MHI Email Tracking

Jae Yong (Francisco) Lee

2018-11-13

Master of Health Informatics Cohort Email Tracking

Data Preparation

Emails received between August 8th, 2018 and November 13th, 2018 were imported from the author's email account. 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 into the sender(s) for the majority of listservs.

From this analysis, the author explored 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_original.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:115
##  Class :character
##  Mode  :character
##
##
##
```

```
##
##
##  MHI Student Orientation/Bio Book 2018
##  Notice of Varsity Publications Board of Directors By-Election
##  *UGRENT* Response needed - Save the Date - November 8, 2018
##  IHPME GSU Peer Support Program
##  Student Seminar Series 2018-19
##  **Reminder: TODAY** Thomas Rice NAO and CCHE Special Lecture/ November 12: Improving Consumer Dec
##  (Other)
##               type
##                 : 3
##  email          :17
##  listserv       :61
##  listserv, email:33
##  noreply        : 1
##
##
##                                                                    from
##  ihpme-mhi-2018-l@listserv.utoronto.ca, ihpme.events@utoronto.ca     :24
##  ihpme-mhi-2018-l@listserv.utoronto.ca, ihpme.mhi.grad@utoronto.ca   :15
##  student_society_utgsu-l@listserv.utoronto.ca, utgsu@utgsu.ca        :15
##  ihpme-mhi-2018-l@listserv.utoronto.ca, ihpmegsu@utoronto.ca         :10
##  ihpme.mhi.grad@utoronto.ca                                          : 7
##  student_society_varsity-l@listserv.utoronto.ca, editor@thevarsity.ca: 6
##  (Other)                                                             :38
##                                              to
##  ihpme-mhi-2018-l@listserv.utoronto.ca         :59
##  student_society_utgsu-l@listserv.utoronto.ca  :17
##  jaeyongf.lee@mail.utoronto.ca                 :16
##                                                :14
##  student_society_varsity-l@listserv.utoronto.ca: 5
##  aquatics-l@listserv.utoronto.ca               : 2
##  (Other)                                       : 2
##                           cc
##                            :110
##  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':    115 obs. of  6 variables:
##  $ date   : chr  "08/08" "08/27" "08/30" "09/04" ...
##  $ subject: Factor w/ 108 levels "**Reminder: TODAY** Thomas Rice NAO and CCHE Special Lecture/ No
##  $ type   : Factor w/ 5 levels "","email","listserv",..: 2 2 3 2 2 2 2 2 2 3 ...
##  $ from   : Factor w/ 24 levels "aquatics-l@listserv.utoronto.ca",..: 14 14 22 15 14 14 14 14 19 5
##  $ to     : Factor w/ 8 levels "","amra.das@mail.utoronto.ca, howardw.wong@mail.utoronto.ca, jaeyo
##  $ cc     : Factor w/ 3 levels "","ihpme.mhi.grad@utoronto.ca",..: 1 1 1 1 1 1 2 2 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'
```

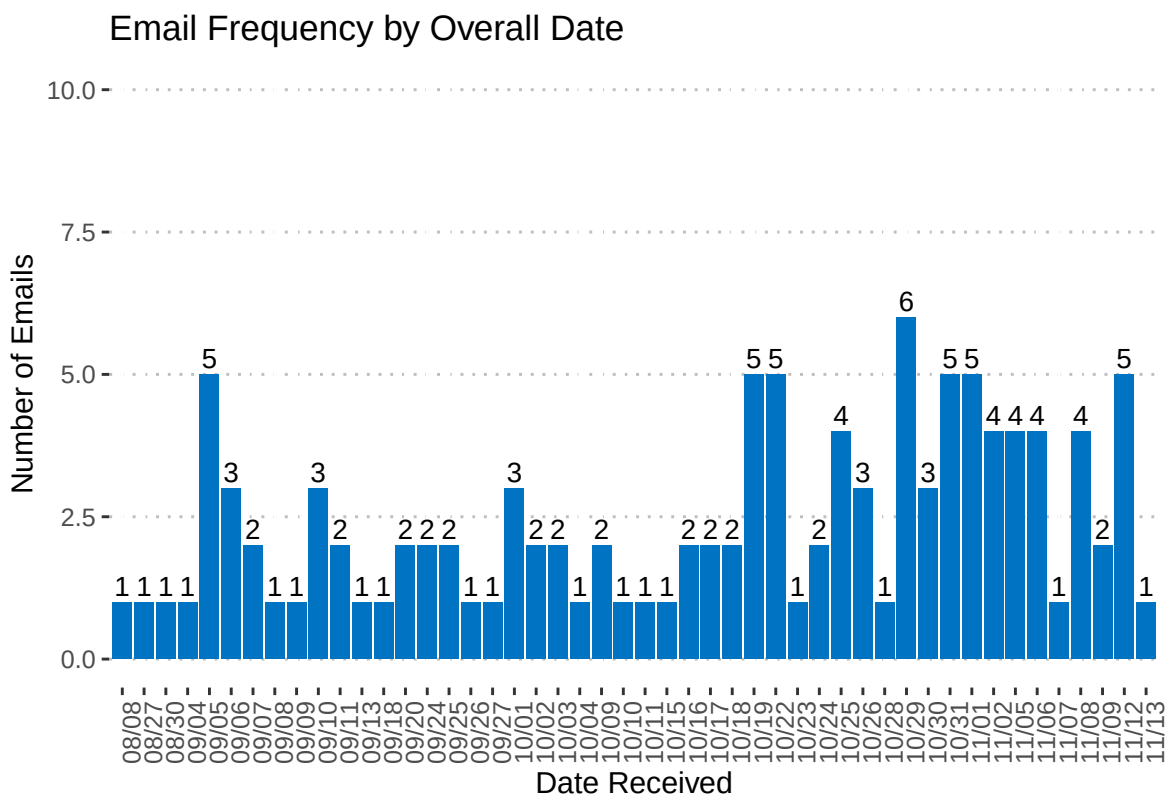
```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

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

```
df <- mhi_email %>% group_by(date) %>% summarise(counts = n())
df
```

```
## # A tibble: 48 x 2
##    date  counts
##    <chr>  <int>
##  1 08/08      1
##  2 08/27      1
##  3 08/30      1
##  4 09/04      1
##  5 09/05      5
##  6 09/06      3
##  7 09/07      2
##  8 09/08      1
##  9 09/09      1
## 10 09/10      3
## # ... with 38 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") + scale_y_continuous(limits = c(0,
    10)) + 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: 48 x 2
##    date  counts
##    <chr>  <int>
##  1 08/08      1
##  2 08/27      1
##  3 08/30      1
##  4 09/04      1
##  5 09/05      5
##  6 09/06      3
##  7 09/07      2
##  8 09/08      1
##  9 09/09      1
## 10 09/10      3
## # ... with 38 more rows
```

```
# August
df8 <- df[grep("08/", df$date, perl = TRUE, value = FALSE), ]
g8 <- ggplot(df8, aes(date, counts)) + scale_y_continuous(limits = c(0,
    10)) + geom_bar(fill = "#0073C2FF", stat = "identity") +
    geom_text(aes(label = counts), vjust = -0.3) + theme_pubclean() +
    theme(axis.title.x = element_blank(), axis.text.x = element_text(angle = 90))

# September
df9 <- df[grep("09/", df$date, perl = TRUE, value = FALSE), ]
g9 <- ggplot(df9, aes(date, counts)) + scale_y_continuous(limits = c(0,
```

```
    10)) + geom_bar(fill = "#0073C2FF", stat = "identity") +
    geom_text(aes(label = counts), vjust = -0.3) + theme_pubclean() +
    theme(axis.title.x = element_blank(), axis.text.x = element_text(angle = 90))

# October
df10 <- df[grep("10/", df$date, perl = TRUE, value = FALSE),
    ]
g10 <- ggplot(df10, aes(date, counts)) + scale_y_continuous(limits = c(0,
    10)) + geom_bar(fill = "#0073C2FF", stat = "identity") +
    geom_text(aes(label = counts), vjust = -0.3) + theme_pubclean() +
    theme(axis.title.x = element_blank(), axis.text.x = element_text(angle = 90))

# November
df11 <- df[grep("11/", df$date, perl = TRUE, value = FALSE),
    ]
g11 <- ggplot(df11, aes(date, counts)) + scale_y_continuous(limits = c(0,
    10)) + geom_bar(fill = "#0073C2FF", stat = "identity") +
    geom_text(aes(label = counts), vjust = -0.3) + theme_pubclean() +
    theme(axis.title.x = element_blank(), axis.text.x = element_text(angle = 90))

# Stack the months gridExtra::grid.arrange(grobs = c(g8, g9),
# ncol=2)
```

(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")
```

Table 1: Abbreviation of Senders

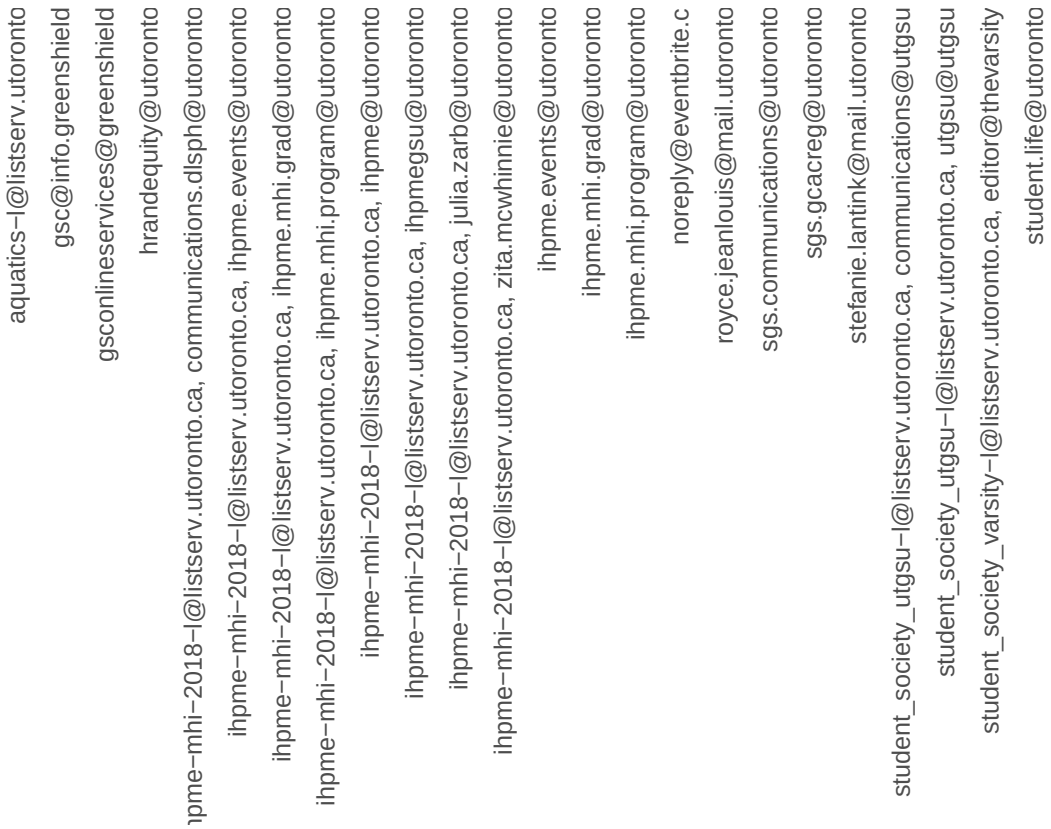
x
ihpme.mhi.grad@utoronto.ca
student_society_utgsu-l@listserv.utoronto.ca, utgsu@utgsu.ca
ihpme.mhi.program@utoronto.ca
sgs.gcacreg@utoronto.ca
ihpme-mhi-2018-l@listserv.utoronto.ca, communications.dlsph@utoronto.ca
ihpme-mhi-2018-l@listserv.utoronto.ca, ihpme.mhi.grad@utoronto.ca
aquatics-l@listserv.utoronto.ca
noreply@eventbrite.com
royce.jeanlouis@mail.utoronto.ca
ihpme-mhi-2018-l@listserv.utoronto.ca, ihpme@utoronto.ca
stefanie.lantink@mail.utoronto.ca
ihpme-mhi-2018-l@listserv.utoronto.ca, ihpme.events@utoronto.ca
gsc@info.greenshield.ca
ihpme-mhi-2018-l@listserv.utoronto.ca, ihpme.mhi.program@utoronto.ca
student_society_utgsu-l@listserv.utoronto.ca, communications@utgsu.ca
student_society_varsity-l@listserv.utoronto.ca, editor@thevarsity.ca
student.life@utoronto.ca
gsconlineservices@greenshield.ca
ihpme-mhi-2018-l@listserv.utoronto.ca, julia.zarb@utoronto.ca
ihpme-mhi-2018-l@listserv.utoronto.ca, ihpmegsu@utoronto.ca
hrandequity@utoronto.ca

x
ihpme-mhi-2018-l@listserv.utoronto.ca, zita.mcwhinnie@utoronto.ca
ihpme.events@utoronto.ca
sgs.communications@utoronto.ca

```
# Compute the frequency.
library(dplyr)
sf <- mhi_email %>% group_by(from) %>% summarise(counts = n())
sf
```

```
## # A tibble: 24 x 2
##    from                                                          counts
##    <fct>                                                          <int>
##  1 aquatics-l@listserv.utoronto.ca                                    2
##  2 gsc@info.greenshield.ca                                            2
##  3 gsconlineservices@greenshield.ca                                   2
##  4 hrandequity@utoronto.ca                                            1
##  5 ihpme-mhi-2018-l@listserv.utoronto.ca, communications.dlsph@uto~     4
##  6 ihpme-mhi-2018-l@listserv.utoronto.ca, ihpme.events@utoronto.ca     24
##  7 ihpme-mhi-2018-l@listserv.utoronto.ca, ihpme.mhi.grad@utoronto.~    15
##  8 ihpme-mhi-2018-l@listserv.utoronto.ca, ihpme.mhi.program@utoron~     3
##  9 ihpme-mhi-2018-l@listserv.utoronto.ca, ihpme@utoronto.ca             2
## 10 ihpme-mhi-2018-l@listserv.utoronto.ca, ihpmegsu@utoronto.ca         10
## # ... with 14 more rows
```

```
# 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

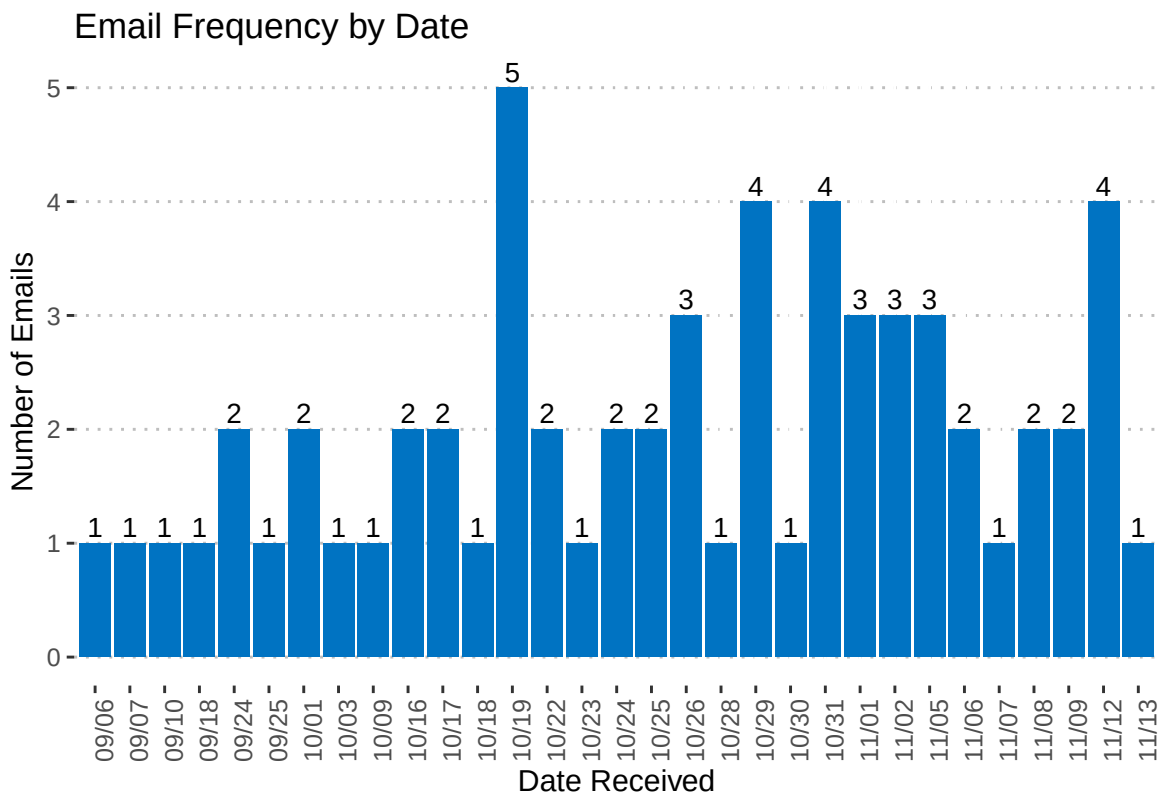
The following graph displays the composition of listserv senders. Given that the total number of listservs received outside of IHPME was three, they were excluded from the selection (two emails were from Aquatics Schedule and one email was from Varsity Magazine).

```
# 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.
```

Discussion

Based on the data exploration, there are redundancies in the content of listservs. The frequency at which event reminders are sent overloads the receiving end.

In addition, the _____