Brain Science Study Part № 1

Nirupam Bidikar-1878058, Akshit Tandon-1792038, Rahul Raj Mogili-1900425 03/08/2020

Analysis 1

Construct the probability distribution function of the year of first publication of faculty.

Listing 1: Probability distribution function calculated

```
1 author_data <- read.csv("brain_author.csv")
2 pub_year <- author_data %>% filter( min_pub_year > 1960) %>% select(↩
    min_pub_year)
3
4 g <- ggplot( pub_year,aes(min_pub_year)) + ggtitle("PDF for First ↩
    Publication year of Faculty") + labs(y = "PDF",x = "Year" ) + ↩
    geom_histogram(aes(y = ..density..),binwidth = 2, colour = "black", ↩
    fill = "white") + geom_density(alpha=.2, fill="#FF6666")
5 g
```

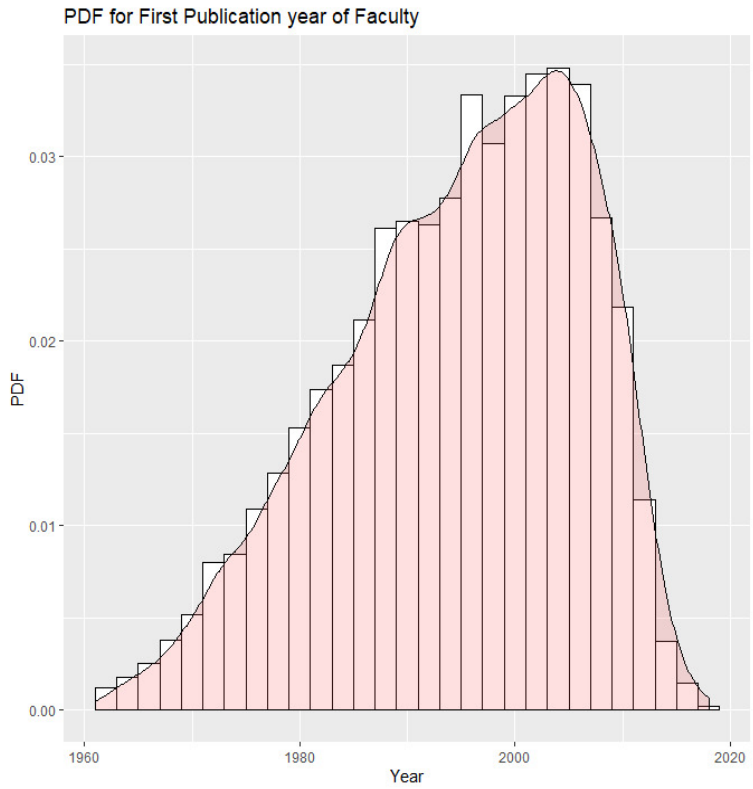

Figure 1: Probability Distribution Function of first year publications

The results indicate that we find a few outliers before the year 1960 which is statistically non relevant in this case study.

The number of publications from the above Probability Distribution Function gradually increases from the year 1960 up and until 2006, peaking at the year 2004. After the year 2006, the number of publications plummets till the time the records were updated. We also observe that the distribution is slightly skewed to the right with the later half of the distribution has more weight than the former.

The implementation of it can be found in the code snippet above. We load the author data and filter it to select minimum publishing year as 1960. We then use the histogram and density functions from ggplot to plot the PDF.

Analysis 2

Construct the probability distribution function of the total citations of faculty.

Listing 2: PDF calculated of the total citations of faculty

```
1 total_citations <- author_data %>% filter(min_pub_year > 1960 && citations↩
     != 0) %>% select(citations)
2 transform <- log(1+total_citations)
3 g <- ggplot(transform,aes(citations)) + ggtitle("PDF for Total Citations ↩
    of Faculty") + labs(y = "PDF",x = "Total Citations" )+ geom_histogram(↩
    aes(y = ..density..), binwidth = 0.5, colour = "black", fill = "white"↩
    ) + geom_density(alpha=.2, fill="#FF6666")
4 g
```

We are using a log operator for this analysis to help plot the Probability Distribution Function. The Probability Distribution Function is almost normally distributed. The implementation can be found in the code snippet above. We load the author data and filter it avoiding all rows with 0 citations and minimum publishing year as 1960. We use the previously used plotting functions to plot the graph.

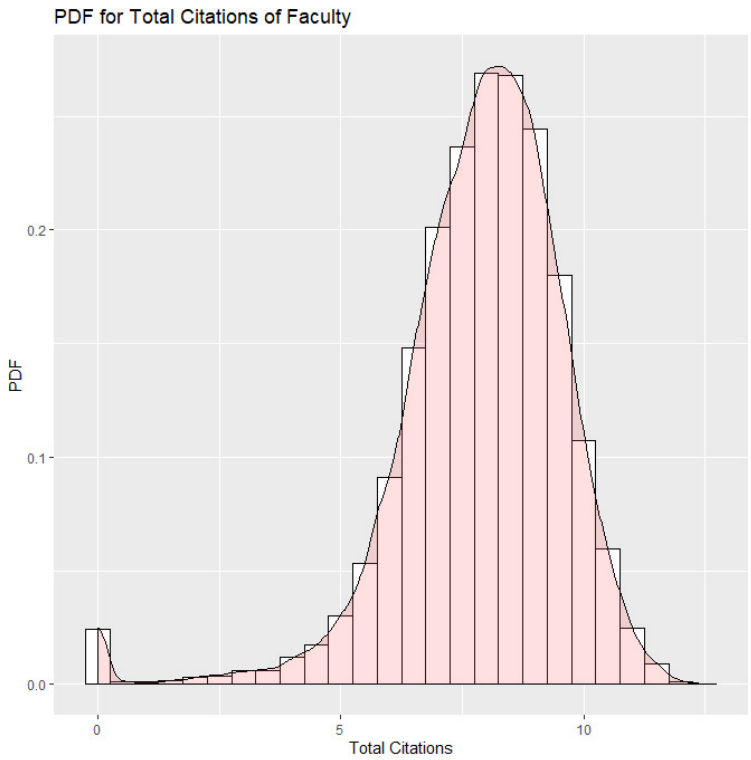

Figure 2: Probability Density Function of total citations of faculty

Analysis 3

Cluster the subject areas of faculty using the Louvain algorithm (k=5)

Listing 3: Cluster using Louvain algorithm

```
publication_cat <- read.csv("brain_publication_areas.csv")
publication_cat <- unique(publication_cat)

#forming the edge list for the graph
combs <- inner_join(publication_cat,publication_cat, by="eid")
combs_alt <- combs %>% filter(!area.x == area.y)

#creatnig a graph from the edge list
g <- graph_from_data_frame(d = combs %>% select(area.x,area.y), directed =↩
    FALSE)

#clustering
clusters <- cluster_louvain(g)

area_to_cluster_map <- cbind(V(g)$name,clusters$membership)
colnames(area_to_cluster_map) <- c("area", "cluster")
area_to_cluster_map <- as.data.frame(area_to_cluster_map)
allclusters <- area_to_cluster_map
```

```
area_to_freq_map <- as.data.frame(table(publication_cat$area))
colnames(area_to_freq_map) <- c("area", "freq")

# selecting top 5 clusters
y <- clusterdata %>% group_by(cluster) %>% add_tally(sort=TRUE) %>% select(cluster,n)
y <- unique(y)
top5 <- c(7,3,1,2,5)

merged_areas <- merge(area_to_freq_map, area_to_cluster_map, by = "area")
merged_areas <- merged_areas %>% filter(merged_areas$cluster %in% top5)
write.csv(merged_areas, "final_clusters_w_freq.csv")

# to make a word cloud -->
# you need to change these 3 variables for every cluster based on your personal choice
cluster_num = 5
min_freq = 1
max_size = 1
clrs = "black"
particular_cluster <- filter(merged_areas, cluster == cluster_num)
wordcloud(words = particular_cluster$area,
          freq = particular_cluster$freq,
          scale = c(max_size,0.5),
          min.freq = min_freq,
          max.words = Inf,
          random.order = FALSE,
          rot.per = .0,
          ordered.colors = TRUE,

          use.r.layout = FALSE)
```

Out of 330 topics and 11 clusters, the top 5 clusters we get are:

1. Neurology
2. Electrical and Electronic Engineering
3. Environmental Chemistry
4. Immunology
5. Cognitive Neuroscience

We encountered difficulty in making the graph for the clustering algorithm and we noticed that a main cluster was getting split into two separate clusters which was the biochemistry cluster.

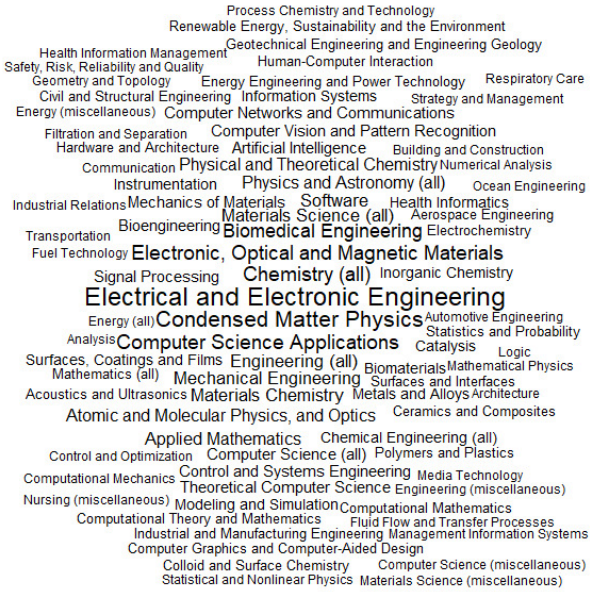

Figure 3: Word Cloud 1

Figure 4: Word Cloud 2

Figure 5: Word Cloud 3

Figure 6: Word Cloud 4

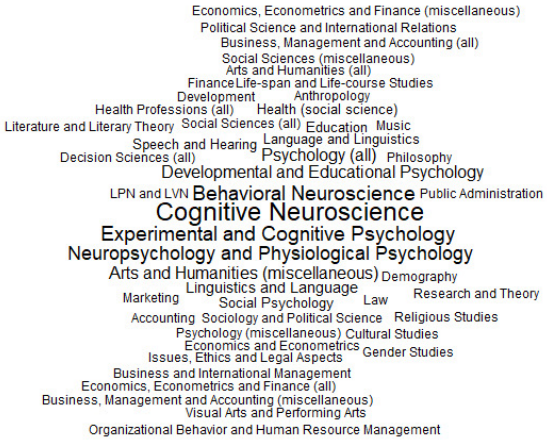

Figure 7: Word Cloud 5

Analysis 4

Generate a bar plot of the number of publications per subject area per continent pre-2014, post-2014, and their difference.

Listing 4: Generating bar plot of the number of publications per subject

```
clusterdata <- read.csv("final_clusters_w_freq.csv")
author <- read.csv("brain_author.csv")
pub_details <- read.csv("brain_publication_details.csv")
publication_cat <- read.csv("brain_publication_areas.csv")

combined_frame_1 <- merge(author, pub_details, by="scopus_id")
combined_frame_2 <- merge(combined_frame_1, publication_cat, by.x="eids",↩
    by.y="eid")
mergedData <- merge(combined_frame_2, clusterdata, by.x = "area",by.y ="↩
    area")
clusterHead <- sqldf("select area,cluster,max(Freq)from clusterdata group ↩
    by cluster")
clusterHead
finalDataset <- merge(mergedData,clusterHead,by="cluster")
#this is a 2.6 GB file
write.csv(finalDataset,"finalset.csv")

# Final Dataset to use after all joins

finalClusterSet <- read.csv("finalset.csv")
finalClusterSet <- finalClusterSet %>% select(eids,area.y,cip_title,↩
    scopus_id,pub_year,region,)

after_2014_q4 <- filter(finalClusterSet, pub_year > 2014)
before_2014_q4 <- filter(finalClusterSet, pub_year <= 2014)


#filtering values before and after 2014 and grouping based on subject area↩
     and region
values_after_2014 <- after_2014_q4 %>% group_by(after_2014_q4$area.y, ↩
    region) %>% filter(!is.na(region))  %>% add_tally()

values_before_2014 <- before_2014_q4 %>% group_by(before_2014_q4$area.y, ↩
    region) %>% filter(!is.na(region)) %>% add_tally()


#plotting after 2014
g1 <- ggplot(values_after_2014, aes(x = region, y = n, fill =↩
    values_after_2014$area.y ))+
```

```
32     geom_bar(stat = "identity",position=position_dodge())+ coord_flip() + ↩
          ggtitle("Number of publications per subject area per continent post↩
          -2014") +
33     labs(x = "region", y= "publications", fill="Subject Area")
34
35  #plotting beofre 2014
36  g2 <- ggplot(values_before_2014, aes(x = region, y = n, fill =↩
       values_before_2014$area.y ))+
37     geom_bar(stat = "identity",position=position_dodge())+ coord_flip() + ↩
          ggtitle("Number of publications per subject area per continent pre↩
          -2014") +
38     labs(x = "region", y= "publications", fill="Subject Area")
39
40
41  grid.arrange(g1,g2)
```

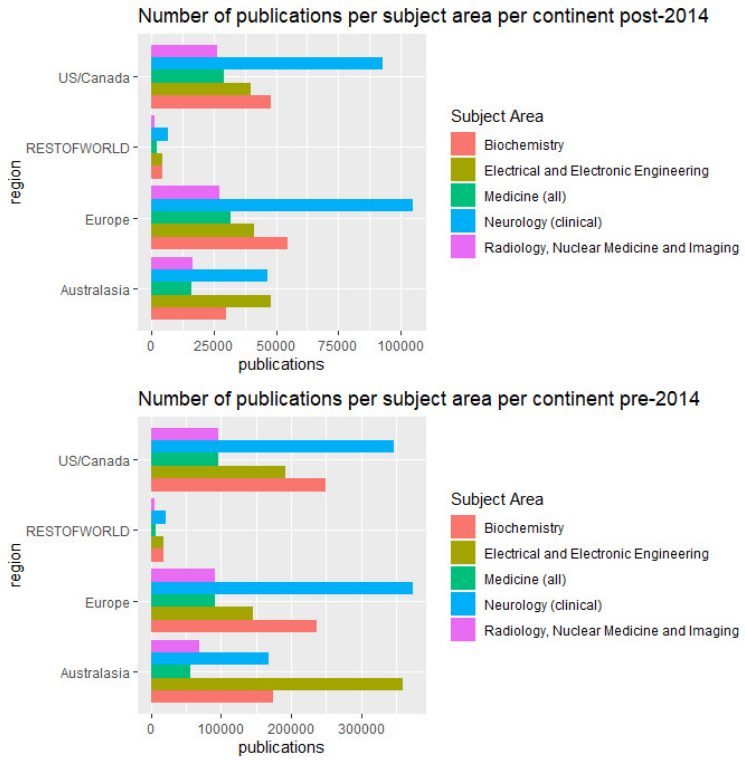

Figure 8: Number of Publications per continent pre and post 2014

1. Neurology has most publications in all the regions except for Australia in both pre-2014 and post-2014.

2. Meanwhile, Australasia has most number of publication in the subject area of Electrical and Electronic Engineering than any other region in both pre and post 2014

3. Europe has the most number of publications for a given subject area which is Neurology compared to any other continent.

4. The Rest of the World too has Neurology being the most number of publications.

Analysis 5

Generate a bar plot of the number of publications per CIP category per continent pre-2014, post-2014, and their difference.

Listing 5: Generating bar plot of the number of publications per CIP category per continent

```
author_data <- read.csv("brain_author.csv")
pub_details <- read.csv("brain_publication_details.csv")

combined_frame <- merge(author_data, pub_details, by="scopus_id")
freq_cip <- as.data.frame(table(combined_frame$cip_title))
colnames(freq_cip) <- c("Cip_Title","Freq")
freq_cip_top6 <- sqldf("select * from freq_cip order by Freq Desc limit 6")
dd <- merge(combined_frame,freq_cip_top6 , by.x = "cip_title",by.y = "Cip_Title")
new_dd <- sqldf("select * from dd where region <> 'NA' ")

after_2014 <- filter(new_dd, pub_year > 2014) %>% filter(!is.na(region)) %>% filter(!is.na(cip_title))
grouped_after_2014 <- after_2014%>% group_by(after_2014$region, after_2014$cip_title) %>% add_tally()

before_2014 <-  filter(new_dd, pub_year < 2014) %>% filter(!is.na(region)) %>% filter(!is.na(cip_title))
grouped_before_2014 <- before_2014%>% group_by(before_2014$region, before_2014$cip_title) %>% add_tally()

comb_data <- merge(grouped_after_2014,grouped_before_2014, by.x="cip_title" , by.y="region")


# plot for after 2014
g1 <- ggplot(grouped_after_2014, aes(x = region, y = n, fill = cip_title)) +
  geom_bar(stat = "identity",position=position_dodge())+ coord_flip() +
    ggtitle("Number of publications per CIP per continent post-2014") + labs(x = "Region", y = "Publications", fill ="CIP")

#plot for before 2014
g2 <- ggplot(grouped_before_2014, aes(x = region, y = n, fill = cip_title))+
```

```
27     geom_bar(stat = "identity",position=position_dodge())+ coord_flip() +
28     ggtitle("Number of publications per CIP per continent pre-2014")  + ↩
           labs(x = "Region", y = "Publications", fill ="CIP")
29
30 grid.arrange(g1,g2)
```

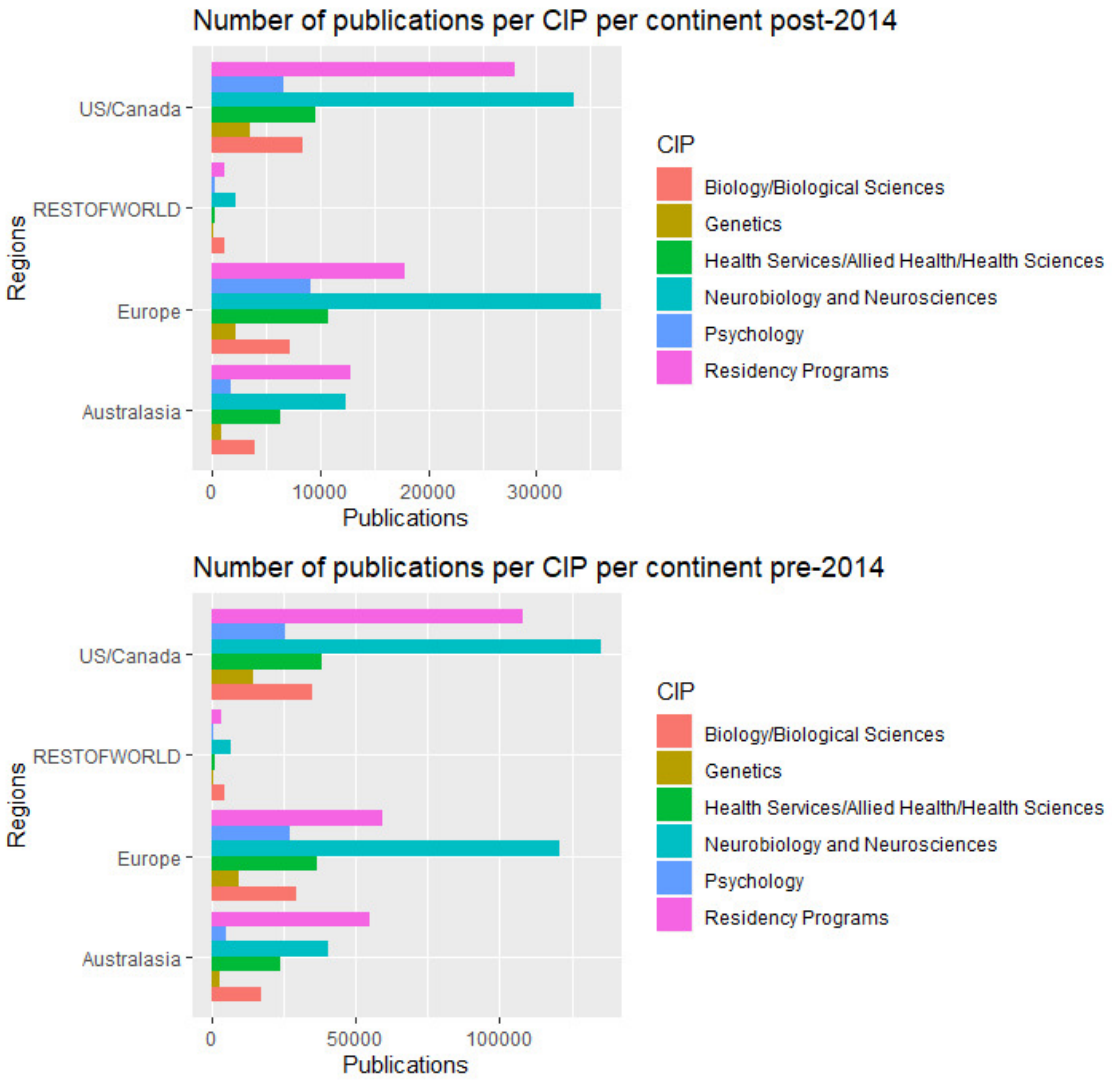

Figure 9: Number of Publications Pre and Post 2014

1. The sum of all the CIP publications in US/Canada is greater than any other continent since the year 1960.

2. Neurobiology and Neurosciences has the highest number of publications not only in Europe but all over the world pre-2014 whereas US/Canada takes the lead post-2014 in the same field overall.