MHI Email Tracking

Jae Yong (Francisco) Lee

2018-11-10

## Master of Health Informatics Cohort Email Tracking

## Data Preparation

Email was imported from the author’s email account between August 8th, 2018 and October 31st, 2018. From the email, the following attributes were extracted: (1) Date, (2) Subject, (3) Type, (4) From, (5) To, and (6) CC. Note that Outlook client does not have the listserv parent (i.e., listserv sent on behalf of …) as an attribute; therefore, it was manually extracted. This distinction is necessary because it provides insight on the sender of majority of listservs.

From this analysis, the author is interested in the exploring the following attributes:

* Areas of Interest
  + From the email,
    - is there a particular time of month?
    - is there a particular account?
  + Out of the listservs,
    - who sends the majority of them?
    - is there an overlap between them and other email senders?

# Set working directory  
setwd("~/Education/UofToronto (2018-2019)/")  
  
# Include libraries  
library(ggplot2)  
library(reshape2)  
library(ggpubr)

## Loading required package: magrittr

library(knitr)  
  
# Set figure theme  
theme\_set(theme\_pubr())  
  
# Import dataset  
mhi\_email <- read.csv(file = "mhi-email.csv", header = TRUE)  
  
# Sanitize date  
mhi\_email$date <- format(as.Date(mhi\_email$date, "%Y-%m-%d"),   
 format = "%m/%d")  
  
# Check missing data and data description  
summary(mhi\_email)

## date   
## Length:63   
## Class :character   
## Mode :character   
##   
##   
##   
##   
## subject   
## MHI Student Orientation/Bio Book 2018 : 3   
## \*UGRENT\* Response needed - Save the Date - November 8, 2018: 2   
## IHPME GSU Peer Support Program : 2   
## \*Reminder\* Professional Master's Bursary 2018 : 1   
## 2019 Vernissage Health Student Applications Now Open : 1   
## 9 new things that you might not know: Oct 3 - 9, 2018 : 1   
## (Other) :53   
## type   
## : 3   
## email :16   
## listserv :26   
## listserv, email:17   
## noreply : 1   
##   
##   
## from   
## ihpme-mhi-2018-l@listserv.utoronto.ca, ihpme.events@utoronto.ca :15   
## ihpme-mhi-2018-l@listserv.utoronto.ca, ihpme.mhi.grad@utoronto.ca :10   
## ihpme.mhi.grad@utoronto.ca : 7   
## ihpme-mhi-2018-l@listserv.utoronto.ca, communications.dlsph@utoronto.ca: 4   
## ihpme.mhi.program@utoronto.ca : 4   
## ihpme-mhi-2018-l@listserv.utoronto.ca, ihpme.mhi.program@utoronto.ca : 3   
## (Other) :20   
## to   
## :14   
## amra.das@mail.utoronto.ca, howardw.wong@mail.utoronto.ca, jaeyongf.lee@mail.utoronto.ca : 1   
## aquatics-l@listserv.utoronto.ca : 2   
## ihpme-emhi-2017-l@listserv.utoronto.ca, ihpme-emhi-2018-l@listserv.utoronto.ca, ihpme-mhi-2017-l@listserv.utoronto.ca: 1   
## ihpme-mhi-2018-l@listserv.utoronto.ca :37   
## jaeyongf.lee@mail.utoronto.ca : 7   
## student\_society\_varsity-l@listserv.utoronto.ca : 1   
## cc   
## :58   
## ihpme.mhi.grad@utoronto.ca: 3   
## julia.zarb@utoronto.ca : 2   
##   
##   
##   
##

sapply(mhi\_email, function(x) sum(is.na(x)))

## date subject type from to cc   
## 0 0 0 0 0 0

str(mhi\_email)

## 'data.frame': 63 obs. of 6 variables:  
## $ date : chr "09/07" "10/01" "10/22" "09/07" ...  
## $ subject: Factor w/ 59 levels "\*Reminder\* Professional Master's Bursary 2018",..: 55 56 57 19 51 6 10 52 46 3 ...  
## $ type : Factor w/ 5 levels "","email","listserv",..: 3 4 2 2 3 4 3 4 3 4 ...  
## $ from : Factor w/ 19 levels "aquatics-l@listserv.utoronto.ca",..: 1 1 2 5 3 6 3 3 3 4 ...  
## $ to : Factor w/ 7 levels "","amra.das@mail.utoronto.ca, howardw.wong@mail.utoronto.ca, jaeyongf.lee@mail.utoronto.ca",..: 3 3 6 1 5 5 5 5 5 5 ...  
## $ cc : Factor w/ 3 levels "","ihpme.mhi.grad@utoronto.ca",..: 1 1 1 1 1 1 1 1 1 1 ...

## Frequency based on:

# (1) Date

The following graph displays the number of emails received by a student on the corresponding date.

# Compute the frequency.  
library(dplyr)

##   
## Attaching package: 'dplyr'

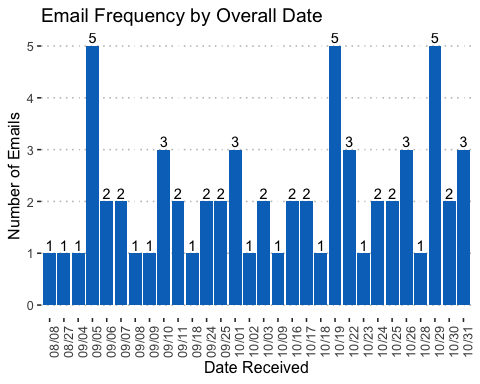
## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

df <- mhi\_email %>% group\_by(date) %>% summarise(counts = n())  
df

## # A tibble: 30 x 2  
## date counts  
## <chr> <int>  
## 1 08/08 1  
## 2 08/27 1  
## 3 09/04 1  
## 4 09/05 5  
## 5 09/06 2  
## 6 09/07 2  
## 7 09/08 1  
## 8 09/09 1  
## 9 09/10 3  
## 10 09/11 2  
## # ... with 20 more rows

# Create bar plot.  
ggplot(data = df, aes(x = date, y = counts)) + ggtitle("Email Frequency by Overall Date") +   
 theme(plot.title = element\_text(hjust = 0.5)) + labs(x = "Date Received",   
 y = "Number of Emails") + geom\_bar(fill = "#0073C2FF", stat = "identity") +   
 geom\_text(aes(label = counts), vjust = -0.3) + theme\_pubclean() +   
 theme(axis.text.x = element\_text(angle = 90, hjust = 1))

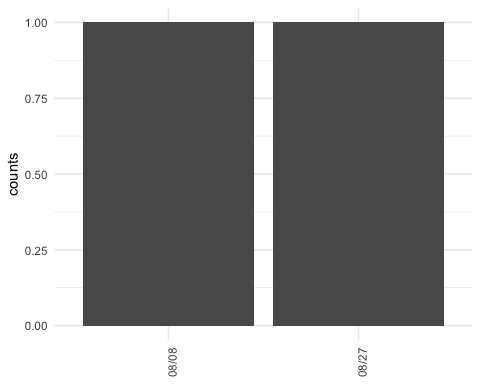


The following graph displays the number of emails received by a student

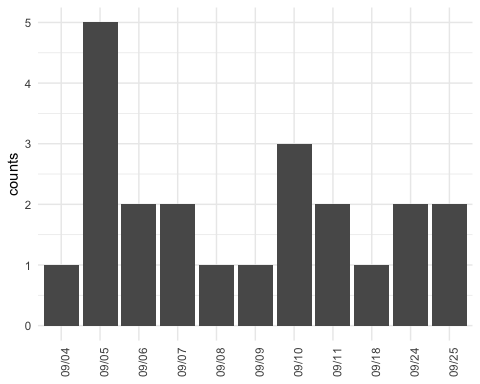
# Compute the frequency.  
library(dplyr)  
df <- mhi\_email %>% group\_by(date) %>% summarise(counts = n())  
df

## # A tibble: 30 x 2  
## date counts  
## <chr> <int>  
## 1 08/08 1  
## 2 08/27 1  
## 3 09/04 1  
## 4 09/05 5  
## 5 09/06 2  
## 6 09/07 2  
## 7 09/08 1  
## 8 09/09 1  
## 9 09/10 3  
## 10 09/11 2  
## # ... with 20 more rows

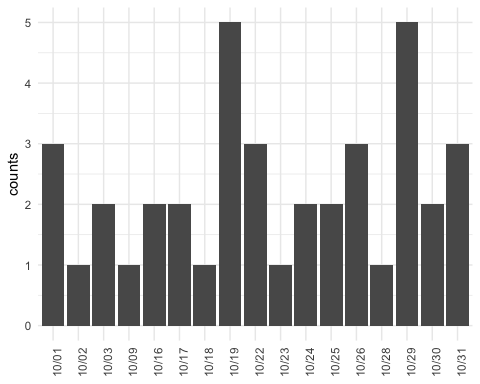
# August  
df8 <- df[grep("08/", df$date, perl = TRUE, value = FALSE), ]  
ggplot(df8, aes(date, counts)) + geom\_bar(stat = "identity") +   
 theme\_minimal() + theme(axis.title.x = element\_blank(), axis.text.x = element\_text(angle = 90))



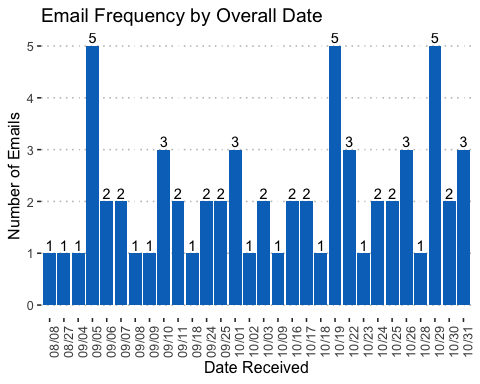
# September  
df9 <- df[grep("09/", df$date, perl = TRUE, value = FALSE), ]  
ggplot(df9, aes(date, counts)) + geom\_bar(stat = "identity") +   
 theme\_minimal() + theme(axis.title.x = element\_blank(), axis.text.x = element\_text(angle = 90))



# October  
df10 <- df[grep("10/", df$date, perl = TRUE, value = FALSE),   
 ]  
ggplot(df10, aes(date, counts)) + geom\_bar(stat = "identity") +   
 theme\_minimal() + theme(axis.title.x = element\_blank(), axis.text.x = element\_text(angle = 90))



# Create bar plot.  
ggplot(data = df, aes(x = date, y = counts)) + ggtitle("Email Frequency by Overall Date") +   
 theme(plot.title = element\_text(hjust = 0.5)) + labs(x = "Date Received",   
 y = "Number of Emails") + geom\_bar(fill = "#0073C2FF", stat = "identity") +   
 geom\_text(aes(label = counts), vjust = -0.3) + theme\_pubclean() +   
 theme(axis.text.x = element\_text(angle = 90, hjust = 1))



# (2) Sender

The following graph displays the number of emails sent by an account on the corresponding date. Note that the compilation of sender was abbreviated in order to reduce the length of labels.

# Table of abbreviations  
knitr::kable(unique(mhi\_email$from), caption = "Abbreviation of Senders")

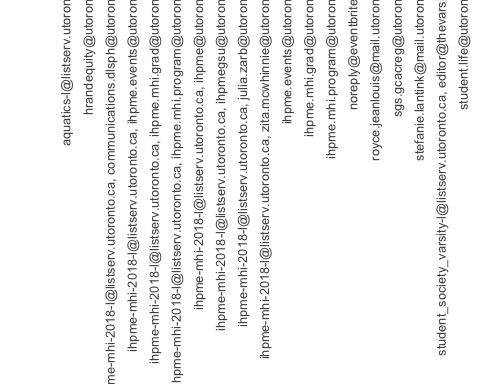
Abbreviation of Senders

|  |
| --- |
| x |
| [aquatics-l@listserv.utoronto.ca](mailto:aquatics-l@listserv.utoronto.ca) |
| [hrandequity@utoronto.ca](mailto:hrandequity@utoronto.ca) |
| [ihpme-mhi-2018-l@listserv.utoronto.ca](mailto:ihpme-mhi-2018-l@listserv.utoronto.ca), [ihpme.mhi.grad@utoronto.ca](mailto:ihpme.mhi.grad@utoronto.ca) |
| [ihpme-mhi-2018-l@listserv.utoronto.ca](mailto:ihpme-mhi-2018-l@listserv.utoronto.ca), [communications.dlsph@utoronto.ca](mailto:communications.dlsph@utoronto.ca) |
| [ihpme-mhi-2018-l@listserv.utoronto.ca](mailto:ihpme-mhi-2018-l@listserv.utoronto.ca), [ihpme.mhi.program@utoronto.ca](mailto:ihpme.mhi.program@utoronto.ca) |
| [ihpme-mhi-2018-l@listserv.utoronto.ca](mailto:ihpme-mhi-2018-l@listserv.utoronto.ca), [ihpme.events@utoronto.ca](mailto:ihpme.events@utoronto.ca) |
| [ihpme-mhi-2018-l@listserv.utoronto.ca](mailto:ihpme-mhi-2018-l@listserv.utoronto.ca), [ihpme@utoronto.ca](mailto:ihpme@utoronto.ca) |
| [ihpme-mhi-2018-l@listserv.utoronto.ca](mailto:ihpme-mhi-2018-l@listserv.utoronto.ca), [ihpmegsu@utoronto.ca](mailto:ihpmegsu@utoronto.ca) |
| [ihpme-mhi-2018-l@listserv.utoronto.ca](mailto:ihpme-mhi-2018-l@listserv.utoronto.ca), [julia.zarb@utoronto.ca](mailto:julia.zarb@utoronto.ca) |
| [ihpme-mhi-2018-l@listserv.utoronto.ca](mailto:ihpme-mhi-2018-l@listserv.utoronto.ca), [zita.mcwhinnie@utoronto.ca](mailto:zita.mcwhinnie@utoronto.ca) |
| [ihpme.events@utoronto.ca](mailto:ihpme.events@utoronto.ca) |
| [ihpme.mhi.grad@utoronto.ca](mailto:ihpme.mhi.grad@utoronto.ca) |
| [ihpme.mhi.program@utoronto.ca](mailto:ihpme.mhi.program@utoronto.ca) |
| [noreply@eventbrite.com](mailto:noreply@eventbrite.com) |
| [royce.jeanlouis@mail.utoronto.ca](mailto:royce.jeanlouis@mail.utoronto.ca) |
| [sgs.gcacreg@utoronto.ca](mailto:sgs.gcacreg@utoronto.ca) |
| [stefanie.lantink@mail.utoronto.ca](mailto:stefanie.lantink@mail.utoronto.ca) |
| [student\_society\_varsity-l@listserv.utoronto.ca](mailto:student_society_varsity-l@listserv.utoronto.ca), [editor@thevarsity.ca](mailto:editor@thevarsity.ca) |
| [student.life@utoronto.ca](mailto:student.life@utoronto.ca) |

# Compute the frequency.  
library(dplyr)  
sf <- mhi\_email %>% group\_by(from) %>% summarise(counts = n())  
sf

## # A tibble: 19 x 2  
## from counts  
## <fct> <int>  
## 1 aquatics-l@listserv.utoronto.ca 2  
## 2 hrandequity@utoronto.ca 1  
## 3 ihpme-mhi-2018-l@listserv.utoronto.ca, communications.dlsph@uto… 4  
## 4 ihpme-mhi-2018-l@listserv.utoronto.ca, ihpme.events@utoronto.ca 15  
## 5 ihpme-mhi-2018-l@listserv.utoronto.ca, ihpme.mhi.grad@utoronto.… 10  
## 6 ihpme-mhi-2018-l@listserv.utoronto.ca, ihpme.mhi.program@utoron… 3  
## 7 ihpme-mhi-2018-l@listserv.utoronto.ca, ihpme@utoronto.ca 1  
## 8 ihpme-mhi-2018-l@listserv.utoronto.ca, ihpmegsu@utoronto.ca 3  
## 9 ihpme-mhi-2018-l@listserv.utoronto.ca, julia.zarb@utoronto.ca 3  
## 10 ihpme-mhi-2018-l@listserv.utoronto.ca, zita.mcwhinnie@utoronto.… 1  
## 11 ihpme.events@utoronto.ca 1  
## 12 ihpme.mhi.grad@utoronto.ca 7  
## 13 ihpme.mhi.program@utoronto.ca 4  
## 14 noreply@eventbrite.com 1  
## 15 royce.jeanlouis@mail.utoronto.ca 2  
## 16 sgs.gcacreg@utoronto.ca 1  
## 17 stefanie.lantink@mail.utoronto.ca 1  
## 18 student\_society\_varsity-l@listserv.utoronto.ca, editor@thevarsi… 1  
## 19 student.life@utoronto.ca 2

# Create bar plot.  
ggplot(sf, aes(x = from, y = counts)) + geom\_bar(fill = "#0073C2FF",   
 stat = "identity") + geom\_text(aes(label = counts), vjust = -0.3) +   
 theme\_pubclean() + theme(axis.text.x = element\_text(angle = 90,   
 hjust = 1))



# (3) Sender and Date

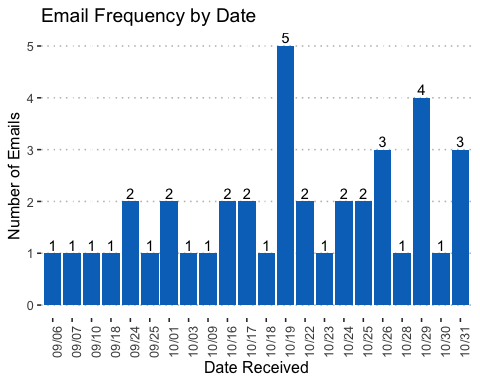
The following graph displays the number of emails sent by an account on the corresponding date.

# Compute the sender frequency by date.  
library(dplyr)  
sfd <- mhi\_email %>% group\_by(date, from) %>% summarise(n = n())  
  
# Create bar plot.

# (4) Listserv Majority

The following graph displays the composition of listserv senders. Given that the total number of listservs received outside of IHPME was three, they were excluded from the selection (two emails were from Aquatics Schedule and one email was from Varsity Magazine).

# Extract listservs  
listservs <- mhi\_email[grep("ihpme-mhi-2018-l@listserv.utoronto.ca",   
 mhi\_email$from, perl = TRUE, value = FALSE), ]  
  
# Replace listserv address with empty string from 'From'  
listservs$from <- sub("(ihpme-mhi-2018-l@listserv.utoronto.ca, )",   
 "", listservs$from)  
  
# Compute the date frequency  
lfd <- listservs %>% group\_by(date) %>% summarise(counts = n())  
  
# Create bar plot.  
ggplot(data = lfd, aes(x = date, y = counts)) + ggtitle("Email Frequency by Date") +   
 theme(plot.title = element\_text(hjust = 0.5)) + labs(x = "Date Received",   
 y = "Number of Emails") + geom\_bar(fill = "#0073C2FF", stat = "identity") +   
 geom\_text(aes(label = counts), vjust = -0.3) + theme\_pubclean() +   
 theme(axis.text.x = element\_text(angle = 90, hjust = 1))



# Compute the sender frequency  
lfs <- listservs %>% group\_by(from) %>% summarise(counts = n())  
  
# Create bar plot.