
6    ## Remove Usernames
7    corpus <- tm_map(corpus, content_transformer(function(x)
  ↪ gsub("@\\w+", "", x)))
8    ## Misc
9    corpus <- tm_map(corpus, FUN = removeNumbers)
10   corpus <- tm_map(corpus, FUN = removePunctuation)
11   corpus <- tm_map(corpus, FUN = stripWhitespace)
12   corpus <- tm_map(corpus, FUN = tolower)
13   corpus <- tm_map(corpus, FUN = removeWords, mystop)
14   ## stopwords() returns characters and is fead as second argument
15   corpus <- tm_map(corpus, FUN = stemDocument)
16   return(corpus)
17 }
18
19 tweet_corpus_clean <- clean_corp(tweet_corpus)

```

Listing 23: Use the tm\_map function to clean the tweets

### ▪ *Pre-Processing*

- “Today was the first time in over a month that I have gone 24 hours without checking the coronavirus death toll. Thanks @Ubisoft.”
- “@btwimskrank @TheDivisionGame @UbiMassive @Ubisoft @jgerighty @hamishbode @Tideman92 @janeyo\_jane @slimjd Very odd... I'll even post a video about it.”

### ▪ *Post-Processing*

- “today first time month gone hour without check coronavirus death toll thank”
- “odd ill even post video”

```
1 tweet_corpus_raw[[1]]$content
2 tweet_corpus_clean[[1]]$content
3 tweet_corpus_raw[[2]]$content
4 tweet_corpus_clean[[2]]$content
5 tweet_corpus_raw[[3]]$content
6 tweet_corpus_clean[[3]]$content
7
8
9 ## [1] NA
10 ## [1] "NA"
11 ## [1] "Today was the first time in over a month that I have gone 24
   ↳ hours without checking the coronavirus death toll. Thanks @Ubisoft."
12 ## [1] "today first time month gone hour without check coronavirus
   ↳ death toll thank"
13 ## [1] "@btwimskrank @TheDivisionGame @UbiMassive @Ubisoft @jgerighty
   ↳ @hamishbode @Tideman92 @janeyo_jane @slimjd Very odd... I'll even
   ↳ post a video about it."
14 ## [1] "odd ill even post video"
```

Listing 24: Display the tweets before and after cleaning

## 8.2.13 Create a Term Document Matrix

### Apply Weighting Manually

#### 1. Create Term Document Matrix

A term Document matrix (and its transpose) can be constructed from a corpus using the `tm:TermDocumentMatrix` function as shown in listing 25.

#### 2. Apply TF-IDF Weighting

Weighted term frequency is defined as shown in equation (2), where:

```

1 tweet_matrix_tdm <- TermDocumentMatrix(tweet_corpus_clean)
2 tweet_matrix_dtm <- DocumentTermMatrix(tweet_corpus_clean)

```

Listing 25: Create the *Term Document Matrix* and its transpose using built-in functions.

- $f_{d,t}$  is the frequency of a given term  $t$  in the a document  $d$
- $w_{d,t}$  is the weight of a given term  $t$  in the a document  $d$
- $N$  is the number of documents
- $f_t$  is the number of documents containing  $t$

$$\begin{aligned}
 w_{d,t} &= \text{TF}_t \times \text{IDF}_{d,t} \\
 &= \log_e (f_{d,t} + 1) \log_e \left( \frac{N}{f_t} \right)
 \end{aligned} \tag{2}$$

This would require multiplying each term of each row of the  $\text{TF}_t$  matrix by the corresponding vector element of  $\text{IDF}_{d,t}$ , this can be implemented by taking the matrix product of a diagonalized matrix, this is shown in listing 26.

```

1 N <- nrow(as.matrix(tweet_matrix_dtm)) # Number of Documents
2 ft=colSums(as.matrix(tweet_matrix_dtm) > 0) #in how many documents term
   ↪ t appeared in,
3
4 TF <- log(as.matrix(tweet_matrix_dtm) + 1) # built in uses log2()
5 IDF <- log(N/ft)
6
7 tweet_weighted <- TF %*% diag(IDF)
8 colnames(tweet_weighted) <- colnames(tweet_matrix_dtm)

```

Listing 26: Apply TF-IDF Weigting

There is however a function built into the `tm` package that will weight term-document matrices and this will instead be implemented to analyze the data because it will produce more maintainable code.

### Apply Weighting with Built-in Function

In order to create a term document matrix (and its transpose) with TF-IDF weighted values, the weighting argument may be specified as `weightTfIdf` by passing an appropriate list to the `control` argument of the `TermDocumentMatrix` function, as shown in listing 27

```

1 tweet_weighted_tdm <- tm::TermDocumentMatrix(x = tweet_corpus_clean,
  ↪ control = list(weighting = weightTfIdf))
2 tweet_weighted_dtm <- as.DocumentTermMatrix() %>%
3   as.matrix()

```

Listing 27: Create a Document Term Matrix by transforming a Term Document Matrix

### Remove Empty Documents

Empty Documents may be removed from the matrix by a logical test as shown in the listing 28<sup>4</sup> this provides that 328 documents were empty following the processing. A summary of the first rows and columns of this matrix, following the removal of empty documents, is provided in table 6 of the appendix.

```

1 null = which(rowSums(as.matrix(tweet_weighted_dtm)) == 0)
2 rowSums(as.matrix(tweet_weighted_dtm)==0)
3
4 if(length(null)!=0){
5   tweet_weighted_dtm = tweet_weighted_dtm[-null,]
6 }
7
8 length(null)
9
10 ## [1] 328

```

Listing 28: Remove empty documents from the Matrix, observe that the [-null] filter is wrapped in an if statement to prevent an empty vector bug.

## 8.2.14 How many Clusters are there

### Use Cosine Distance

To consider clustering, it can be more effective to consider the distance between the weighted documents in terms of cosine distance, the cosine distance can be calculated from the euclidean distance using the identity shown in (11), and this can be implemented in **R** by taking the matrix product of a diagonalized matrix as shown in listing 29.

<sup>4</sup>It is important not to filter based on the logic of an empty vector, because otherwise an empty vector will returned, hence the if statement in listing 28.

$$\text{dist}(\mathbf{X}, \mathbf{Y}) = \|\mathbf{X} - \mathbf{Y}\| \quad (3)$$

$$= \sqrt{\sum_{i=1}^n [(x_i - y_i)^2]} \quad (4)$$

$$\text{dist}(\mathbf{X}, \mathbf{Y})^2 = \sum_{i=1}^n [(x_i - y_i)^2] \quad (5)$$

$$= \sum_{i=1}^n (x_i^2) + \sum_{i=1}^n (y_i^2) + 2 \sum_{i=1}^n (x_i y_i) \quad (6)$$

$$= 1 + 1 + 2 \times \frac{\sum_{i=1}^n (x_i y_i)}{(1)} \quad (7)$$

$$= 2 + 2 \times \frac{\sum_{i=1}^n (x_i y_i)}{\|\mathbf{X}\| \times \|\mathbf{Y}\|} \quad (8)$$

$$= 2 + 2 \cos(\mathbf{X}, \mathbf{Y}) \quad (9)$$

$$(10)$$

$$\Rightarrow (1 - \cos(\mathbf{X}, \mathbf{Y})) = \frac{\text{dist}(\mathbf{X}, \mathbf{Y})^2}{2} \quad (11)$$

```

1 norm.tweet_weighted_dtm = diag(
2     1/sqrt(rowSums(tweet_weighted_dtm^2))
3     ) %*% tweet_weighted_dtm
4 D =dist(norm.tweet_weighted_dtm, method = "euclidean")^2/2

```

Listing 29: Calculate the cosine distance by implementing the identity from equation (11)

## Project into Euclidean Space

The cosine distance, however, is not an appropriate distance measurement with which to perform  $k$  means clustering and so the data instead should be projected back into euclidean space, this can be achieved using *Multi-Dimensional Scaling* via the `cmdscale` function as shown in listing 30.

The distance is a measure of between document distance so the number of dimensions should correspond to the number of documents, however, if there are zero-value eigenvalues, these dimensions won't help explain the data in the projection, hence the number of eigenvalues have been used as the dimension of projection in this case.

## Measure Within Cluster Variance

To determine the appropriate number of clusters, the within cluster variance can be measured, the number of clusters at which this value ceases to decrease is indicative of a potentially appropriate number of clusters. This is implemented in listing 31 and shown in figure 2.

Figure 2 Indicates a sudden stop of decrease in variance at 7 clusters and following that the within cluster variance begins to decrease at a slightly slower rate. For this reason, 7 could be an appropriate



```

1 l <- min(nrow(tweet_weighted_dtm),
2         ncol(tweet_weighted_dtm))
3 ev <- eigen(tweet_weighted_dtm[1:l, 1:l])
4 k <- (ev$values != 0) %>% sum()
5
6 mds.tweet_weighted_dtm <- cmdscale(D, k=k) #TODO What should K be? see
   ↪ issue #10

```

Listing 30: Project the *DocumentTermMatrix* into *Euclidean Space*, identifying the number of dimensions from the number of non-zero eigen vectors.

```

1 set.seed(271)
2 n = 15 # Assume it bends at 7 clusters
3 SSW = rep(0, n)
4 for (a in 1:n) {
5   K = kmeans(mds.tweet_weighted_dtm, a, nstart = 20)
6   SSW[a] = K$tot.withinss
7   paste(a*100/n, "%") %>% print()
8 }
9 SSW
10
11
12 SSW_tb <- tibble::enframe(SSW)
13 ggplot(SSW_tb, aes(x = name, y = value)) +
14   geom_point(col = "#Cd5b45", size = 5) +
15   geom_line(col = "#Da70d6") +
16   geom_vline(xintercept = 7, lty = 3, col = "blue") +
17   theme_bw() +
18   labs(x = "Number of Clusters",
19        y = "Within Cluster Sum of Square Distance",
20        title = "Within Cluster Variance across Clusters")

```

Listing 31: Use a loop to evaluate the performace of various cluster models, plot this with *ggplot2*

candidate for the number of clusters, however, the minimal amount of change in the within-cluster variance indicates that the data is most likely not clustered at all.

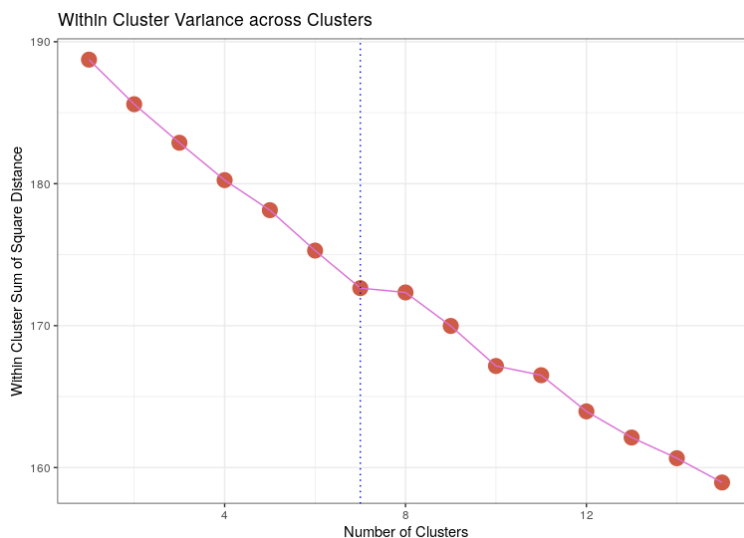


Figure 2: Plot of the Within Cluster Variance of the tweets using cosine distance (projected into Euclidean Space)

## 8.2.15 Find the Number of tweets in each cluster

Moving forward 3 clusters will be used to categorize the tweets, 7 is too large and a smaller number will likely be more effective at categorizing the data (particularly given that the stratification of the data appears to be quite limited as shown by figure 2).

The number of tweets in a cluster may be measured by using the `table` function as shown in listing 32 and table 2.

```
1 K = kmeans(mds.tweet_weighted_dtm, 3, nstart = 20)
2 table(K$cluster)
```

Listing 32: The `table` function can count the number of tweets per cluster.

Table 2: Number of tweets in each cluster identified by  $k$  means clustering

Cluster	1	2	3
# of tweets	77	58	537

## 8.2.16 Visualise the Clusters

The clusters can be projected into 2D *Euclidean-Space* using *Multi-Dimensional Scaling*, these dimensions would represent the first two principal components of the cosine distance of the weighted tweets. This is demonstrated in listings 33. A plot of the friend count-status mapped to shape and clusters mapped to colour is demonstrated in listing 34 shown in 3.

```
1 MDS_Euclid_2D <- cmdscale(D, k=2) #TODO What should K be? see issue #10
2 mds_data$Cluster <- factor(mds_data$Cluster)
3
4 if (nrow(MDS_Euclid_2D[,1:2]) == length(K$cluster)
5     && length(K$cluster) == nrow(tweets[-null,])) {
6   mds_data <- cbind(MDS_Euclid_2D[,1:2], "Cluster" = K$cluster,
7     ↪ tweets[-null,])
8 }
9 mds_data$Cluster <- factor(mds_data$Cluster)
10 names(mds_data)[1:2] <- c("MDS1", "MDS2")
```

Listing 33: Use *Multi-Dimensional* scaling to project the data into 2 dimensions

```
1 ggplot(pca_data, aes(x = MDS1, y = MDS2, col = Cluster)) +
2   geom_point(aes(shape = Friend_Status), size = 2) +
3   stat_ellipse(level = 0.9) +
4   theme_classic() +
5   labs(main = "Principal Components of Twitter Data",
6     x = TeX("MDS_1"), y = TeX("MDS_2")) +
7   scale_shape_discrete(label = c("High Friends", "Low Friends")) +
8   guides(shape = guide_legend("Friend Count \n Status"))
```

Listing 34: Create a plot of the distribution of Friends Count Status as shown in figure 3

## 8.2.17 Comment on the Visualisation

Figure 3 indicates that although there isn't a clear distinction between clusters there is a separation of the tweets in such a way that does allow some degree of classification to occur. Figure 3 also indicates that having many or few friends is reasonably independent of the cluster that the tweet belongs to but that there is possibly also other factors.

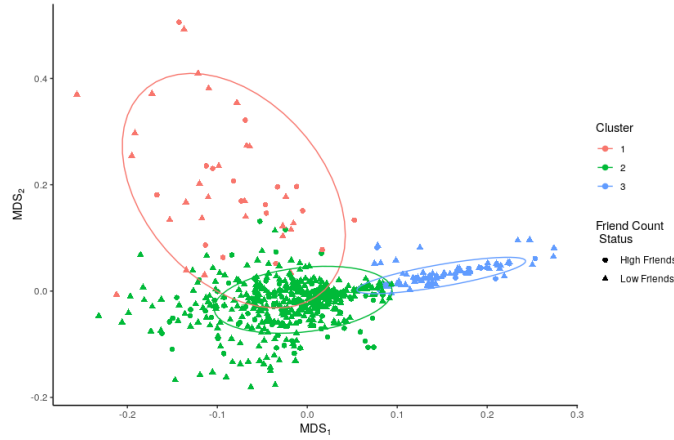


Figure 3: Distribution of Friend Count Status over Clusters

### 8.2.18 Cluster with Highest Number Friends

The Number of above or below friends corresponding to a given cluster can be tabulated by using the table function as shown in listing 35 and table 3, this indicates that there is a difference in the proportion of users with above-average friend counts between clusters with cluster #1 having the highest proportion of users with above-average friend counts.

```

1 (clust_friend <- table(
2   pca_data[,names(pca_data) %in% c("Cluster", "Friend_Status")])
3 )
4
5 cbind(clust_friend,
6        "Proportion" = signif(
7          (clust_friend[,1] / rowSums(clust_friend)),
8          2)
9      )

```

Listing 35: Tabulate the distribution of friends in

Table 3: Above Average Friend Counts Across Clusters

<b>Cluster</b>	<b>Above Average</b>	<b>Below Average</b>	<b>Proportion</b>
1	17	27	0.39
2	141	407	0.26
3	16	64	0.20

## 8.2.18 Sample Tweets from the cluster

Tweets can be sampled from the clusters by using the `sample` function with a logical test, this can be combined with a `for` loop as shown in the listing 36, this provides the following output.

- **Cluster 1**
  - @BelovedOfBayek @BayekOfSiwa @assassinscreed @Ubisoft @UbisoftMTL @Captured\Collec @GamerGram\GG @\GameScreenshot And those smiles - and love - are contagious. Believe me, it makes me very happy! What you have is special! ~ Steffi
  - @Ubisoft why is it wen I headshot sumone why dont it register a headshot is a 1 shot kill fix your game sir
  - @BikiniBodhi I think we need another movement going. Obviously Ubisoft already has plans for the next few releases of ops but we really need an Op whose ability is to reinforce more walls than others. Especially another castle on the team but for walls. Like 4 walls instead of 2.
  - The United Arab Emirates logged into my Ubisoft account. For why?
  - @Atalagummy Está gratis este finde en lo de ubisoft creo
- **Cluster 2**
  - Here's everything you need to know about Ubisoft's Watch Dogs Legion in Hindi - Release date, Story - Everything we know about it till now. <https://t.co/B4lMshJdqw> via @YouTube
  - @Rainbow6Game @TheGodlyNoob I have never seen a game company ruin their reputation so fast and so careless as Ubisoft
  - tiltei com a ubisoft, dei block na minha conta sem querer, to mt puto, real
  - My first game that really hyped me was a game called rolling thunder back in the 80s <https://t.co/mPWim2hwVY>
  - @videogamemorals @PartisanClown Two more remakes ofăLunar: The Silver Star,ăLunar LegendăandăLunar: Silver Star Harmony, were released in 2002 byăMedia RingsăandăUbisoftăand in 2009 byăGungHo Online EntertainmentăandăXseed Games, respectively.\
- **Cluster 3**
  - @Ubisoft @UbisoftSupport crossplay between xbox and pc for Division 2 please?
  - @tornado\raphi @Ubisoft @UbisoftDE Haha wollte auch einmal schlau sein :(
  - @Ubisoft I was in the middle in a game waiting for us to spawn in and it took forever and it somehow kicked me from the game for inactivity. Anyway to fix this <https://t.co/Uz1yZ73R4M>
  - @Operatedleech87 Yo lo veia al revez algo de Ubisoft en Girls Frontline, pero igual un juego en consola jalaria.
  - @VGPNetwork @GamerGram\GG @Ubisoft Haha.. I love that film!

## 8.2.19 Identify themes in the Tweets

### Highest Friends Count

The themes in the cluster with the highest friend count are:

```

1  set.seed(8923)
2  for(i in 1:3) {
3    n <- sample(which(pca_data$Cluster == i), size = 5)
4    print(tweets$text[n])
5    print("=====")
6    print("=====")
7  }

```

Listing 36: Sample Tweets from the Individual Clusters

- Discussion about upcoming titles
  - For example the *Sunborn-Ubisoft* collaboration
    - \* This is where the use of *collab* comes from in figure 4
  - The upcoming *Division 2* release
- Multiplayer games.

### Lowest Friends Count

The themes in the cluster with the lowest friends count are:

- Discussion about older titles
  - e.g. mincemeat refers to *Farcry*, mozzi is a *Rainbow 6* character
  - the recurring use of *remember* and *earliest*
- Console games and crossplay between them
- Feature Requests

## 8.2.20 Create WordClouds

Creating a word cloud from the stemmed words will make for a poor visualization while creating a word cloud based merely on the frequency of words will poorly identify words central to the theme of the cluster.

In light of this, the most appropriate method to generate word clouds for the purpose of visualizing the themes of the clusters is to create a document term matrix based on a corpus cleaned without word stemming as shown in the listing 37. Following that word clouds can then be generated by indexing for the cluster, this is demonstrated in listing 38 and shown in figures 4 and 6.

## 8.2.21 Use a Dendrogram to Display the Themes

A Dendrogram can be used to display the themes of the highest and lowest clusters. To do this the most frequent words of a cluster need to be filtered out to reduce the dimensions of the dendrogram,

```

1 clean_corp_ns <- function(corpus) {
2   ## Remove URL's
3   corpus <- tm_map(corpus, content_transformer(
4     function(x) gsub("(f|ht)tp(s?)://\\S+", "", x)))
5   ## Remove Usernames
6   corpus <- tm_map(corpus, content_transformer(function(x)
7     ↪ gsub("@\\w+", "", x)))
8   ## Misc
9   corpus <- tm_map(corpus, FUN = removeNumbers)
10  corpus <- tm_map(corpus, FUN = removePunctuation)
11  corpus <- tm_map(corpus, FUN = stripWhitespace)
12  corpus <- tm_map(corpus, FUN = tolower)
13  corpus <- tm_map(corpus, FUN = removeWords, mystop)
14  ## stopwords() returns characters and is feed as second argument
15  return(corpus)
16 }
17
18 tweet_corpus_clean_ns <- clean_corp(tweet_corpus)
19
20 tweet_raw_dtm <- tm::TermDocumentMatrix(x = tweet_corpus_ns,
21   control = list(weighting = weightTfIdf)) %>%
22   as.DocumentTermMatrix() %>%
23   as.matrix()
24
25 null = which(rowSums(as.matrix(tweet_matrix_dtm)) == 0)
26 length(null)
27
28 (rowSums(as.matrix(tweet_matrix_dtm))) %>% table()
29 if(length(null)!=0){
30   tweet_matrix_dtm = tweet_matrix_dtm[-null,]
31 }

```

Listing 37: Apply *TF-IDF* weighting to an unstemmed corpus and then use a for loop to create wordclouds corresponding to each cluster.

```

1 i <- 1
2
3 for (i in c(1,3)) {
4   n <- which(pca_data$Cluster == i)
5
6   (relevant <- sort(apply(tweet_raw_dtm[n,], 2, mean),
7     decreasing = TRUE)[1:30]) %>% head()
8
9   p <- brewer.pal(n = 5, name = "Set2")
10  wordcloud(
11    words = names(relevant),
12    freq = relevant,
13    colors = p,
14    random.color = FALSE
15  )
16
17 }

```

Listing 38: Apply *TF-IDF* weighting to an unstemmed corpus and then use a for loop to create wordclouds corresponding to each cluster.

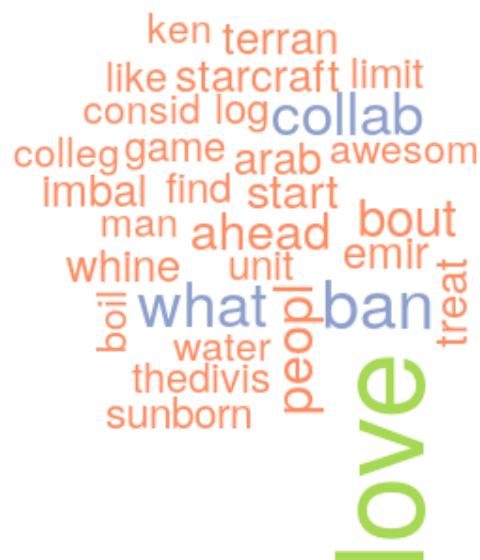


Figure 4: Wordcloud of Cluster # 1 using *TF-IDF* weighting





Figure 5: Wordcloud of Cluster # 2 using *TF-IDF* weighting



Figure 6: Wordcloud of Cluster # 3 using *TF-IDF* weighting

following that a dendrogram may be made to model the cosine distance between the terms. To achieve this in **R** the identity from equation (11) can be used and then a dendrogram modeled over the term document matrix.

### Highest Friend Count

Any word that occurs more than once was a candidate for the dendrogram visualizing the themes of the cluster with the highest number of friends. This dendrogram is shown in figure 7 and the corresponding code to produce it is provided in the listing 39.

### Lowest Friend Count

Any word that occurs more than once was a candidate for the dendrogram visualizing the themes of the cluster with the lowest number of friends. This dendrogram is shown in figure 8 and the corresponding code to produce it is provided in the listing 40.

```
1  ## Filter the Data To match the Cluster
2  tweet_weighted_dtm_c1 <- tweet_weighted_dtm[pca_data$Cluster==1, ]
3
4  ## Choose terms of a given frequency to reduce the dimensions
5  frequent.words = which(colSums(tweet_weighted_dtm_c1 > 0) > 1)
6  term.matrix = tweet_weighted_dtm_c1[,frequent.words]
7
8  ## In order to use the Cosine Distance Make each vector have a
9  ## magnitude of 1
10 unit_term.matrix = term.matrix %*% diag(1/sqrt(colSums(term.matrix^2)))
11
12 ## Preserve the column Names
13 colnames(unit_term.matrix) = colnames(term.matrix)
14 colnames(unit_term.matrix)
15
16 ## Find the Cosine Distance between the Terms
17 ## (Distance between terms so transpose)
18 t(unit_term.matrix)
19 D = dist(t(unit_term.matrix), method = "euclidean")^2/2
20
21 ## Perform Heirarchical Clustering
22 h = hclust(D, method="average")
23 plot(h, main = "Themes of Cluster with Highest Friend Count")
```

Listing 39: Create a dendrogram of the terms in the cluster with the highest friends count, average linkage was used.

```

1  ## Filter the Data To match the Cluster
2  tweet_weighted_dtm_c1 <- tweet_weighted_dtm[pca_data$Cluster==3, ]
3
4  ## Choose terms of a given frequency to reduce the dimensions
5  frequent.words = which(colSums(tweet_weighted_dtm_c1 > 0) > 3)
6  term.matrix = tweet_weighted_dtm_c1[,frequent.words]
7
8  ## In order to use the Cosine Distance Make each vector have a
9  ## magnitude of 1
10 unit_term.matrix = term.matrix %*% diag(1/sqrt(colSums(term.matrix^2)))
11
12 ## Preserve the column Names
13 colnames(unit_term.matrix) = colnames(term.matrix)
14 colnames(unit_term.matrix)
15
16 ## Find the Cosine Distance between the Terms
17 ## (Distance between terms so transpose)
18 t(unit_term.matrix)
19 D = dist(t(unit_term.matrix), method = "euclidean")^2/2
20
21 ## Perform Heirarchical Clustering
22 h = hclust(D, method="complete")
23 plot(h, main = "Themes of Cluster with Lowest Friend Count")

```

Listing 40: Create a dendrogram of the terms in the cluster with the highest friends count, average linkage was used.

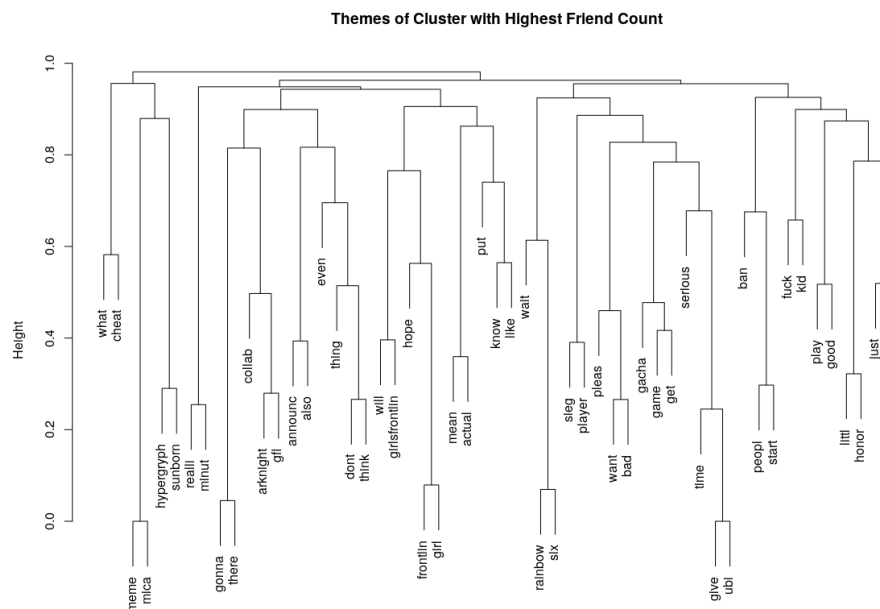


Figure 7: Dendrogram of Terms in the cluster with the highest friend count using average Linkage.

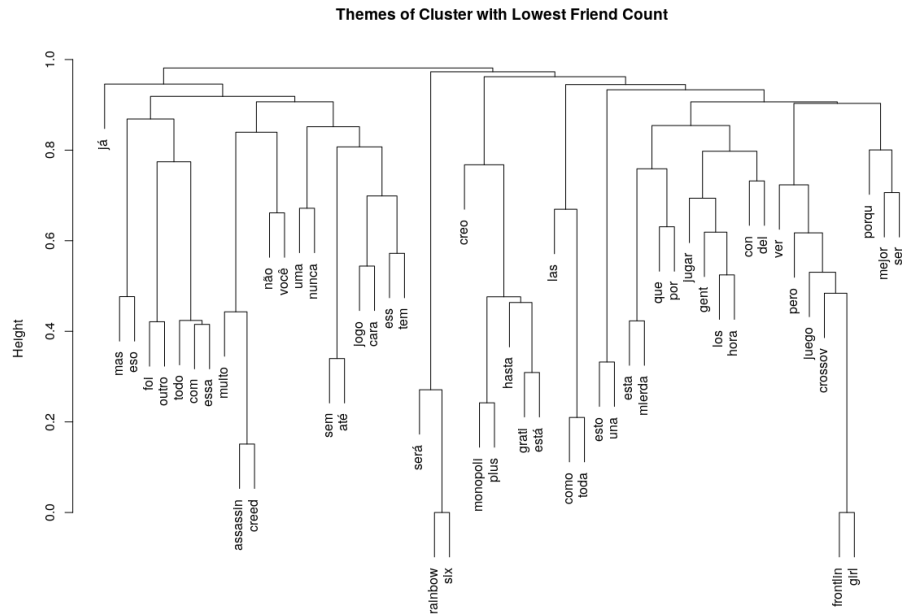


Figure 8: Dendrogram of Terms in the cluster with the highest friend count using average Linkage.

## 8.2.23 Conclusion

It appears that users who discuss newer titles and upcoming collaborations belong to a cluster with a higher proportion of users with an above average friend count, whereas users that discuss older titles (such as *Assasins Creed* as demonstrated in figure 8) tend to belong to a cluster with a lower proportion of users with high friends count.

This could imply that users discussing new titles are more likely to have more friends or it could imply that Twitter users with a high friend count are more likely to show interest in newer titles and material.

## 8.3 Building Networks

### 8.3.24 Find the 10 Most Popular Friends of the Twitter Handle

The `get_friends` function can be used to discover the friends of the [@ubisoft](#) account, these user *ID* values can then be passed as an argument to the `lookup_users` function, this will return the features of those given users and these results can be sorted by friend count to show the most popular friends, as shown in the listing 41, this provides the following list of friends in order of the number of there popularity:

- |                 |                |             |                 |                |
|-----------------|----------------|-------------|-----------------|----------------|
| 1. BarackObama  | 2. XboxSupport | 3. DEADLINE | 4. gameinformer | 5. Arbys       |
| 6. PeterHollens | 7. HarleyPlays | 8. G4C      | 9. Xbox         | 10. GailSimone |

```
1  ## * 8.2.24 Find 10 Most Popular Friends of the Twitter Handle
2  ## ## ** Get the User ID of Friends of Ubisoft
3  t <- get_friends("ubisoft", token = tk)
4  ## *** Get More Information of Friends
5  friends = lookup_users(t$user_id, token = tk)
6  ## **** Inspect the friends
7  dim(friends)
8  names(friends)
9
10 friends$screen_name[1] #name of friend at index 1
11 friends$followers_count[1] #examine the follower count of the first
   ↪ friend
12 friends$screen_name[2]
13
14 ## ** Find the 10 Most Popular Friends
15 friendPosition = order(friends$friends_count, decreasing = TRUE)[1:10]
16 topFriends = friends[friendPosition,] #ids of top 10 friends
17 ## *** Print the top 10 most popular friends
18 topFriends$screen_name
19
20 ## [1] "BarackObama" "XboxSupport" "DEADLINE" "gameinformer"
   ↪ "Arbys"
21 ## [6] "PeterHollens" "HarleyPlays" "G4C" "Xbox"
   ↪ "GailSimone"
```

Listing 41: Use `rtweet` to obtain the friends of *Ubisoft* with the most friends

### 8.3.25 Obtain a 2-degree egocentric graph centred at *Ubisoft*

#### Download Friends of Friends

The information of friends of friends can be downloaded by using the `get_friends` and `lookup_users` functions as shown in listing 42.

```
1  if (!file.exists("./AllGraphData.RData")) {  
2    for (i in 1:10) {  
3      ## Get friends of Each Friend  
4      t <- get_friends(topFriends$user_id[i], token <- tk)  
5  
6      ## Get the Data from Each Friend  
7      further_friends[[i]] <- lookup_users(t$user_id, token <- tk)  
8  
9      ## Save the Data  
10     save(list = ls(), file = "AllGraphData.RData")  
11   }  
12 } else {  
13   load(file = "AllGraphData.RData")  
14 }
```

Listing 42: Download the information of friends of friends with a loop, an `if` statement is implemented so this is only performed once.

#### Limit the number of Nodes for Want of Visualisation

To effectively visualize this graph, the number of nodes will be limited to 13, this can be achieved by filtering the list for more than 13 observations as shown in the listing 43, without performing this the visualization will be dominated by edges and difficult to interpret.

```
1  n <- 13  
2  for (a in 1:10) {  
3    if (nrow(further_friends[[a]]) > n) {  
4      further_friends[[a]] <- further_friends[[a]][1:n, ]  
5    }  
6  }
```

Listing 43: Limit the number of nodes in order to aid in visualisation of the graph.

## Build the Edge List

By building a matrix of screen names a description of the edges of the graph can be built, this is shown in the listing 44.

```
1 Edges <- cbind(rep(user$screen_name, 10),  
2               topFriends$screen_name[1:10]) # bind the columns to create  
               ↪ a matrix
```

Listing 44: Combine the friends into an Edge matrix.

## Build the Graph

the `igraph` package can then be used to construct a graph object using the `graph.edgelist` function, this can then be passed to the `plot` function to produce a plot of the graph as demonstrated in the listing 45 and shown in figure 9.

```
1 g <- graph.edgelist(Edges)  
2 plot(g, layout = layout.fruchterman.reingold,  
3      vertex.size = 7)
```

Listing 45: Create a graph from the edges and plot it.

## Use GGPLOT2

This Graph could also be produced using `ggplot2` and with an alternate algorithm for arranging the nodes, to do this a *tidy* data frame needs to be produced as demonstrated in the listing 46 and then the `ggplot` function can be used to build a plot as demonstrated in listing 47, this produces an alternative visualization of the graph as shown in figure 10.

### 8.3.26 Compute the closeness centrality score for each user

To compute a closeness centrality score it is necessary to rebuild the graph as shown in *Question 8.3.25*, it is crucial, however, that the data not be reduced for meaningful values to be provided, i.e. listing 43 should be omitted when re-producing the graph object to measure centrality.

The closeness centrality score can be determined by using the `closeness` function, the names of the nodes can be taken from the graph and then this can be sorted as shown in the listing 48.

### 8.3.27 List the Most Central People in the graph according to the closeness centrality

The most central people in order of closeness centrality are:

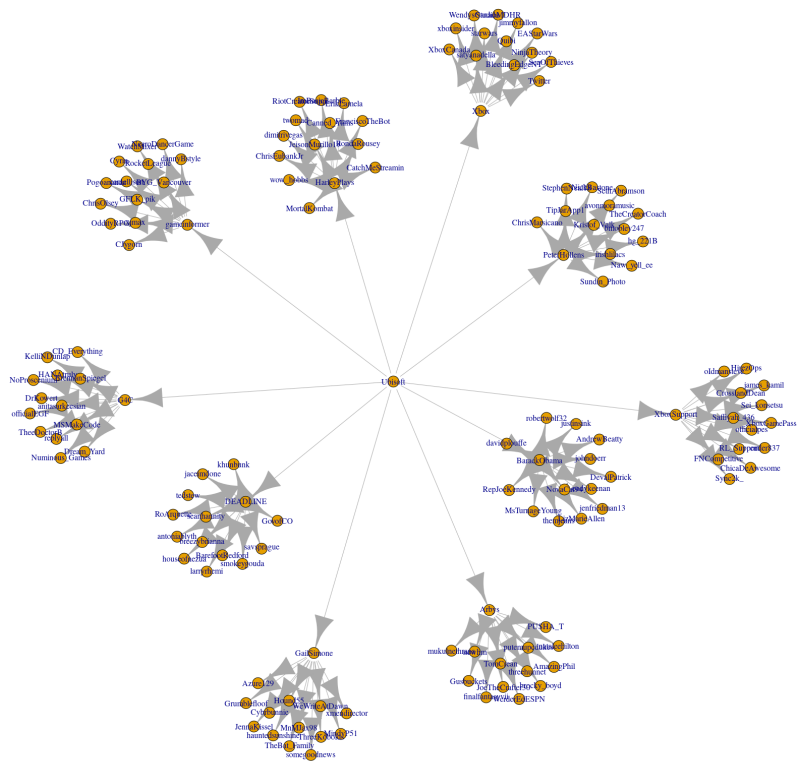


Figure 9: Base Plot of Graph using *Fruchterman-Reingold* Alorighm

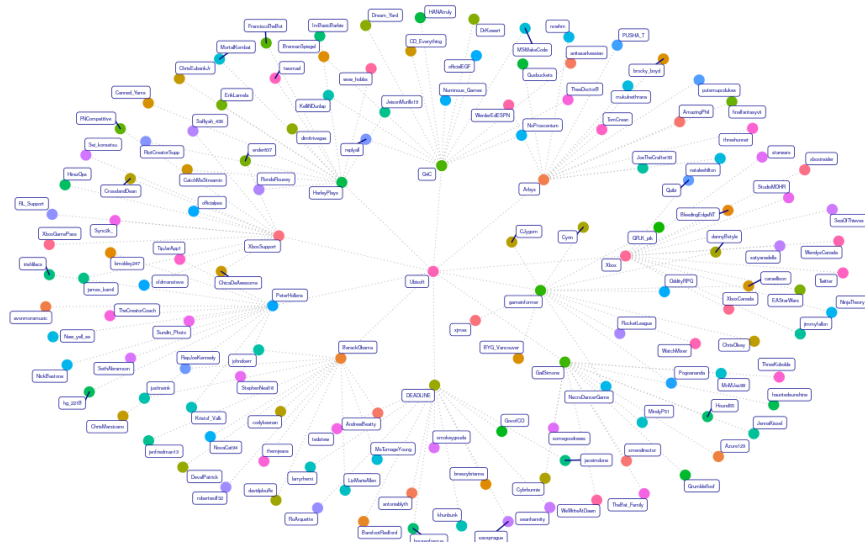


Figure 10: Graph Visualised using *Kamada-Kawai* algorithm and built with ggplot2



```

1  ## *** Get Edges #####
2  Edges <- as.data.frame(Edges)
3  names(Edges) <- c("Source", "Target")
4
5  ## *** Build the Data Frame #####
6  ne <- nrow(Edges) # Number of Edges
7  laytg <- as.data.frame(igraph::layout.kamada.kawai(g, dim = 2))
8  laytg$node <- vertex_attr(g)[[1]]
9  names(laytg) <- c("xval", "yval", "node")
10
11  laytg$xval[laytg$node==Edges$Source[2]]
12
13  ys <- xe <- ye <- xs <- vector(length = ne)
14  for (i in seq_len(length(xs))) {
15    xs[i] <- laytg$xval[laytg$node==Edges$Source[i]]
16    ys[i] <- laytg$yval[laytg$node==Edges$Source[i]]
17    xe[i] <- laytg$xval[laytg$node==Edges$Target[i]]
18    ye[i] <- laytg$yval[laytg$node==Edges$Target[i]]
19  }
20
21  starts <- data.frame("xval" = xs, "yval" = ys, edgenum = 1:ne) #
    ↪ TODO Make a factor
22  ends <- data.frame("xval" = xe, "yval" = ye, edgenum = 1:ne)
23  Edges_val <- as_tibble(rbind(starts,ends))

```

Listing 46: Create a tidy data frame describing the graph in order to use ggplot2.

```

1  ggplot(Edges_val, aes(x = xval, y = yval)) +
2  geom_label_repel(data = laytg,
3    mapping = aes(x = xval, y = yval, label = node),
4    col = "darkblue", size = 1.5, nudge_x = 0, nudge_y = 0) +
5  geom_line(aes(group = edgenum), col = "grey", lty = 2) +
6  geom_point(data = laytg,
7    aes(x = xval,
8      y = yval,
9      col = node),
10    size = 2) +
11  theme_classic() +
12  theme(axis.line = element_blank(),
13    axis.text.y=element_blank(),axis.ticks=element_blank(),
14    axis.text.x=element_blank())
15  guides(col = FALSE)

```

Listing 47: Use ggplot2 to construct a visualisation of the graph

```

1  ## Inspect the Closeness
2  closeness(g)[1:20]
3
4  ## Sort the Closeness
5  index=order(closeness(g), decreasing = TRUE)
6  centrality_vector <- c(names(g[index][,1]), closeness(g)[index])
7
8  ## Print the 3 most Central Users
9  centrality_vector[1:4]
10
11 ## "Ubisoft" "BarackObama" "XboxSupport"      "DEADLINE"

```

Listing 48: Calculating the *Closeness Centrality Score*

1. Ubisoft
2. Arbys
3. Xbox
4. GailSimone

Unsurprisingly *Ubisoft* is the most central person, this is expected because the graph was produced from the perspective of *Ubisoft* and the nodes had a very limited degree of interconnectivity.

### 8.3.28 Comment on the Results

*Ubisoft* has many connections as shown in figures 9 and 10, these connections, however, have a limited amount of interconnectivity.

Surprisingly, however, Gail Simone, a comic book author involved in works such as *Batgirl* and *Birds of Prey*, is a very central character, in combination with the central connection to *Arbys* this is indicative of the interests and lifestyle of the typical demographic of friends of the *Ubisoft* account.

## Appendix

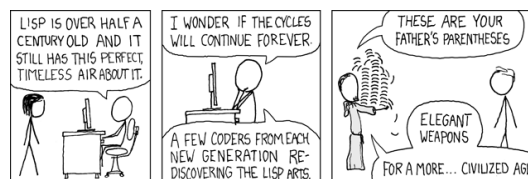


Figure 11: *xkcd* # 297

Table 4: User ID and Friend Count of users with above highest friend count in sample

<i><b>User ID</b></i>	<i><b>Friend Count</b></i>
274488119	8752
743771665	5002
1036014247	4999
2281452613	4992
1554453560	4958
981233818408570880	4944
931765564388921344	4836
807405140	4710
1112579152970842112	4514
2441577446	4322
552692862	4229
956297007127252992	3976
22493896	3675
255922782	3500
1067409881332936709	3312
27998570	3210
715118521555017728	3099
2356170174	2885
2372688230	2880
1868357425	2719

Table 5: User ID and Friend Count of users with above highest friend count in sample

<i><b>User ID</b></i>	<i><b>Friend Count</b></i>
1254280995592966145	0
875126772978913280	0
1254256124217319425	0
1250219450210480128	0
1214921087328411648	0
1254115699628421120	0
1217600080376520704	0
1253480062453600257	0
1254178435502571521	0
1251955545092718592	0
1106864828700712960	0
1160744587620524032	0
1254256536710504448	1
1129040408384868352	1
1254121201871589376	1
1248687797755658243	2
1210265263867932675	2
3380784928	3
1177274165239275520	3
54645521	3

Table 6: Document Term Matrix of first 6 documents and first 6 Words

check	coronavirus	death	first	gone	hour
0.615	0.747	0.830	0.596	0.747	0.698
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0

## References

- [1] James Carpenter and John Bithell. "Bootstrap Confidence Intervals: When, Which, What? A Practical Guide for Medical Statisticians". en. In: *Statistics in Medicine* 19.9 (2000), pp. 1141–1164. ISSN: 1097-0258. DOI: [10.1002/\(SICI\)1097-0258\(20000515\)19:9<1141::AID-SIM479>3.0.CO;2-F](https://doi.org/10.1002/(SICI)1097-0258(20000515)19:9<1141::AID-SIM479>3.0.CO;2-F). URL: [https://doi-org.ezproxy.uws.edu.au/10.1002/\(SICI\)1097-0258\(20000515\)19:9%3C1141::AID-SIM479%3E3.0.CO;2-F](https://doi-org.ezproxy.uws.edu.au/10.1002/(SICI)1097-0258(20000515)19:9%3C1141::AID-SIM479%3E3.0.CO;2-F) (visited on 04/27/2020) (cit. on p. 7).
- [2] A. C. Davison and D. V. Hinkley. *Bootstrap Methods and Their Application*. Cambridge ; New York, NY, USA: Cambridge University Press, 1997. ISBN: 978-0-521-57391-7 978-0-521-57471-6 (cit. on pp. 6, 7).
- [3] Tim C. Hesterberg. "What Teachers Should Know About the Bootstrap: Resampling in the Undergraduate Statistics Curriculum". In: *The American Statistician* 69.4 (Oct. 2015), pp. 371–386. ISSN: 0003-1305. DOI: [10.1080/00031305.2015.1089789](https://doi.org/10.1080/00031305.2015.1089789). URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4784504/> (visited on 04/26/2020) (cit. on p. 7).
- [4] Gareth James et al., eds. *An Introduction to Statistical Learning: With Applications in R*. Springer Texts in Statistics 103. OCLC: ocn828488009. New York: Springer, 2013. ISBN: 978-1-4614-7137-0 (cit. on p. 7).
- [5] Michael Kearney. *Get Tweets Data on Statuses Identified via Search Query*. *Search\_tweets*. en. Manual. 2019. URL: [https://rtweet.info/reference/search\\_tweets.html](https://rtweet.info/reference/search_tweets.html) (visited on 04/26/2020) (cit. on p. 2).
- [6] Travis LeCompte and Jianhua Chen. "Sentiment Analysis of Tweets Including Emoji Data". In: *2017 International Conference on Computational Science and Computational Intelligence (CSCI)*. Dec. 2017, pp. 793–798. DOI: [10.1109/CSCI.2017.137](https://doi.org/10.1109/CSCI.2017.137) (cit. on p. 17).
- [7] NIST. *1.3.3.14.6. Histogram Interpretation: Skewed (Non-Normal) Right*. Oct. 2013. URL: <https://www.itl.nist.gov/div898/handbook/eda/section3/histogr6.htm> (visited on 04/26/2020) (cit. on p. 5).
- [8] Brian Ripley. *Boot.Ci Function | R Documentation*. Apr. 2020. URL: <https://www.rdocumentation.org/packages/boot/versions/1.3-25/topics/boot.ci> (visited on 04/27/2020) (cit. on p. 7).
- [9] Mohammed O. Shiha and Serkan Ayvaz. "The Effects of Emoji in Sentiment Analysis". In: *International Journal of Computer and Electrical Engineering* 9.1 (2017), pp. 360–369. ISSN: 17938163. DOI: [10.17706/IJCEE.2017.9.1.360-369](https://doi.org/10.17706/IJCEE.2017.9.1.360-369). URL: <http://www.ijcee.org/vol9/943-T048.pdf> (visited on 04/29/2020) (cit. on p. 17).

- [10] Matt Wiley and Joshua Wiley. *Advanced R Statistical Programming and Data Models*. New York, NY: Springer Berlin Heidelberg, 2019. ISBN: 978-1-4842-2871-5 (cit. on p. 7).