BRAIN SCIENCE STUDY

Nirupam Bidikar-1878058 | Akshit Tandon-1792038 | Rahul Raj Mogili-1900425

Construction of Author Networks

We aim to visualise author networks for USA, Europe, and Australasia. In these networks, each node is a scholar and edges connecting scholars represent joint publications. The purpose of creating these networks is to compute pagerank which is a fundamental factor in evaluating an author's work in research. Below we show our approach and methodology with explanation of the code.

The Approach

For this part of the project, we import dplyr, plyr and data.table libraries. Initially, let's load the brain author csv file and factor it according to the `scopus_id`. We conduct the factorization for `scopus_id` in brainAuthor and `scopus_id`, `co_author_scopus_id` in brainPubAuthor data frames. Later, we load the CIP_category csv file given.

```
1  library(dplyr)
2  library(plyr)
3  library(data.table)
4
5  brainAuthor <- read.csv("brain_author.csv")
6  brainAuthor$scopus_id <- as.factor(brainAuthor$scopus_id)
7  brainPubAuthors <- read.csv("brain_publication_authors.csv")
8  brainPubAuthors$scopus_id <- as.factor(brainPubAuthors$scopus_id)
9  brainPubAuthors$co_author_scopus_id <- as.factor(↩
       brainPubAuthors$co_author_scopus_id)
10 #brainPubDetails <- read.csv("brain_publication_details.csv")
11 cipCatFile <- read.csv("CIP_category.csv")
12 colnames(cipCatFile) <- c("cip_title", "cip_category")
```

We create a node list by creating an inner join from the brainAuthor and CIP_category by filtering out the minimum publication year that has to be greater than 1960 and number of publications to be greater than 10, as considering less than 10 would make the author look like a PhD student.

```
nodeListInit <- distinct(inner_join(brainAuthor,cipCatFile,by="cip_title")
    ) %>% filter(min_pub_year > 1960) %>% filter(num_publications > 10)
    %>% select(scopus_id,region,min_pub_year,cip_category,
    total_deflated_dollar_2010)
```

We initiate with the USA node. We filter out the region to be "US/Canada". Since gephi requires an argument of source in the input, we use the 'rn' from the csv file and rename it to 'Source'.

```
#USA NETWORK
usaNode <- unique(nodeListInit %>% filter(region == "US/Canada"))
setDT(usaNode,keep.rownames = TRUE)[]
names(usaNode)[names(usaNode) == 'rn'] <- 'Source'
write.csv(usaNode,"usanodelist2.csv",row.names = FALSE)
```

To create an edge list 'edgeListInit', we apply an inner join for the usaNode and brainPubAuthor. Now, we select the important columns scopus_id, region, co_author_scopus_id and Source and filter out any redundancies present in the edge list. We now rename the scopus_id in the usaNode to co_author_scopus_id for using that in the inner join of the usaNode and edgeListInit. To differentiate the two columns 'Source' we use 'Source.x' for 'Target' and 'Source.y' for 'Source'. We now plug these two inputs to gephi and get a visualisation of usaNode with edges. We write the 'Source' and 'Target' columns in a csv format with page ranks. In gephi, we make use of the Giant Component filter which returns the single most highly connected component in the network which is what shown in the figures for all regions as the original network generated is too complex.

```
edgeListInit <- inner_join(usaNode,brainPubAuthors,by="scopus_id") %>%
    filter(co_author_scopus_id %in% usaNode$scopus_id) %>% filter(
    scopus_id %in% usaNode$scopus_id) %>% filter(scopus_id !=
    co_author_scopus_id)
edgeListInit <- edgeListInit %>% select(scopus_id,region,
    co_author_scopus_id,Source)
edgeListInit <- unique(edgeListInit)

names(usaNode)[names(usaNode) == 'scopus_id'] <- 'co_author_scopus_id'
edgeListInit <- inner_join(usaNode,edgeListInit,by="co_author_scopus_id")

names(edgeListInit)[names(edgeListInit) == 'Source.y'] <- 'Target'
names(edgeListInit)[names(edgeListInit) == 'Source.x'] <- 'Source'

```

```
usaEdgeList <- edgeListInit  %>% select(Source,Target)
write.csv(usaEdgeList,"usaedgelist2.csv",row.names = FALSE)
```

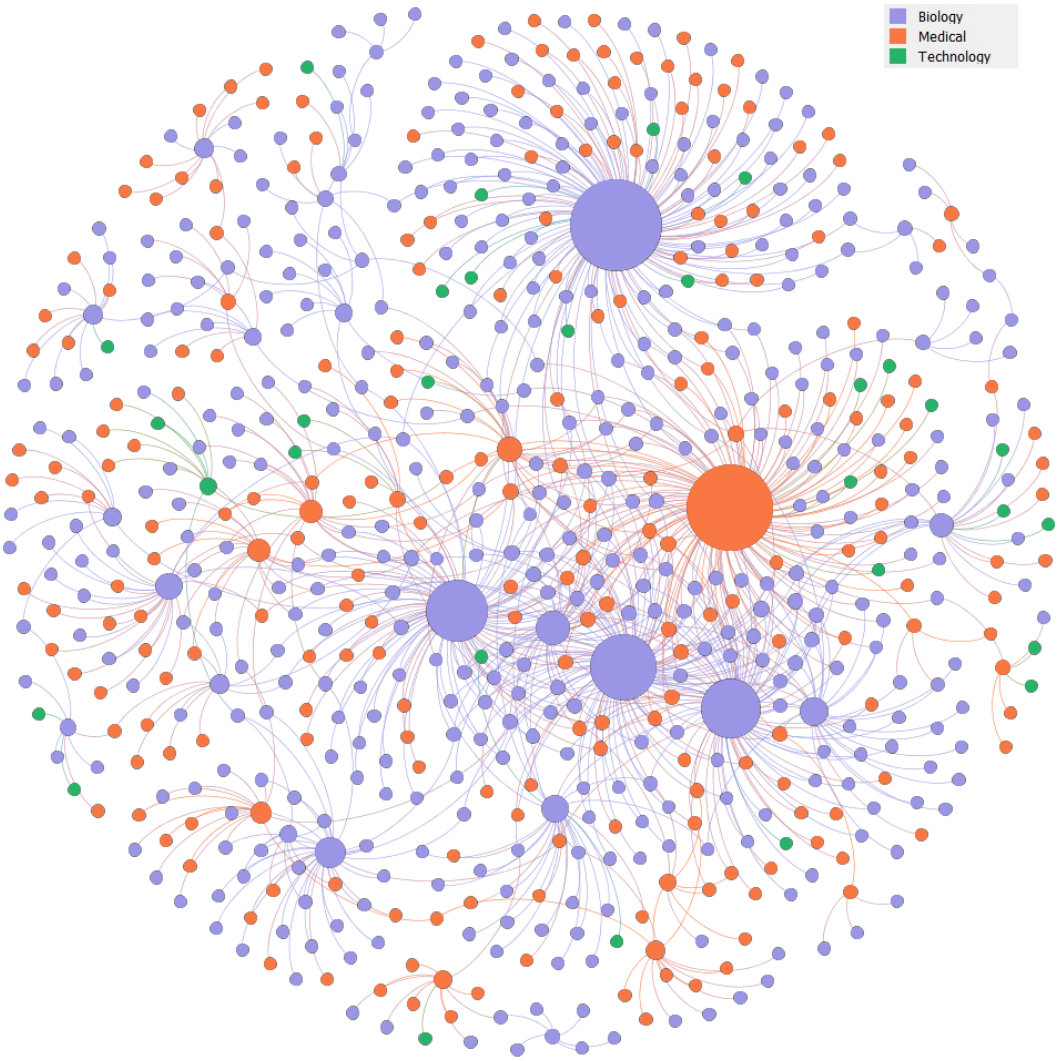

Figure 1: USA Network

With respect to the fig 1, we can observe two authors with highest number of co authors each from Biology and Medical. Apart from that, Biology has more than half number of nodes followed by Medical and the least being Technology.

We do similar implementations for Europe by replacing the region filter to 'Europe'.

```
#EUROPE NETWORK
europeNode <- nodeListInit %>% filter(region == "Europe")
setDT(europeNode,keep.rownames = TRUE)[]
names(europeNode)[names(europeNode) == 'rn'] <- 'Source'
write.csv(europeNode,"europeNodelist.csv",row.names = FALSE)
```

```
6
7  europeEdgeList <- inner_join(europeNode,brainPubAuthors,by="scopus_id") ↩
       %>% filter(co_author_scopus_id %in% europeNode$scopus_id) %>% filter(↩
       scopus_id %in% europeNode$scopus_id) %>% filter(scopus_id != ↩
       co_author_scopus_id)
8  europeEdgeList <- europeEdgeList %>% select(scopus_id,region,↩
       co_author_scopus_id,Source)
9  europeEdgeList <- unique(europeEdgeList)
10
11 names(europeNode)[names(europeNode) == 'scopus_id'] <- '↩
       co_author_scopus_id'
12 europeEdgeList <- inner_join(europeNode,europeEdgeList,by="↩
       co_author_scopus_id")
13
14 names(europeEdgeList)[names(europeEdgeList) == 'Source.y'] <- 'Target'
15 names(europeEdgeList)[names(europeEdgeList) == 'Source.x'] <- 'Source'
16 europeEdgeList <- europeEdgeList %>% select(Source,Target)
17 write.csv(europeEdgeList,"europeEdgeList.csv",row.names = F)
```

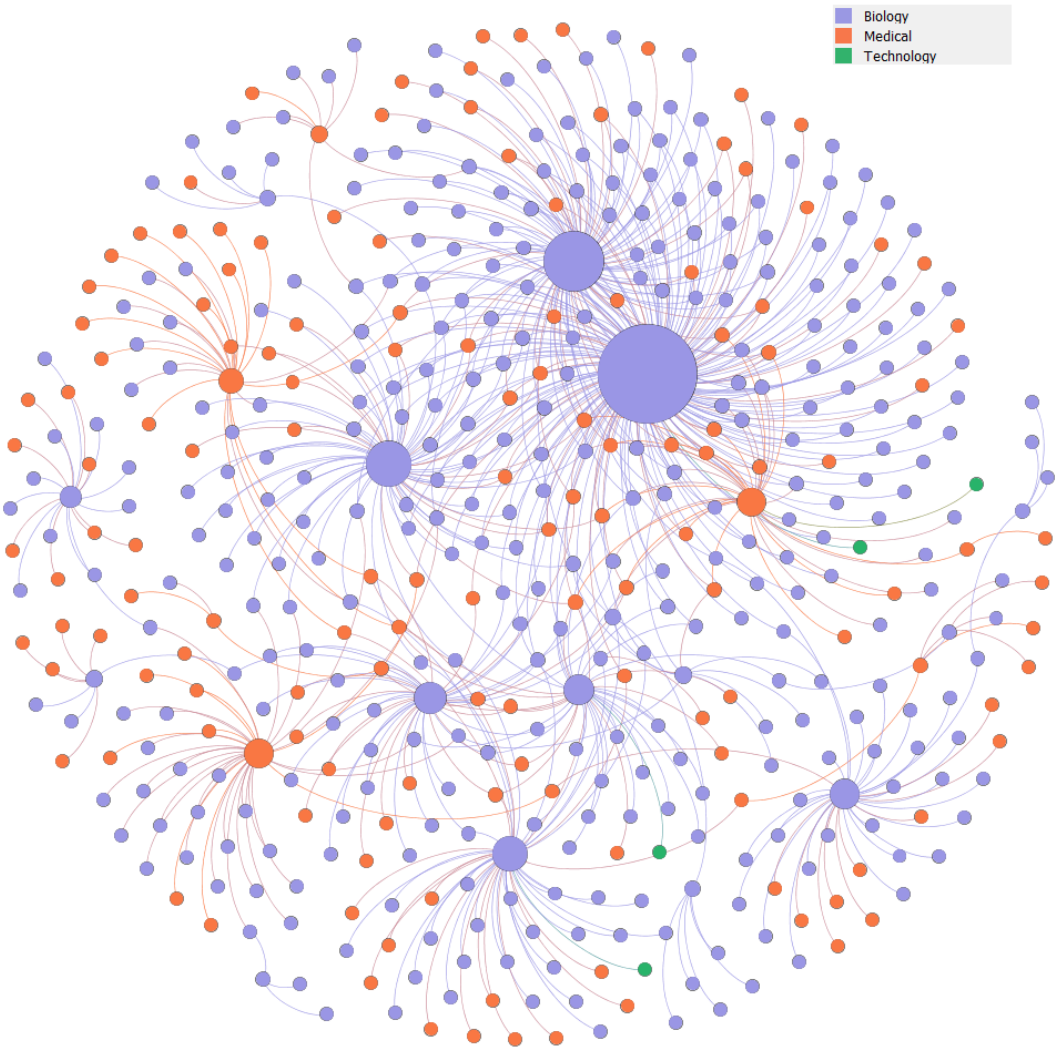

Figure 2: EUROPE Network

From fig 2, we can observe that, Biology unanimously outnumbered Medical and Technology in terms of networking. Technology department has very few edges.

We do similar implementations for Australasia by replacing the region filter to 'Australasia'.

```
#AUSTRALASIA NETWORK
australasiaNodeList <- nodeListInit %>% filter(region == "Australasia")
setDT(australasiaNodeList,keep.rownames = TRUE)[]
names(australasiaNodeList)[names(australasiaNodeList) == 'rn'] <- 'Source'
write.csv(australasiaNodeList,"australasiaNodelist.csv",row.names = FALSE)

australasiaEdgeList <- inner_join(australasiaNodeList,brainPubAuthors,by="↩
    scopus_id") %>% filter(co_author_scopus_id %in% ↩
    australasiaNodeList$scopus_id) %>% filter(scopus_id %in% ↩
```

```
    australasiaNodeList$scopus_id) %>% filter(scopus_id != ↩
    co_author_scopus_id)
australasiaEdgeList <- australasiaEdgeList %>% select(scopus_id,region,↩
    co_author_scopus_id,Source)
australasiaEdgeList <- unique(australasiaEdgeList)

names(australasiaNodeList)[names(australasiaNodeList) == 'scopus_id'] <- '↩
    co_author_scopus_id'
australasiaEdgeList <- inner_join(australasiaNodeList,australasiaEdgeList,↩
    by="co_author_scopus_id")

names(australasiaEdgeList)[names(australasiaEdgeList) == 'Source.y'] <- '↩
    Target'
names(australasiaEdgeList)[names(australasiaEdgeList) == 'Source.x'] <- '↩
    Source'
australasiaEdgeList <- australasiaEdgeList %>% select(Source,Target)
write.csv(australasiaEdgeList,"australasiaEdgeList.csv",row.names = F)
```

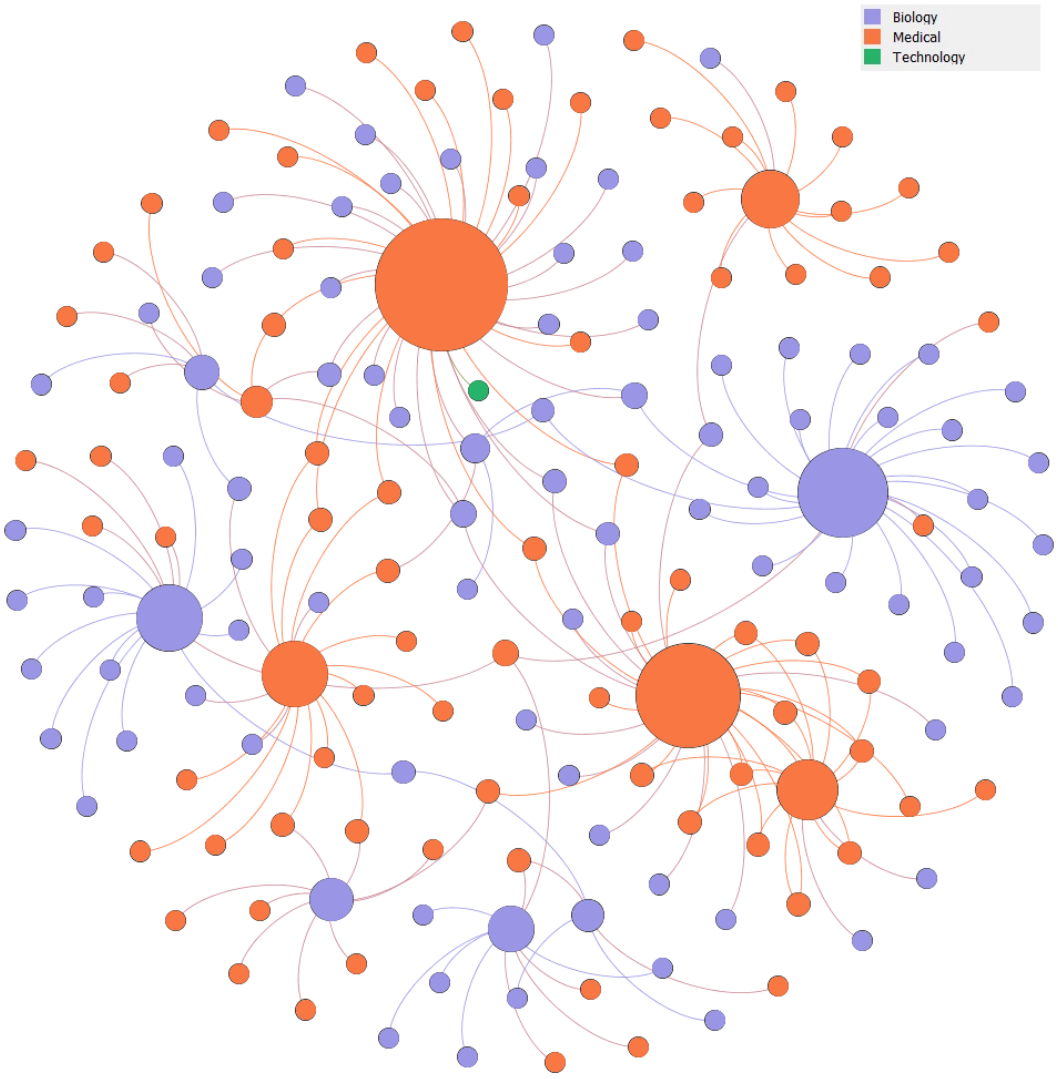

Figure 3: AUSTRALASIA Network

From fig 3, we can observe that Biology and Medical have almost same number of edges. Medical has bigger nodes compared to Biology. Even in this region, Technology has least number of edges.

With the observations from 1, 2 and 2, we can say that USA has the most number of authors followed by Europe and Australasia region being the least. We can also see from the networks that researchers mostly collaborate with other researches from the same field.

Fitting Linear Models

Here, we aim to develop 3 multiple regressions models, one for each geographic area - USA, Europe, and Australasia which help us predict the success of a researcher based on several attributes. The key predictor variable is cross-disciplinarity and the response variable being total citations. Our primary idea is to find the relationship between cross disciplinarity and success.

We also investigate if cross disciplinary research translates to greater success. We use the following base equation to construct the models and choose varying attributes for different regions due to unavailability of data.

$$log(citations) = log(1 + deptRank) + log(1 + funding) + log(1 + pagerank) + cipCategory + AgeVariation$$

We also use a stricter model with an author being considered if he has 4 or more links with authors from different fields. The construction and subsequent analysis is elaborated below.

Developing Linear Models

For this part of analysis, we import dplyr, plyr, data.table and GGally libraries. Initially let's start with the USA model. We generate the page rank file from gephi and load it along with the USA node csv file and do an inner_join between the page rank and the node file with respect to scopus_id.

```
1 library(dplyr)
2 library(plyr)
3 library(data.table)
4 library("GGally")
5
6 #USA MODEL
7 pagerank <- read.csv("usapagerank2.csv")
8 cdfile <- read.csv("usa-nodes.csv")
9 pagerankwithcd <- inner_join(pagerank,cdfile,by="scopus_id")
```

Now, within the data frame, let us remove additional columns like Label, timeset, cip_category.x, minimum publication year and the number of funds. Factor the cip_category and y_05 columns and mutate them.

```
1 new <- within(pagerankwithcd, rm(Label,timeset,cip_category.x,min_pub_year
    ,X,num_of_fund))
2 new$cip_category <- as.factor(new$cip_category.y)
3 new$y_05 <- as.factor(new$y_05)
4 new <- new %>% mutate(y_05 = relevel(y_05, ref = "1960"))
```

We set 'NA' values in the data frame to 0 for funding and ignore the rows if their department rank is null. Now, let us run a linear model on the citations, cross disciplinarity, dept_rank, total_deflated_dollar_2010, pageranks, cip_category, y_05 corresponding with the 'new' data frame. Summarize the results. Finally, set up a matrix of plots using 'ggpairs'.

From fig 4, we can infer the following interpretations. We observe the total_deflated_dollar_2010, citations and pageranks to be exponential. We also see that Biology has most number of publications followed by Medical and Technology. In spite of Medical having less number of publications compared to Biology, it's dept_rank is more significant compared to the other two departments.

```
#usa model
new$y_05 <- as.factor(new$y_05)
new <- new %>% mutate(y_05 = relevel(y_05, ref = "2015"))
new <- new %>% filter(!is.na(dept_rank))
new$total_deflated_dollar_2010[is.na(new$total_deflated_dollar_2010)] = 0
usamodel <- lm(log(1 +citations) ~  factor(CD) + log(1+dept_rank) +log(1+total_deflated_dollar_2010) + log(1+pageranks)+  cip_category + y_05 ,data=new)
summary(usamodel)

graphframe <- new %>% select (total_deflated_dollar_2010, pageranks, citations, cip_category, dept_rank) %>% filter(!is.na(dept_rank))
ggpairs(graphframe,aes(col = cip_category, alpha=0.4))

#usa  strict model

new$cip_category <- as.factor(new$cip_category.y)
new <- new %>% mutate(cip_category = relevel(cip_category, ref = "Technology"))
new$y_05 <- as.factor(new$y_05)
new <- new %>% mutate(y_05 = relevel(y_05, ref = "2015"))
new <- new %>% filter(!is.na(dept_rank))
new$total_deflated_dollar_2010[is.na(new$total_deflated_dollar_2010)] = 0
usamodel <- lm(log(1 +citations) ~  factor(Stricter_CD) + log(1+dept_rank)+log(1+total_deflated_dollar_2010) + log(1+pageranks)+  cip_category + y_05 ,data=new)
summary(usamodel)
graphframe <- new %>% select (total_deflated_dollar_2010, pageranks, citations, cip_category, dept_rank) %>% filter(!is.na(dept_rank))
ggpairs(graphframe,aes(col = cip_category, alpha=0.4))
```

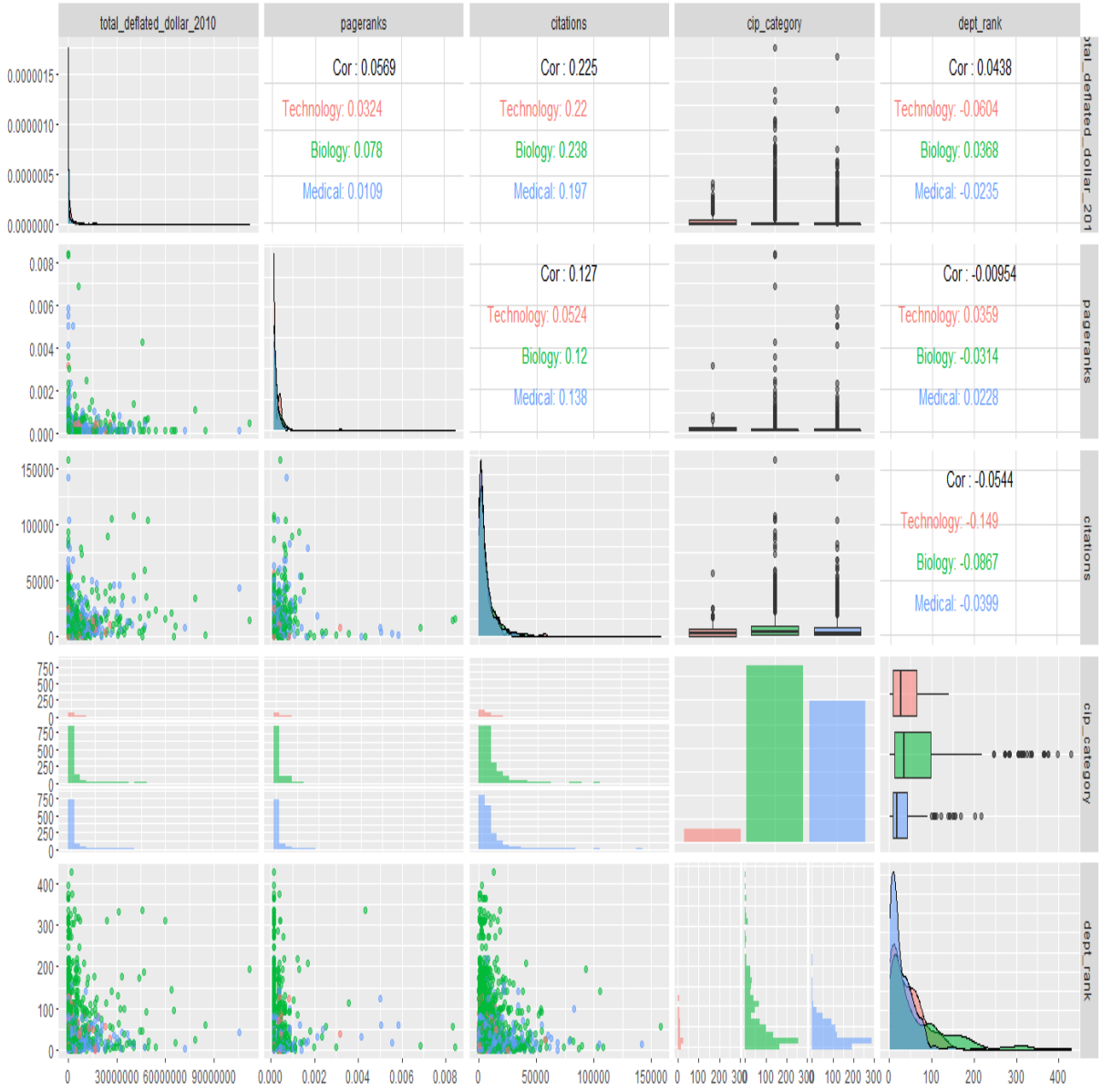

Figure 4: USA CORRELATION

We also implemented a stricter cross disciplinary model for the USA region, as shown in fig 5. We observe that there are minute changes in the coefficients with the estimates increasing in the stricter cross disciplinary model.

```
Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                       5.494614   0.875234   6.278 4.16e-10 ***
factor(CD)1                       0.576694   0.069221   8.331  < 2e-16 ***
log(1 + dept_rank)               -0.108681   0.016027  -6.781 1.55e-11 ***
log(1 + total_deflated_dollar_2010)  0.015232   0.002737   5.564 2.97e-08 ***
log(1 + pageranks)              239.106755  41.674738   5.737 1.10e-08 ***
cip_categoryBiology               0.398306   0.094724   4.205 2.72e-05 ***
cip_categoryMedical               0.224540   0.095635   2.348 0.018974 *
y_051960                          3.474757   0.902927   3.848 0.000122 ***
y_051965                          3.433557   0.877846   3.911 9.47e-05 ***
y_051970                          3.364749   0.872482   3.857 0.000118 ***
y_051975                          3.129403   0.870637   3.594 0.000333 ***
y_051980                          2.922965   0.869623   3.361 0.000790 ***
y_051985                          2.861087   0.869598   3.290 0.001018 **
y_051990                          2.503382   0.868977   2.881 0.004007 **
y_051995                          2.136561   0.868914   2.459 0.014017 *
y_052000                          1.645789   0.868710   1.895 0.058294 .
y_052005                          1.008708   0.868715   1.161 0.245714
y_052010                          0.158831   0.870472   0.182 0.855235
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 5: USA BASE MODEL

In both base model (fig 5) and the stricter model (fig 6), we see pagerank being extremely significant. We can also observe a decreasing trend in the coefficient values of the age factor which translates to the fact that authors who have recently started publishing research are less likely to be more successful. Department rank has a negative coefficient showing greater the rank lesser the significance. Biology and Medical fields show significance which indicates that an author belonging to these fields is more likely to be successful. Cross disciplinarity is another significant factor but has a lesser value than pagerank and age factor. Funding turns out to be a significant attribute but the coefficient suggests it's not a big influence on the response variable.

```
Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                       6.06372    0.85350   7.104 1.65e-12 ***
factor(Stricter_CD)1              0.54623    0.04258  12.828  < 2e-16 ***
log(1 + dept_rank)               -0.10200    0.01569  -6.500 1.00e-10 ***
log(1 + total_deflated_dollar_2010)  0.01633    0.00268   6.093 1.32e-09 ***
log(1 + pageranks)              210.51095   40.84883   5.153 2.80e-07 ***
cip_categoryBiology               0.40749    0.09265   4.398 1.15e-05 ***
cip_categoryMedical               0.22840    0.09357   2.441 0.014734 *
y_051960                          2.99192    0.88424   3.384 0.000729 ***
y_051965                          2.93416    0.85998   3.412 0.000657 ***
y_051970                          2.87008    0.85475   3.358 0.000800 ***
y_051975                          2.66921    0.85283   3.130 0.001773 **
y_051980                          2.43069    0.85192   2.853 0.004371 **
y_051985                          2.36463    0.85193   2.776 0.005559 **
y_051990                          2.02828    0.85121   2.383 0.017269 *
y_051995                          1.66681    0.85111   1.958 0.050317 .
y_052000                          1.18572    0.85087   1.394 0.163607
y_052005                          0.55820    0.85073   0.656 0.511802
y_052010                         -0.25096    0.85235  -0.294 0.768454
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 6: USA MODEL WITH STRICT CD

We do similar implementations for Europe with Page rank csv file of Europe. Here we drop the department rank attribute due to unavailability of data.

```
#europe model

eucomb$cip_category.y <- as.factor(eucomb$cip_category.y)
eucomb$y_05 <- as.factor(eucomb$y_05)
eucomb <- eucomb %>% mutate(cip_category.y = relevel(cip_category.y, ref =↩
    "Technology"))
eucomb <- eucomb %>% mutate(y_05 = relevel(y_05, ref = "2015"))
europemodel <-  lm(log(1 +citations) ~  factor(CD) +log(1+↩
   total_deflated_dollar_2010) + log(1+pageranks)+  cip_category.y + y_05↩
    ,data=eucomb)
summary(europemodel)

graphframe <- eucomb %>% select (total_deflated_dollar_2010, pageranks, ↩
   citations, cip_category.y)
ggpairs(graphframe,aes(col = cip_category.y, alpha=0.4))
#europe stricter model

eucomb$cip_category.y <- as.factor(eucomb$cip_category.y)
eucomb$y_05 <- as.factor(eucomb$y_05)
eucomb <- eucomb %>% mutate(cip_category.y = relevel(cip_category.y, ref =↩
    "Technology"))
```

```
eucomb <- eucomb %>% mutate(y_05 = relevel(y_05, ref = "2015"))
eucomb$total_deflated_dollar_2010[is.na(eucomb$total_deflated_dollar_2010)] = 0
europemodel <-  lm(log(1 +citations) ~  factor(Stricter_CD) +log(1+total_deflated_dollar_2010) + log(1+pageranks)+  factor(cip_category.y) + factor(y_05) ,data=eucomb)
summary(europemodel)
graphframe <- eucomb %>% select (total_deflated_dollar_2010, pageranks, citations, cip_category.y)
ggpairs(graphframe,aes(col = cip_category.y, alpha=0.4))
```

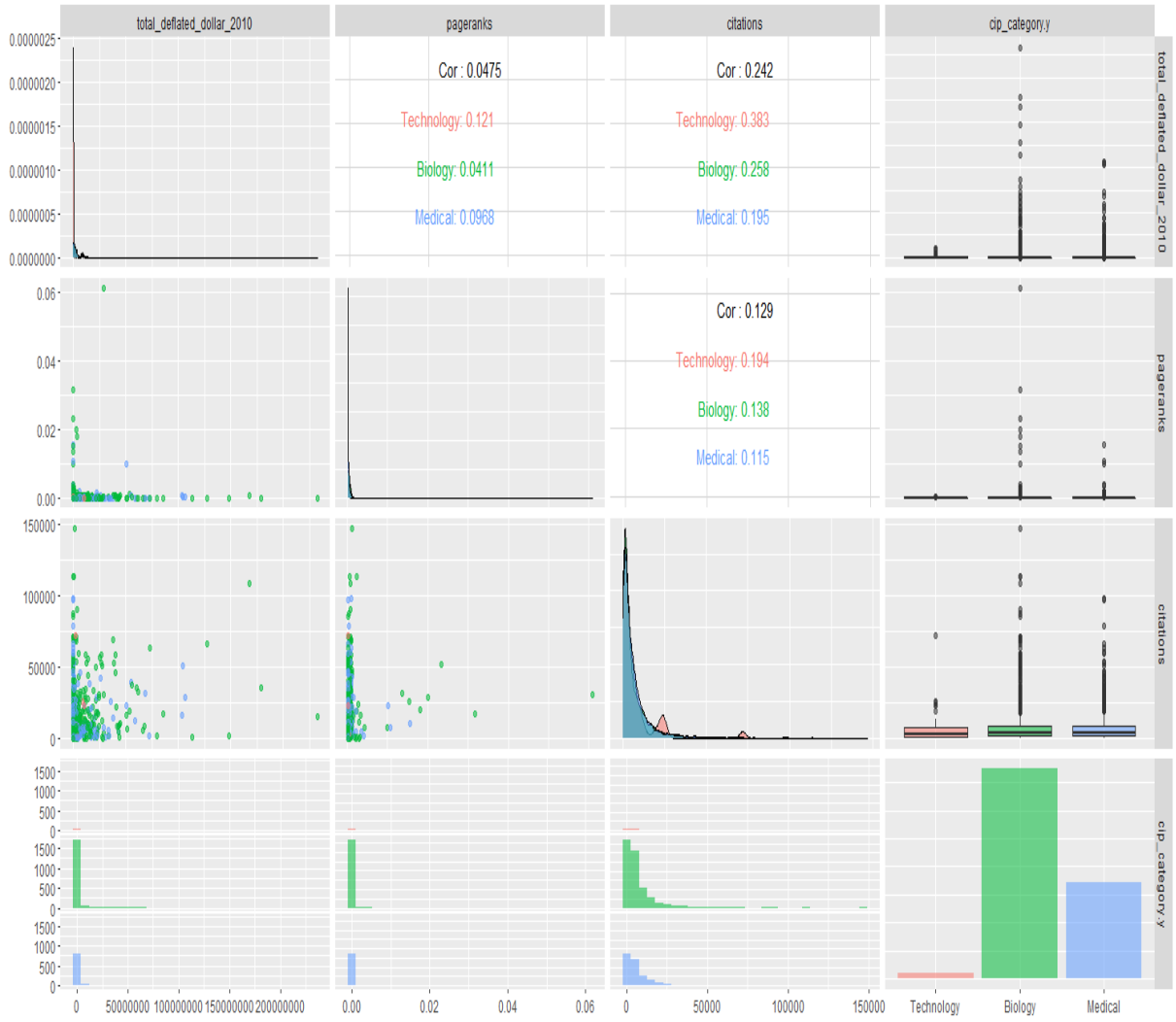

Figure 7: EUROPE CORRELATION

```
Coefficients:
                                 Estimate Std. Error t value     Pr(>|t|)
(Intercept)                      3.666398   0.878792   4.172 0.0000311126 ***
factor(CD)1                      0.541634   0.047228  11.469      < 2e-16 ***
log(1 + total_deflated_dollar_2010)  0.029433   0.002928  10.051      < 2e-16 ***
log(1 + pageranks)              59.268637  10.608454   5.587 0.0000000254 ***
cip_category.yBiology            0.343920   0.129855   2.648     0.008131 **
cip_category.yMedical            0.212887   0.131528   1.619     0.105654
y_051960                         5.024013   0.927902   5.414 0.0000000668 ***
y_051965                         4.804794   0.878552   5.469 0.0000000493 ***
y_051970                         4.743576   0.873291   5.432 0.0000000606 ***
y_051975                         4.632957   0.870881   5.320 0.0000001122 ***
y_051980                         4.532280   0.869913   5.210 0.0000002027 ***
y_051985                         4.291559   0.869256   4.937 0.0000008403 ***
y_051990                         3.981849   0.869058   4.582 0.0000048143 ***
y_051995                         3.748904   0.868861   4.315 0.0000165406 ***
y_052000                         3.273078   0.868815   3.767     0.000168 ***
y_052005                         2.611235   0.868942   3.005     0.002679 **
y_052010                         1.881962   0.870867   2.161     0.030779 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 8: EUROPE BASE MODEL

```
Coefficients:
                                 Estimate Std. Error t value    Pr(>|t|)
(Intercept)                      4.224011   0.854417   4.944 0.000000812 ***
factor(Stricter_CD)1             0.592610   0.034807  17.026     < 2e-16 ***
log(1 + total_deflated_dollar_2010)  0.031213   0.002853  10.942     < 2e-16 ***
log(1 + pageranks)              52.589815  10.342685   5.085 0.000000393 ***
factor(cip_category.y)Biology    0.393128   0.126488   3.108    0.001903 **
factor(cip_category.y)Medical    0.198024   0.128069   1.546    0.122163
factor(y_05)1960                 4.469888   0.904047   4.944 0.000000810 ***
factor(y_05)1965                 4.267682   0.856002   4.986 0.000000656 ***
factor(y_05)1970                 4.267865   0.850788   5.016 0.000000560 ***
factor(y_05)1975                 4.111904   0.848521   4.846 0.000001330 ***
factor(y_05)1980                 4.014076   0.847598   4.736 0.000002292 ***
factor(y_05)1985                 3.764374   0.846975   4.444 0.000009157 ***
factor(y_05)1990                 3.463017   0.846742   4.090 0.000044405 ***
factor(y_05)1995                 3.225299   0.846551   3.810    0.000142 ***
factor(y_05)2000                 2.794233   0.846327   3.302    0.000974 ***
factor(y_05)2005                 2.122419   0.846446   2.507    0.012218 *
factor(y_05)2010                 1.355541   0.848235   1.598    0.110141
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 9: EUROPE MODEL WITH STRICT CD

Similar to USA, we can see a decreasing trend in the coefficients as well. The page rank values for both the base model (fig 8) and strict CD model (fig 9) are highly significant. We can also observe there are no negative coefficients, like the USA model. Biology and Medical show a good amount of significance stating that authors who worked for these fields are likely to be successful. Additionally, funding has a contributing factor on the significance but has less weight on the response variable. Cross-disciplinarity is on the positive side and we can observe it increases from the base model to stricter model but, like the funding, has no high influence on the model.

We do similar implementation for Australasia with Page rank csv file of Australasia.

```
#australasia model


auscomb$y_05 <- as.factor(auscomb$y_05)
auscomb <- auscomb %>% mutate(cip_category.y = relevel(cip_category.y, ref = "Technology"))
auscomb <- auscomb %>% mutate(y_05 = relevel(y_05, ref = "2015"))
auscomb$total_deflated_dollar_2010[is.na(auscomb$total_deflated_dollar_2010)] = 0
ausmodel <- lm(log(1 +citations) ~  factor(CD)  + log(1+pageranks)+  cip_category.y + y_05 ,data=auscomb)
summary(ausmodel)
graphframe <- auscomb %>% select  (pageranks, citations, cip_category.y)
ggpairs(graphframe,aes(col = cip_category.y, alpha=0.4))

#australia strict model

auscomb$cip_category.y <- as.factor(auscomb$cip_category.y)
auscomb$y_05 <- as.factor(auscomb$y_05)
auscomb <- auscomb %>% mutate(cip_category.y = relevel(cip_category.y, ref = "Technology"))
auscomb <- auscomb %>% mutate(y_05 = relevel(y_05, ref = "2015"))
auscomb$total_deflated_dollar_2010[is.na(auscomb$total_deflated_dollar_2010)] = 0
ausmodel <- lm(log(1 +citations) ~  factor(Stricter_CD) +log(1+total_deflated_dollar_2010) + log(1+pageranks)+  factor(cip_category.y) + factor(y_05) ,data=auscomb)
summary(ausmodel)
graphframe <- auscomb %>% select  (pageranks, citations, cip_category.y)
ggpairs(graphframe,aes(col = cip_category.y, alpha=0.4))
```

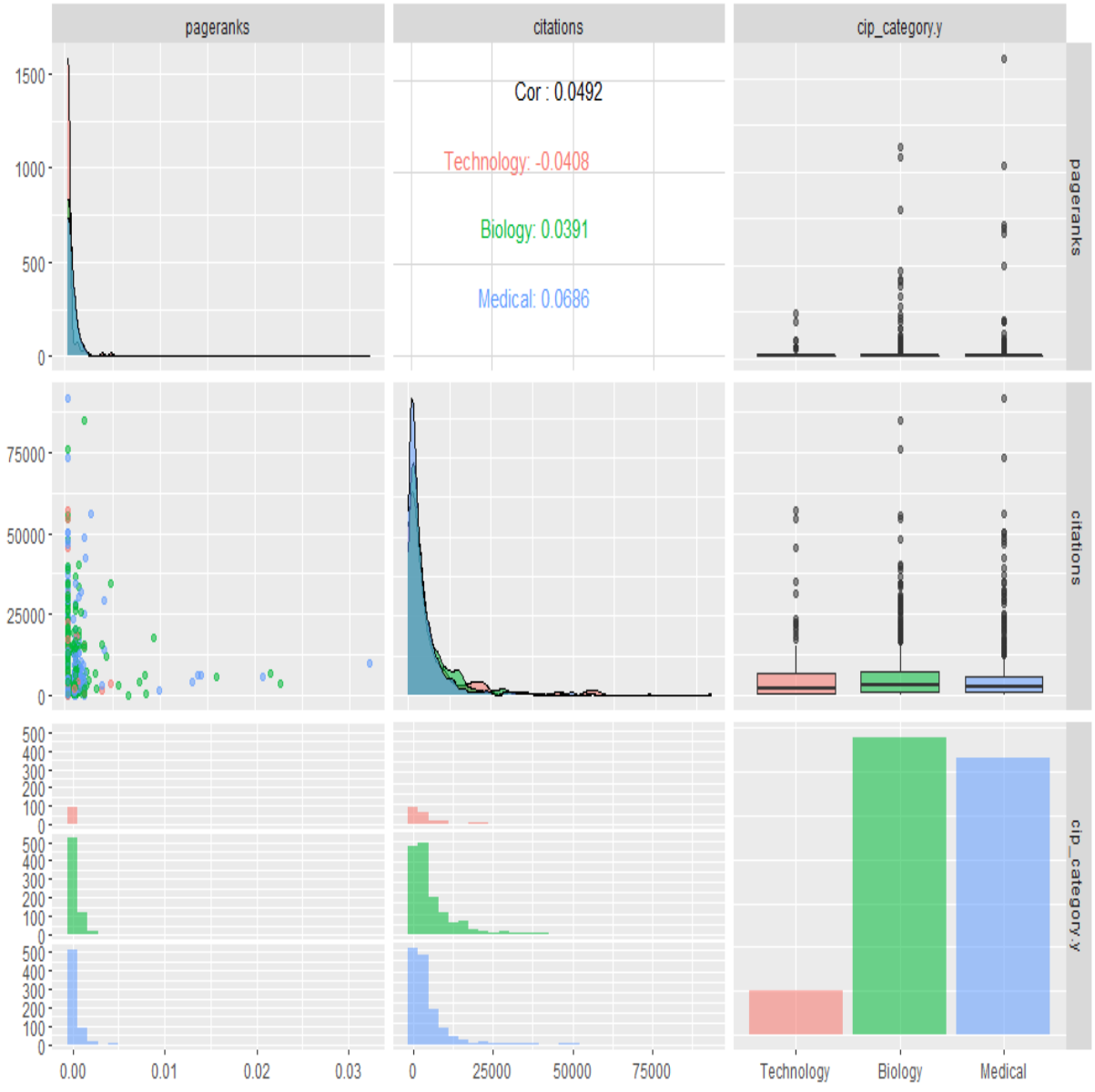

Figure 10: AUSTRALASIA CORRELATION

```
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           4.29351    0.45384   9.460  < 2e-16 ***
factor(CD)1           0.44597    0.05805   7.682 2.93e-14 ***
log(1 + pageranks)   33.90436   14.52974   2.333  0.01977 *
cip_category.yBiology 0.26753    0.09541   2.804  0.00512 **
cip_category.yMedical 0.04955    0.09559   0.518  0.60428
y_051960              5.33069    0.76919   6.930 6.40e-12 ***
y_051965              4.78386    0.55798   8.574  < 2e-16 ***
y_051970              4.35117    0.47432   9.173  < 2e-16 ***
y_051975              4.23185    0.45991   9.201  < 2e-16 ***
y_051980              4.04454    0.45324   8.924  < 2e-16 ***
y_051985              3.94853    0.45028   8.769  < 2e-16 ***
y_051990              3.55205    0.44968   7.899 5.65e-15 ***
y_051995              3.21366    0.44883   7.160 1.30e-12 ***
y_052000              2.55444    0.44822   5.699 1.47e-08 ***
y_052005              1.97231    0.44881   4.395 1.19e-05 ***
y_052010              1.49930    0.45512   3.294  0.00101 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 11: AUSTRALASIA BASE MODEL

```
Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                      4.254081   0.440730   9.652  < 2e-16 ***
factor(Stricter_CD)1             0.426248   0.046599   9.147  < 2e-16 ***
log(1 + total_deflated_dollar_2010) 0.034012   0.003961   8.586  < 2e-16 ***
log(1 + pageranks)              29.933687  14.122161   2.120 0.034214 *
factor(cip_category.y)Biology    0.292223   0.092686   3.153 0.001651 **
factor(cip_category.y)Medical    0.088139   0.092969   0.948 0.343269
factor(y_05)1960                 5.074180   0.747329   6.790 1.66e-11 ***
factor(y_05)1965                 4.848679   0.541343   8.957  < 2e-16 ***
factor(y_05)1970                 4.339899   0.460340   9.428  < 2e-16 ***
factor(y_05)1975                 4.265628   0.445969   9.565  < 2e-16 ***
factor(y_05)1980                 4.066182   0.439400   9.254  < 2e-16 ***
factor(y_05)1985                 3.950620   0.436680   9.047  < 2e-16 ***
factor(y_05)1990                 3.574830   0.436224   8.195 5.62e-16 ***
factor(y_05)1995                 3.260047   0.435427   7.487 1.24e-13 ***
factor(y_05)2000                 2.620077   0.434905   6.024 2.17e-09 ***
factor(y_05)2005                 2.061754   0.435466   4.735 2.42e-06 ***
factor(y_05)2010                 1.597919   0.441707   3.618 0.000308 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 12: AUSTRALASIA MODEL WITH STRICT CD

For the Australasia region, we notice that the pagerank is less significant when compared to US, Europe. We again see a decreasing trend in coefficient values as the year values increase which coincides with our findings from the previous models. Medical field turns out to be insignificant in this region and cross disciplinarity continues to have a positive impact.

Conclusion

After looking through six different models - two for each region, we have seen that pagerank had the most significant impact on the success factor. We have seen commonalities in models with the trends in years, cross disciplinarity and, importance of Biology field.
In the stricter model, we observe a slight reduction in the coefficient values for cross discipline with the exception of Europe, but the residual errors value have dropped compared to the base models telling us that the stricter model is a slightly better fit for the data. Cross disciplinary research produces greater technological advancements and accelerates the growth of all the fields involved in the research. This will lead us to believe that it has to be an extremely significant factor in the success of a researcher. From our models we find that it is a significant attribute but is overshadowed by pagerank and the amount of years the author has spent in research. Given this is a multi linear regression model, there is interplay between many factors which may be causing cross disciplinarity to have a lesser significance.
Secondly, we see a decreasing trend in Age i.e. the estimate value decreases as the year increase starting from 1960 to 2015. Hence, it can be concluded that researchers who have recently started publishing are likely to be less successful than researchers who started back in the 1960's.