Performance of linking graduates

Flavio & Christoph & Mona

23 July, 2023

Contents

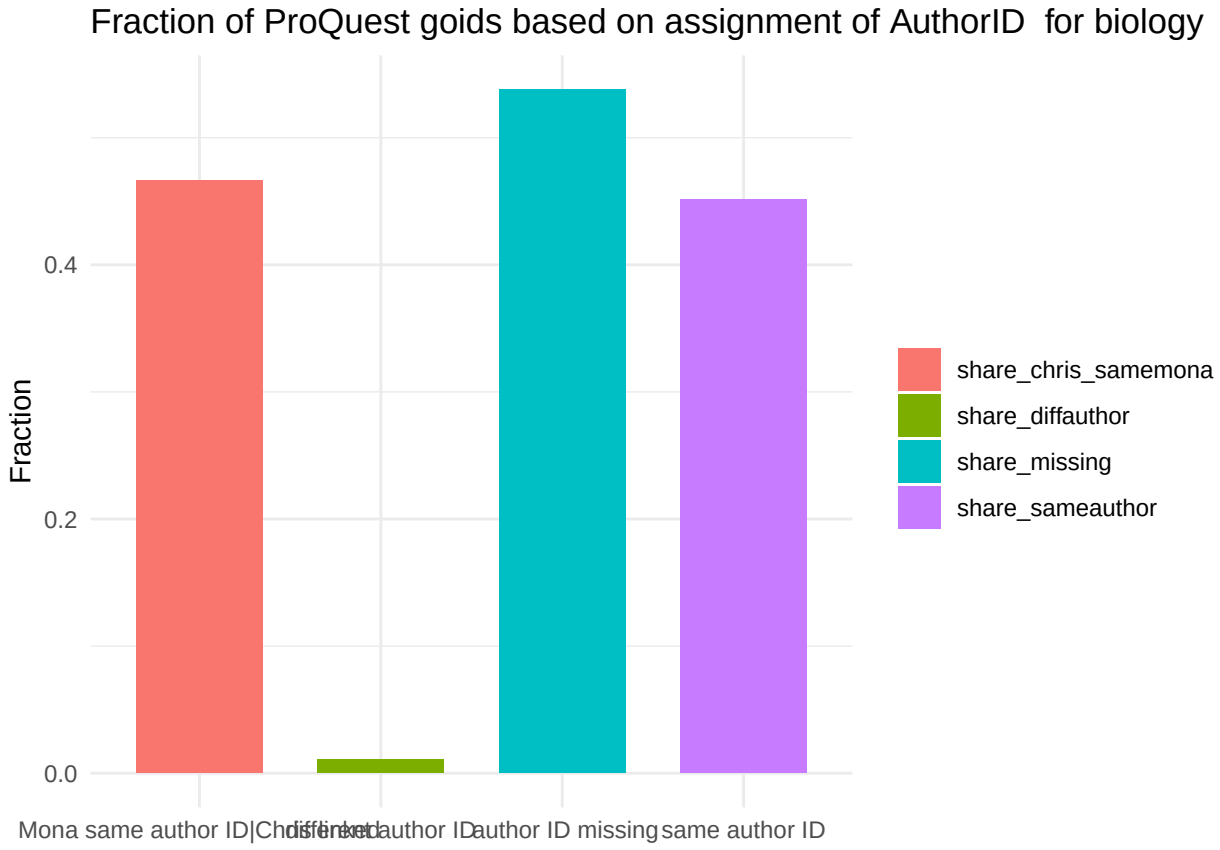
Now do the same just for physics but for christoph, mona and flavio 17

```
 # Function to process the data for a specific field

process_data <- function(field) {

  # Read the data for the specified field

  if (field %in% c("history")) {
  links_graduates_mona <- read.csv(paste0(datapath,"links_graduates_", field, "_mona_degree0_19852015.cs
    filter(link_score>0.7) %>%
    rename(authorid_mona = AuthorId) %>%
    rename(linkscore_mona=link_score)
  } else {
    links_graduates_mona <- read.csv(paste0(datapath,"/links_graduates_", field, "_mona_degree0_1985201
    filter(link_score>0.7) %>%
    rename(authorid_mona = grantid_authorposition) %>%
    rename(goid = AuthorId) %>%
    rename(linkscore_mona=link_score)
}

  if (field %in% c("biology", "computer science","economics", "engineering", "environmental science", "g
  links_graduates_christoph <- read.csv(paste0(datapath,"links_graduates_", field, "_christoph_fielddegr
    filter(link_score>0.7) %>%
    rename(authorid_christoph = AuthorId) %>%
    rename(linkscore_christoph=link_score)
  } else {
  links_graduates_christoph <- read.csv(paste0(datapath,"links_graduates_", field, "_christoph_degree0_1
    filter(link_score>0.7) %>%
    rename(authorid_christoph = AuthorId) %>%
    rename(linkscore_christoph=link_score)
  }


links_graduates_mona <- collect(links_graduates_mona)
links_graduates_christoph <- collect(links_graduates_christoph)

# Performs the full join: bothlink=1 if same authorID assigned in both, 0 if different authorID assigne
# Then calculates the share of links found by Christoph also found by Mona (number links found by both
```

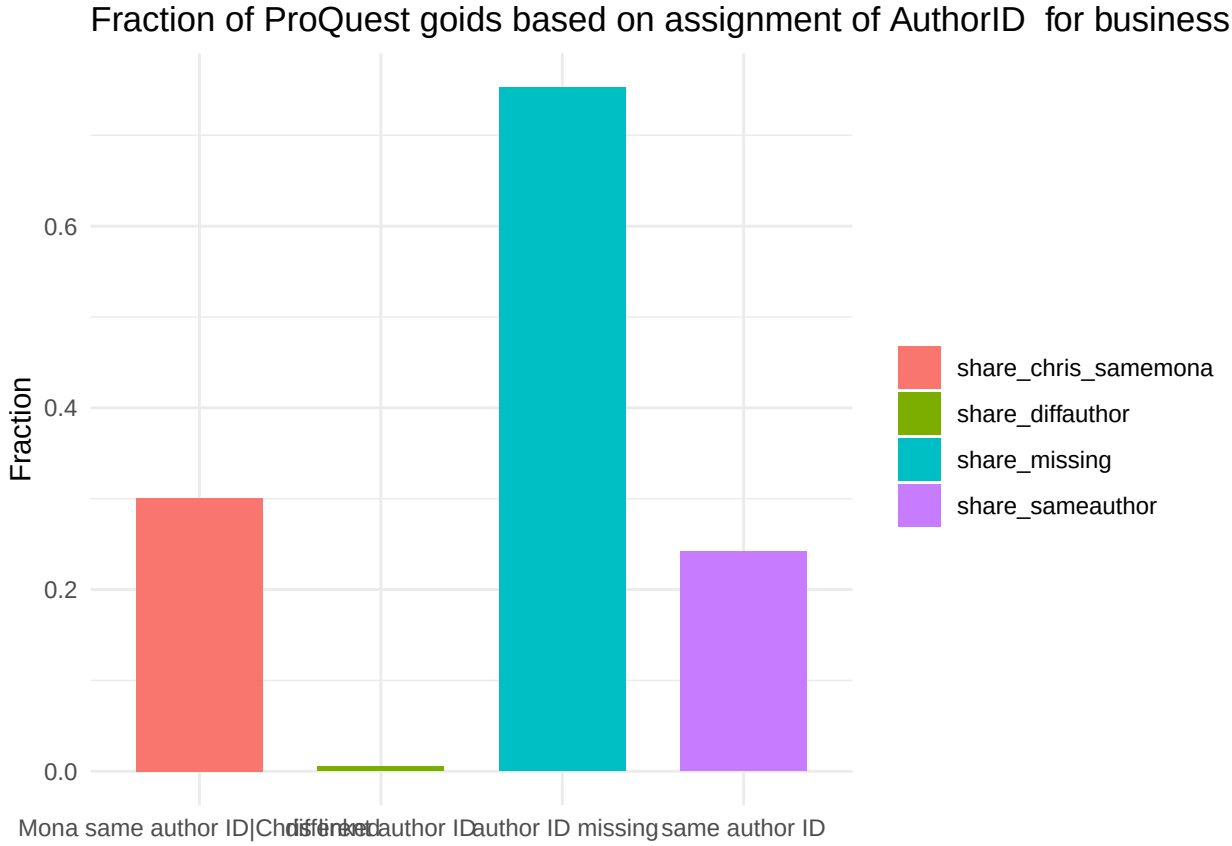
```r
links_graduates <- links_graduates_mona %>%
  full_join(links_graduates_christoph, by = c("goid")) %>%
  mutate(
    field = field,
    monalink = ifelse(!is.na(authorid_mona), 1, 0),
    chrislink = ifelse(!is.na(authorid_christoph), 1, 0),
    bothlink = ifelse(is.na(authorid_christoph) | is.na(authorid_mona),
                      NA,
                      ifelse(authorid_christoph == authorid_mona, 1, 0)),
     share_bothlink = sum(bothlink == 1 & chrislink == 1, na.rm = TRUE) / sum(chrislink == 1, na.rm = TR
  )

# Look closer at link differences:
# share of ProQuest goids assigned to same AuthorId (share_sameauthor), distinct AuthorId (share_diffau


links_graduates <- links_graduates %>%
 mutate(
    share_sameauthor = sum(bothlink == 1, na.rm = TRUE) / n_distinct(goid),
    share_diffauthor = sum(bothlink == 0 & !is.na(bothlink), na.rm = TRUE) / n_distinct(goid),
    share_missing = sum(is.na(bothlink)) / n_distinct(goid),
    share_missing_mona = sum(is.na(bothlink) & monalink == 0) / n_distinct(goid),
    share_missing_chris = sum(is.na(bothlink) & chrislink == 0) / n_distinct(goid),
    share_chris_samemona = sum(chrislink & bothlink==1, na.rm=TRUE) / sum(chrislink, na.rm=TRUE)
 )

# Create table with the shares by field
# Problem: not shown in pdf, ugly table here

shares_table  <- links_graduates %>%
  select(field, share_bothlink, share_sameauthor, share_diffauthor, share_missing, share_chris_samemona
  group_by(field) %>%
  summarize(
    share_bothlink = mean(share_bothlink, na.rm = TRUE),
    share_sameauthor = mean(share_sameauthor, na.rm = TRUE),
    share_diffauthor = mean(share_diffauthor, na.rm = TRUE),
    share_missing = mean(share_missing, na.rm = TRUE),
    share_chris_samemona = mean(share_chris_samemona, na.rm = TRUE)
  )

# Select the shares for the bar chart: total number of goids as base

shares_data <- links_graduates %>%
  summarise(
    share_sameauthor = mean(share_sameauthor, na.rm = TRUE),
    share_diffauthor = mean(share_diffauthor, na.rm = TRUE),
    share_missing = mean(share_missing, na.rm = TRUE),
    share_chris_samemona = mean(share_chris_samemona, na.rm = TRUE)
  ) %>%
  gather(variable, value)

# Create the bar chart
bar_chart <- ggplot(shares_data, aes(x = variable, y = value, fill = variable)) +
```

```
  geom_bar(stat = "identity", width = 0.7) +
  theme_minimal() +
  labs(
    x = NULL,
    y = "Fraction",
    title = paste("Fraction of ProQuest goids based on assignment of AuthorID  for",field),
    fill= NULL
  ) +
  scale_x_discrete(labels = c("Mona same author ID|Chris linked","different author ID", " author ID miss


# Print the bar chart
print(bar_chart)

# Print the summary table

shares_table %>%
  kable(format = "html",
        align = c("l", "c", "c", "c", "c", "c"),
        digits = 2,
        caption = "Share Statistics by Field",
        booktabs = TRUE)
}
```

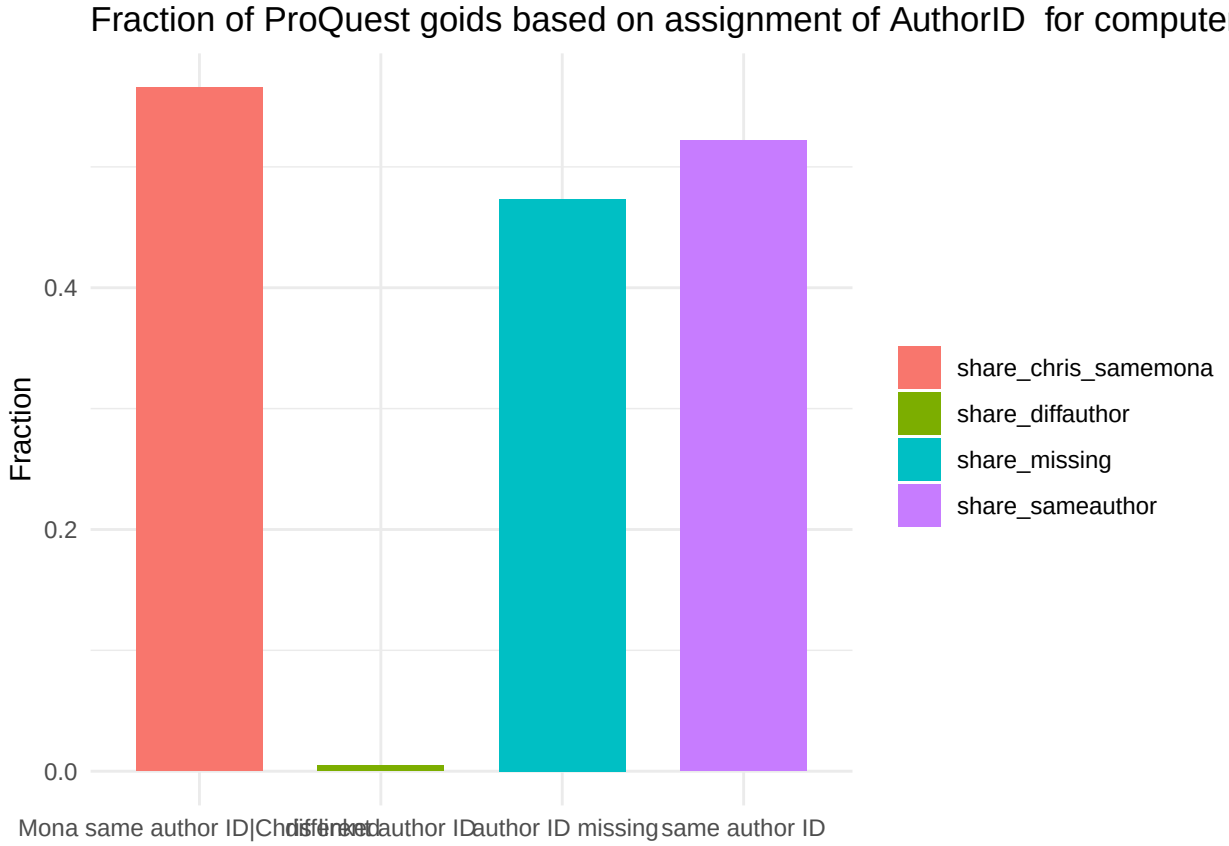
```
# Fields to process
# missing fields: ""art, "chemistry", "geography", "history", "mathematics","medicine", "sociology"


fields_to_process <- c("biology", "business", "computer science", "economics", "engineering", "environme

# Loop through the fields

for (field in fields_to_process) {
  table <- process_data(field)
  print(table)
  cat("\n\n")
}
```



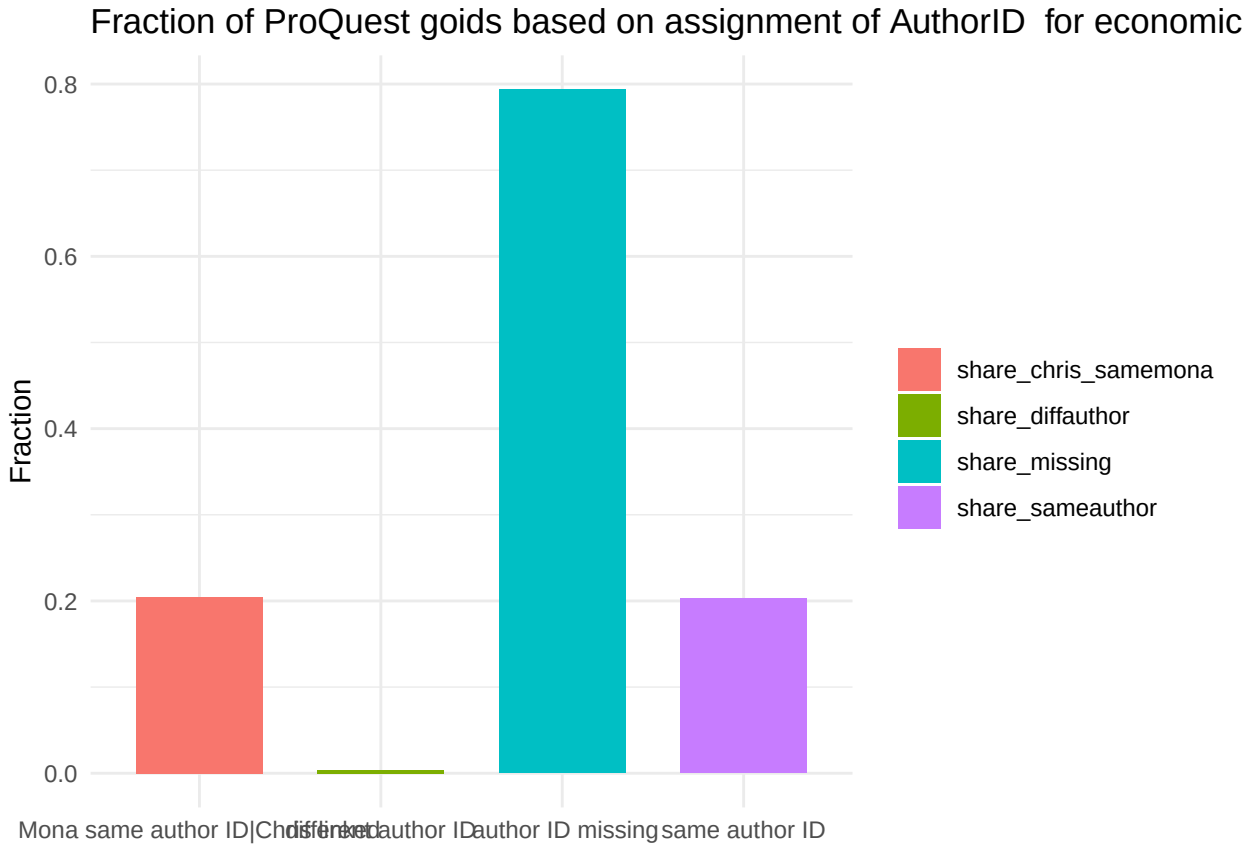

```
## <table>
## <caption>Share Statistics by Field</caption>
##  <thead>
##   <tr>
##    <th style="text-align:left;"> field </th>
##    <th style="text-align:center;"> share_bothlink </th>
##    <th style="text-align:center;"> share_sameauthor </th>
##    <th style="text-align:center;"> share_diffauthor </th>
##    <th style="text-align:center;"> share_missing </th>
##    <th style="text-align:center;"> share_chris_samemona </th>
##   </tr>
##  </thead>
## <tbody>
##   <tr>
##    <td style="text-align:left;"> biology </td>
##    <td style="text-align:center;"> 0.47 </td>
##    <td style="text-align:center;"> 0.45 </td>
##    <td style="text-align:center;"> 0.01 </td>
##    <td style="text-align:center;"> 0.54 </td>
##    <td style="text-align:center;"> 0.47 </td>
##   </tr>
## </tbody>
## </table>
```



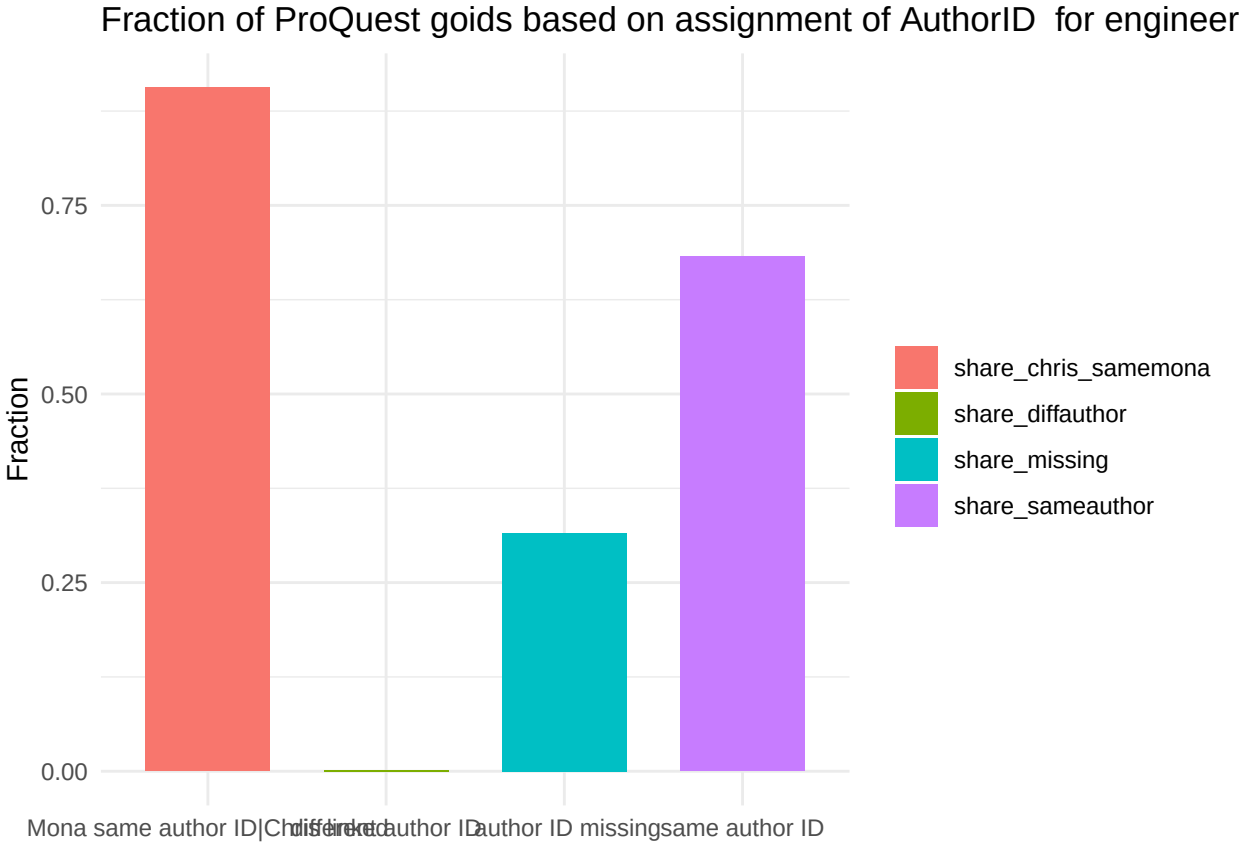

```
## <table>
## <caption>Share Statistics by Field</caption>
##  <thead>
##   <tr>
##    <th style="text-align:left;"> field </th>
##    <th style="text-align:center;"> share_bothlink </th>
##    <th style="text-align:center;"> share_sameauthor </th>
##    <th style="text-align:center;"> share_diffauthor </th>
##    <th style="text-align:center;"> share_missing </th>
##    <th style="text-align:center;"> share_chris_samemona </th>
##   </tr>
##  </thead>
## <tbody>
##   <tr>
##    <td style="text-align:left;"> business </td>
##    <td style="text-align:center;"> 0.3 </td>
##    <td style="text-align:center;"> 0.24 </td>
##    <td style="text-align:center;"> 0.01 </td>
##    <td style="text-align:center;"> 0.75 </td>
##    <td style="text-align:center;"> 0.3 </td>
##   </tr>
## </tbody>
## </table>
```



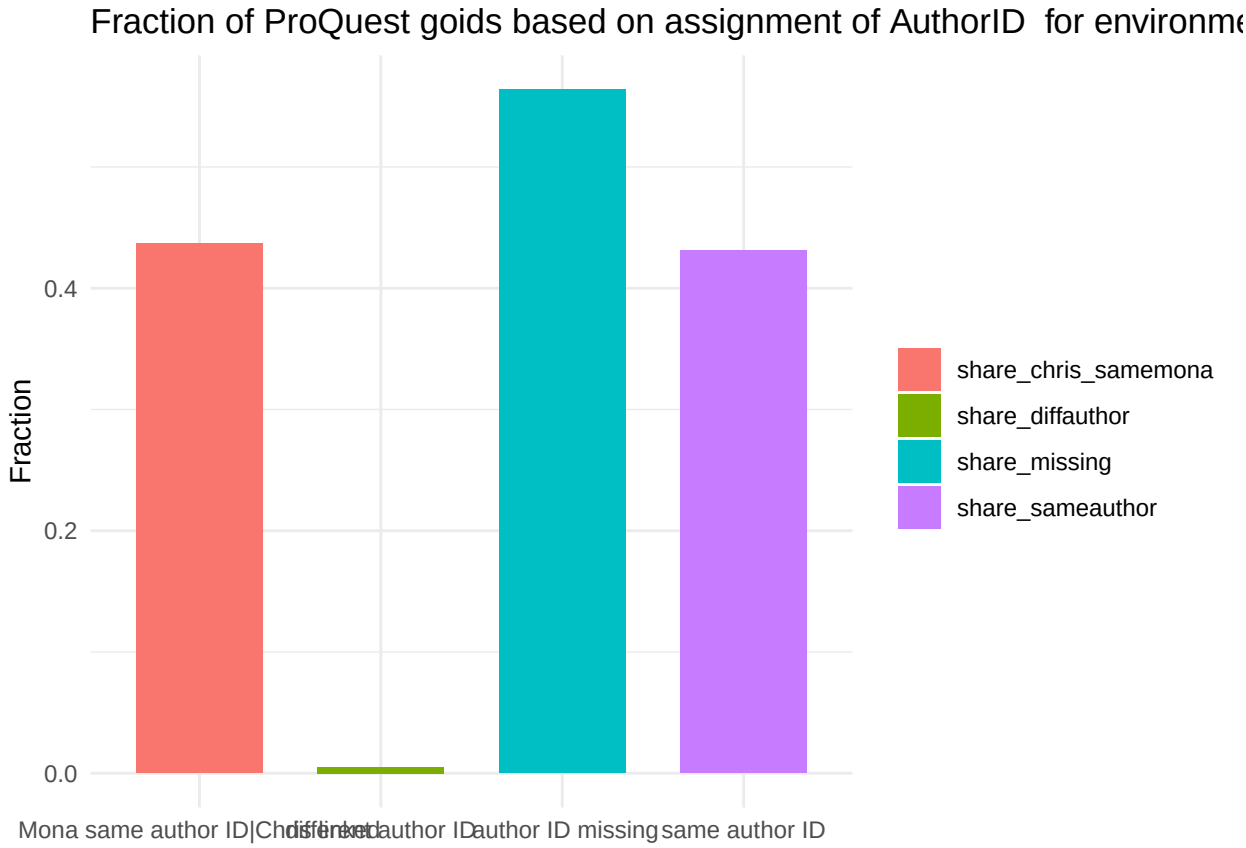

```
## <table>
## <caption>Share Statistics by Field</caption>
##  <thead>
##   <tr>
##    <th style="text-align:left;"> field </th>
##    <th style="text-align:center;"> share_bothlink </th>
##    <th style="text-align:center;"> share_sameauthor </th>
##    <th style="text-align:center;"> share_diffauthor </th>
##    <th style="text-align:center;"> share_missing </th>
##    <th style="text-align:center;"> share_chris_samemona </th>
##   </tr>
##  </thead>
## <tbody>
##   <tr>
##    <td style="text-align:left;"> computer science </td>
##    <td style="text-align:center;"> 0.57 </td>
##    <td style="text-align:center;"> 0.52 </td>
##    <td style="text-align:center;"> 0 </td>
##    <td style="text-align:center;"> 0.47 </td>
##    <td style="text-align:center;"> 0.57 </td>
##   </tr>
## </tbody>
## </table>
```



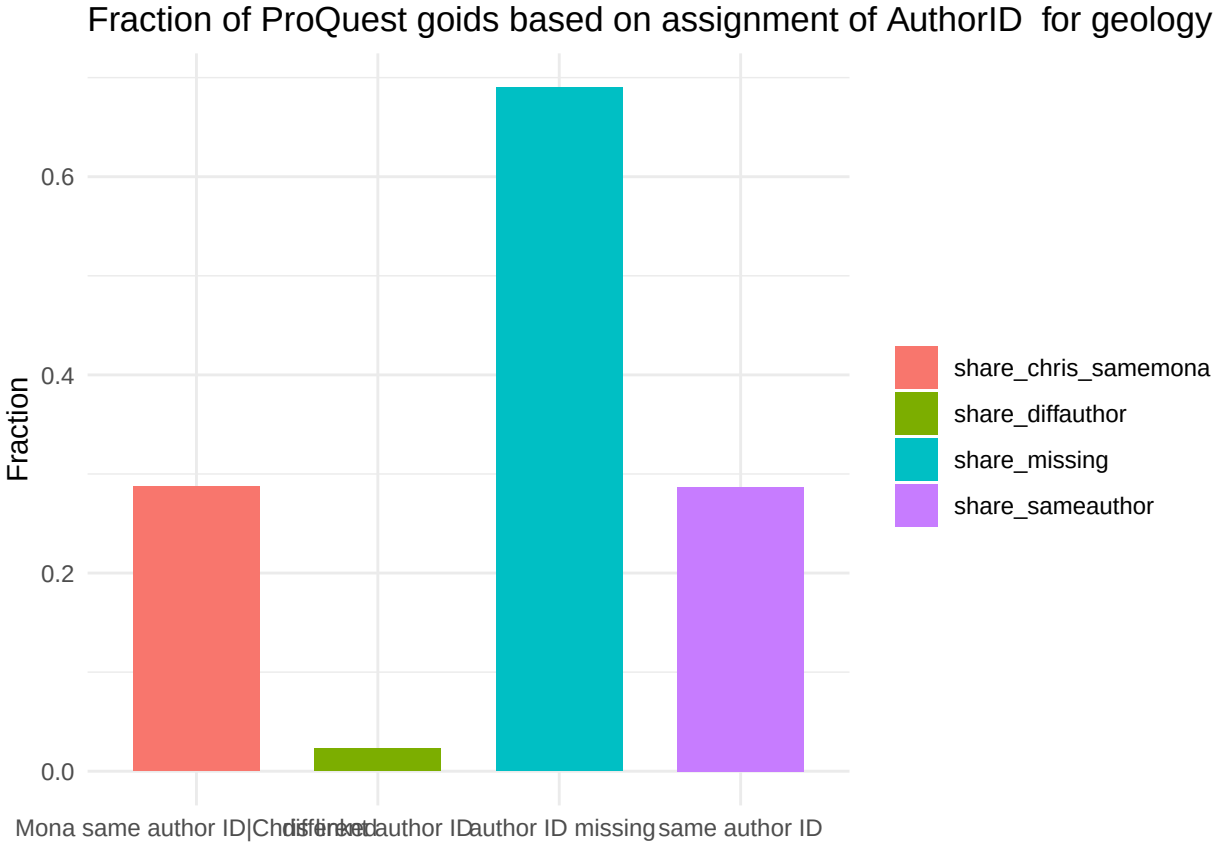

```
## <table>
## <caption>Share Statistics by Field</caption>
##  <thead>
##   <tr>
##    <th style="text-align:left;"> field </th>
##    <th style="text-align:center;"> share_bothlink </th>
##    <th style="text-align:center;"> share_sameauthor </th>
##    <th style="text-align:center;"> share_diffauthor </th>
##    <th style="text-align:center;"> share_missing </th>
##    <th style="text-align:center;"> share_chris_samemona </th>
##   </tr>
##  </thead>
## <tbody>
##   <tr>
##    <td style="text-align:left;"> economics </td>
##    <td style="text-align:center;"> 0.2 </td>
##    <td style="text-align:center;"> 0.2 </td>
##    <td style="text-align:center;"> 0 </td>
##    <td style="text-align:center;"> 0.79 </td>
##    <td style="text-align:center;"> 0.2 </td>
##   </tr>
## </tbody>
## </table>
```



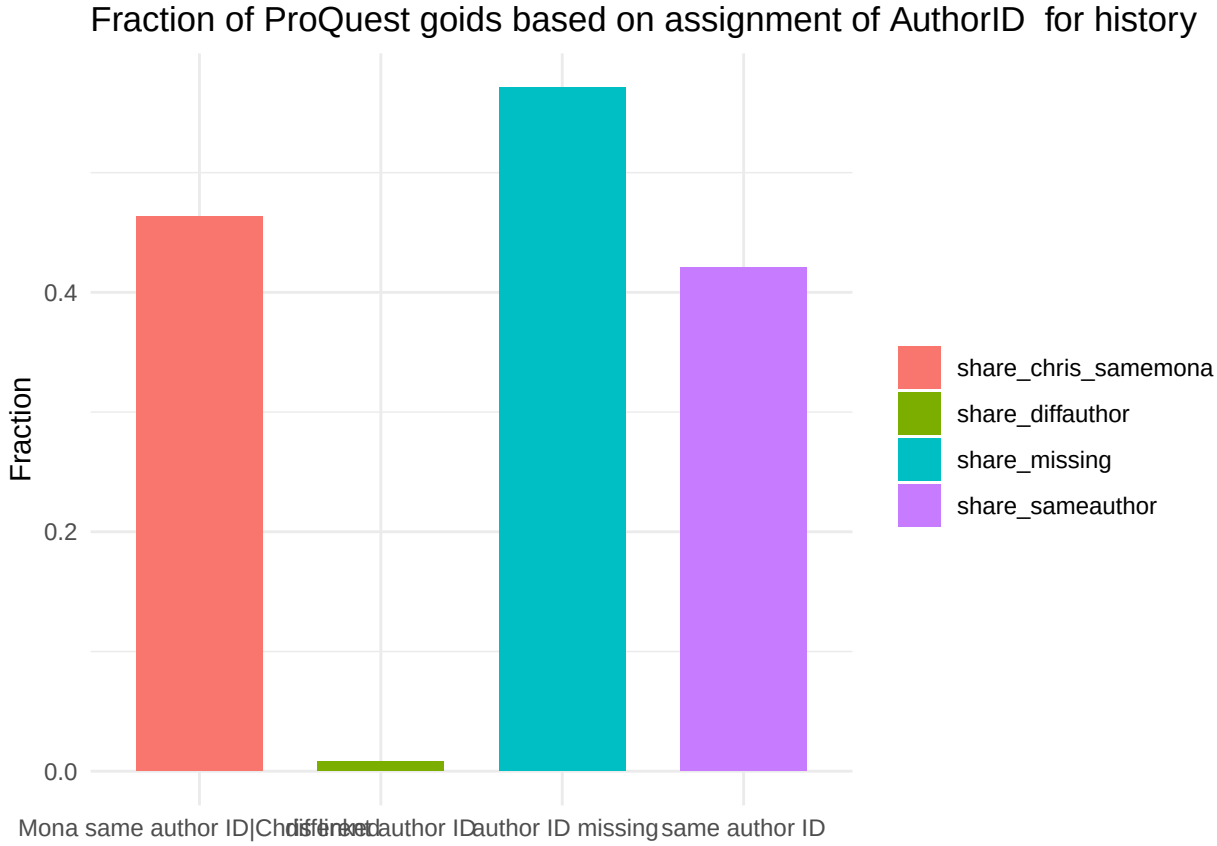

```
## <table>
## <caption>Share Statistics by Field</caption>
##  <thead>
##   <tr>
##    <th style="text-align:left;"> field </th>
##    <th style="text-align:center;"> share_bothlink </th>
##    <th style="text-align:center;"> share_sameauthor </th>
##    <th style="text-align:center;"> share_diffauthor </th>
##    <th style="text-align:center;"> share_missing </th>
##    <th style="text-align:center;"> share_chris_samemona </th>
##   </tr>
##  </thead>
## <tbody>
##   <tr>
##    <td style="text-align:left;"> engineering </td>
##    <td style="text-align:center;"> 0.91 </td>
##    <td style="text-align:center;"> 0.68 </td>
##    <td style="text-align:center;"> 0 </td>
##    <td style="text-align:center;"> 0.32 </td>
##    <td style="text-align:center;"> 0.91 </td>
##   </tr>
## </tbody>
## </table>
```



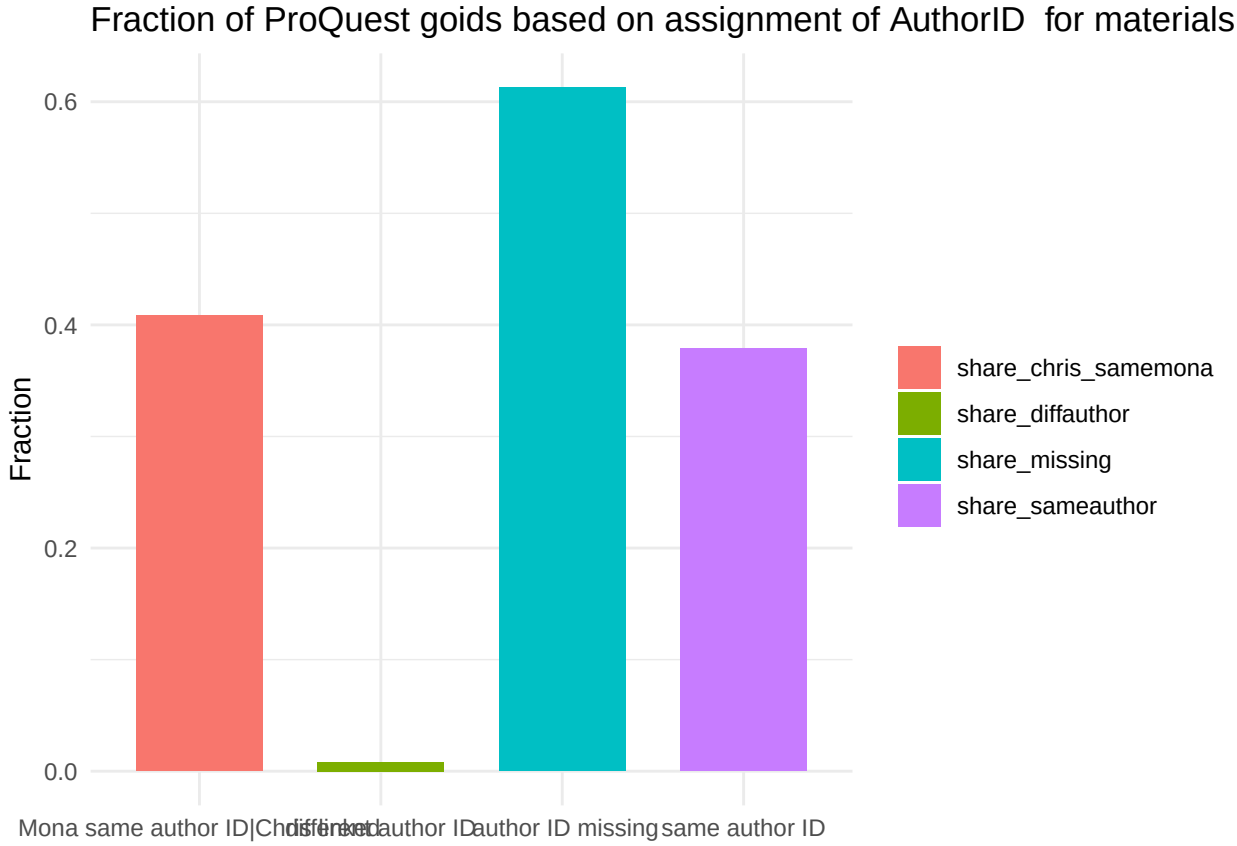

```
## <table>
## <caption>Share Statistics by Field</caption>
##  <thead>
##   <tr>
##    <th style="text-align:left;"> field </th>
##    <th style="text-align:center;"> share_bothlink </th>
##    <th style="text-align:center;"> share_sameauthor </th>
##    <th style="text-align:center;"> share_diffauthor </th>
##    <th style="text-align:center;"> share_missing </th>
##    <th style="text-align:center;"> share_chris_samemona </th>
##   </tr>
##  </thead>
## <tbody>
##   <tr>
##    <td style="text-align:left;"> environmental science </td>
##    <td style="text-align:center;"> 0.44 </td>
##    <td style="text-align:center;"> 0.43 </td>
##    <td style="text-align:center;"> 0.01 </td>
##    <td style="text-align:center;"> 0.56 </td>
##    <td style="text-align:center;"> 0.44 </td>
##   </tr>
## </tbody>
## </table>
```



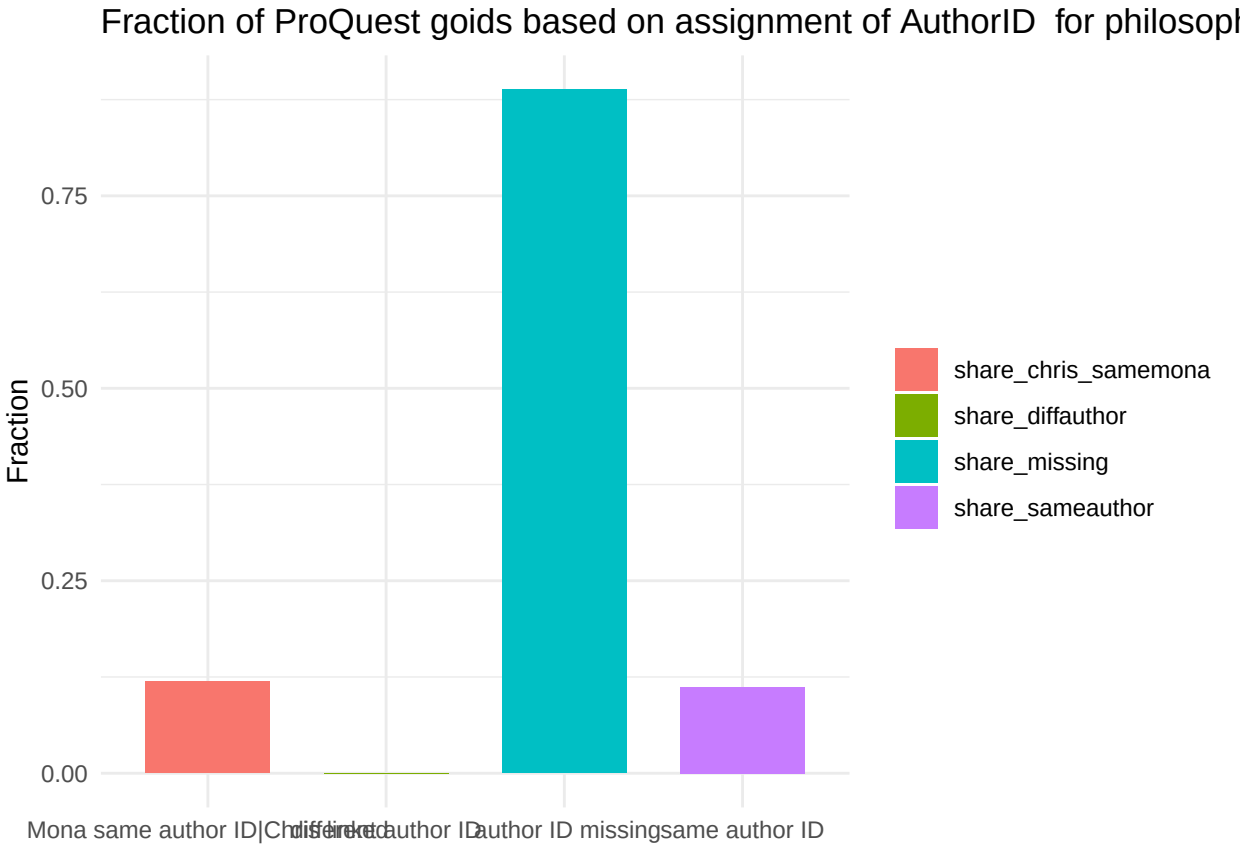

```
## <table>
## <caption>Share Statistics by Field</caption>
##  <thead>
##   <tr>
##    <th style="text-align:left;"> field </th>
##    <th style="text-align:center;"> share_bothlink </th>
##    <th style="text-align:center;"> share_sameauthor </th>
##    <th style="text-align:center;"> share_diffauthor </th>
##    <th style="text-align:center;"> share_missing </th>
##    <th style="text-align:center;"> share_chris_samemona </th>
##   </tr>
##  </thead>
## <tbody>
##   <tr>
##    <td style="text-align:left;"> geology </td>
##    <td style="text-align:center;"> 0.29 </td>
##    <td style="text-align:center;"> 0.29 </td>
##    <td style="text-align:center;"> 0.02 </td>
##    <td style="text-align:center;"> 0.69 </td>
##    <td style="text-align:center;"> 0.29 </td>
##   </tr>
## </tbody>
## </table>
```



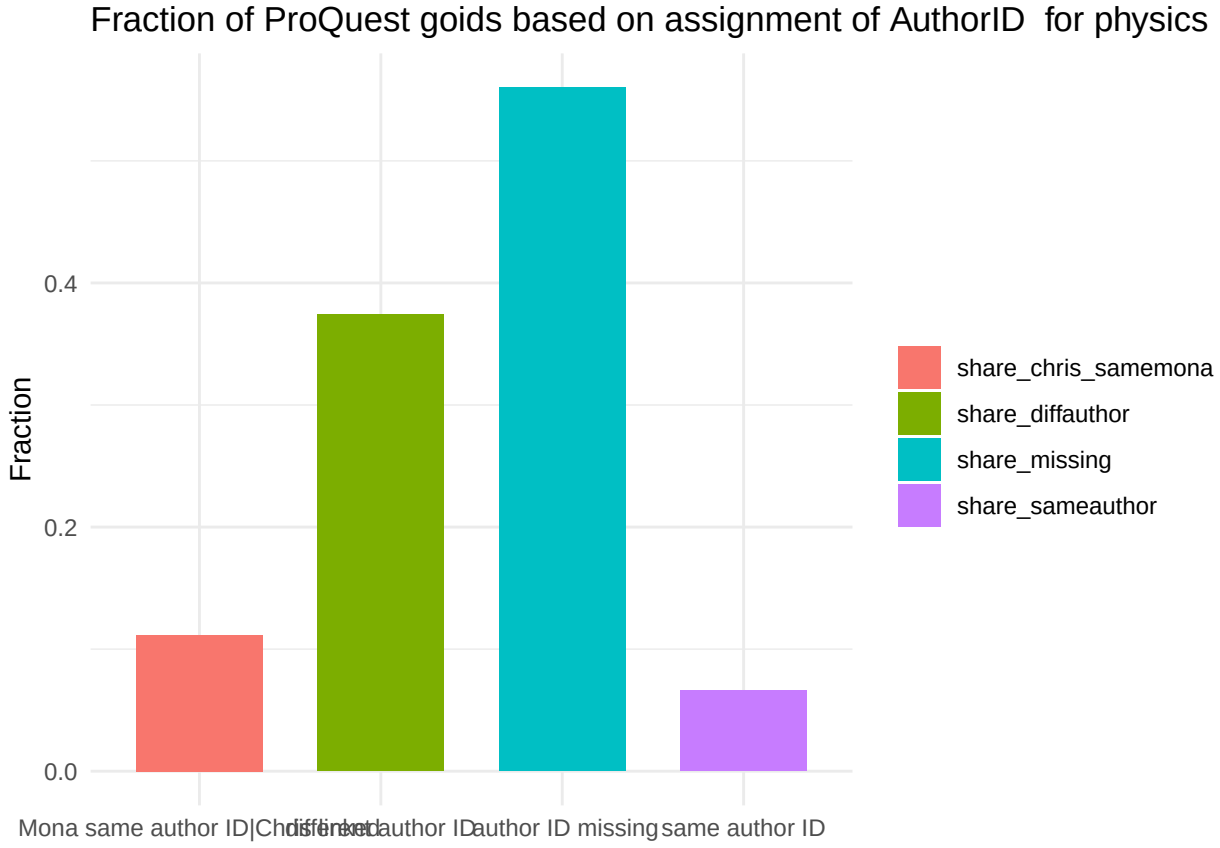

```
## <table>
## <caption>Share Statistics by Field</caption>
##  <thead>
##   <tr>
##    <th style="text-align:left;"> field </th>
##    <th style="text-align:center;"> share_bothlink </th>
##    <th style="text-align:center;"> share_sameauthor </th>
##    <th style="text-align:center;"> share_diffauthor </th>
##    <th style="text-align:center;"> share_missing </th>
##    <th style="text-align:center;"> share_chris_samemona </th>
##   </tr>
##  </thead>
## <tbody>
##   <tr>
##    <td style="text-align:left;"> history </td>
##    <td style="text-align:center;"> 0.46 </td>
##    <td style="text-align:center;"> 0.42 </td>
##    <td style="text-align:center;"> 0.01 </td>
##    <td style="text-align:center;"> 0.57 </td>
##    <td style="text-align:center;"> 0.46 </td>
##   </tr>
## </tbody>
## </table>
```



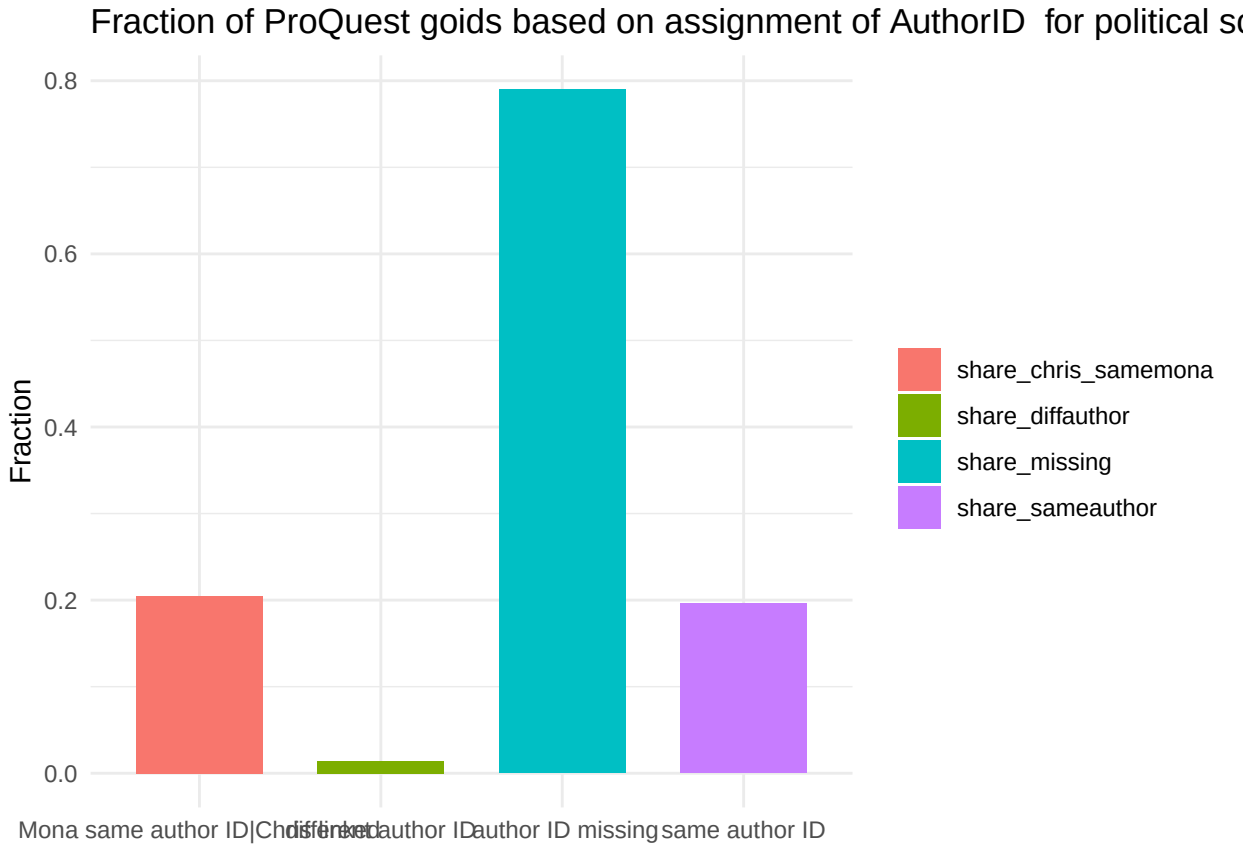

```
## <table>
## <caption>Share Statistics by Field</caption>
##  <thead>
##   <tr>
##    <th style="text-align:left;"> field </th>
##    <th style="text-align:center;"> share_bothlink </th>
##    <th style="text-align:center;"> share_sameauthor </th>
##    <th style="text-align:center;"> share_diffauthor </th>
##    <th style="text-align:center;"> share_missing </th>
##    <th style="text-align:center;"> share_chris_samemona </th>
##   </tr>
##  </thead>
## <tbody>
##   <tr>
##    <td style="text-align:left;"> materials science </td>
##    <td style="text-align:center;"> 0.41 </td>
##    <td style="text-align:center;"> 0.38 </td>
##    <td style="text-align:center;"> 0.01 </td>
##    <td style="text-align:center;"> 0.61 </td>
##    <td style="text-align:center;"> 0.41 </td>
##   </tr>
## </tbody>
## </table>
```



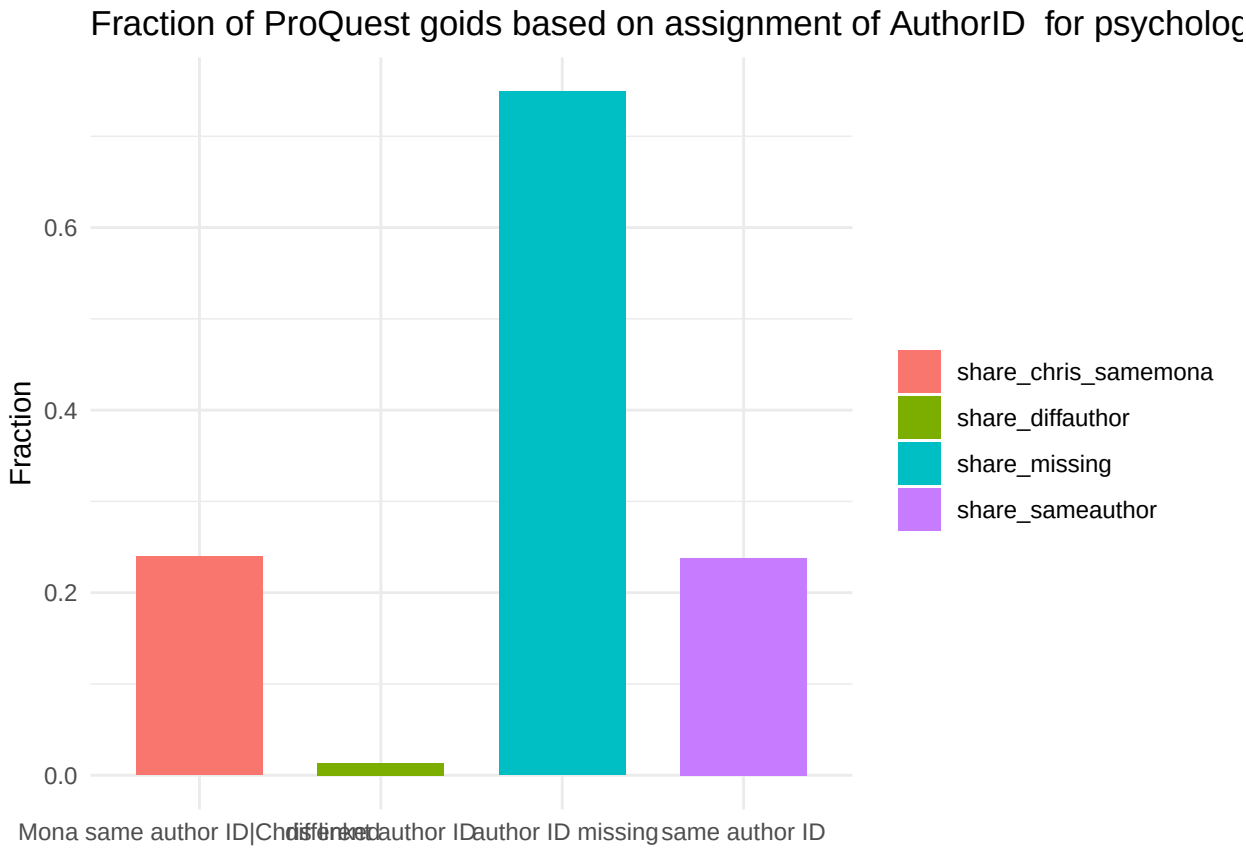

```
## <table>
## <caption>Share Statistics by Field</caption>
##  <thead>
##   <tr>
##    <th style="text-align:left;"> field </th>
##    <th style="text-align:center;"> share_bothlink </th>
##    <th style="text-align:center;"> share_sameauthor </th>
##    <th style="text-align:center;"> share_diffauthor </th>
##    <th style="text-align:center;"> share_missing </th>
##    <th style="text-align:center;"> share_chris_samemona </th>
##   </tr>
##  </thead>
## <tbody>
##   <tr>
##    <td style="text-align:left;"> philosophy </td>
##    <td style="text-align:center;"> 0.12 </td>
##    <td style="text-align:center;"> 0.11 </td>
##    <td style="text-align:center;"> 0 </td>
##    <td style="text-align:center;"> 0.89 </td>
##    <td style="text-align:center;"> 0.12 </td>
##   </tr>
## </tbody>
## </table>
```




```
## <table>
## <caption>Share Statistics by Field</caption>
##  <thead>
##   <tr>
##    <th style="text-align:left;"> field </th>
##    <th style="text-align:center;"> share_bothlink </th>
##    <th style="text-align:center;"> share_sameauthor </th>
##    <th style="text-align:center;"> share_diffauthor </th>
##    <th style="text-align:center;"> share_missing </th>
##    <th style="text-align:center;"> share_chris_samemona </th>
##   </tr>
##  </thead>
## <tbody>
##   <tr>
##    <td style="text-align:left;"> physics </td>
##    <td style="text-align:center;"> 0.11 </td>
##    <td style="text-align:center;"> 0.07 </td>
##    <td style="text-align:center;"> 0.37 </td>
##    <td style="text-align:center;"> 0.56 </td>
##    <td style="text-align:center;"> 0.11 </td>
##   </tr>
## </tbody>
## </table>
```




```
## <table>
## <caption>Share Statistics by Field</caption>
##  <thead>
##   <tr>
##    <th style="text-align:left;"> field </th>
##    <th style="text-align:center;"> share_bothlink </th>
##    <th style="text-align:center;"> share_sameauthor </th>
##    <th style="text-align:center;"> share_diffauthor </th>
##    <th style="text-align:center;"> share_missing </th>
##    <th style="text-align:center;"> share_chris_samemona </th>
##   </tr>
##  </thead>
## <tbody>
##   <tr>
##    <td style="text-align:left;"> political science </td>
##    <td style="text-align:center;"> 0.2 </td>
##    <td style="text-align:center;"> 0.2 </td>
##    <td style="text-align:center;"> 0.01 </td>
##    <td style="text-align:center;"> 0.79 </td>
##    <td style="text-align:center;"> 0.2 </td>
##   </tr>
## </tbody>
## </table>
```




```
## <table>
## <caption>Share Statistics by Field</caption>
##  <thead>
##   <tr>
##    <th style="text-align:left;"> field </th>
##    <th style="text-align:center;"> share_bothlink </th>
##    <th style="text-align:center;"> share_sameauthor </th>
##    <th style="text-align:center;"> share_diffauthor </th>
##    <th style="text-align:center;"> share_missing </th>
##    <th style="text-align:center;"> share_chris_samemona </th>
##   </tr>
##  </thead>
## <tbody>
##   <tr>
##    <td style="text-align:left;"> psychology </td>
##    <td style="text-align:center;"> 0.24 </td>
##    <td style="text-align:center;"> 0.24 </td>
##    <td style="text-align:center;"> 0.01 </td>
##    <td style="text-align:center;"> 0.75 </td>
##    <td style="text-align:center;"> 0.24 </td>
##   </tr>
## </tbody>
## </table>
```

- many missings between Christoph's and Mona's links
- share of Mona's links compared to Christoph's links low (share_bothlink) but mostly due to missings

and different author assignment
- in most fields, goids linked to the same authors, only few that were linked to different ones
- exception in physics, most authors linked differently, why? (no obvious mistakes when renaming variables and joining the datasets)

Now do the same just for physics but for christoph, mona and flavio

```
process_data <- function(field) {
  # Reads data for specified field

  # Read the data for Mona
  links_graduates_mona <- read.csv(paste0(datapath,"links_graduates_", field, "_mona_degree0_19852015.cs
    filter(link_score>0.7) %>%
    rename(authorid_mona = grantid_authorposition) %>%
    rename(goid = AuthorId) %>%
    rename(linkscore_mona=link_score)
  # Read the data for Flavio
  links_graduates_flavio <- read.csv(paste0(datapath,"links_graduates_", field, "_flavio_degree0_198520
    filter(link_score>0.7) %>%
    rename(authorid_flavio = AuthorId) %>%
    rename(linkscore_flavio=link_score)

  # Read the data for Christoph
  links_graduates_christoph <- read.csv(paste0(datapath,"links_graduates_", field, "_christoph_degree0_
    filter(link_score>0.7) %>%
    rename(authorid_christoph = AuthorId) %>%
    rename(linkscore_christoph=link_score)

  links_graduates_mona <- collect(links_graduates_mona)
  links_graduates_flavio <- collect(links_graduates_flavio)
  links_graduates_christoph <- collect(links_graduates_christoph)

  # Join the data
  links_graduates <- links_graduates_mona %>%
    full_join(links_graduates_flavio, by = c("goid")) %>%
    full_join(links_graduates_christoph, by = c("goid")) %>%
    mutate(
      field = field,

      mona_same_as_chris = ifelse(authorid_mona == authorid_christoph, 1, 0),
      flavio_same_as_chris = ifelse(authorid_flavio == authorid_christoph, 1, 0),

      mona_diff_from_chris = ifelse(authorid_mona != authorid_christoph & !is.na(authorid_mona) & !is.na
      flavio_diff_from_chris = ifelse(authorid_flavio != authorid_christoph & !is.na(authorid_flavio) &

    ) %>%
    replace_na(list(mona_same_as_chris = 0, flavio_same_as_chris = 0, mona_diff_from_chris = 0, flavio_
    mutate(
      # Calculate shares
      share_mona_same_as_chris = mean(mona_same_as_chris, na.rm = TRUE),
      share_flavio_same_as_chris = mean(flavio_same_as_chris, na.rm = TRUE),
      share_mona_diff_from_chris = mean(mona_diff_from_chris, na.rm = TRUE),
      share_flavio_diff_from_chris = mean(flavio_diff_from_chris, na.rm = TRUE)
```

```
    )

  # Create table with the shares by field
  shares_table  <- links_graduates %>%
    select(field, share_mona_same_as_chris, share_flavio_same_as_chris, share_mona_diff_from_chris, shar
    group_by(field) %>%
    summarize(
      share_mona_same_as_chris = mean(share_mona_same_as_chris, na.rm = TRUE),
      share_flavio_same_as_chris = mean(share_flavio_same_as_chris, na.rm = TRUE),
      share_mona_diff_from_chris = mean(share_mona_diff_from_chris, na.rm = TRUE),
      share_flavio_diff_from_chris = mean(share_flavio_diff_from_chris, na.rm = TRUE)
    )

  # Return the table
  shares_table %>%
    kable(format = "latex",
          align = c("l", "c", "c", "c", "c"),
          digits = 2,
          caption = "Share Statistics by Field",
          booktabs = TRUE)
}
```

```
# Specify the field to process
field_to_process <- "physics"

# Process the data
table <- process_data(field_to_process)

# Print the table
print(table)
```

```
## \begin{table}
##
## \caption{\label{tab:unnamed-chunk-5}Share Statistics by Field}
## \centering
## \begin{tabular}[t]{lcccc}
## \toprule
## field & share\_mona\_same\_as\_chris & share\_flavio\_same\_as\_chris & share\_mona\_diff\_from\_chri
## \midrule
## physics & 0.06 & 0.5 & 0.36 & 0.03\\
## \bottomrule
## \end{tabular}
## \end{table}
```