Performance of linking graduates

Flavio & Christoph & Mona

21 July, 2023

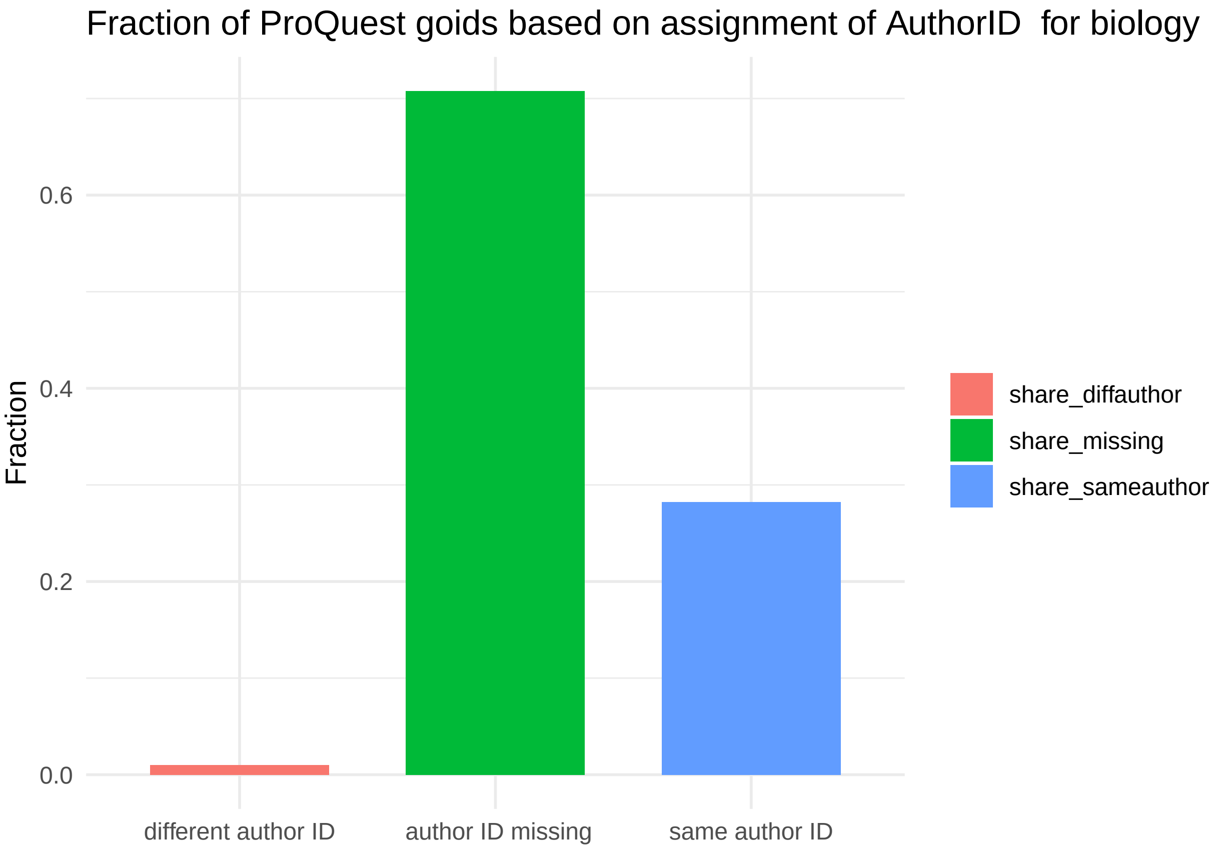
Contents

This script makes some plots of the links. But not all fields complete

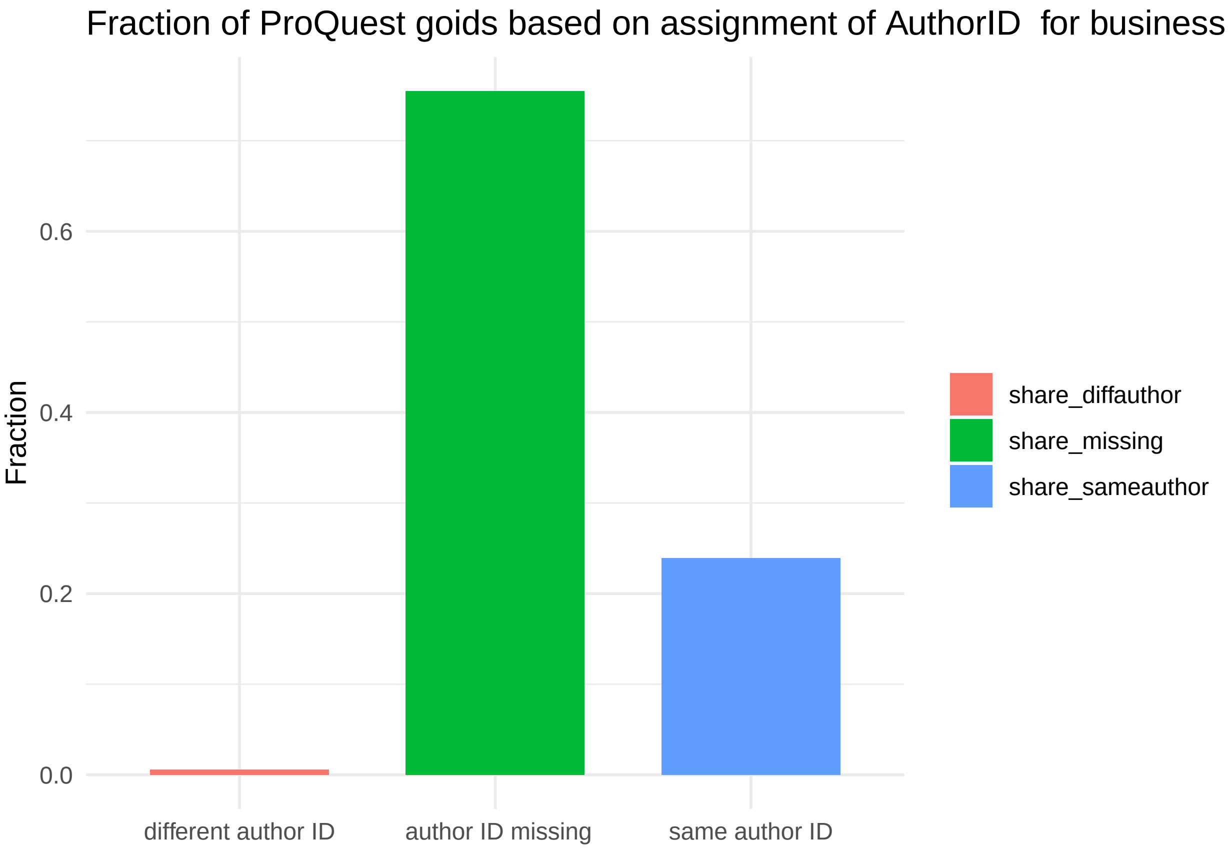
```
 # Function to process the data for a specific field
# Fields to process
# missing fields: ""art, "chemistry", "geography", "history", "mathematics","medicine", "sociology"


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

# Loop through the fields

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.c
    rename(authorid_mona = AuthorId) %>%
    rename(linkscore_mona=link_score)
  } else {
    links_graduates_mona <- read.csv(paste0(datapath,"/links_graduates_", field, "_mona_degree0_1985201
    rename(authorid_mona = grantid_authorposition) %>%
    rename(goid = AuthorId) %>%
    rename(linkscore_mona=link_score)
}
  if (field %in% c("biology", "computer science","economics", "engineering", "environmental science", "
  links_graduates_christoph <- read.csv(paste0(datapath,"links_graduates_", field, "_christoph_fielddeg
    rename(authorid_christoph = AuthorId) %>%
    rename(linkscore_christoph=link_score)
  } else {
  links_graduates_christoph <- read.csv(paste0(datapath,"links_graduates_", field, "_christoph_degree0_
    rename(authorid_christoph = AuthorId) %>%
    rename(linkscore_christoph=link_score)
  }


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

```r
# 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

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 = T
  )

# 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)
 )

# 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) %>%
  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)
  )

# Print the summary table in Markdown format (not shown in pdf, ugly here)

cat(kable(shares_table, format = "markdown",
           align = c("l", "c", "c", "c", "c"), # Align columns (left, center, center, center, center)
           caption = "Share Statistics by Field", # Table caption
           digits = 4, # Number of digits to display for numeric values
           booktabs = TRUE # Use booktabs style for the table
           ))

# 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)
  ) %>%
  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("different author ID", " author ID missing", "same author ID"))


# Print the bar chart
print(bar_chart)
}

for (field in fields_to_process) {
  process_data(field)
  cat("\n\n")
}
```

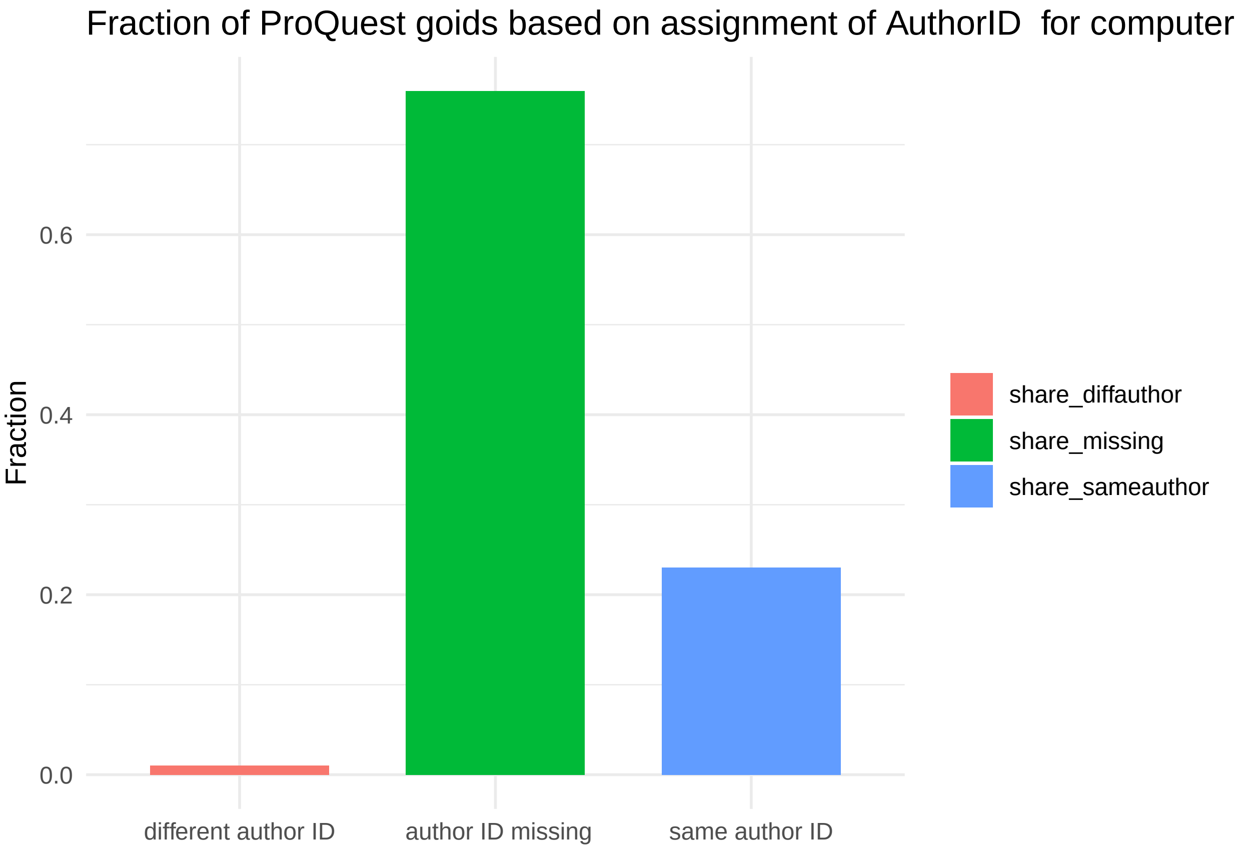
```
## Table: Share Statistics by Field  |field   | share_bothlink | share_sameauthor | share_diffauthor |
```



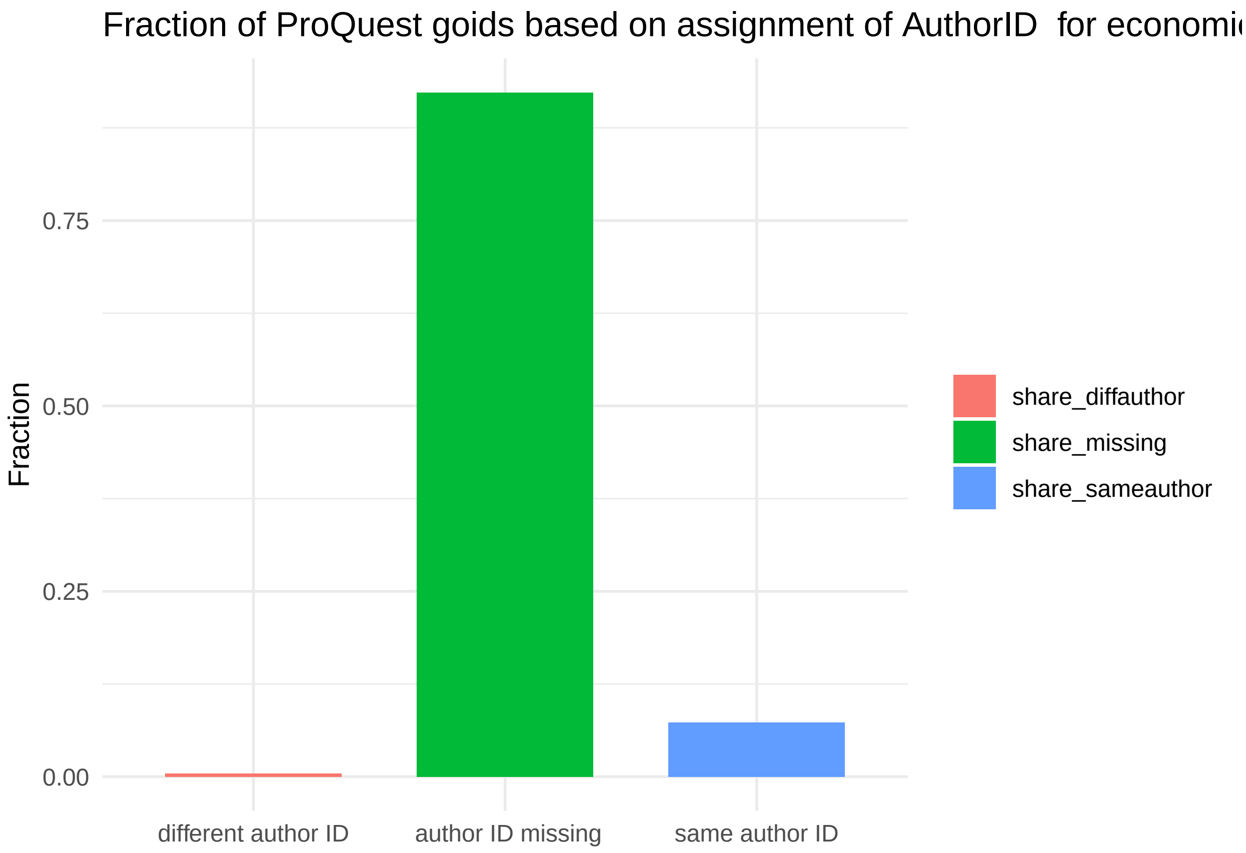
```
##
##
## Table: Share Statistics by Field  |field    | share_bothlink | share_sameauthor | share_diffauthor |
```



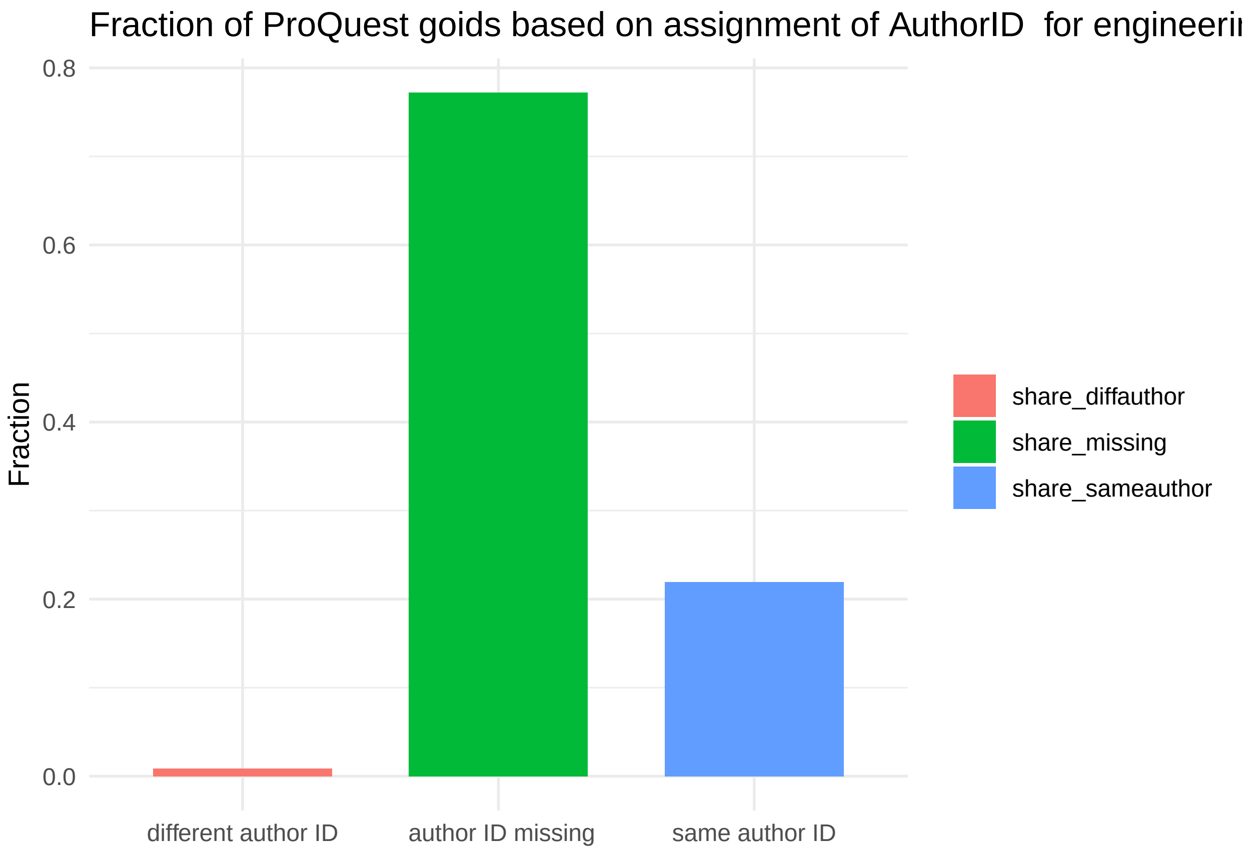
```
##
##
## Table: Share Statistics by Field  |field          | share_bothlink | share_sameauthor | share_diffa
```



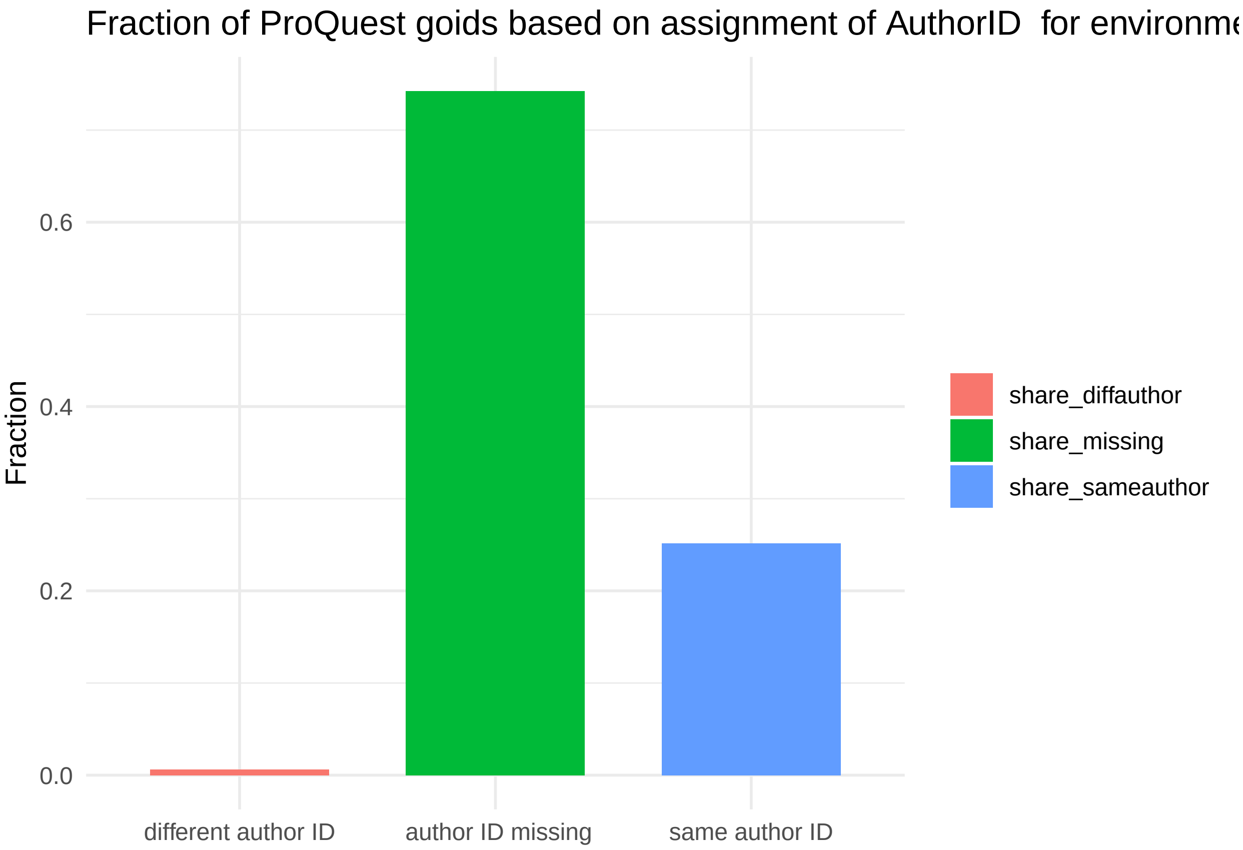
```
##
##
## Table: Share Statistics by Field  |field     | share_bothlink | share_sameauthor | share_diffauthor
```



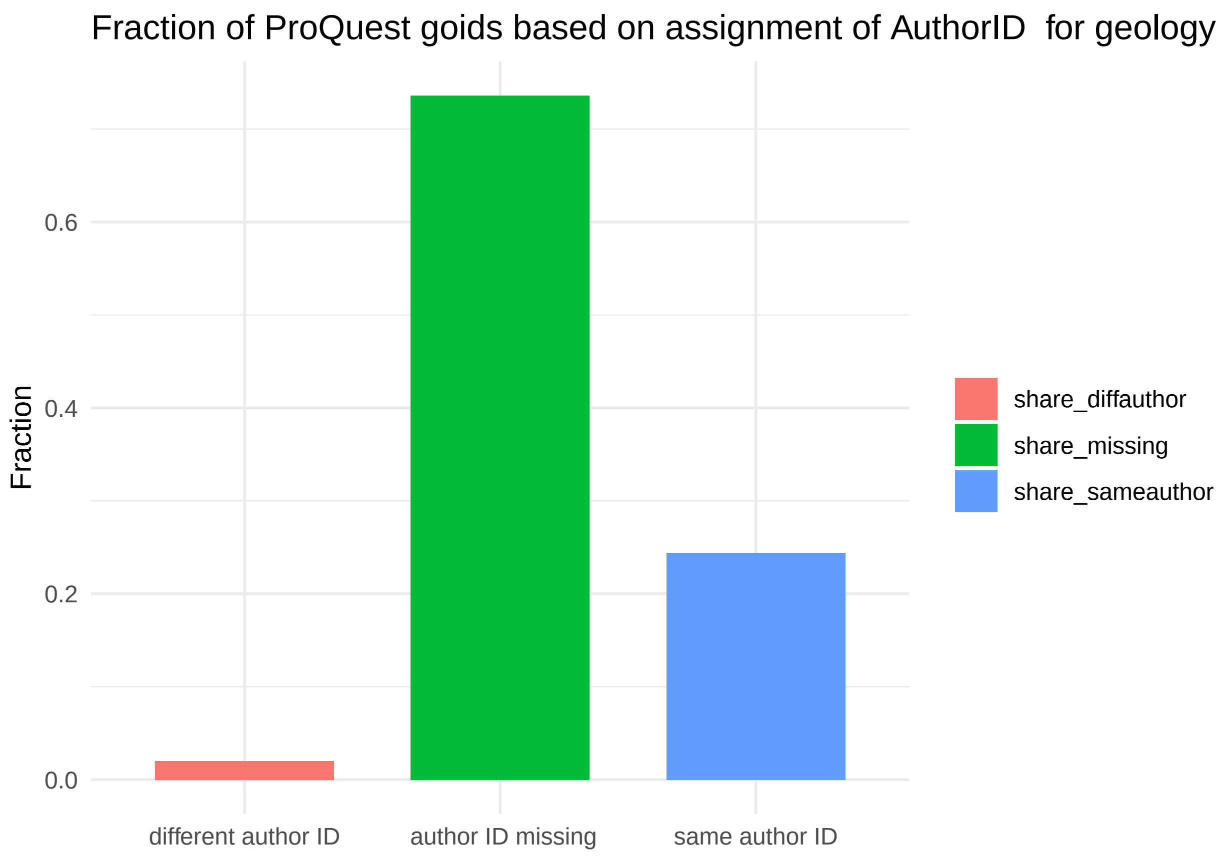
```
##
##
## Table: Share Statistics by Field  |field       | share_bothlink | share_sameauthor | share_diffautho
```



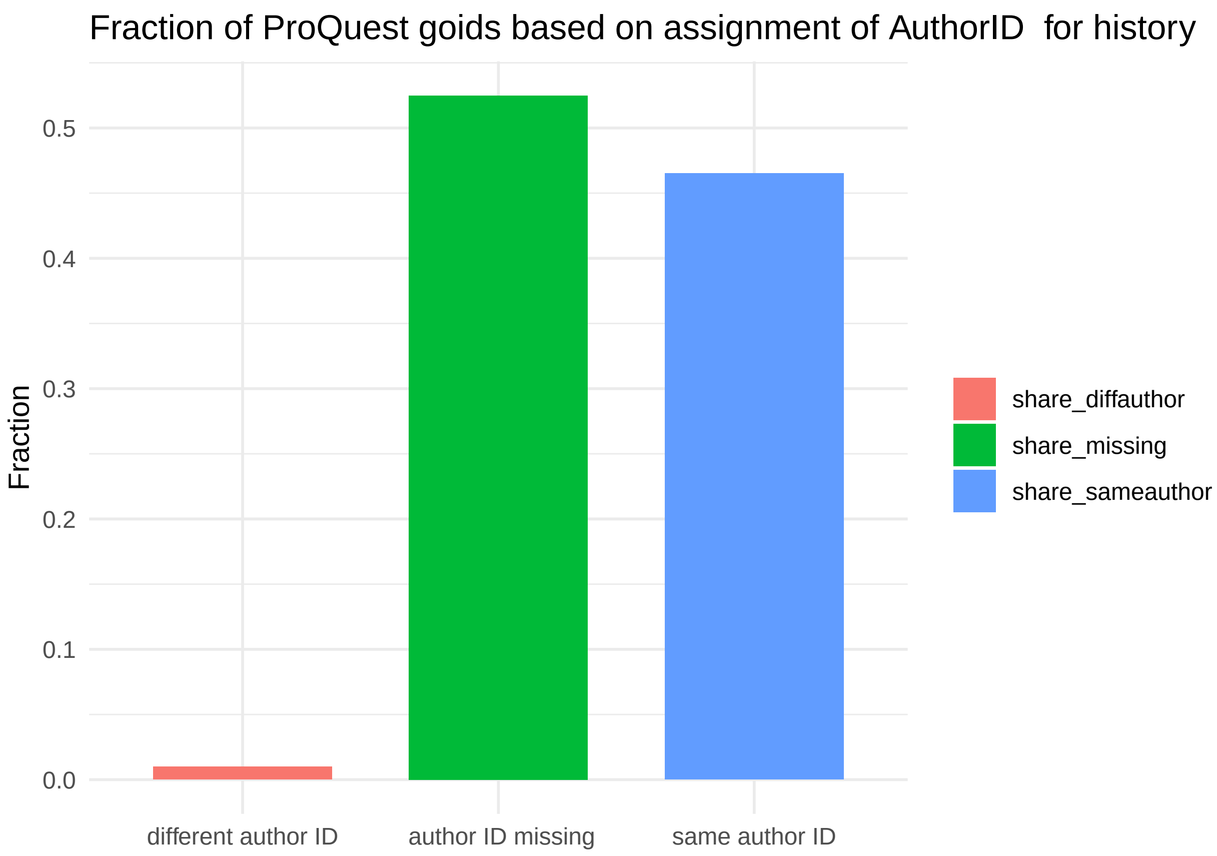
```
##
##
## Table: Share Statistics by Field  |field                  | share_bothlink | share_sameauthor | share_
```



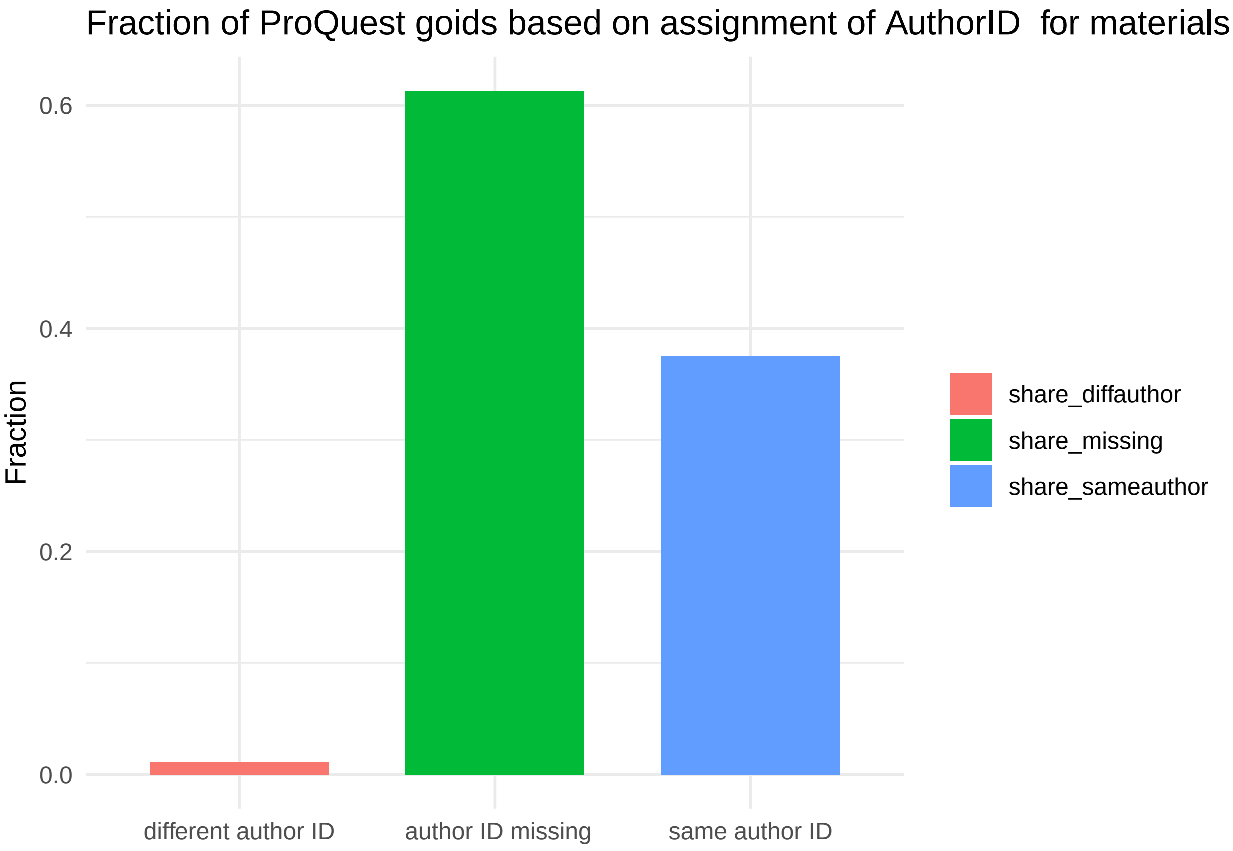
```
##
##
## Table: Share Statistics by Field  |field   | share_bothlink | share_sameauthor | share_diffauthor | s
```



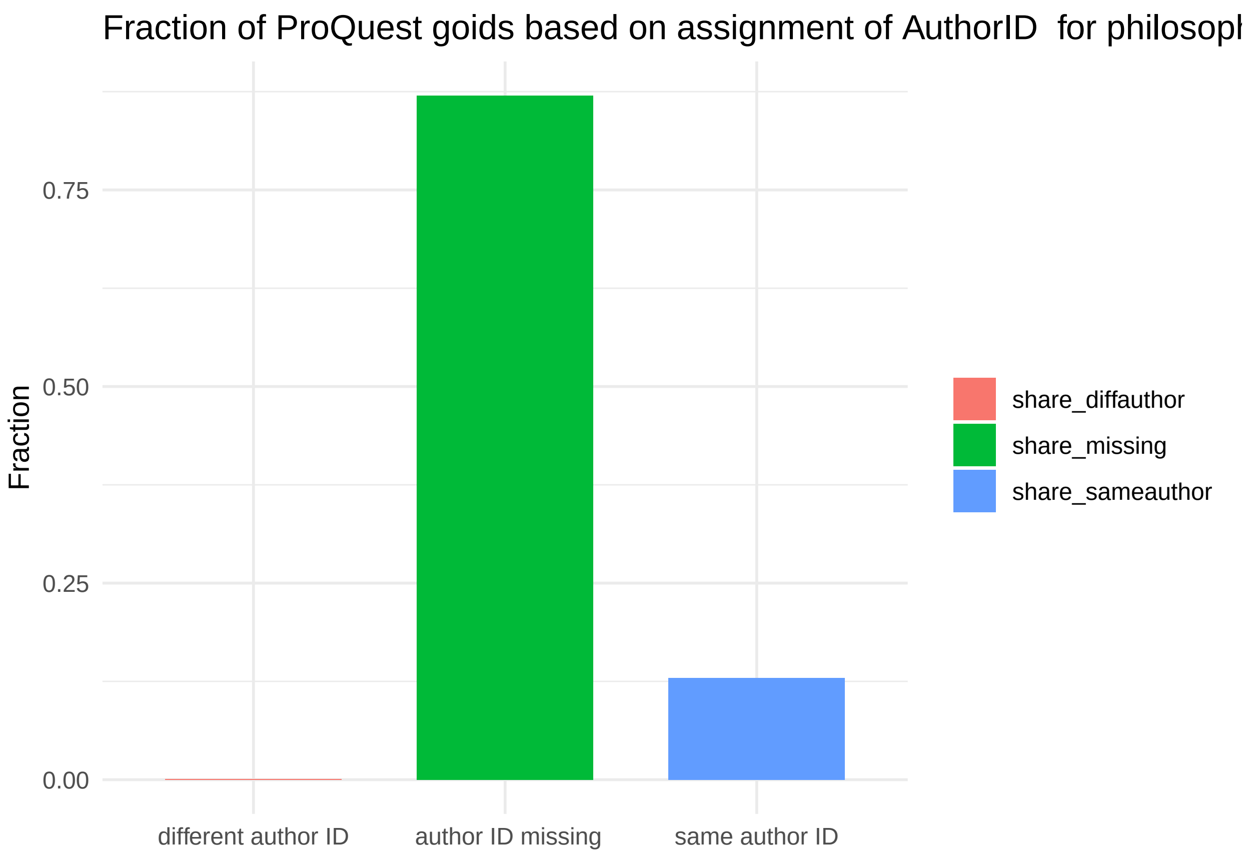
```
##
##
## Table: Share Statistics by Field  |field   | share_bothlink | share_sameauthor | share_diffauthor | s
```


Fraction of ProQuest goids based on assignment of AuthorID for history

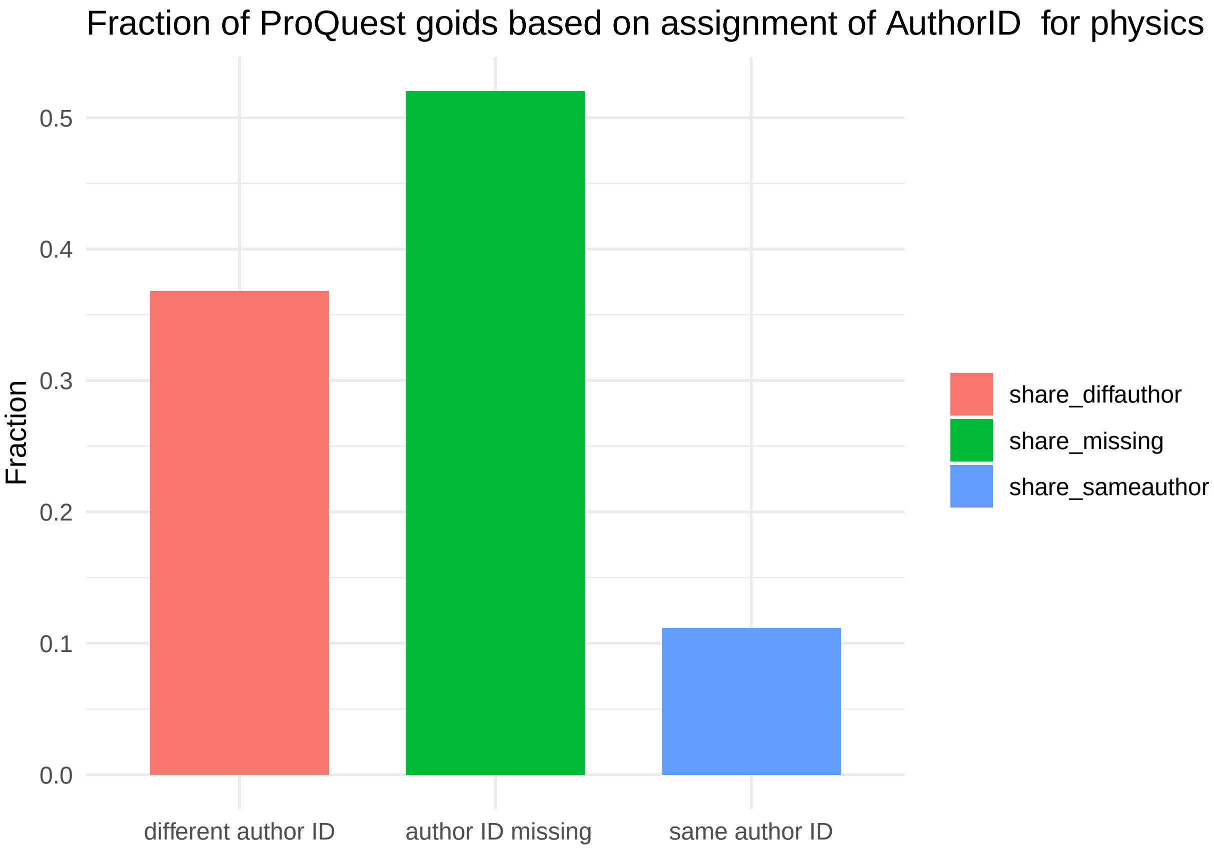
```
##
##
## Table: Share Statistics by Field  |field            | share_bothlink | share_sameauthor | share_dif
```



```
##
##
## Table: Share Statistics by Field  |field      | share_bothlink | share_sameauthor | share_diffauthor
```

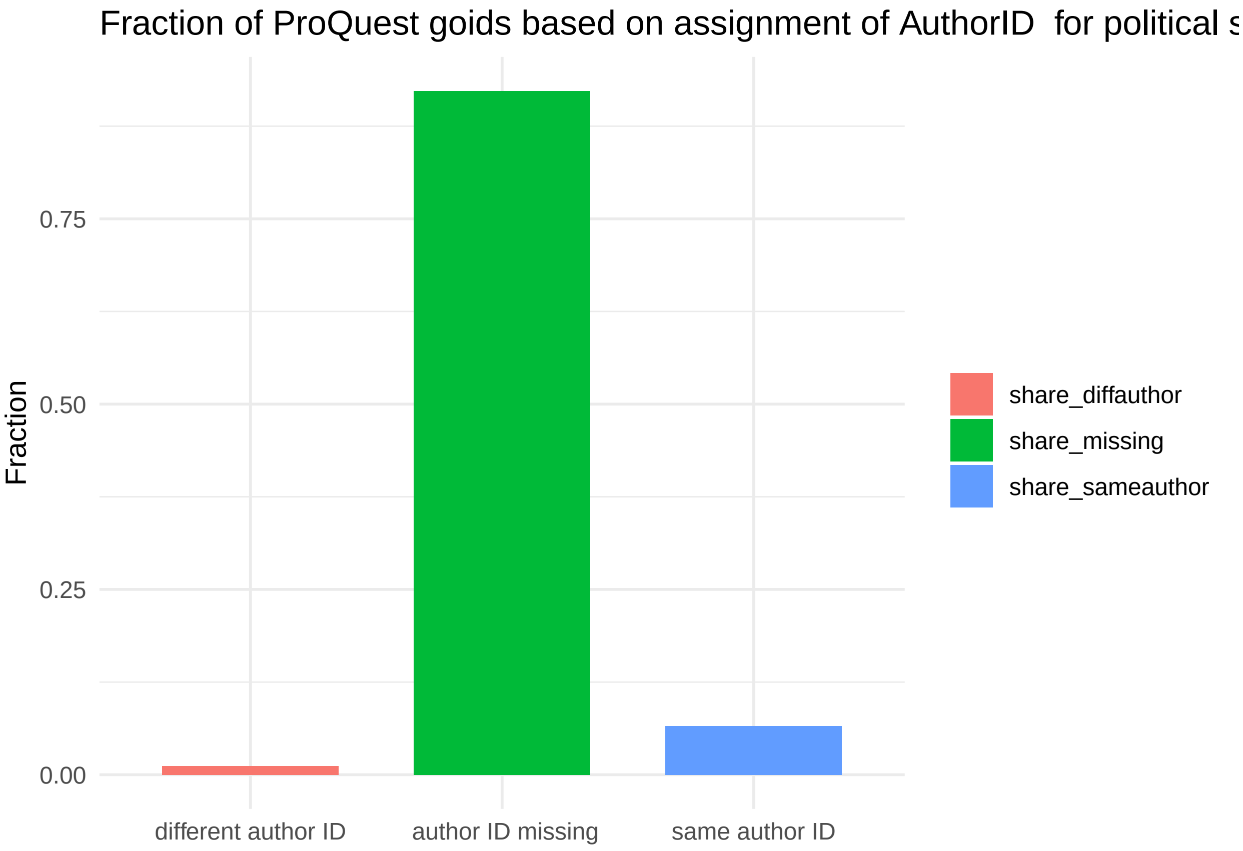


```
##
##
## Table: Share Statistics by Field  |field   | share_bothlink | share_sameauthor | share_diffauthor | s
```

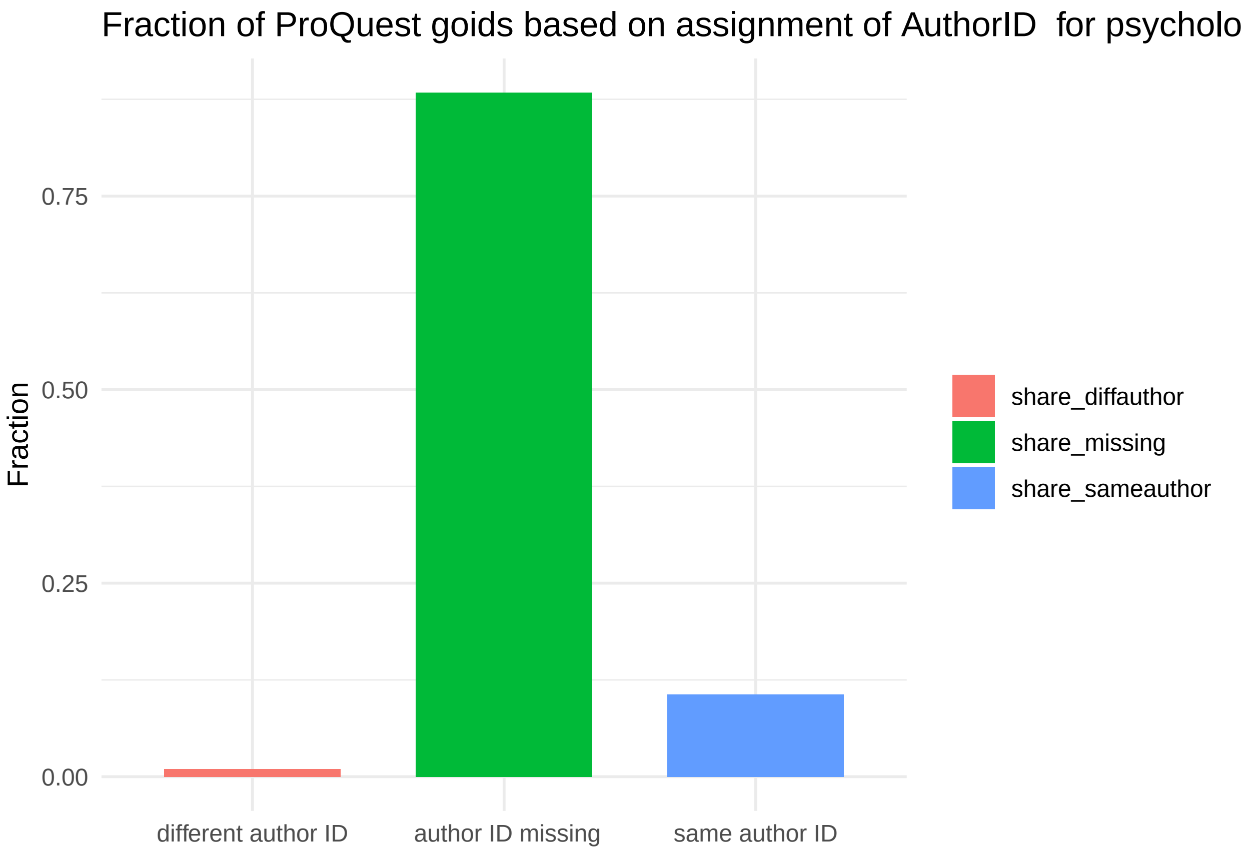


Fraction of ProQuest goids based on assignment of AuthorID for physics

```
##
##
## Table: Share Statistics by Field  |field             | share_bothlink | share_sameauthor | share_dif
```



```
##
##
## Table: Share Statistics by Field  |field      | share_bothlink | share_sameauthor | share_diffauthor
```



- 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)