Performance of linking researchers to theses

Flavio & Christoph

04 November, 2022

Contents

Overview . . . 1
Linking scores . . . 1
Link performance by graduation year . . . 2
Notes . . . 4
Number of linked advisors . . . 5
Compare number of links across iterations within fields . . . 6
Fraction of theses with at least 1 supervisor linked to MAG . . . 8
Notes . . . 9
Note: the "usable" links are saved to the db in src/dataprep/main/link/prep_linked_data.py . . . 9

Check whether the entities exist at all in the underlying data from MAG **9**
Why are so many not in the sample? note they are all in author_sample . . . 12

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
```

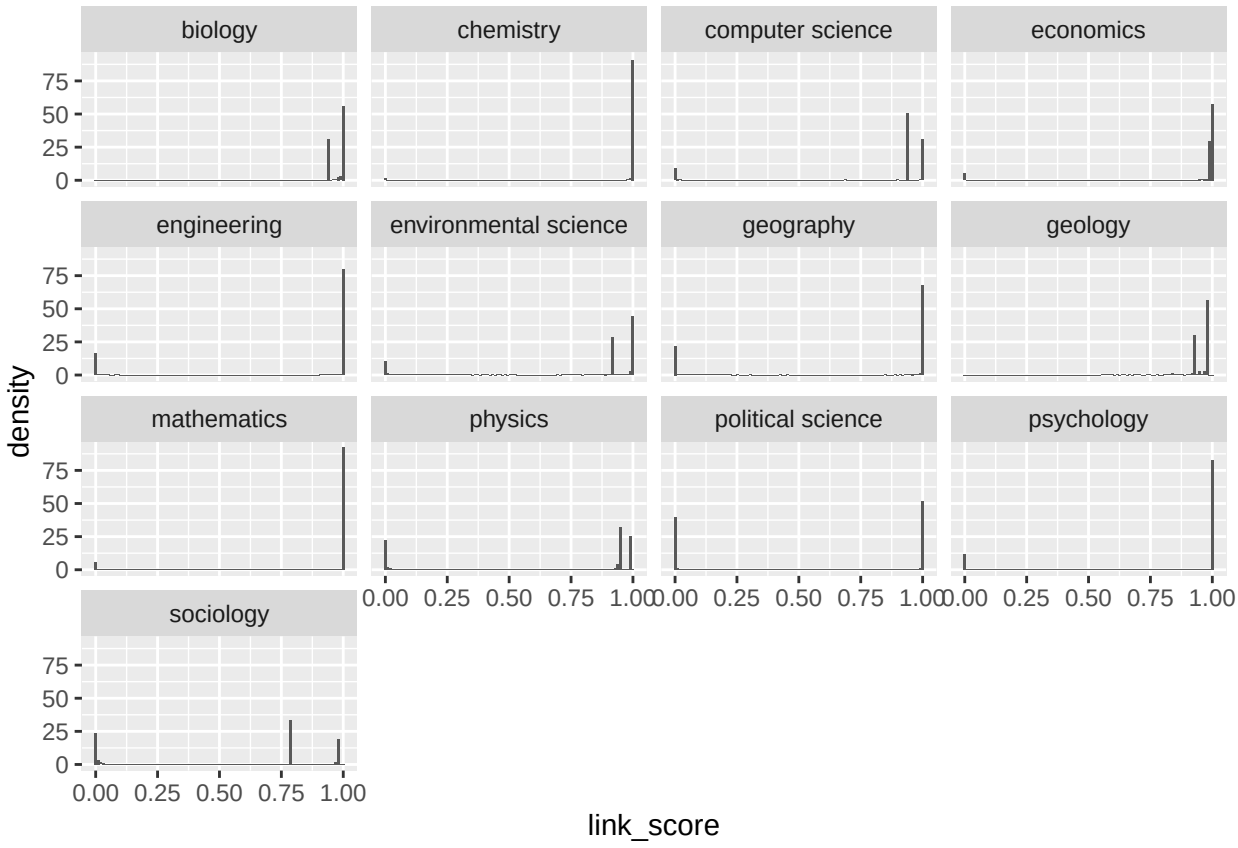
Overview

```
linked_advisors <- collect(linked_advisors)
theses <- collect(theses)
linking_info <- collect(linking_info)
pq_fields_mag <- collect(pq_fields_mag)
```

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)
```



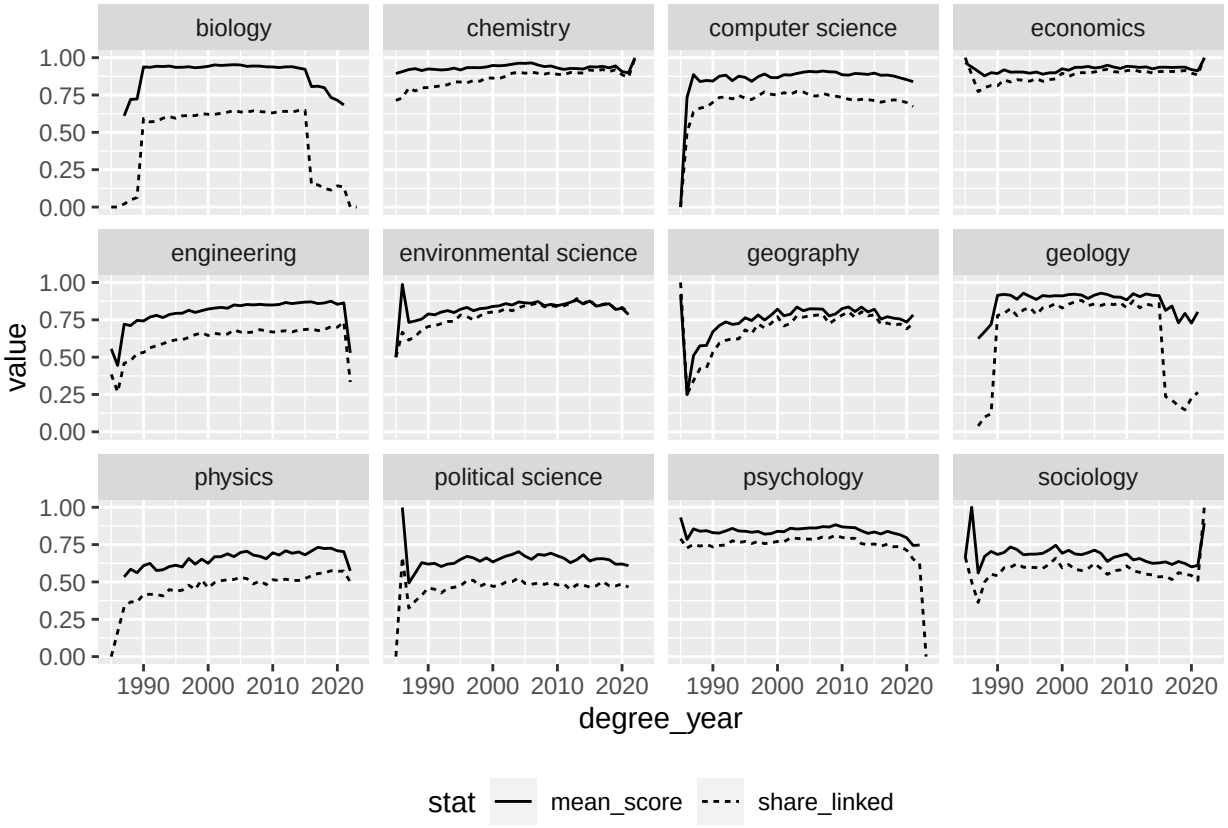
Link performance by graduation year

- fraction of listed advisors where the link_score is above the treshold
- the mean link score for advisors where dedupe finds a link (link_score is not NA)
- NOTE: the field here is assigned based on the first reported in the dissertation, and the crosswalked to the MAG field
  - in the figure above, we used the field from iteration_id, but this only works for advisors that dedupe suggests to be a link

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

score_by_year <- theses %>%
  filter(degree_year >= 1985) %>%
  left_join(linked_advisors,
            by = "relationship_id") %>%
  left_join(pq_fields_mag, by = "goid") %>%
  filter(field %in% keep_fields)

# %>%
#   left_join(linking_info,
#             by = "iteration_id")

score_by_year %>%
  mutate(link_score_adj = ifelse(is.na(link_score), -1, link_score)) %>%
```

```
  group_by(degree_year, field) %>%
  summarise(mean_score = mean(link_score, na.rm = TRUE),
            #p50_score = quantile(link_score, probs = 0.5),
            share_linked = mean(link_score_adj > min_score_advisors),
            .groups = "drop") %>%
  pivot_longer(cols = all_of(c("mean_score", "share_linked")),
               names_to = "stat") %>%
  ggplot(aes(x = degree_year, y = value)) +
  geom_line(aes(linetype = stat)) +
  facet_wrap(~field) +
  theme(legend.position = "bottom")
```

```
## Warning: Removed 2 row(s) containing missing values (geom_path).
```




```
score_by_year %>%
  filter(field == "biology") %>%
  filter(is.na(link_score)) %>%
  filter(degree_year < 2000) %>%
  group_by(firstname, lastname, uni_name, degree_year) %>%
  summarise(nb = n(),
            .groups = "drop") %>%
  arrange(desc(nb)) %>%
  head(10)
```

```
## # A tibble: 10 x 5
##    firstname  lastname  uni_name                                degre~1    nb
##    <chr>      <chr>     <chr>                                     <int> <int>
##  1 john a     gerlt     university of maryland college park        1994     6
```

```
##  2 mingdaw    tsai      ohio state university                          1997     5
##  3 naba k     gupta     university of nebraska lincoln                 1997     5
##  4 paul f     cook      university of north texas                      1993     5
##  5 c brent    theurer   university of arizona                          1997     4
##  6 c channa   reddy     pennsylvania state university                  1995     4
##  7 chawnshang chang     university of wisconsin madison                1996     4
##  8 daniel     rittschof duke university                                1997     4
##  9 david b    wake      university of california berkeley              1993     4
## 10 eric n     olson     university of texas graduate school of bi~     1995     4
## # ... with abbreviated variable name 1: degree_year
```

```
# score_by_year %>% filter(lastname == "dasgupta" & firstname == "asim" & !is.na(iteration_id)) # never
# score_by_year %>% filter(lastname == "freeling" & firstname == "michael") # never linked

# scale this up? check all the main fields of the authors with such names? -- tedious
```

Notes

- Reasons for why advisor not linked
  - they are not sampled for linking either in the mag or proquest data
    - most plausibly because they are assigned to different fields
  - institution names do not overlap
  - dedupe does not find a link even though it should
    - but how can it explain the time trend?
- Comparing fields in MAG and ProQuest dissertations
  - General
    - not linking an advisor in biology does not mean do not link them in chemistry if the thesis is also classified in chemistry
    - in the data above, this happens if biology is listed at position 0
  - Biology
    - main field biology: dasgupta, freeling
    - at least one of the dissertations of freeling are sampled for the linking
  - Sociology
    - different main field: ishisaka, coulton (medicine), howell (geography), mindel (psychology)
    - not in MAG, but findable on google: khleif, gullerud
    - not in MAG, not findable on google: liff
- Next steps
  - widen the sampled field in MAG
  - re-train and re-check

Here is some python code to look at the learned settings, based on

- https://github.com/dedupeio/rlr/blob/master/rlr/lr.py (new dedupe does not use this anymore I think)
- https://github.com/dedupeio/dedupe/blob/5742efc7fc696c06d3327e038541532e584551a8/dedupe/api.py
- Note: The predicates are similar for all three fields I looked at. I do not know how the weights correspond to the logit regression coefficients

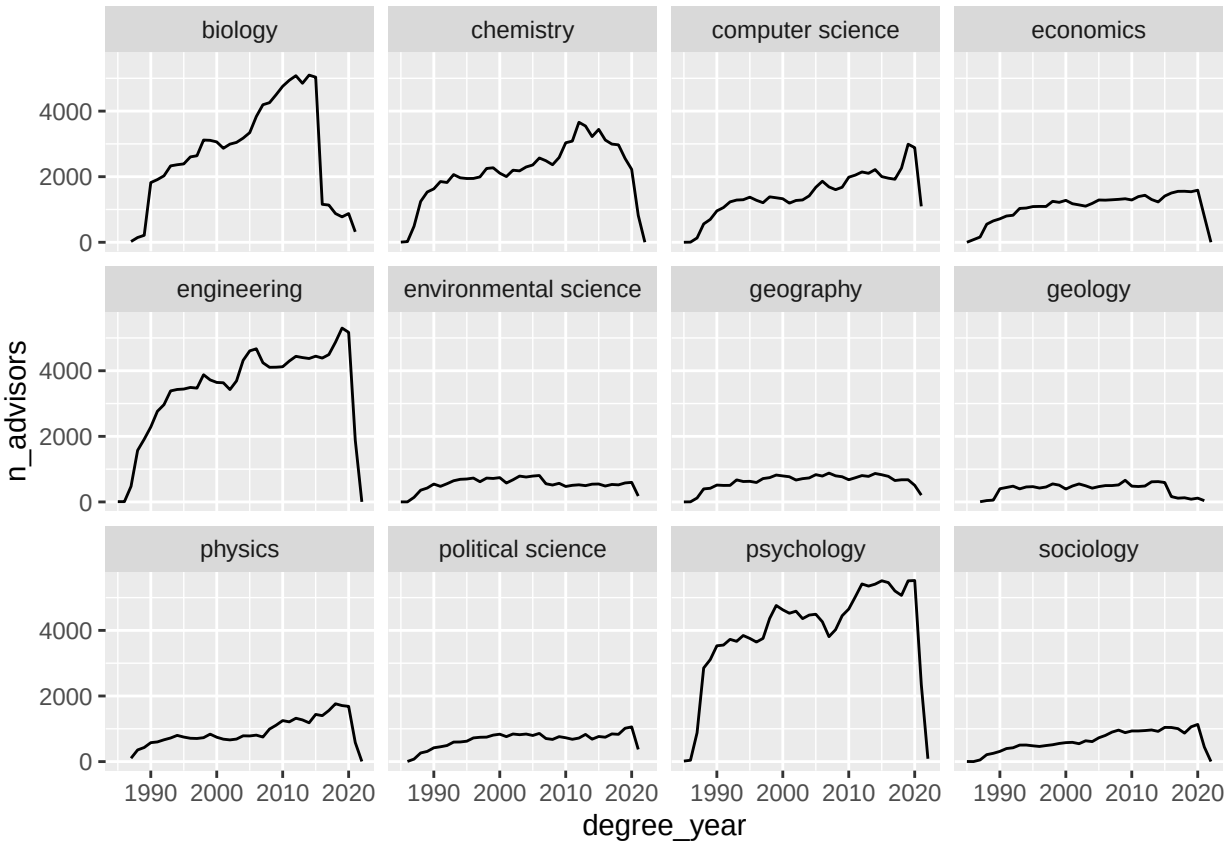
```
sf_biology = "/mnt/ssd/DedupeFiles/advisors/settings_biology_1985_2022_institutionTrue_fieldofstudy_cat
sf_chemistry = "/mnt/ssd/DedupeFiles/advisors/settings_chemistry_1985_2022_institutionTrue_fieldofstudy
sf_cs = "/mnt/ssd/DedupeFiles/advisors/settings_computer_science_1985_2022_institutionTrue_fieldofstudy

with open(sf_biology, "rb") as sf:
    linker_biology = dedupe.StaticRecordLink(sf)
```

```
with open(sf_chemistry, "rb") as sf:
    linker_chemistry = dedupe.StaticRecordLink(sf)


with open(sf_cs, "rb") as sf:
    linker_cs = dedupe.StaticRecordLink(sf)

linker_biology.predicates
linker_chemistry.predicates
linker_cs.predicates

linker_biology.classifier.weights
linker_chemistry.classifier.weights
linker_cs.classifier.weights
```

Number of linked advisors

- not sure this is still relevant?

```
score_by_year %>%
  filter(!is.na(link_score)
         & field %in% keep_fields) %>%
  group_by(degree_year, field) %>%
  summarise(n_advisors = n(),
            .groups = "drop") %>%
  ggplot(aes(x = degree_year, y = n_advisors)) +
  geom_line() +
  facet_wrap(~field)
```



old comments

- 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

Compare number of links across iterations within fields

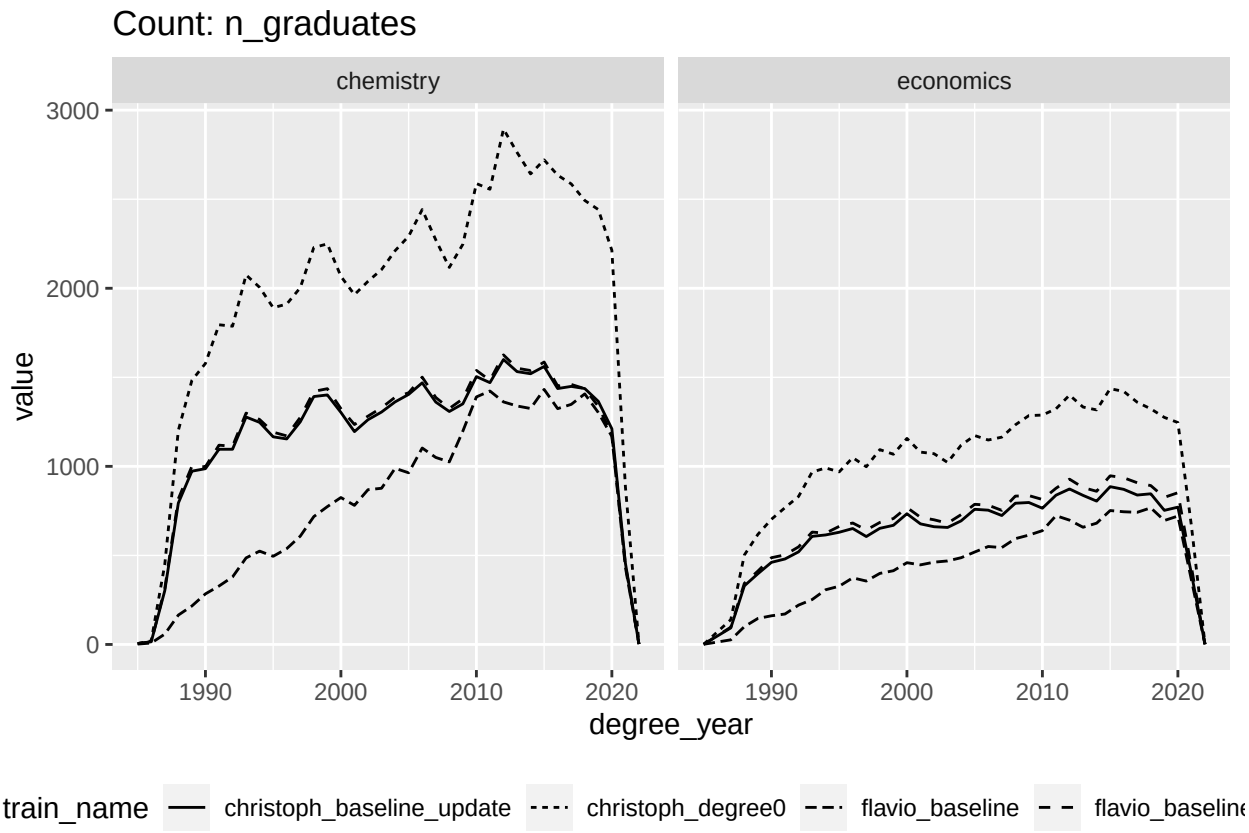
```
fields_iter_compare <- c("economics", "chemistry")
min_score <- 0.8

keep_iter_ids <- tbl(con, "linking_info_advisors") %>%
  filter(field %in% fields_iter_compare) %>%
  filter(testing == 0) %>%
  collect() %>%
  group_by(field, train_name) %>%
  arrange(iteration_id) %>%
  mutate(nb = n(),
         id = row_number()) %>%
  ungroup() %>%
  filter(id == nb) %>%
  select(iteration_id, field, train_name)

linked_ids_to_compare <- tbl(con, "linked_ids_advisors") %>%
  inner_join(
    tbl(con, "linking_info_advisors") %>%
      filter(field %in% fields_iter_compare),
    by = "iteration_id"
```

```
  ) %>%
  inner_join(
    tbl(con, "pq_advisors") %>%
      select(relationship_id, goid),
    by = "relationship_id"
  ) %>%
  inner_join(
    tbl(con, "pq_authors") %>%
      select(goid, degree_year),
    by = "goid"
  ) %>%
  collect() %>%
  filter(iteration_id %in% keep_iter_ids$iteration_id)
```

Number of graduates with at least 1 advisor

```
d_sum <- linked_ids_to_compare %>%
  filter(link_score >= min_score) %>%
  group_by(train_name, field, degree_year) %>%
  summarise(n_advisors = n(),
            n_graduates = n_distinct(goid),
            .groups = "drop") %>%
  pivot_longer(cols = starts_with("n_"), names_to = "variable")

plotvars <- c("n_graduates")

map(.x = plotvars,
    .f = ~d_sum %>%
      filter(variable == .x) %>%
      ggplot(aes(x = degree_year, y = value)) +
      geom_line(aes(linetype = train_name)) +
      facet_wrap(~field) +
      theme(legend.position = "bottom") +
      labs(title = paste0("Count: ", .x))
    )
```

```
## [[1]]
```

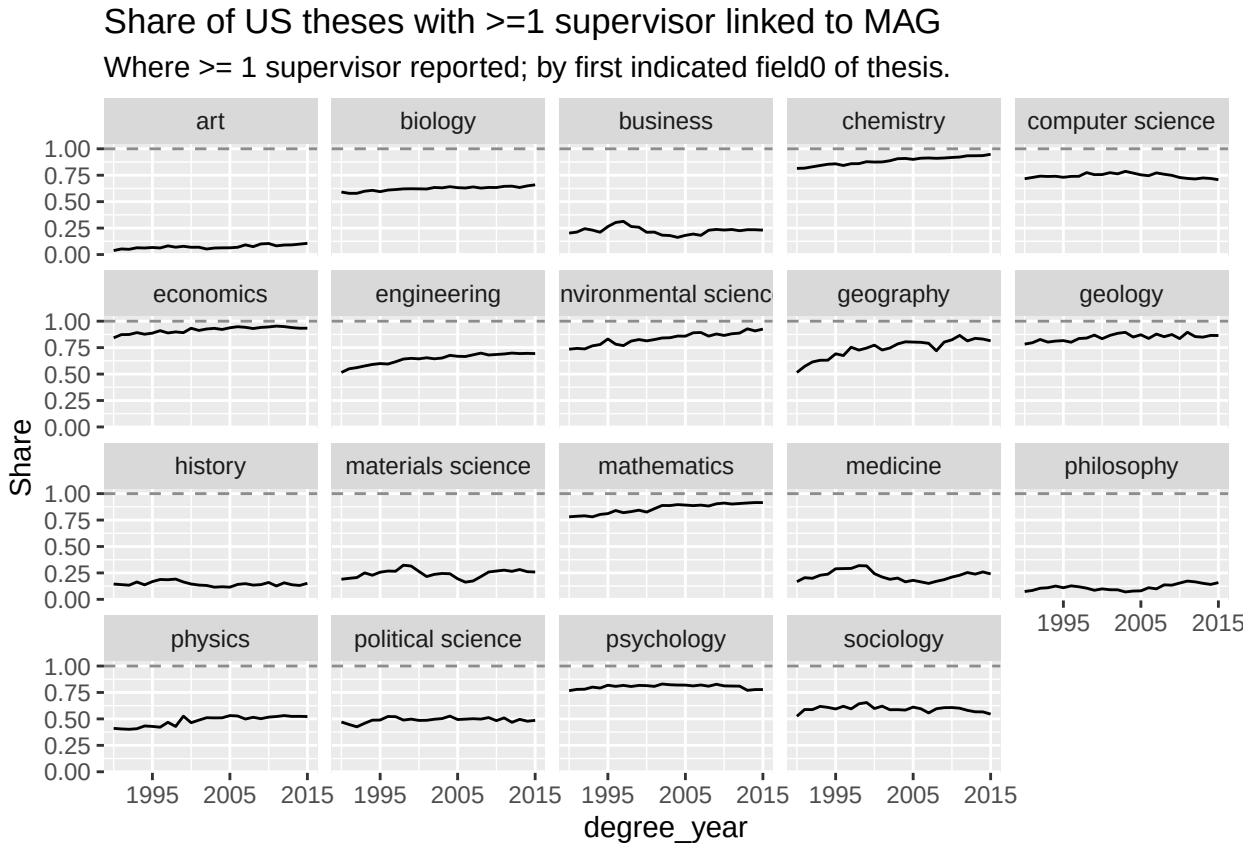


Fraction of theses with at least 1 supervisor linked to MAG

```
s_thesis_advisor_link <- theses %>%
  filter(degree_year %in% 1990:2015) %>%
  left_join(linked_advisors %>%
              filter(link_score > min_score_advisors) %>%
              select(relationship_id) %>%
              mutate(linked = 1),
            by = "relationship_id") %>%
  mutate(linked = ifelse(is.na(linked), 0, linked)) %>%
  group_by(goid) %>%
  mutate(any_link = max(linked)) %>%
  ungroup() %>%
  filter(!duplicated(goid)) %>%
  group_by(degree_year, any_link, fieldname0_mag) %>%
  summarise(n_theses = n(),
            .groups = "drop") %>%
  group_by(degree_year, fieldname0_mag) %>%
  mutate(s = n_theses / sum(n_theses)) %>%
  ungroup() %>%
  filter(any_link == 1)
```

```
s_thesis_advisor_link %>%
  ggplot(aes(x = degree_year, y = s)) +
  geom_line() +
  facet_wrap(~fieldname0_mag) +
  scale_x_continuous(breaks = c(1995, 2005, 2015)) +
```

```
  geom_hline(yintercept = 1, color = "grey55", linetype = "dashed") +
  labs(y = "Share", title = "Share of US theses with >=1 supervisor linked to MAG",
       subtitle = "Where >= 1 supervisor reported; by first indicated field0 of thesis.")
```



Notes

- Idea: since supervisors tend to be established researchers and publish regularly, we should find a large fraction of supervisors reported in ProQuest in the MAG data.
- The split by field is not exact because the link may have been found using a different reported field0.
- The close to 100% is reassuring of the MAG data quality on affiliations in these fields.
- Fields of concern: physics, sociology, poli science, biology (the level, the break and the trend).

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

Check whether the entities exist at all in the underlying data from MAG

Copy-paste the sql query for now

```
fieldofstudy_id <- 86803240
query_mag <- paste0(
  "SELECT f.AuthorId
        , f.year
        , f.YearLastPub
        , f.firstname
```

```
    , f.lastname
    , CASE TRIM(SUBSTR(f.middle_lastname, 1, f.l_fullname - f.l_firstname - f.l_lastname - 1))
        WHEN
            '' THEN NULL
            ELSE TRIM(SUBSTR(f.middle_lastname, 1, f.l_fullname - f.l_firstname - f.l_lastname - 1)
        END as middlename
    , f.fieldofstudy
    , g.keywords
    , g.coauthors
    , g.institution
    , g.main_us_institutions_year
FROM (
    SELECT a.AuthorId
        , a.YearFirstPub AS year
        , a.YearLastPub
        , a.FirstName AS firstname
        , REPLACE(b.NormalizedName, RTRIM(b.NormalizedName, REPLACE(b.NormalizedName, ' ', '')), ''
        , TRIM(SUBSTR(b.NormalizedName, length(a.FirstName) + 1)) AS middle_lastname
        , length(b.NormalizedName) as l_fullname
        , length(a.FirstName) as l_firstname
        , length(REPLACE(b.NormalizedName, RTRIM(b.NormalizedName, REPLACE(b.NormalizedName, ' ', '
        , e.NormalizedName AS fieldofstudy
    FROM author_sample AS a
    INNER JOIN (
        SELECT AuthorId, NormalizedName
        FROM Authors
    ) AS b USING(AuthorId)
    INNER JOIN (
        SELECT AuthorId
        FROM author_field0
        WHERE FieldOfStudyId_lvl0 =", fieldofstudy_id, "
            AND Degree <= 0
    ) USING(AuthorId)
    LEFT JOIN (
        SELECT AuthorId, NormalizedName
        FROM author_fields c
        INNER JOIN (
            SELECT FieldOfStudyId, NormalizedName
            FROM FieldsOfStudy
        ) AS d USING(FieldOfStudyId)
        INNER JOIN (
            SELECT ParentFieldOfStudyId, ChildFieldOfStudyId
            FROM crosswalk_fields
            WHERE ParentLevel = 0
                AND ParentFieldOfStudyId = ", fieldofstudy_id, "
        ) AS e ON (e.ChildFieldOfStudyId = c.FieldOfStudyId)
        WHERE FieldClass = 'first'
    ) AS e USING(AuthorId)
) f
LEFT JOIN (
    SELECT AuthorId
            , main_us_institutions_career as institution
            , coauthors
```

```
                , keywords
                , main_us_institutions_year
                , all_us_institutions_year
        FROM author_info_linking
    ) AS g USING(AuthorId)
    WHERE length(firstname) > 1 AND year >= 1985 AND year <= 2015 + 5 AND institution is not NULL
    AND institution not like '%chinese academy of sciences%'
    "
)

advisor_sample_mag <- tbl(con, sql(query_mag))
advisor_sample_mag <- collect(advisor_sample_mag)
```

```
unlinked_advisors <- score_by_year %>%
  filter(field == "biology"
         & is.na(link_score)) %>%
  group_by(firstname, lastname, uni_name, degree_year) %>%
  summarise(nb = n(),
            .groups = "drop")

dk <- unlinked_advisors %>%
  left_join(advisor_sample_mag %>%
              select(AuthorId, year, firstname, lastname, main_us_institutions_year),
            by = c("firstname", "lastname", "degree_year" = "year"))
```

```
head(dk %>% filter(is.na(AuthorId) & degree_year <= 2015) %>% arrange(desc(nb)), 10)
```

```
## # A tibble: 10 x 7
##    firstname  lastname uni_name                    degre~1    nb Autho~2 main_~3
##    <chr>      <chr>    <chr>                         <int> <int> <int64> <chr>
##  1 andrew     fire     stanford university            2012     9      NA <NA>
##  2 benjamin c stark    illinois institute of tech~    2002     8      NA <NA>
##  3 garry      nolan    stanford university            2010     7      NA <NA>
##  4 calvin     kuo      stanford university            2012     6      NA <NA>
##  5 dagmar     ringe    brandeis university            2000     6      NA <NA>
##  6 john a     gerlt    university of maryland col~    1994     6      NA <NA>
##  7 julien     sage     stanford university            2012     6      NA <NA>
##  8 mark       davis    stanford university            2010     6      NA <NA>
##  9 w          nelson   stanford university            2010     6      NA <NA>
## 10 alexander  dunn     stanford university            2015     5      NA <NA>
## # ... with abbreviated variable names 1: degree_year, 2: AuthorId,
## #   3: main_us_institutions_year
```

Do these entities exist in MAG? - andrew fire 2012 in stanford. yes. not in advisor sample. authorid 683352831. - benjamin c start 2002 at iit. yes. has some duplicates. is in advisor sample. authorid 1992276655. - garry nolan 2010 stanford. yes. not in advisor sample. authorid 1989754750. -

```
head(dk %>% filter(is.na(AuthorId) & degree_year <= 2000) %>% arrange(desc(nb)), 10)
```

```
## # A tibble: 10 x 7
##    firstname  lastname  uni_name                   degre~1    nb Autho~2 main_~3
##    <chr>      <chr>     <chr>                        <int> <int> <int64> <chr>
##  1 dagmar     ringe     brandeis university           2000     6      NA <NA>
##  2 john a     gerlt     university of maryland co~    1994     6      NA <NA>
##  3 mingdaw    tsai      ohio state university         1997     5      NA <NA>
```

```
##  4 naba k     gupta     university of nebraska li~    1997     5      NA <NA>
##  5 paul f     cook      university of north texas     1993     5      NA <NA>
##  6 c brent    theurer   university of arizona         1997     4      NA <NA>
##  7 c channa   reddy     pennsylvania state univer~    1995     4      NA <NA>
##  8 chawnshang chang     university of wisconsin m~    1996     4      NA <NA>
##  9 daniel     rittschof duke university               1997     4      NA <NA>
## 10 david b    wake      university of california ~    1993     4      NA <NA>
## # ... with abbreviated variable names 1: degree_year, 2: AuthorId,
## #   3: main_us_institutions_year
```

- dagmar ringe 2000 brandeis. yes. has some duplicates. not in advisor sample. authorid 2171354986
- naba k gupta 1997 at nebraska lincoln. yes. has some duplicates. not in advisor sample. authorid 2298396241

Why are so many not in the sample? note they are all in author_sample

- not in biology. no, they all have field level 0 biology.
- none of them has chinese academy of sciences affiliation
- other filtering
  - all of the examples above start their publishing career before 1985. Our query for the linking wrongly filters them out.

Here is some sql code that I used for checking the cases of the non-linked biology advisors

```sql
-- 1. they are not in the sample
-- 2. they are not recognized as links: (a) the model is wrong, (b) the data are wrong (ie, dedupe is c
-- the fact that the same entities are not linked even after our improvements suggests perhaps that we

-- from older checks:
-- john a gerlt: actually registered as "j a gerlt" in MAG (authorid 93129757). dedupe links him to aut
-- asim dasgupta: authorid 2150423063;

-- authors: 2298396241, 2171354986, 683352831, 1989754750

-- all of them are in biology
select count(distinct authorid)
from author_field0 a
inner join (
    select fieldofstudyid, normalizedname
    from fieldsofstudy
) b on (a.fieldofstudyid_lvl0 = b.fieldofstudyid)
where authorid in (2298396241, 2171354986, 683352831, 1989754750)
and a.Degree = 0
and normalizedname = "biology"

-- none of them are in chinese academy of sciences
select count(*)
from author_info_linking
where authorid in (2298396241, 2171354986, 683352831, 1989754750)
and main_us_institutions_career not like "chinese academy of sciences"

-- only two have no missing information on the first field. but is this relevant? -> clearly not, as ot
SELECT AuthorId, NormalizedName
FROM author_fields c
INNER JOIN (
```

```
    SELECT FieldOfStudyId, NormalizedName
    FROM FieldsOfStudy
) AS d USING(FieldOfStudyId)
-- ## Condition on fieldofstudy being in the level 0 id_field
INNER JOIN (
    SELECT ParentFieldOfStudyId, ChildFieldOfStudyId
    FROM crosswalk_fields
    WHERE ParentLevel = 0
        AND ParentFieldOfStudyId IN (86803240)
) AS e ON (e.ChildFieldOfStudyId = c.FieldOfStudyId)
WHERE FieldClass = 'first'
and authorid in (2298396241, 2171354986, 683352831, 1989754750)
limit 10;

-- they all start publishing before 1985
select *
from author_sample
where authorid in (2298396241, 2171354986, 683352831, 1989754750)
```