Performance of linking researchers to theses

Flavio & Christoph

05 September, 2022

Contents

This script makes some plots of the advisor links and saves the most plausible links to a table in the database.

```
# parameters for selecting links
min_score_advisors <- 0.7 # minimum score from dedupe
max_year_diff <- 5 # maximum difference between advisory and own publication at institution. 5 is arbit
max_uniname_distance <- 0.02 # keep only links where the jarowinkler distance between the institution n
```

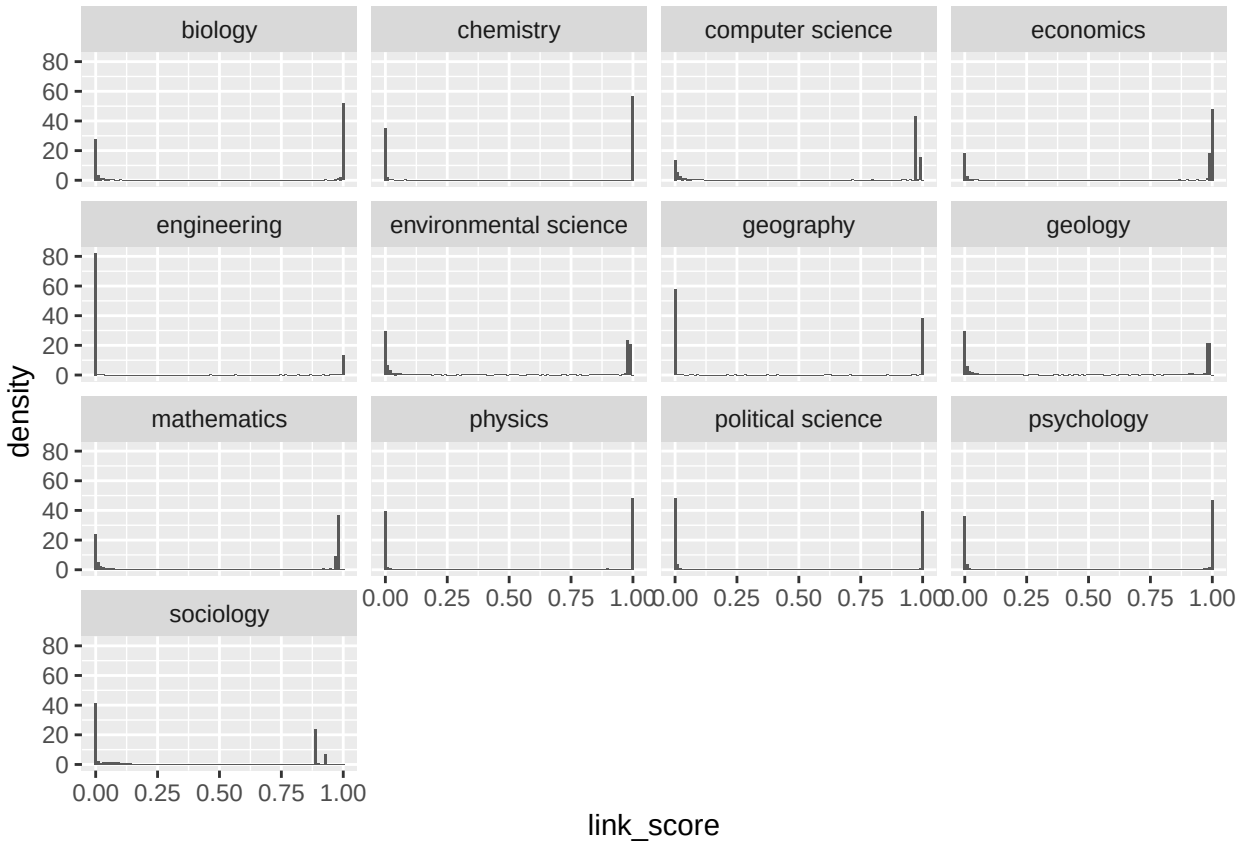
Overview

```
current_links <- collect(current_links)
linked_advisors <- collect(linked_advisors)
theses <- collect(theses)
authors_affiliation <- collect(authors_affiliation)
linking_info <- collect(linking_info)
```

Linking scores

- conditioning on link score > 0.7 is fine

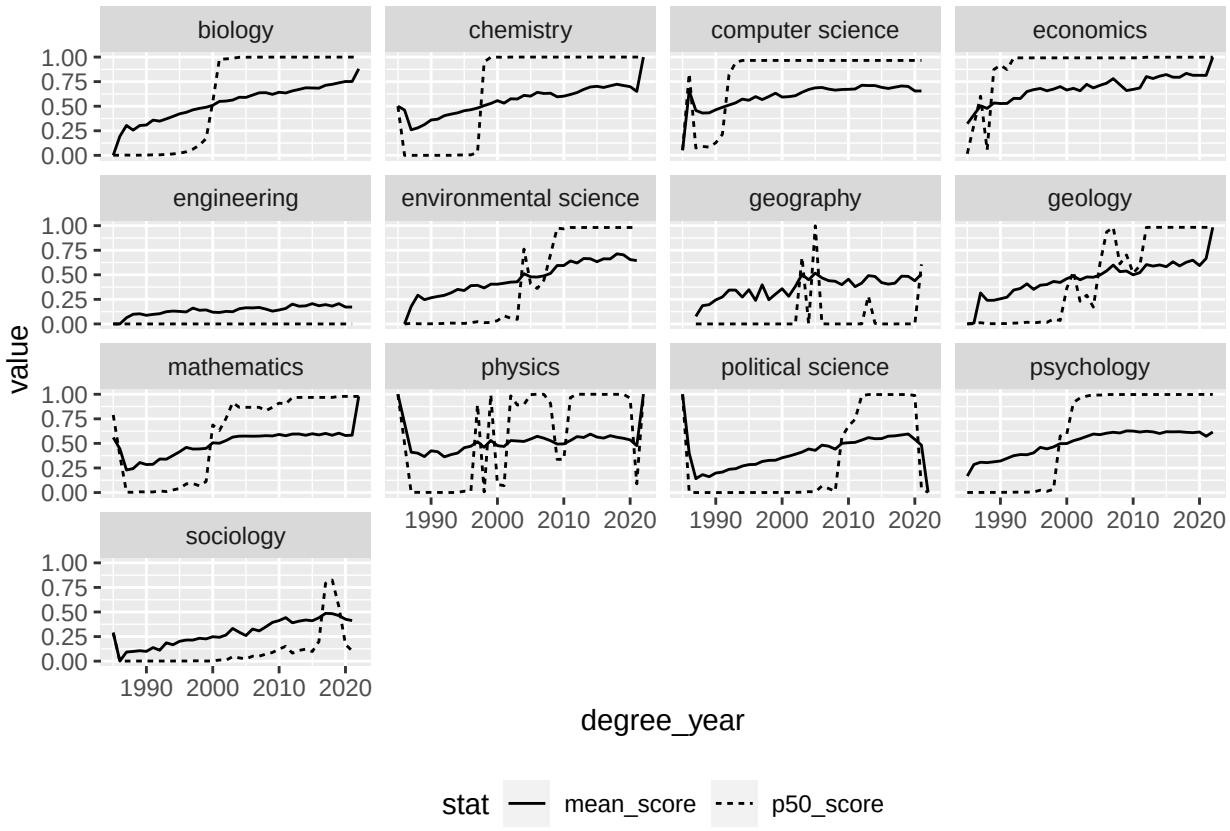
```
linked_advisors %>%
  left_join(linking_info, by = "iteration_id") %>%
  ggplot(aes(x = link_score)) +
  geom_histogram(bins = 100, aes( y = ..density..)) +
  facet_wrap(~field)
```



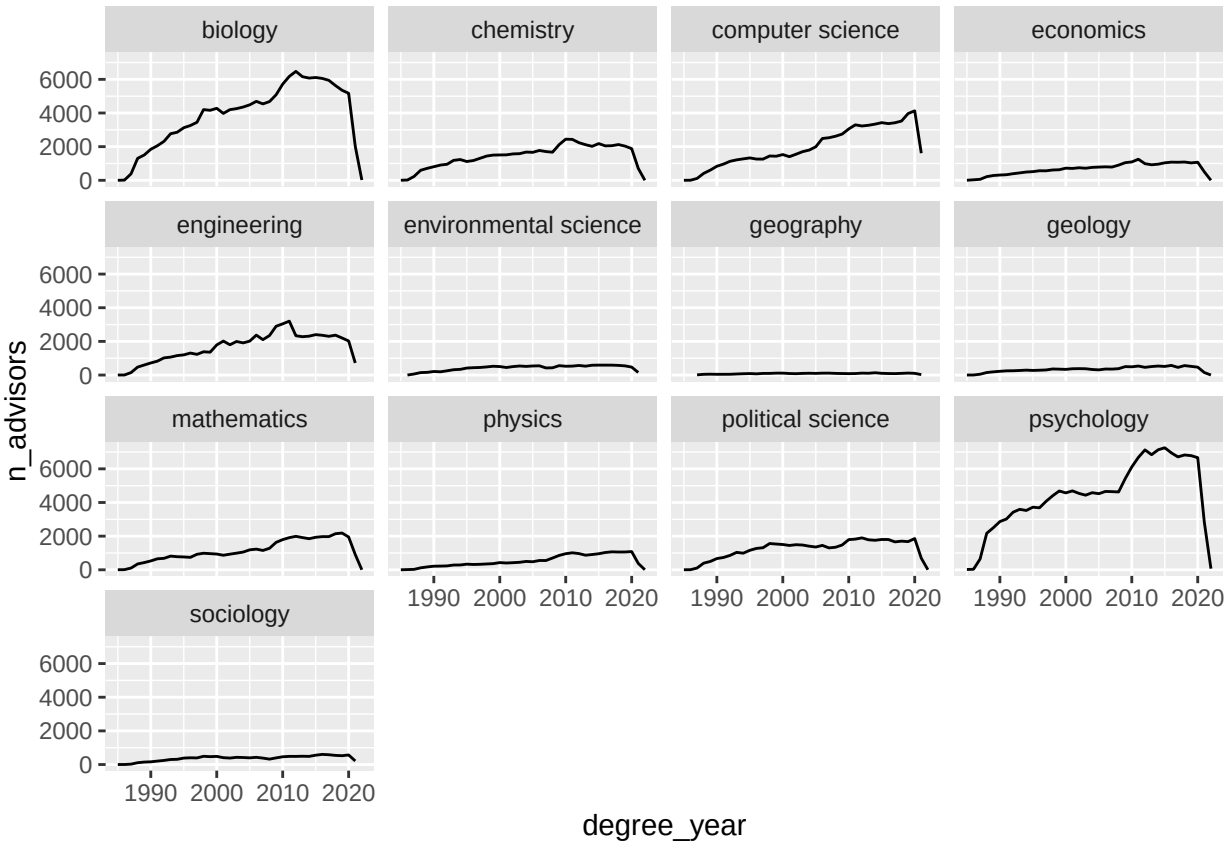
Average link score by graduation year

```
score_by_year <- linked_advisors %>%
  left_join(theses %>%
              select(degree_year, relationship_id),
            by = "relationship_id") %>%
  left_join(linking_info,
            by = "iteration_id")

score_by_year %>%
  group_by(degree_year, field) %>%
  summarise(mean_score = mean(link_score),
            p50_score = quantile(link_score, probs = 0.5),
            .groups = "drop") %>%
  pivot_longer(cols = ends_with("_score"),
               names_to = "stat") %>%
  ggplot(aes(x = degree_year, y = value)) +
  geom_line(aes(linetype = stat)) +
  facet_wrap(~field) +
  theme(legend.position = "bottom")
```




```
score_by_year %>%
  group_by(degree_year, field) %>%
  summarise(n_advisors = n(),
            .groups = "drop") %>%
  ggplot(aes(x = degree_year, y = n_advisors)) +
  geom_line() +
  facet_wrap(~field)
```




```
# another stat
  # linked advisors per linked dissertation
  # versus advisors per dissertation (in the US)

# score_by_year %>%
#   mutate(grp = ifelse(degree_year < 2000, "pre", "post")) %>%
#   ggplot(aes(x = link_score)) +
#   geom_histogram() +
#   facet_wrap(~grp)
```

Comments

- General upward trend in link_score
  - why?
  - use new method on one field, see difference
- What is happening to the median? Where does it come from?
  - The median goes from 0 to 1 mostly as a function of whether more or less than 50 percent of links have a score close to 1.
  - Possible reasons
    - We are not using the degree year for linking
    - ... ?
- for instance, a student of michael j lambert (authorid 2120159045; relationship id 303670971_0 in proquest) from pre-1990 is link score of 0.02, but should be a clear link

Check overlap of institution names and years

```
d_main <- linked_advisors %>%
  filter(link_score > min_score_advisors) %>%
  left_join(theses %>%
              mutate(fullname = paste0(firstname, " ", lastname)) %>%
              select(relationship_id, degree_year, uni_name, fullname),
            by = "relationship_id") %>%
  inner_join(current_links %>%
               select(author_name, AuthorId),
             by = "AuthorId") %>%
  # join on year; filter on max similarity within relationship_id; still need to examine multiple match
  left_join(authors_affiliation,
            by = c("AuthorId")) %>%
  mutate(dist_uni_name = stringdist(uni_name, affil_name, method = "jw"),
         dist_year = abs(degree_year - Year)) %>%
  group_by(relationship_id) %>%
  filter(dist_uni_name == min(dist_uni_name)) %>%
  filter(dist_year == min(dist_year)) %>%
  mutate(nb = n()) %>% # can still have multiple links if e.g. the dissertation is in x, but the affili
  ungroup()

d_main <- d_main %>%
  filter(!duplicated(relationship_id))

cat("Split of links by whether years are >/<", max_year_diff, "apart")
```

```
## Split of links by whether years are >/< 5 apart
```

```
table(d_main$dist_year <= max_year_diff)
```

```
##
## FALSE  TRUE
##  2321 72718
```

```
d_main <- d_main %>%
  filter(dist_year <= max_year_diff & dist_uni_name <= max_uniname_distance)
```

Note: the "usable" links are saved to the db in src/dataprep/main/link/prep_linked_data.py

Fraction with matched advisor status by cohort of own graduation

```
pq_authors <- tbl(con, "pq_authors") %>% collect()
```

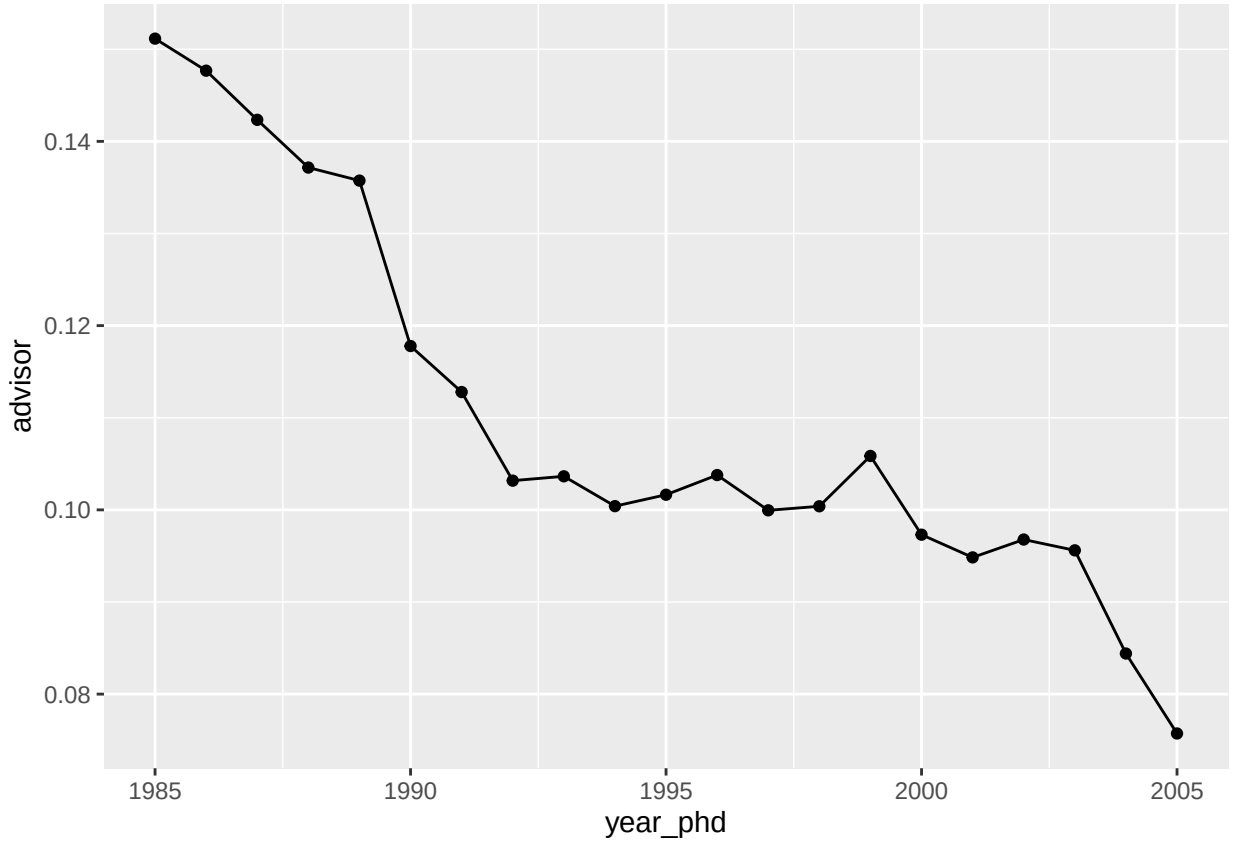
```
d_agg <- d_main %>%
  select(AuthorId, relationship_id, degree_year) %>%
  group_by(AuthorId) %>%
  filter(degree_year == min(degree_year)) %>%
  ungroup() %>%
  rename(year_firstadvisee = degree_year) %>%
  filter(!duplicated(AuthorId))
```

```
d_links <- current_links %>%
  left_join(pq_authors %>%
              select(goid, year_phd = degree_year),
```

```
            by = "goid") %>%
  left_join(d_agg, by = "AuthorId")
```

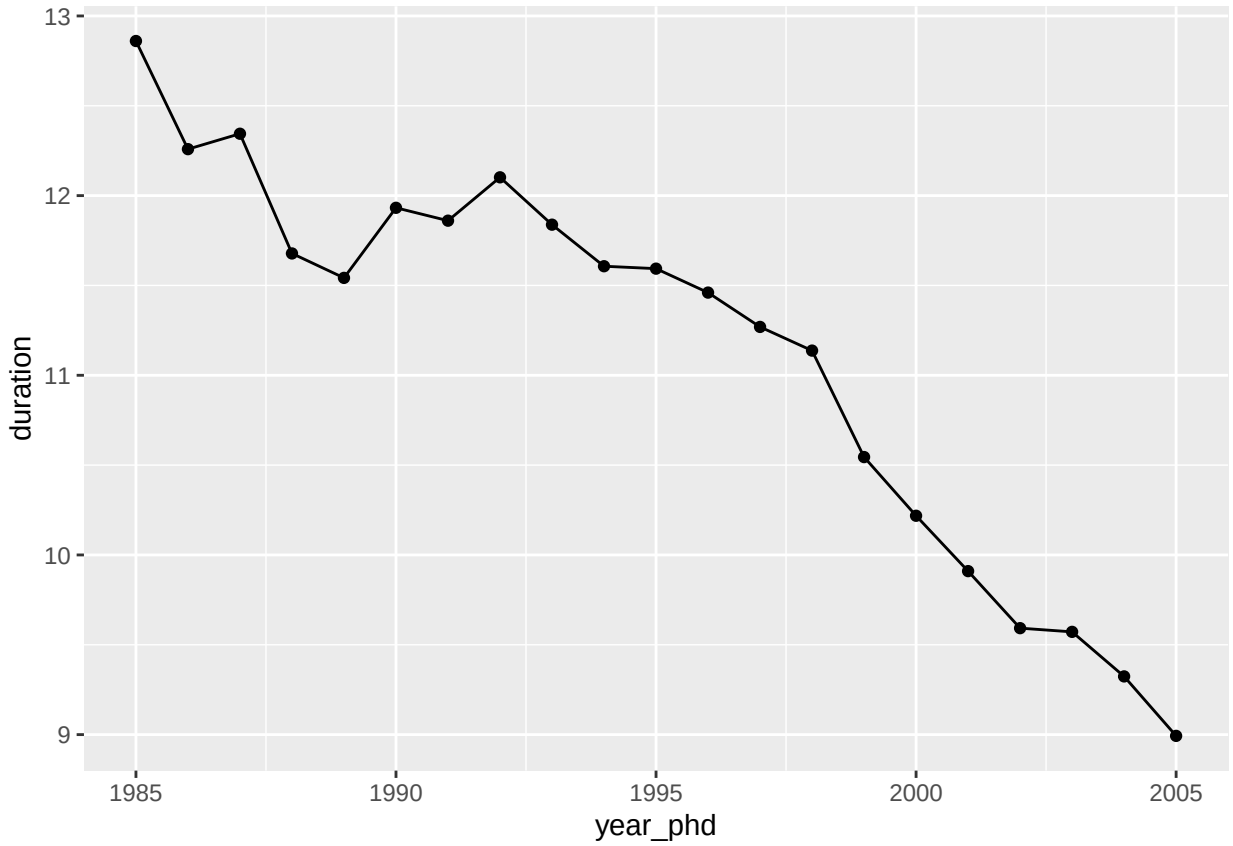
Fraction of authors that eventually becomes advisor

```
d_links %>%
  mutate(advisor = ifelse(is.na(year_firstadvisee), 0, 1)) %>%
  group_by(year_phd) %>%
  summarise(advisor = mean(advisor),
            .groups = "drop") %>%
  ggplot(aes(x = year_phd, y = advisor)) +
  geom_point() +
  geom_line()
```



Duration to advisor

```
d_links %>%
  filter(!is.na(year_firstadvisee)) %>%
  mutate(duration = year_firstadvisee - year_phd) %>%
  group_by(year_phd) %>%
  summarise(duration = mean(duration),
            .groups = "drop") %>%
  ggplot(aes(x = year_phd, y = duration)) +
  geom_point() +
  geom_line()
```


13
12
11
10
9
duration
1985
1990
1995
2000
2005
year_phd