Compare Linking across linking runs

Christoph, Flavio, Mona

2023-07-23

Linking Advisors and Graduates from ProQuest to MAG AuthorIds

To check whether the linking makes sense we compare the links across several independent runs of the labelling.

Which fields and which linking runs to compare?

```
fields_to_process <-c( #"art", <- dropped
  "biology",
  "business",
  "chemistry",
  "computer science" ,
  "economics",
  "engineering",
  "environmental science",
  "geography",
  "geology" ,
  "history",
  "materials science",
  "mathematics",
  #"medicine",   <- dropped
  "philosophy",
  "physics",
  "political science",
  "psychology" ,
  "sociology")

linker1 = "christoph_with_protocol_magkeywords"
linker2 = "flavio_with_protocol_magkeywords"
```

Load additional data from proquest for tables

```
all_graduates <- get_proquest(conn = con, from = "graduates", start_year = 1990, end_year = 2015) |>
  collect()
```

```
## Warning: Missing values are always removed in SQL aggregation functions.
## Use `na.rm = TRUE` to silence this warning
## This warning is displayed once every 8 hours.
```

```
graduate_links <- get_links(conn = con, from = "graduates") |>
  collect() |>
  filter(goid %in% all_graduates$goid)
```

```
advisor_links <- get_links(conn = con, from = "advisors") |>
  left_join(tbl(con, "pq_advisors") |>
              select(goid, relationship_id),
            by = "relationship_id") |>
  collect() |>
  filter(goid %in% all_graduates$goid)
```

```
## Note: At the moment, using a link score below 0.95 for advisors can result in suspiciously many false
```

```
d_links <- list(
  graduates = graduate_links,
  advisors = advisor_links
)

d_links <- map(
  .x = d_links,
  .f = ~.x |>
    left_join(all_graduates |>
                select(goid, fieldname0_mag),
              by = "goid")
)


# some are not loaded in get_proquest b/c they have special degrees (Psy.D). -> filter on goid being in
# others are in get_proquest but have missing field of study. why? -- but they need to have a link in t
```

If some linked graduates in `get_proquest` had missing fields, it would show up here and should be addressed.

```
map(d_links, ~mean(is.na(.x[["fieldname0_mag"]])))
```

```
## $graduates
## [1] 0
##
## $advisors
## [1] 0
```

```
map(d_links, ~.x |> filter(is.na(fieldname0_mag)) |> head())
```

```
## $graduates
## # A tibble: 0 x 4
## # i 4 variables: AuthorId <int64>, goid <int64>, link_score <dbl>,
## #   fieldname0_mag <chr>
##
## $advisors
## # A tibble: 0 x 5
## # i 5 variables: AuthorId <int64>, relationship_id <chr>, link_score <dbl>,
## #   goid <int64>, fieldname0_mag <chr>
```

```
# count is zero for advisors: this means that none of the graduates with missing field above are in the
# Note: a complication is that linked graduates and all_graduates both have units that are not in the o
  # (this is because the sampling for linking is different than the sampling in get_proquest)
# therefore, it's also important to make the sets overlapping, and then compare the links found/check w
```

```
graduate_counts <- list(
  "total" = all_graduates,
```

field	same	only1	only2	diff	diff_rel1	nlink1	nlink2	namedist_pq_1	namedist_pq_2	namedist_diff	namedist_diff_pq_1	namedist_diff_pq_2	namedist_pq_only1	namedist_pq_only2
biology	0.67	0.29	0.03	0.01	0.01	82100	60252	0.09	0.08	0.08	0.09	0.08	0.09	0.80
business	0.67	0.15	0.13	0.05	0.05	18048	17640	0.12	0.06	0.28	0.12	0.06	0.12	0.07
chemistry	0.86	0.12	0.01	0.01	0.01	35806	32036	0.08	0.08	0.04	0.08	0.08	0.08	0.74
computer science	0.88	0.04	0.07	0.02	0.02	34852	35990	0.05	0.05	0.03	0.05	0.05	0.05	0.03
economics	0.76	0.16	0.07	0.01	0.01	12742	11470	0.09	0.08	0.10	0.09	0.08	0.09	0.18
engineering	0.59	0.24	0.14	0.04	0.04	49771	44095	0.06	0.09	0.13	0.06	0.09	0.06	0.15
environmental science	0.79	0.03	0.17	0.01	0.02	11186	13039	0.08	0.08	0.04	0.08	0.08	0.08	0.01
geography	0.66	0.29	0.04	0.01	0.01	6902	5123	0.08	0.08	0.03	0.08	0.08	0.08	0.55
geology	0.68	0.26	0.05	0.01	0.01	8715	6769	0.08	0.10	0.13	0.08	0.10	0.08	0.52
history	0.88	0.07	0.05	0.01	0.01	9564	9337	0.08	0.09	0.05	0.08	0.09	0.08	0.13
materials science	0.46	0.26	0.22	0.05	0.06	20657	19710	0.05	0.11	0.22	0.05	0.11	0.05	0.13
mathematics	0.63	0.32	0.04	0.01	0.01	16970	11928	0.06	0.07	0.03	0.06	0.07	0.06	0.62
philosophy	0.85	0.06	0.08	0.01	0.01	4268	4343	0.08	0.08	0.02	0.08	0.08	0.08	0.07
physics	0.62	0.04	0.31	0.03	0.04	10139	13992	0.08	0.11	0.25	0.08	0.11	0.08	0.01
political science	0.85	0.08	0.06	0.01	0.01	12951	12707	0.08	0.07	0.02	0.08	0.07	0.08	0.09
psychology	0.89	0.06	0.04	0.01	0.01	54914	53708	0.07	0.08	0.05	0.07	0.08	0.07	0.12
sociology	0.71	0.05	0.23	0.01	0.02	6751	8332	0.08	0.08	0.06	0.08	0.08	0.08	0.02

```
  "links" = d_links$graduates
)

graduate_counts <- map(
  graduate_counts,
  ~ .x |>
    filter(fieldname0_mag %in% fields_to_process) |>
    group_by(fieldname0_mag) |>
    summarise(nb = n())
)

advisor_link_counts <- d_links$advisors |>
  filter(fieldname0_mag %in% fields_to_process) |>
  group_by(fieldname0_mag) |>
  summarise(nb = n())
```

Comparison for Graduates

Printing out the dataframe

```
res_combined_graduates <- reduce(res, rbind)

res_combined_graduates %>%
  mutate(across(where(is.numeric), \(x) round(x, digits = 2))) %>%  # this is an updated form of mutate_
  kable(format = "latex", digits = 2, booktabs = TRUE) %>%
  kable_styling(latex_options = c("striped", "scale_down"))
```

Add final number of links to res_combined_graduates

```
res_combined_graduates <- res_combined_graduates |>
  left_join(graduate_counts$links |>
              rename(n_links_final = nb),
            by = c("field" = "fieldname0_mag"))
```

Calculate weighted average of linking stats across fields

```
across_fields_graduates <- res_combined_graduates |>
  left_join(graduate_counts$total |>
              rename(n_graduates = nb),
            by = c("field" = "fieldname0_mag"))

across_fields_graduates <- across_fields_graduates |>
  mutate(share = n_graduates / sum(n_graduates)) |>
  summarise(
```

```
    across(all_of(c("same", "only1", "only2", "diff")),
           ~weighted.mean(.x, w = share)),
    across(all_of(c("n_links_final")),
           ~sum(.x))
  ) |>
  mutate(field = "Total")
```

Make table

```
df_table_graduates <-  res_combined_graduates %>%
  select(field, same, only1, only2, diff, n_links_final) %>%
  bind_rows(across_fields_graduates) |>
  mutate(
    field = stringr::str_to_title((field)),
    across(where(is.numeric), \(x) round(x, digits = 2)) # this is an updated form of mutate_if()
    )
```

```
table_caption <- "Linking the graduates"
table_label <- "compare_linking_graduates"
table_columns <- c("Field", "Same ID", "Only by 1", "Only by 2", "Different ID", "Number of links")
footnote <- c(
  "The table summarises the links found from from ProQuest graduates to MAG authors.",
  # the compares the identified links from ProQuest to MAG across training sets by two different labell
  # "The unit of observation is the graduate in ProQuest.",
  "Graduates are defined as the authors of the dissertations in ProQuest.",
  "First, the columns headed by ``Fraction of links found'' compare the identified links across two dif
  "The columns show the fraction of links found for two training sets constructed by two different label
  "``Same ID'' are graduates for which the models trained on the different training sets find the same
  "``Only by 1'' and ``Only by 2'' are graduates for which only the model trained on either of the trai
  "``Different ID'' are graduates for which both models find links to MAG, but to different identifiers
  "Second, the last column reports the total number of links found for each field, after all post-proces
  "Third, the last row reports the total across fields. The fractions are weighted by the number of gra
)
footnote <- paste0(footnote, collapse = " ")

table_graduates <- df_table_graduates |>
  kable(format = "latex",
        digits = 2,
        booktabs = TRUE,
        caption = table_caption,
        label = table_label,
        col.names = table_columns
        )  |>
  kableExtra::row_spec(row = length(fields_to_process), hline_after = TRUE) |>
  kableExtra::add_header_above(
    header = c(" " = 1, "Fraction of links found" = 4, " " = 1),
    line = FALSE
  ) |>
  kableExtra::kable_styling(font_size = 9) |>
  kableExtra::footnote(
    general = footnote,
    footnote_as_chunk = TRUE,
    threeparttable = TRUE,
```

field	same	only1	only2	diff	diff_rel1	nlink1	nlink2	namedist_pq_1	namedist_pq_2	namedist_diff	namedist_diff_pq_1	namedist_diff_pq_2	namedist_pq_only1	namedist_pq_only2
biology	0.78	0.00	0.22	0.00	0.00	27516	35175	0.02	0.02	0.16	0.02	0.02	0.02	0.00
business	0.69	0.01	0.30	0.00	0.00	8060	11629	0.02	0.02	0.05	0.02	0.02	0.02	0.00
chemistry	0.93	0.00	0.02	0.05	0.05	11713	12138	0.02	0.02	0.02	0.02	0.02	0.02	0.00
computer science	0.89	0.08	0.02	0.01	0.01	17469	16498	0.03	0.03	0.15	0.03	0.03	0.03	0.11
economics	0.96	0.00	0.02	0.02	0.02	8446	8655	0.02	0.02	0.05	0.02	0.02	0.02	0.00
engineering	0.81	0.00	0.19	0.00	0.00	21162	26140	0.04	0.05	0.20	0.04	0.05	0.04	0.00
environmental science	0.73	0.23	0.03	0.01	0.01	9066	7312	0.03	0.03	0.07	0.03	0.03	0.03	0.22
geography	0.62	0.32	0.02	0.03	0.04	7107	5114	0.04	0.03	0.12	0.04	0.03	0.04	0.43
geology	0.84	0.00	0.14	0.01	0.01	4584	5436	0.03	0.03	0.01	0.03	0.03	0.03	0.00
history	0.85	0.01	0.12	0.02	0.03	10129	11817	0.02	0.03	0.12	0.02	0.03	0.02	0.00
materials science	0.74	0.01	0.13	0.12	0.14	14246	16079	0.07	0.05	0.13	0.07	0.05	0.07	0.00
mathematics	0.78	0.15	0.01	0.06	0.06	12759	10924	0.04	0.03	0.11	0.04	0.03	0.04	0.57
philosophy	0.93	0.00	0.06	0.00	0.00	4016	4381	0.02	0.02	0.03	0.02	0.02	0.02	0.00
physics	0.51	0.15	0.03	0.31	0.32	13955	12235	0.11	0.17	0.28	0.11	0.17	0.11	0.85
political science	0.87	0.00	0.08	0.04	0.05	13040	14002	0.02	0.05	0.27	0.02	0.05	0.02	0.00
psychology	0.91	0.00	0.08	0.00	0.00	28878	32740	0.02	0.03	0.09	0.02	0.03	0.02	0.00
sociology	0.46	0.01	0.51	0.02	0.03	8270	16343	0.03	0.02	0.21	0.03	0.02	0.03	0.00

```
    fixed_small_size = TRUE
    )
```

```
output_path <- "../../../../output/tables/"
filename <- paste0(output_path, "compare_linking_graduates.tex")
save_kable(table_graduates, file = filename)
```

Comparison for Advisors

- Compare Christoph and Flavio with Protocol and cleaned institutions

```
fields_to_process <-c( #"art", <- dropped
  "biology",
  "business",
  "chemistry",
  "computer science" ,
  "economics",
  "engineering",
  "environmental science",
  "geography",
  "geology" ,
  "history",
  "materials science",
  "mathematics",
  #"medicine",   <- dropped
  "philosophy",
  "physics",
  "political science",
  "psychology" ,
  "sociology")

linker1 = "flavio_with_protocol_cleaninst" # flavio_with_protocol_cleaninst
linker2 = "christoph_with_protocol_cleaninst" #christoph_degree0_with_protocol_updated
```

Printing out the dataframe

```
res_combined_advisors <- reduce(res, rbind)

res_combined_advisors %>%
  mutate(across(where(is.numeric), \(x) round(x, digits = 2))) %>%  # this is an updated form of mutate
  kable(format = "latex", digits = 2, booktabs = TRUE) %>%
  kable_styling(latex_options = c("striped", "scale_down"))
```

```
# res_combined_advisors %>%
#   mutate(across(where(is.numeric), \(x) round(x, digits = 2))) %>% View()
```

Make table Add total final links

```
res_combined_advisors <- res_combined_advisors |>
  left_join(advisor_link_counts |>
              rename(n_links_final = nb),
            by = c("field" = "fieldname0_mag"))
```

Summarise across fields

```
across_fields_advisors <- res_combined_advisors |>
  left_join(graduate_counts$total |>
              rename(n_graduates = nb),
            by = c("field" = "fieldname0_mag"))

across_fields_advisors <- across_fields_advisors |>
  mutate(share = n_graduates / sum(n_graduates)) |>
  summarise(
    across(all_of(c("same", "only1", "only2", "diff")),
           ~weighted.mean(.x, w = share)),
    across(all_of(c("n_links_final")),
           ~sum(.x))
  ) |>
  mutate(field = "Total")
```

```
df_table_advisors <-  res_combined_advisors %>%
  select(field, same, only1, only2, diff, n_links_final) %>%
  bind_rows(across_fields_advisors) |>
  mutate(
    field = stringr::str_to_title((field)),
    across(where(is.numeric), \(x) round(x, digits = 2)) # this is an updated form of mutate_if()
    )
```

```
table_caption <- "Linking the advisors"
table_label <- "compare_linking_advisors"
table_columns <- c("Field", "Same ID", "Only by 1", "Only by 2", "Different ID", "Number of links")
footnote <- c(
  "The table summarises the links found from from ProQuest advisors to MAG authors.",
  "An advisor is defined as one advisor name on one dissertation.",
  "First, the columns headed by ``Fraction of links found'' compare the identified links across two dif
  "The columns show the fraction of links found for two training sets constructed by two different labe
  "``Same ID'' are advisors for which the models trained on the different training sets find the same M
  "``Only by 1'' and ``Only by 2'' are advisors for which only the model trained on either of the train
  "``Different ID'' are advisors for which both models find links to MAG, but to different identifiers.
  "Second, the last column reports the total number of links found for each field, after all post-proce
  "Third, the last row reports the total across fields. The fractions are weighted by the number of gra
)
footnote <- paste0(footnote, collapse = " ")

table_graduates <- df_table_advisors |>
  kable(format = "latex",
        digits = 2,
        booktabs = TRUE,
```

```
        caption = table_caption,
        label = table_label,
        col.names = table_columns
        )  |>
  kableExtra::row_spec(row = length(fields_to_process), hline_after = TRUE) |>
  kableExtra::add_header_above(
    header = c(" " = 1, "Fraction of links found" = 4, " " = 1),
    line = FALSE
  ) |>
  kableExtra::kable_styling(font_size = 9) |>
  kableExtra::footnote(
    general = footnote,
    footnote_as_chunk = TRUE,
    threeparttable = TRUE,
    fixed_small_size = TRUE
    )
```

```
output_path <- "../../../../output/tables/"
filename <- paste0(output_path, "compare_linking_advisors.tex")
save_kable(table_graduates, file = filename)
```

Compare names manually

```
field="chemistry"
linktype="advisors"
years="19902015"

links <- compare(field, linktype, linker1, linker2, years, inspect=TRUE)
```