Performance of linking researchers to NSF Grants

Flavio & Christoph & Mona

02 February, 2023

Contents

This script makes some plots of the grant links. The second graph only depicts geology so far.

```
# parameters for selecting links
min_score_grants <- 0.7 # minimum score from dedupe
```

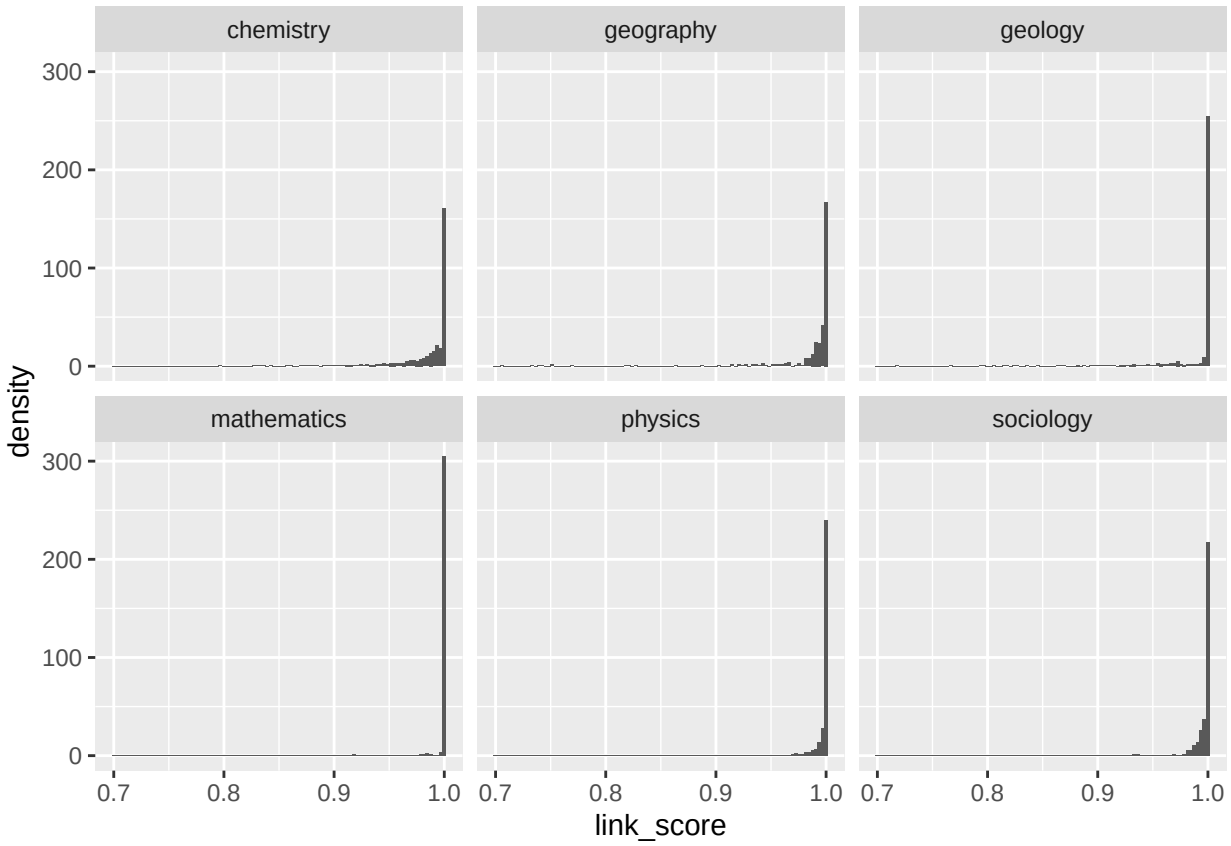
Overview

```
linked_grants_geology <- collect(links_grants_geology_mona_degree0_19902015)
linked_grants <- collect(linked_grants)
grants <- collect(grants)
linking_info <- collect(linking_info)
```

Linking scores

- conditioning on link score > 0.7 is fine

```
linked_grants %>%
  left_join(linking_info, by = "iteration_id") %>%
  filter(link_score>=0.7) %>%
  ggplot(aes(x = link_score)) +
  geom_histogram(bins = 100, aes( y = after_stat(density))) +
  facet_wrap(~field)
```



Link performance by graduation year: so far only for geology

- fraction of grants where the link_score is above the treshold
- the mean link score for grants where dedupe finds a link (link_score is not NA)
- in the figure above, we used the field from iteration_id, but this only works for grants that dedupe suggests to be a link

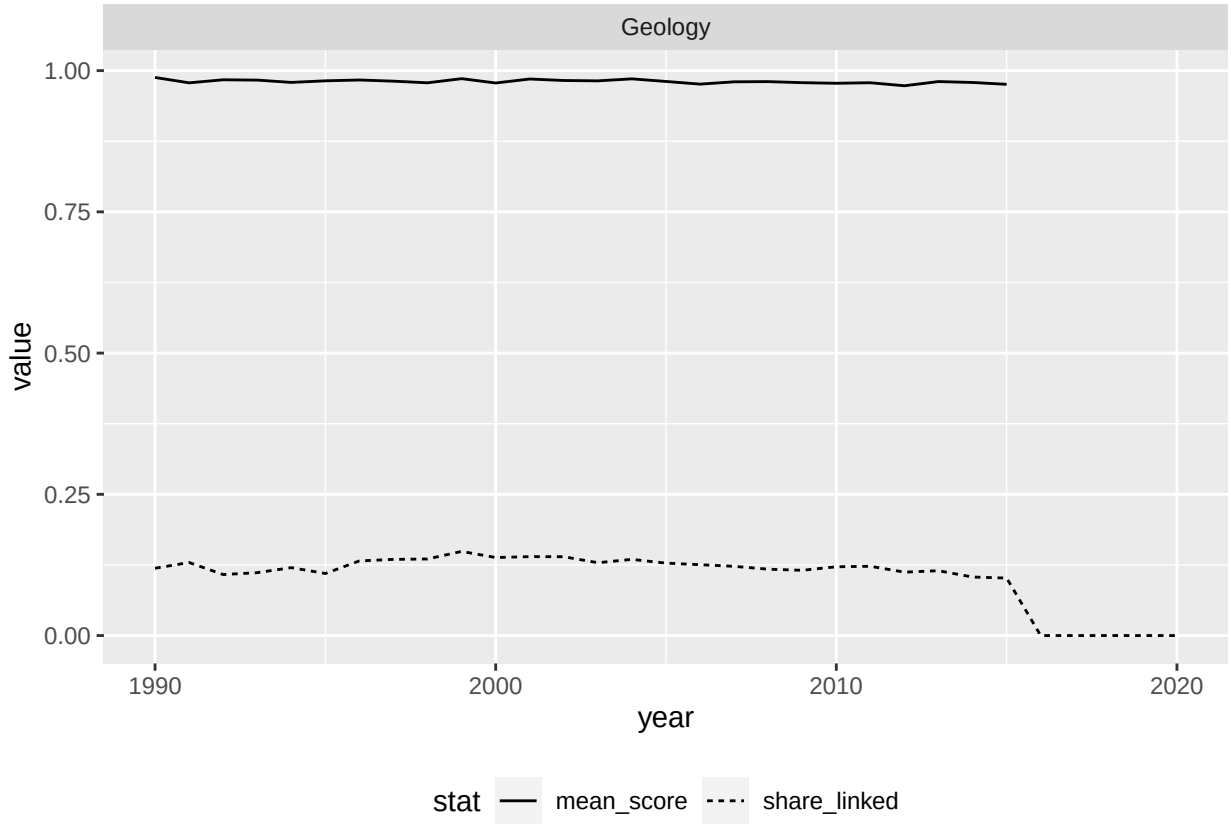
```
keep_fields <- select_fields
# c("biology", "chemistry", "computer science",
#                  "economics", "engineering", "environmental science",
#                  "geography", "geology", "mathetmatics", "physics",
#                  "political science", "psychology", "sociology")


score_by_year <- grants %>%
   left_join(linked_grants_geology,
             by = "grantid_authorposition")%>%
   filter(year >= 1985) #%>%
   #filter(link_score>=0.7) #%>%
   #filter(field %in% keep_fields)


score_by_year %>%
  mutate(link_score_adj = ifelse(is.na(link_score), -1, link_score)) %>%
  group_by(year) %>%
  #p50_score = quantile(link_score, probs = 0.5),
  summarise(mean_score = mean(link_score, na.rm = TRUE),
```

```
            share_linked = mean(link_score_adj > min_score_grants),
            .groups = "drop") %>%
  pivot_longer(cols = all_of(c("mean_score", "share_linked")),
               names_to = "stat") %>%
  ggplot(aes(x = year, y = value)) +
  geom_line(aes(linetype = stat)) +
  facet_wrap(~"Geology") +
  theme(legend.position = "bottom")
```

```
## Warning: Removed 5 rows containing missing values (`geom_line()`).
```



->