Performance of linking graduates to researchers

Flavio & Christoph

02 September, 2024

Contents

This document compares the links we obtain for all fields in the latest iteration. But it does not consider the further processing done in `prep_linked_data.py`. For better information about the final linked sample, see `quality_linking_graduates_chemistry.Rmd`.

Overview

SQL example for sourcing number of authors with same name

```
select *
from author_sample
inner join (
    select authorid, normalizedname, papercount, citationcount
    from authors
    where normalizedname = "lawrence b slobodkin"
) using (authorid)
inner join (
    select authorid, fieldofstudyid
    from author_fields
    where fieldclass = "first"
) using (authorid)
```

Which linking iterations to keep?

```
keep_iter_ids_base <- linking_info %>%
  filter(date <= date_method_change
         & keywords == "False"
         )

keep_iter_ids_revise <- linking_info %>%
  filter(date > date_method_change
        & keywords == "True"
        ) %>%
  # keep only the latest iteration here
  group_by(field) %>%
  filter(iteration_id == max(iteration_id)) %>%
  ungroup()
stopifnot(nrow(keep_iter_ids_revise) == n_distinct(keep_iter_ids_revise$field))

keep_iter_ids <- list(
  base = keep_iter_ids_base,
  revise = keep_iter_ids_revise
)

keep_iter_ids <- map(
  .x = keep_iter_ids,
  .f = ~.x %>%
    filter(field %in% select_fields) %>%
    pull(iteration_id)
)

linked_ids <- map(
  .x = keep_iter_ids,
  .f = ~linked_ids %>%
    filter(iteration_id %in% .x)
)
```

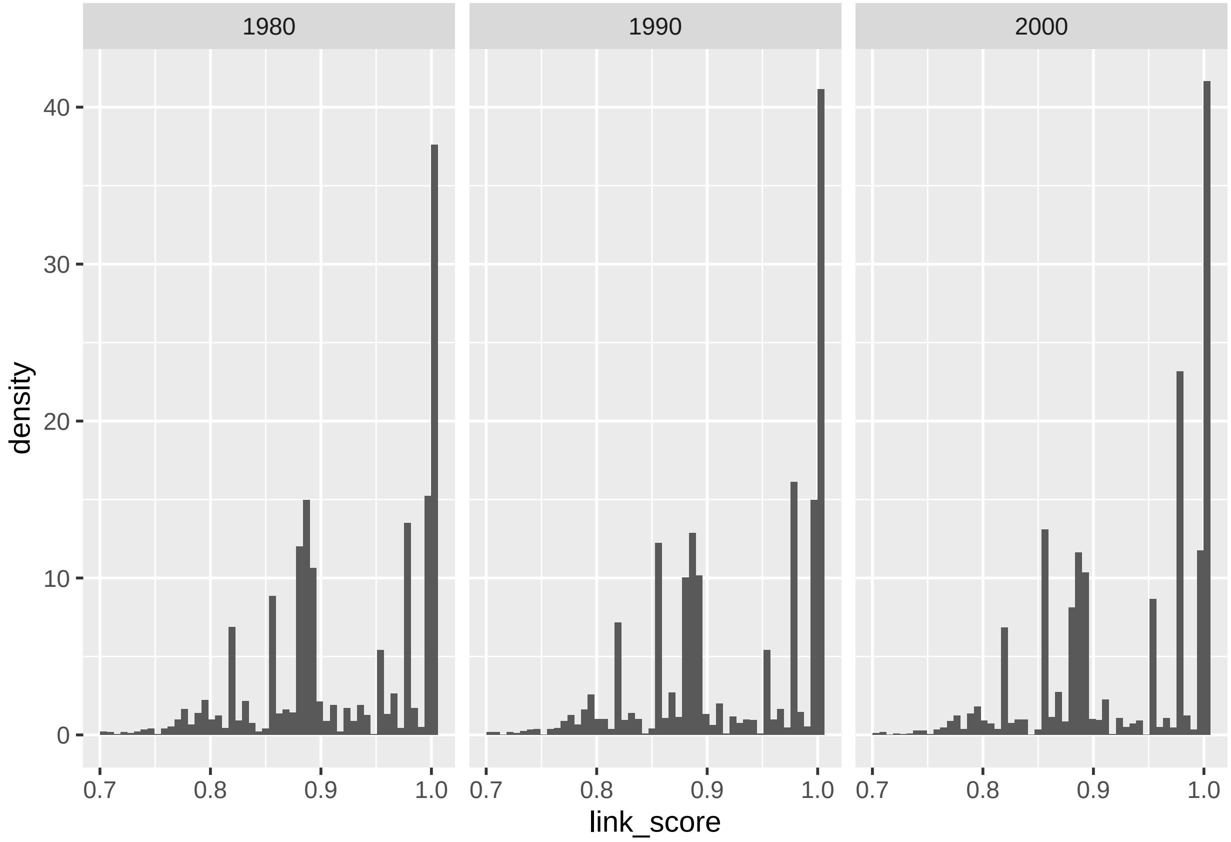
```
d_links <- map(
  .x = linked_ids,
  .f = ~.x %>%
    left_join(mag_authors %>%
                select(AuthorId,
                       year_mag = year,
                       firstname_mag = firstname,
                       lastname_mag = lastname,
                       field_mag = fieldofstudy,
                       field0_mag = mag_field0),
              by = "AuthorId") %>%
    left_join(pq_authors %>%
                select(goid,
                       year_pq = year,
                       firstname_pq = firstname,
                       lastname_pq = lastname,
                       field_pq = fieldofstudy,
                       field0_pq = mag_field0),
              by = "goid") %>%
    mutate(year_diff = year_mag - year_pq,
           same_firstname = ifelse(firstname_mag == firstname_pq, 1, 0),
           same_lastname = ifelse(lastname_mag == lastname_pq, 1, 0)) %>%
    left_join(field_names_id %>%
                rename(main_field = NormalizedName),
              by = c("field0_pq" = "FieldOfStudyId")) %>%
    filter(goid != 305107842)  %>% #  this is some author which was linked but should not have been in f
    filter(link_score > min_link_score
           & abs(year_diff) <= max_year_diff)

  )

d_links$base <- d_links$base %>% filter(year_pq <= 2005)
```

Some histograms

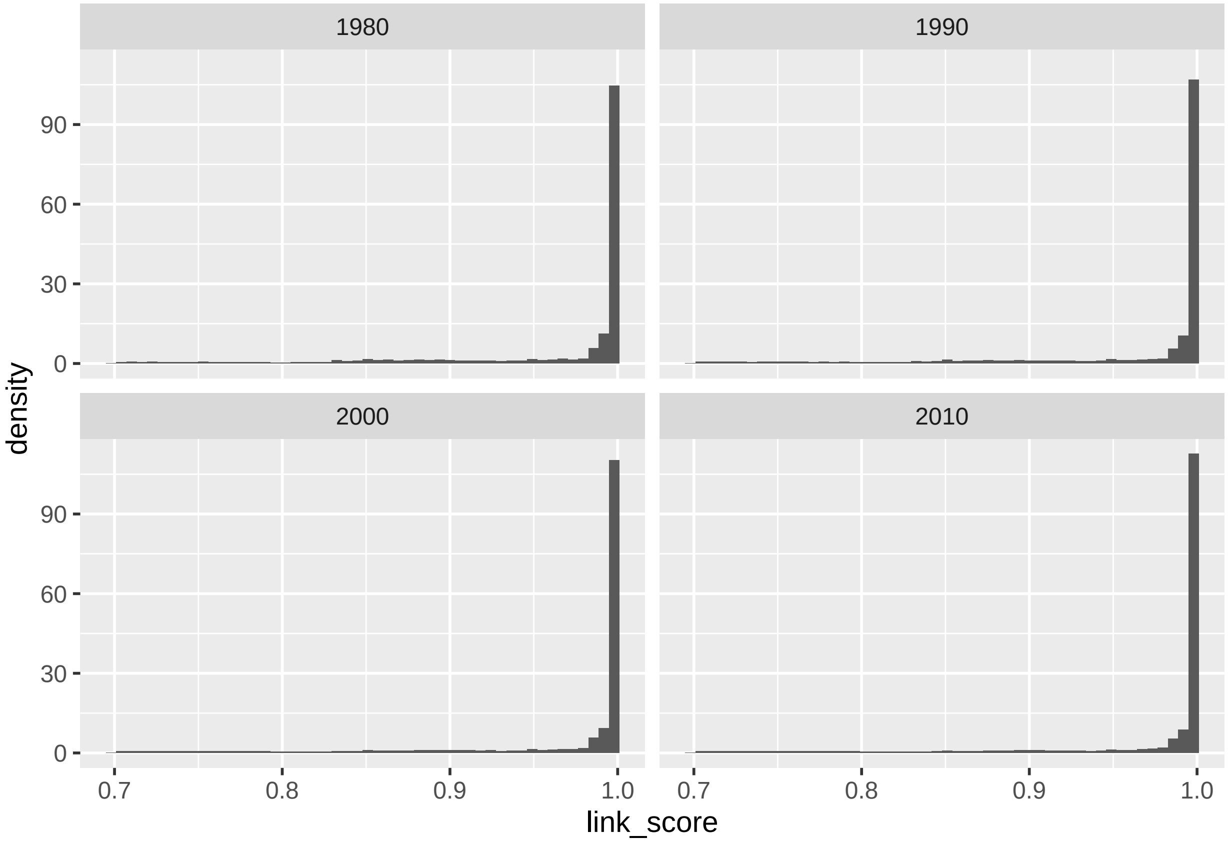
link score by field

```
## $base
```

```
## Warning: The dot-dot notation (`..density..`) was deprecated in ggplot2 3.4.0.
## i Please use `after_stat(density)` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



```
##
## $revise
```

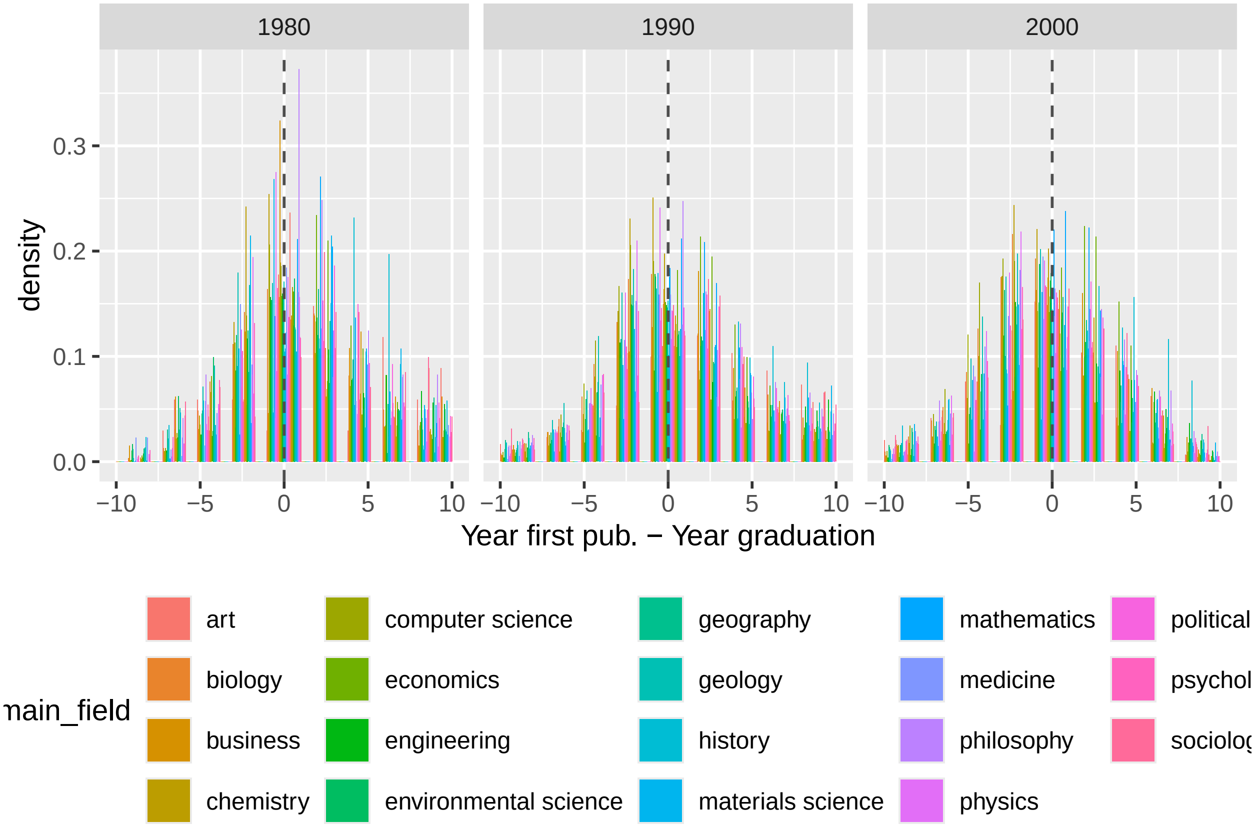


Year between first pub and graduation

- why are there other fields than maths/biology for the following two figures?
- this is because we sample persons whenever they are in any of the linking fields
  - thus, a graduate can be linked in a biology iteration if her first field is chemistry
  - compare this with the advisor links!
  - this also means the join above should take care of this, and indicate the multiplicity of the graduates!

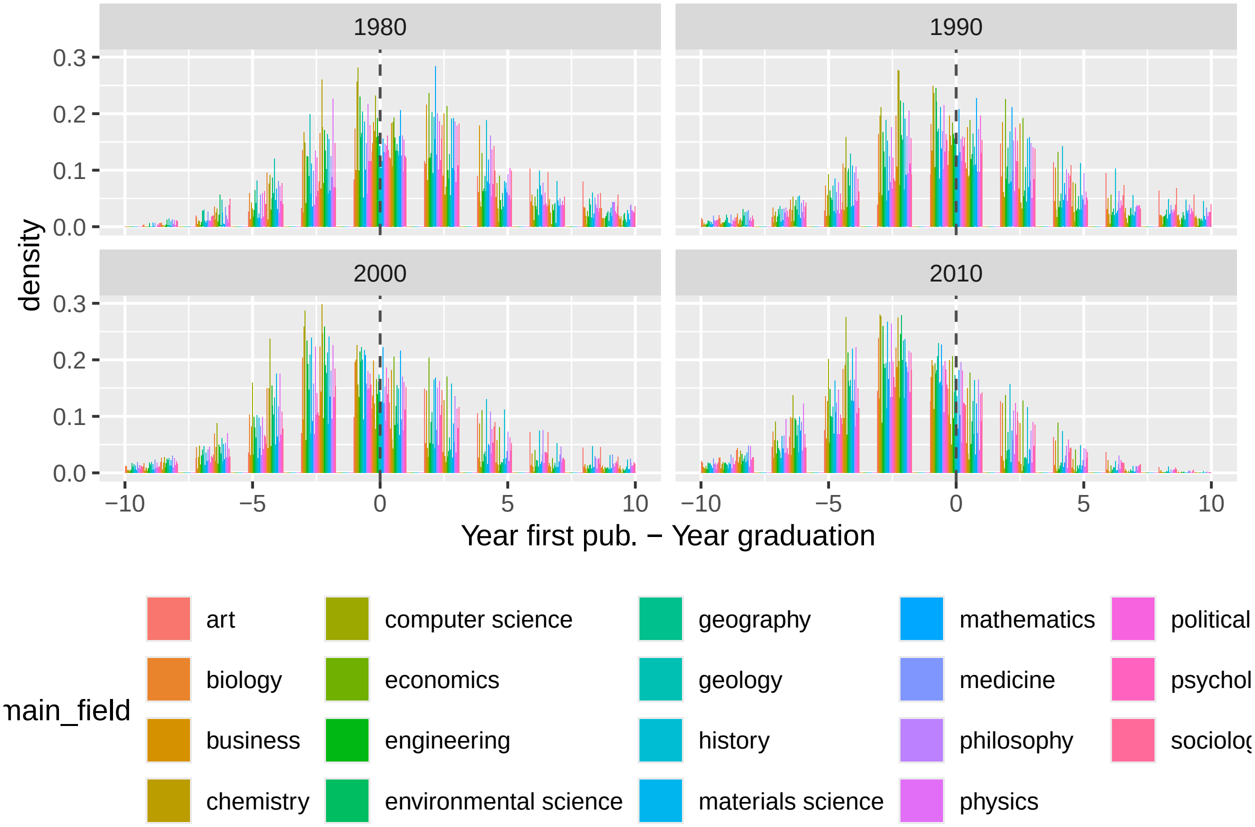
```
## $base
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



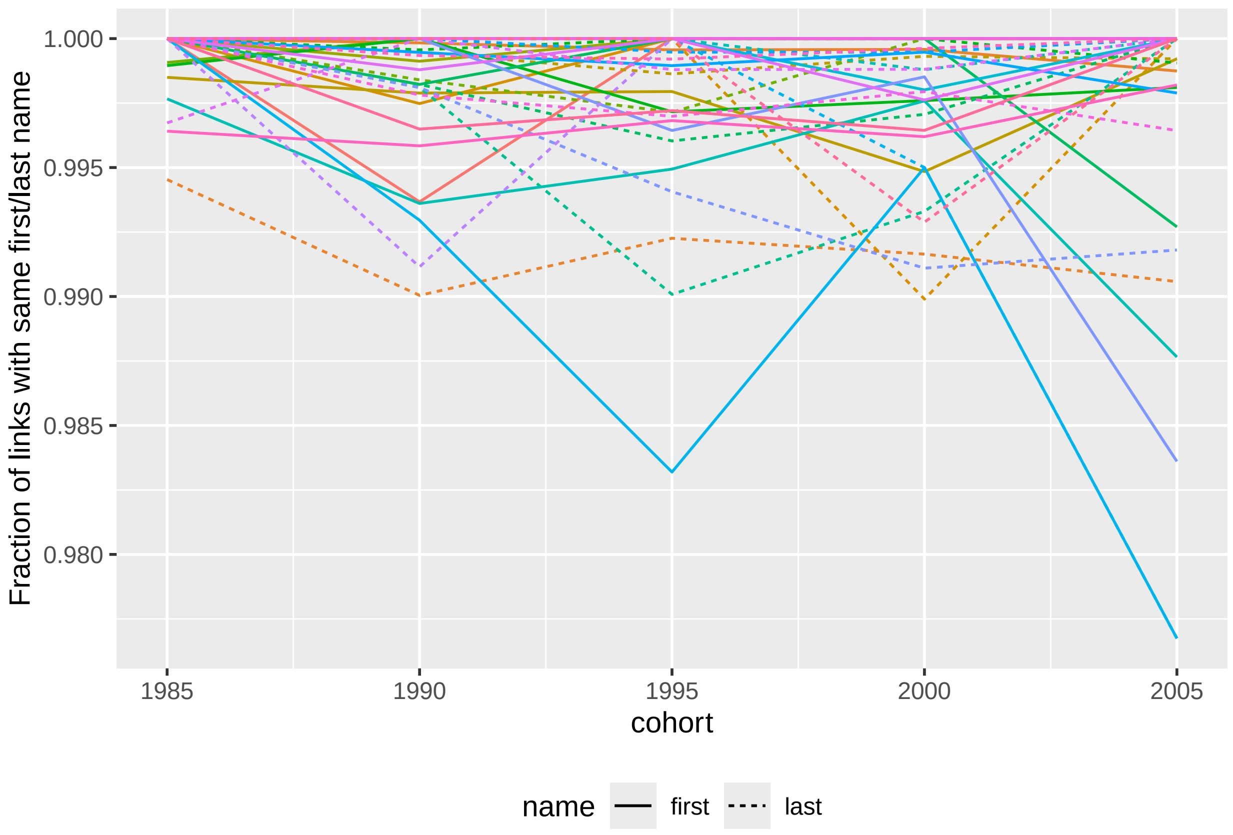
```
##
## $revise
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

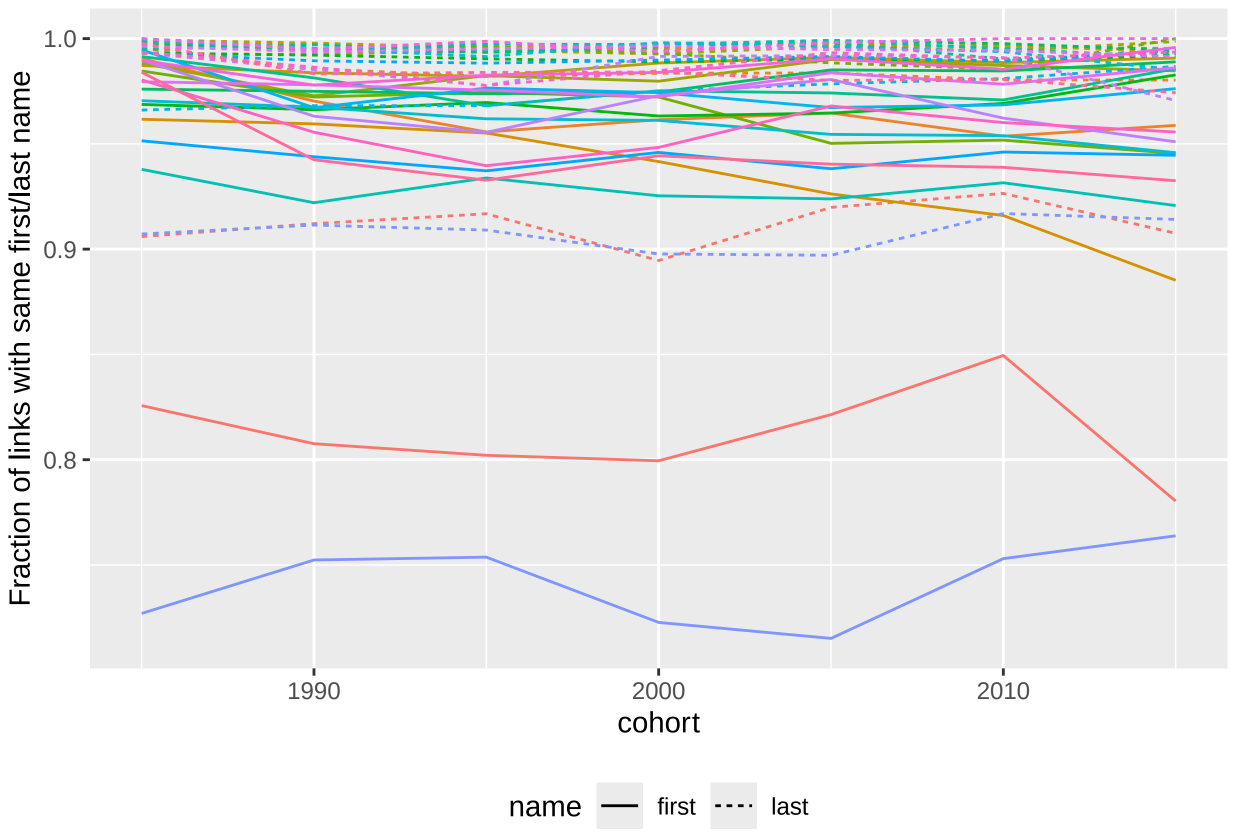


First and last name matches by cohort and field

```
## $base
```

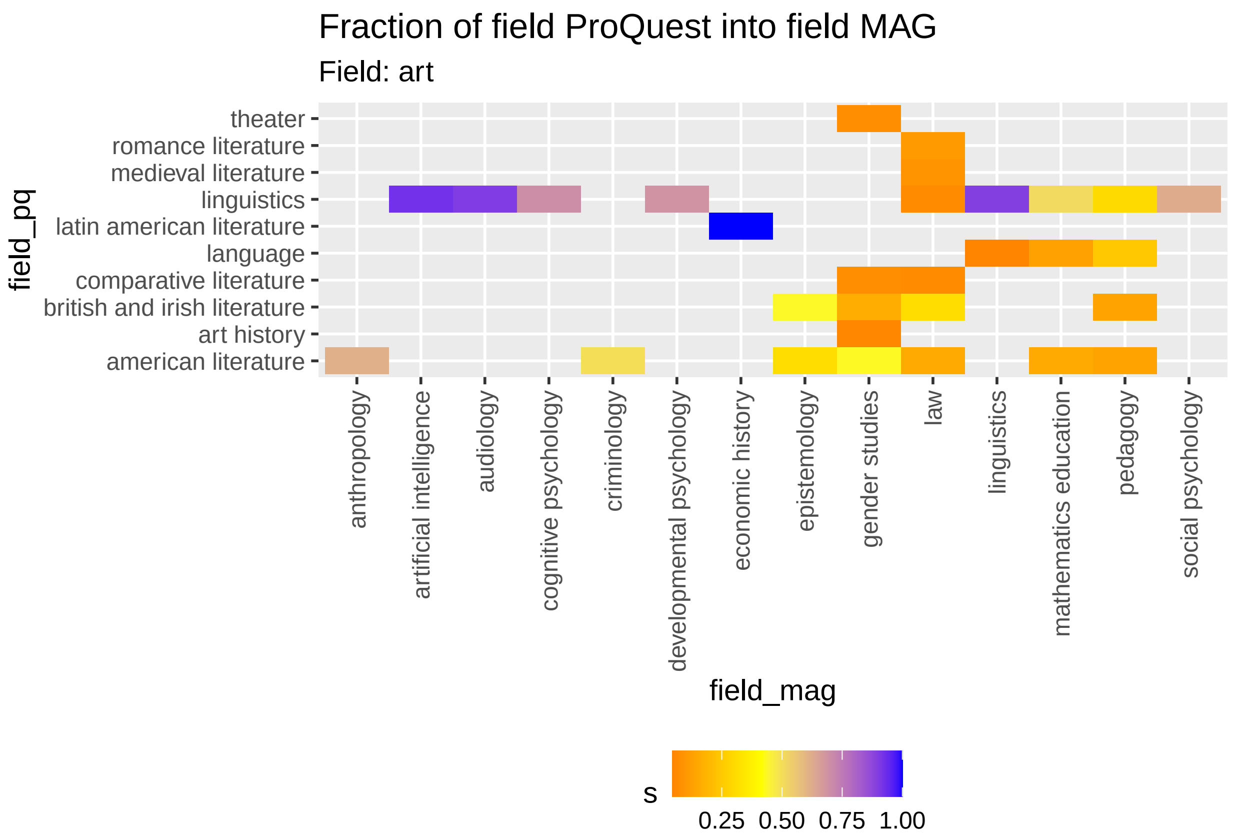

Fraction of links with same first/last name
1.000
0.995
0.990
0.985
0.980
1985
1990
1995
2000
2005
cohort
name
first
last

```
##
## $revise
```

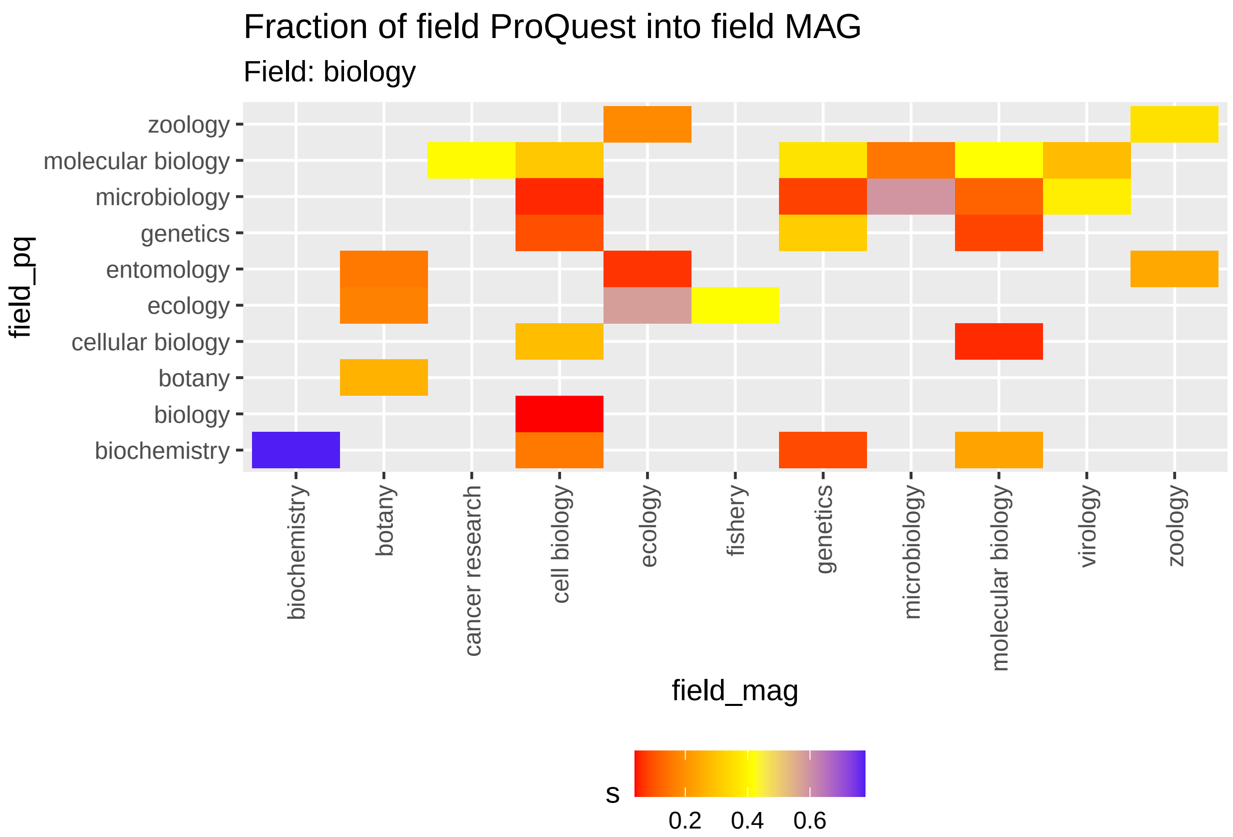


How do fields of ProQuest map into fields in MAG?

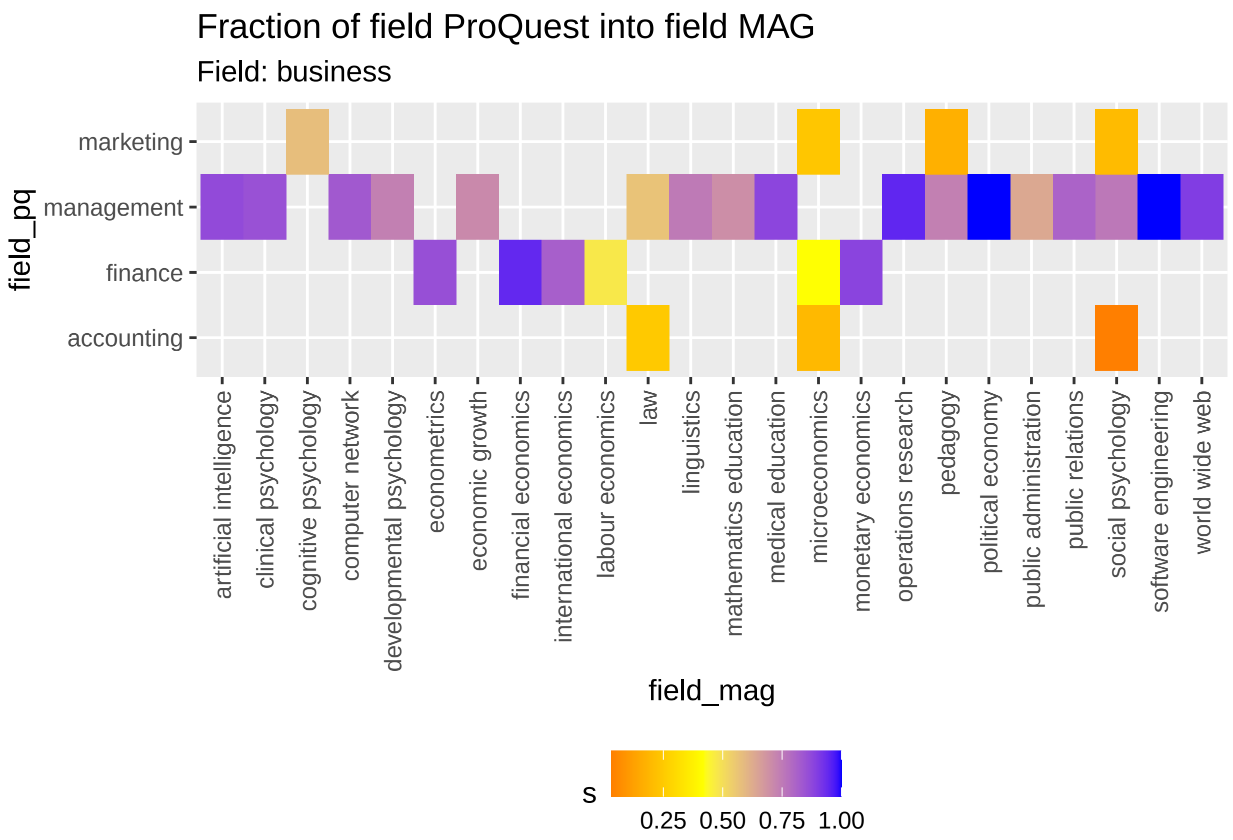
```
## [[1]]
```


Fraction of field ProQuest into field MAG
Field: art
field_pq
theater
romance literature
medieval literature
linguistics
latin american literature
language
comparative literature
british and irish literature
art history
american literature
anthropology
artificial intelligence
audiology
cognitive psychology
criminology
developmental psychology
economic history
epistemology
gender studies
law
linguistics
mathematics education
pedagogy
social psychology
field_mag
s
0.25
0.50
0.75
1.00

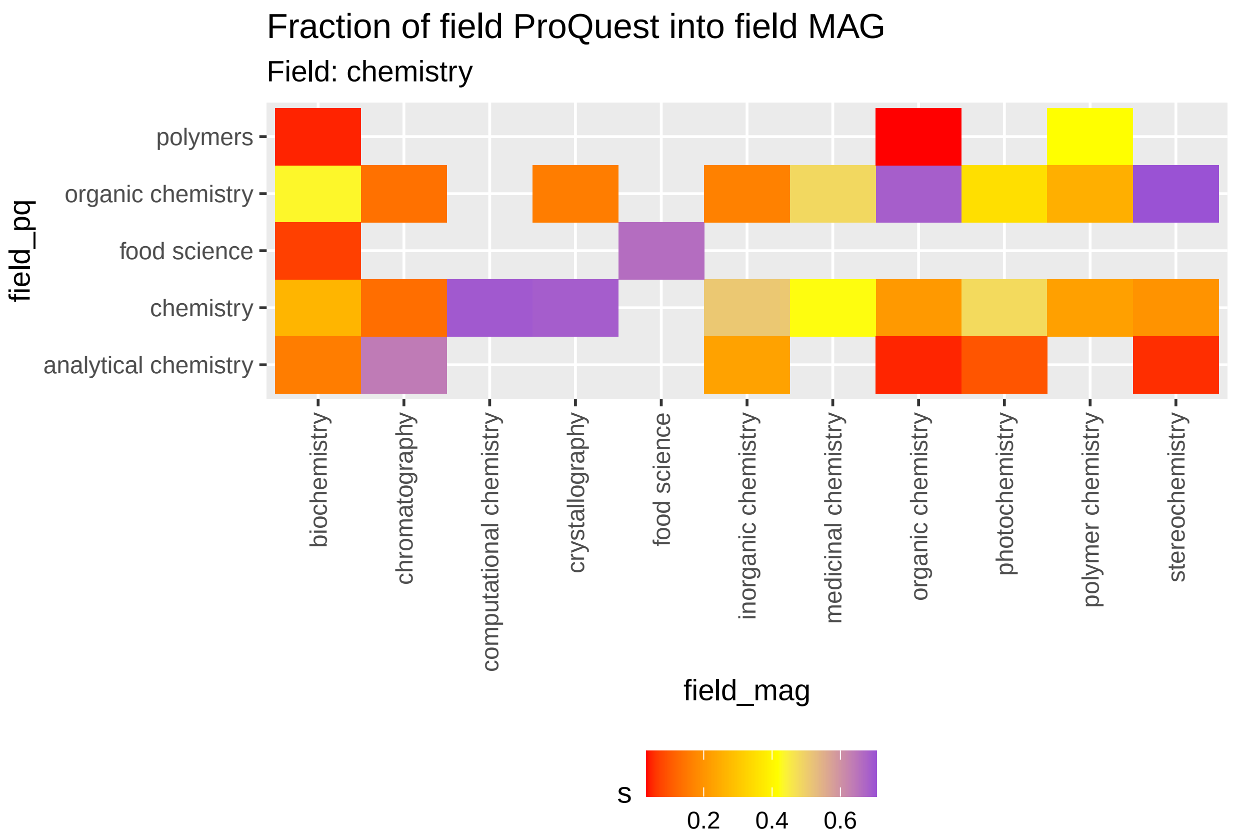
```
##
## [[2]]
```


Fraction of field ProQuest into field MAG
Field: biology
field_pq
zoology
molecular biology
microbiology
genetics
entomology
ecology
cellular biology
botany
biology
biochemistry
biochemistry
botany
cancer research
cell biology
ecology
fishery
genetics
microbiology
molecular biology
virology
zoology
field_mag
s
0.2
0.4
0.6

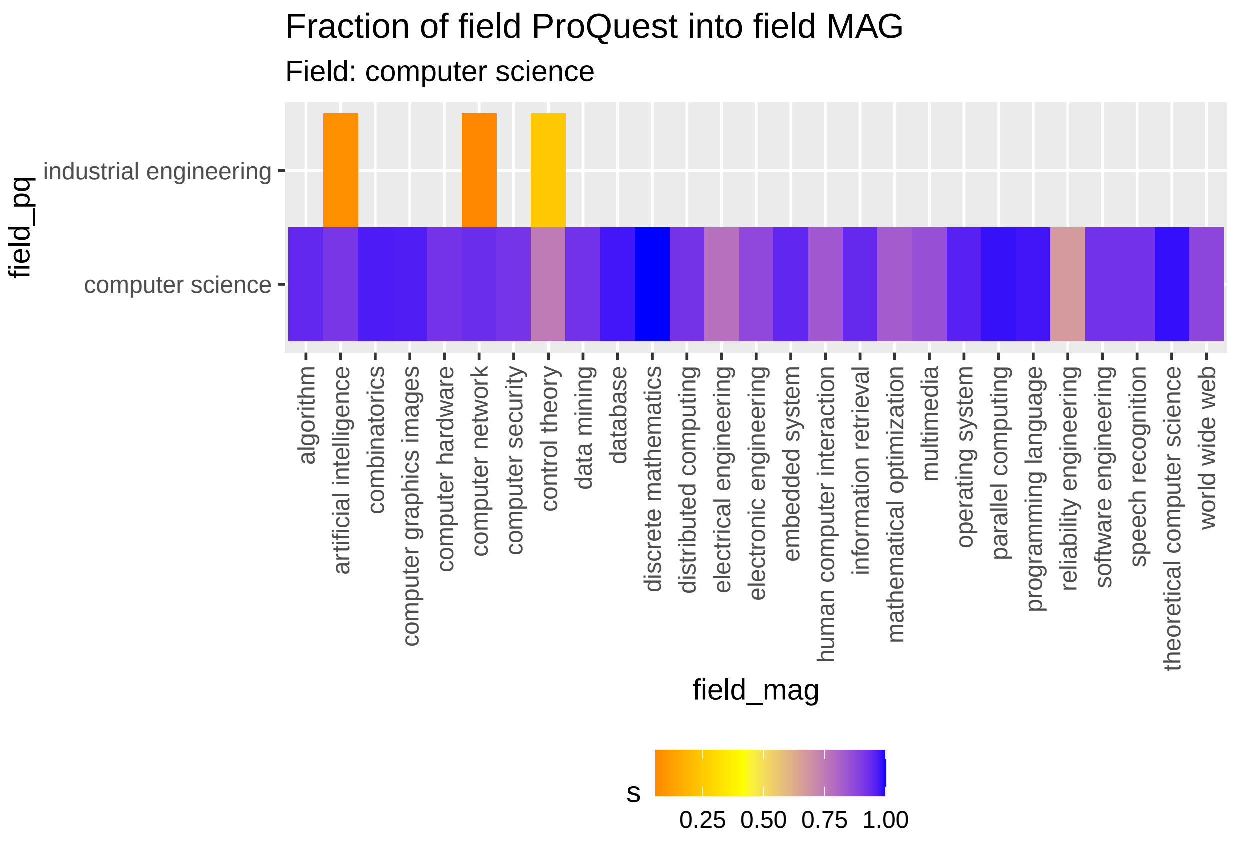
```
##
## [[3]]
```


Fraction of field ProQuest into field MAG
Field: business
field_pq
marketing
management
finance
accounting
artificial intelligence
clinical psychology
cognitive psychology
computer network
developmental psychology
econometrics
economic growth
financial economics
international economics
labour economics
law
linguistics
mathematics education
medical education
microeconomics
monetary economics
operations research
pedagogy
political economy
public administration
public relations
social psychology
software engineering
world wide web
field_mag
s
0.25
0.50
0.75
1.00

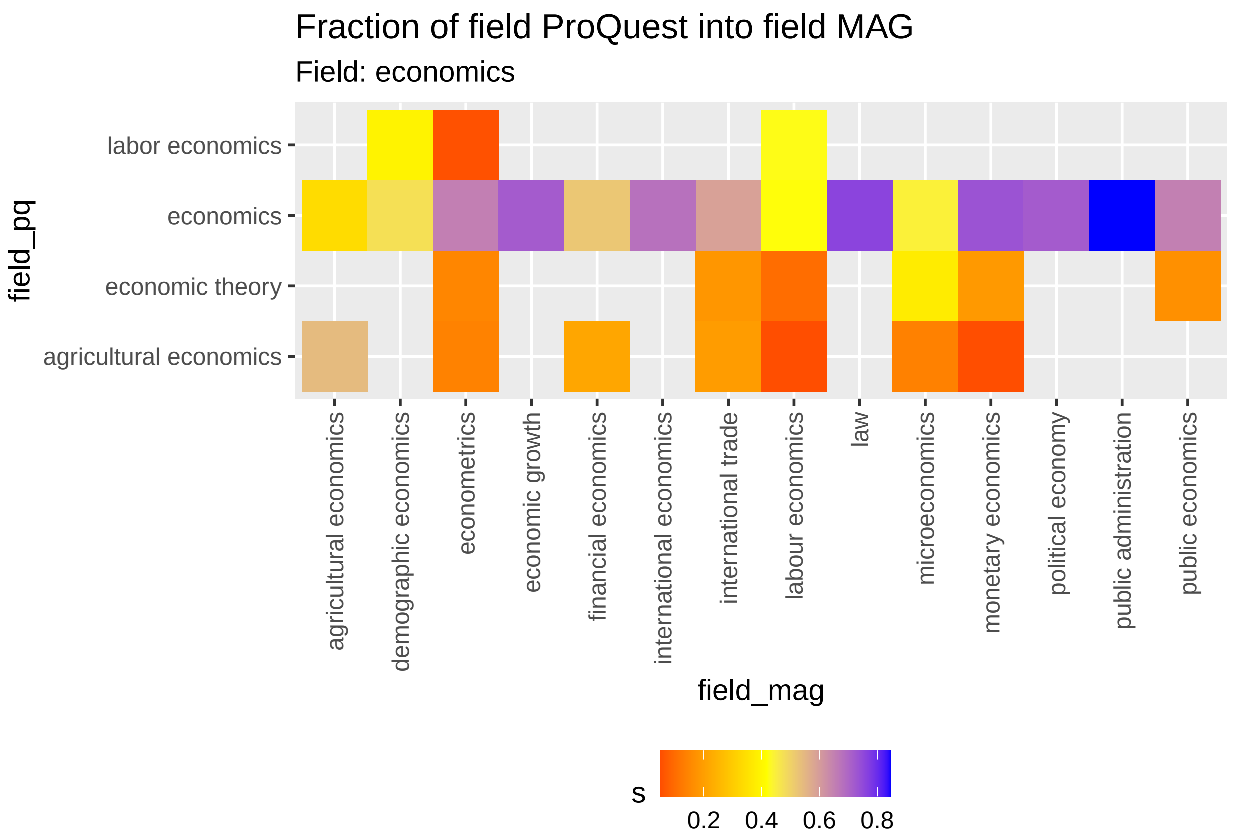
```
##
## [[4]]
```


Fraction of field ProQuest into field MAG
Field: chemistry
field_pq
polymers
organic chemistry
food science
chemistry
analytical chemistry
field_mag
biochemistry
chromatography
computational chemistry
crystallography
food science
inorganic chemistry
medicinal chemistry
organic chemistry
photochemistry
polymer chemistry
stereochemistry
s
0.2
0.4
0.6

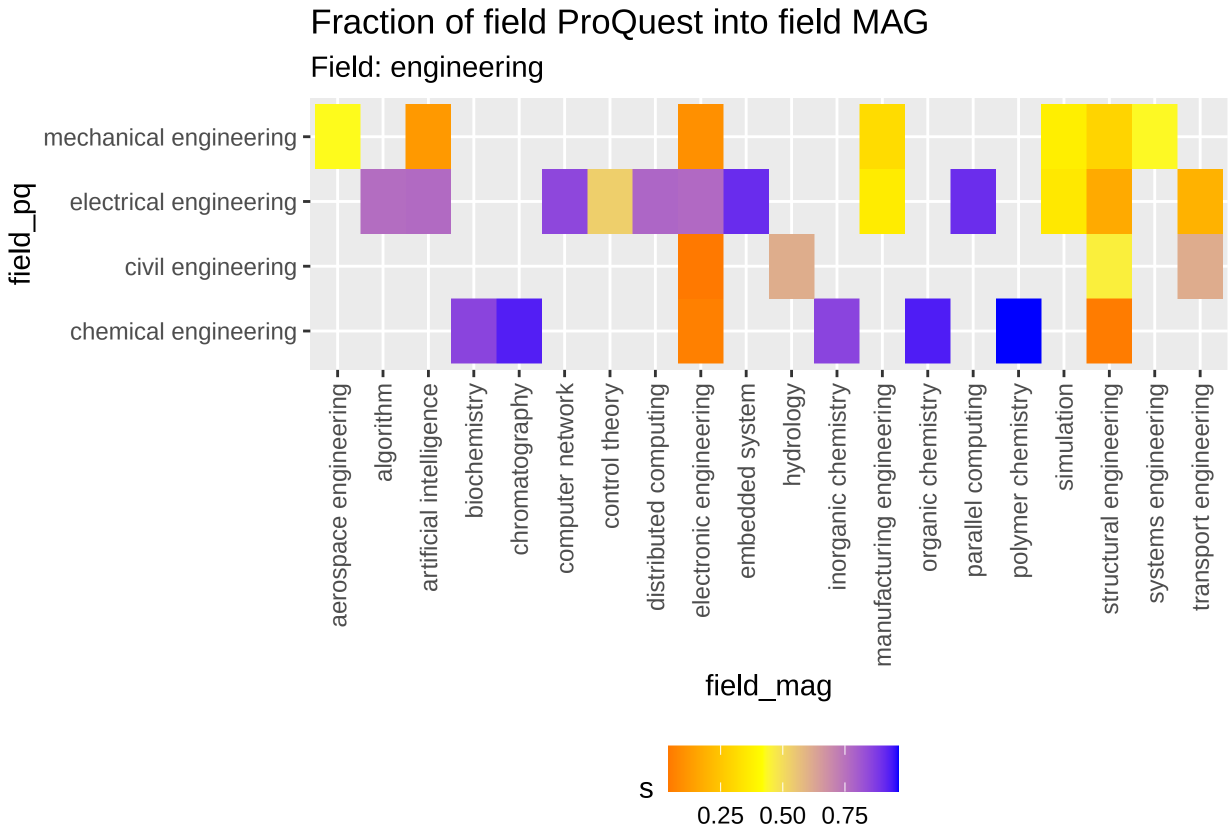
```
##
## [[5]]
```


Fraction of field ProQuest into field MAG
Field: computer science
field_pq
industrial engineering
computer science
algorithm
artificial intelligence
combinatorics
computer graphics images
computer hardware
computer network
computer security
control theory
data mining
database
discrete mathematics
distributed computing
electrical engineering
electronic engineering
embedded system
human computer interaction
information retrieval
mathematical optimization
multimedia
operating system
parallel computing
programming language
reliability engineering
software engineering
speech recognition
theoretical computer science
world wide web
field_mag
s
0.25
0.50
0.75
1.00

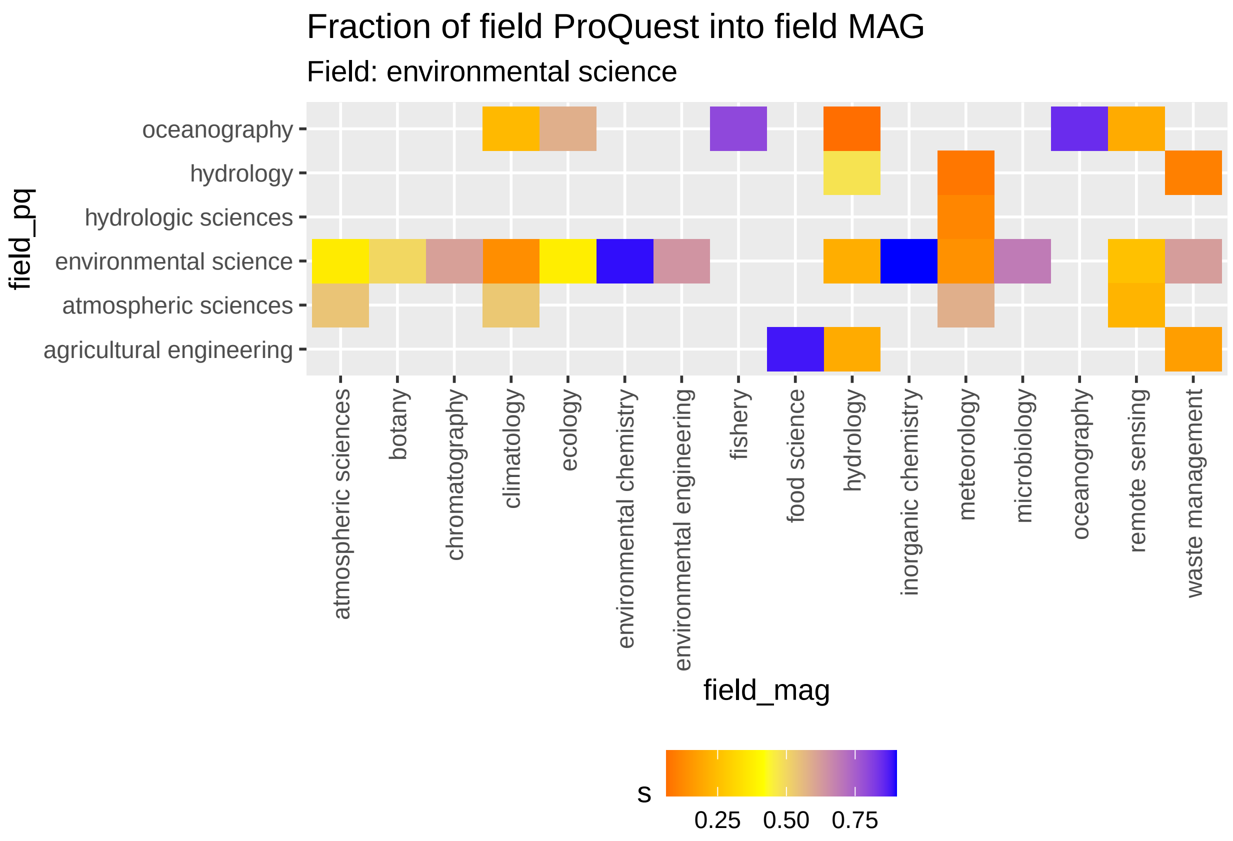
```
##
## [[6]]
```


Fraction of field ProQuest into field MAG
Field: economics
field_pq
labor economics
economics
economic theory
agricultural economics
agricultural economics
demographic economics
econometrics
economic growth
financial economics
international economics
international trade
labour economics
law
microeconomics
monetary economics
political economy
public administration
public economics
field_mag
s
0.2
0.4
0.6
0.8

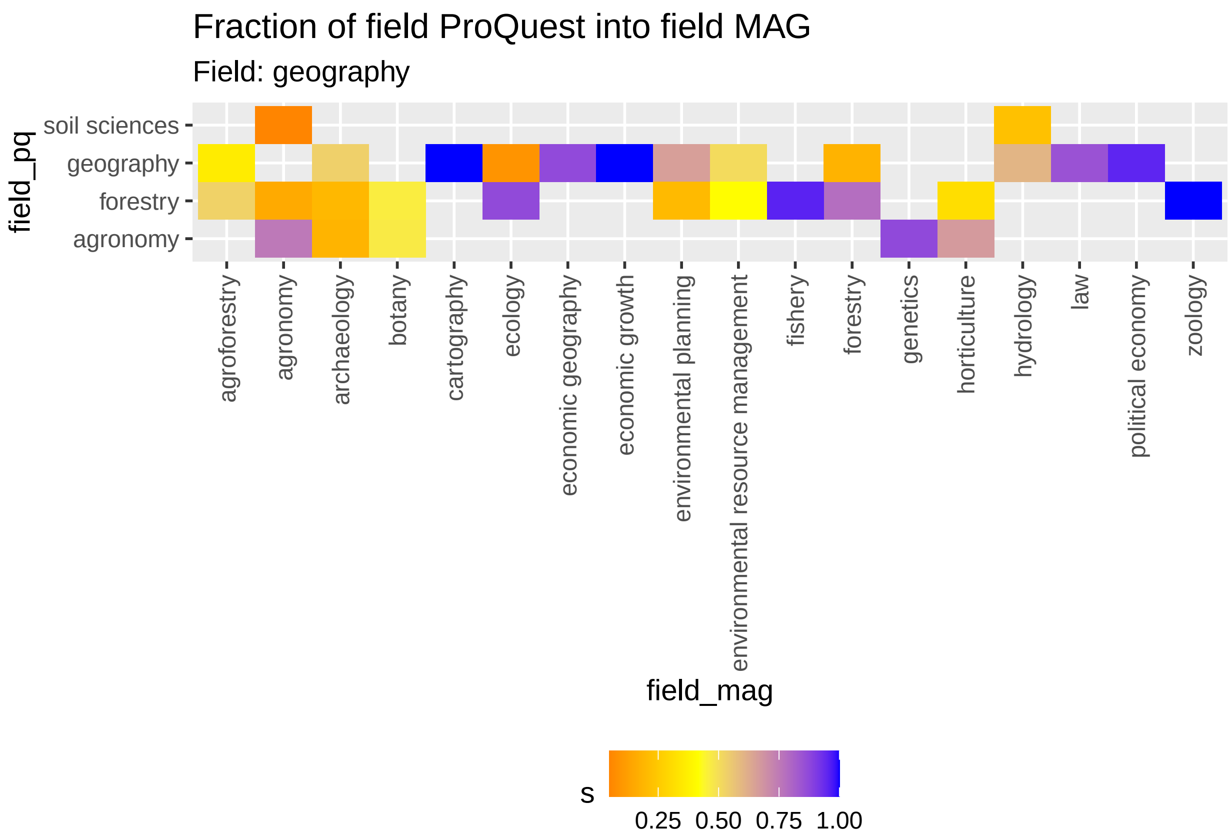
```
##
## [[7]]
```


Fraction of field ProQuest into field MAG
Field: engineering
field_pq
mechanical engineering
electrical engineering
civil engineering
chemical engineering
field_mag
aerospace engineering
algorithm
artificial intelligence
biochemistry
chromatography
computer network
control theory
distributed computing
electronic engineering
embedded system
hydrology
inorganic chemistry
manufacturing engineering
organic chemistry
parallel computing
polymer chemistry
simulation
structural engineering
systems engineering
transport engineering
s
0.25
0.50
0.75

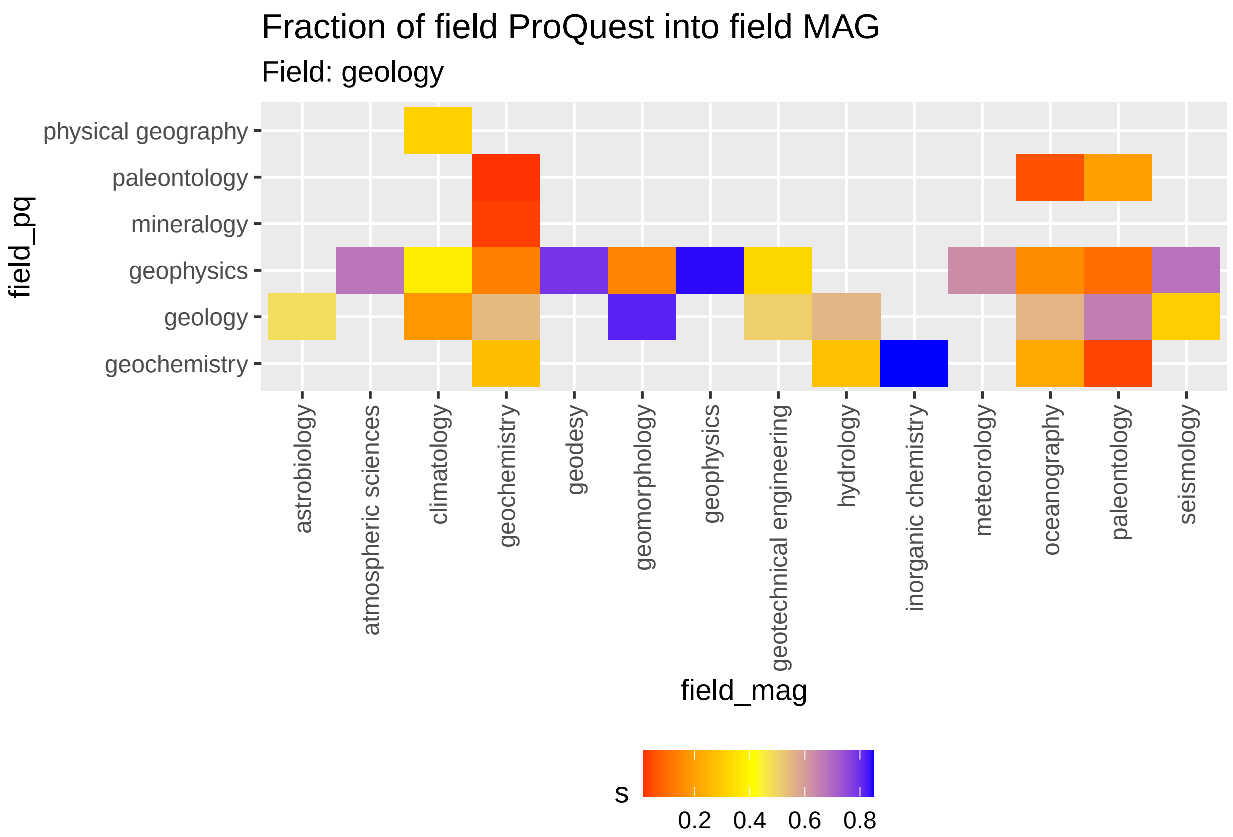
```
##
## [[8]]
```


Fraction of field ProQuest into field MAG
Field: environmental science
field_pq
oceanography
hydrology
hydrologic sciences
environmental science
atmospheric sciences
agricultural engineering
field_mag
atmospheric sciences
botany
chromatography
climatology
ecology
environmental chemistry
environmental engineering
fishery
food science
hydrology
inorganic chemistry
meteorology
microbiology
oceanography
remote sensing
waste management
s
0.25
0.50
0.75

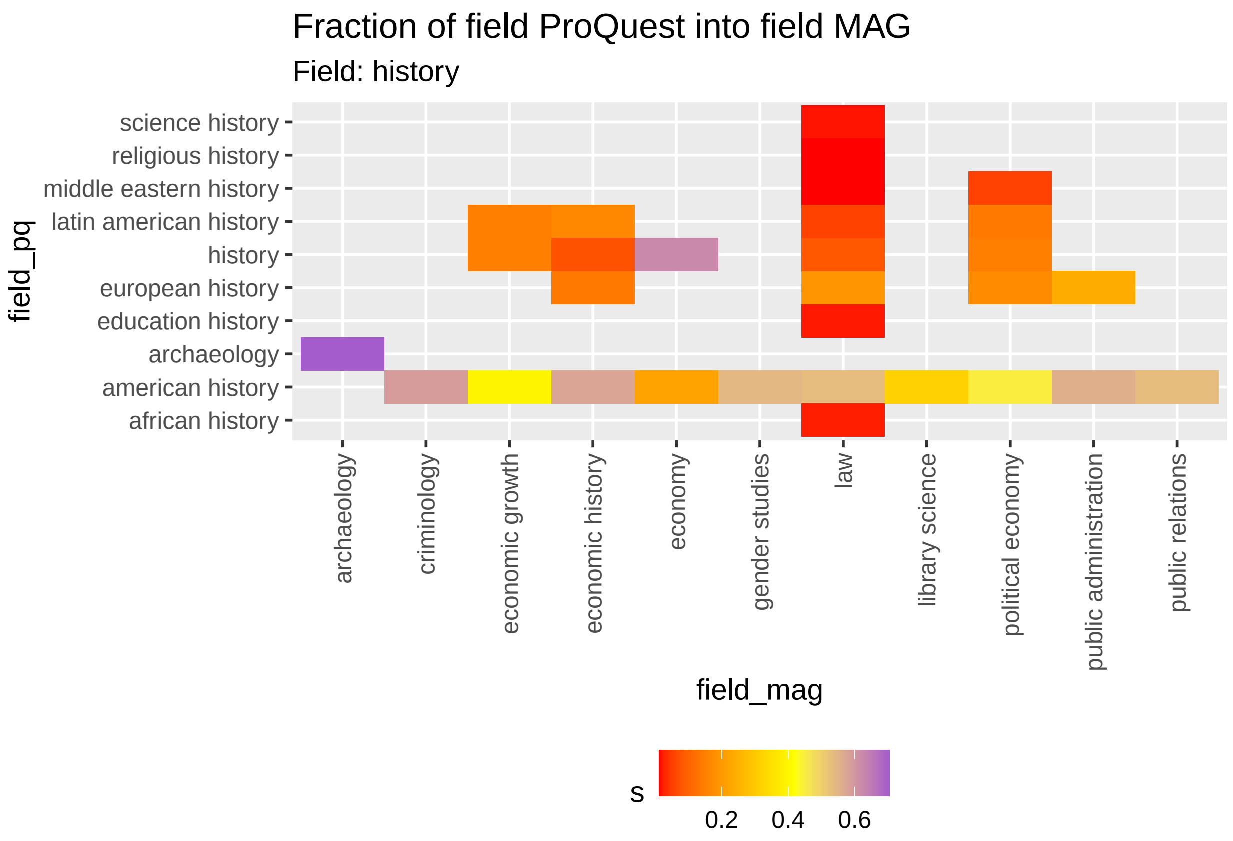
```
##
## [[9]]
```


Fraction of field ProQuest into field MAG
Field: geography
field_pq
soil sciences
geography
forestry
agronomy
agroforestry
agronomy
archaeology
botany
cartography
ecology
economic geography
economic growth
environmental planning
environmental resource management
fishery
forestry
genetics
horticulture
hydrology
law
political economy
zoology
field_mag
s
0.25
0.50
0.75
1.00

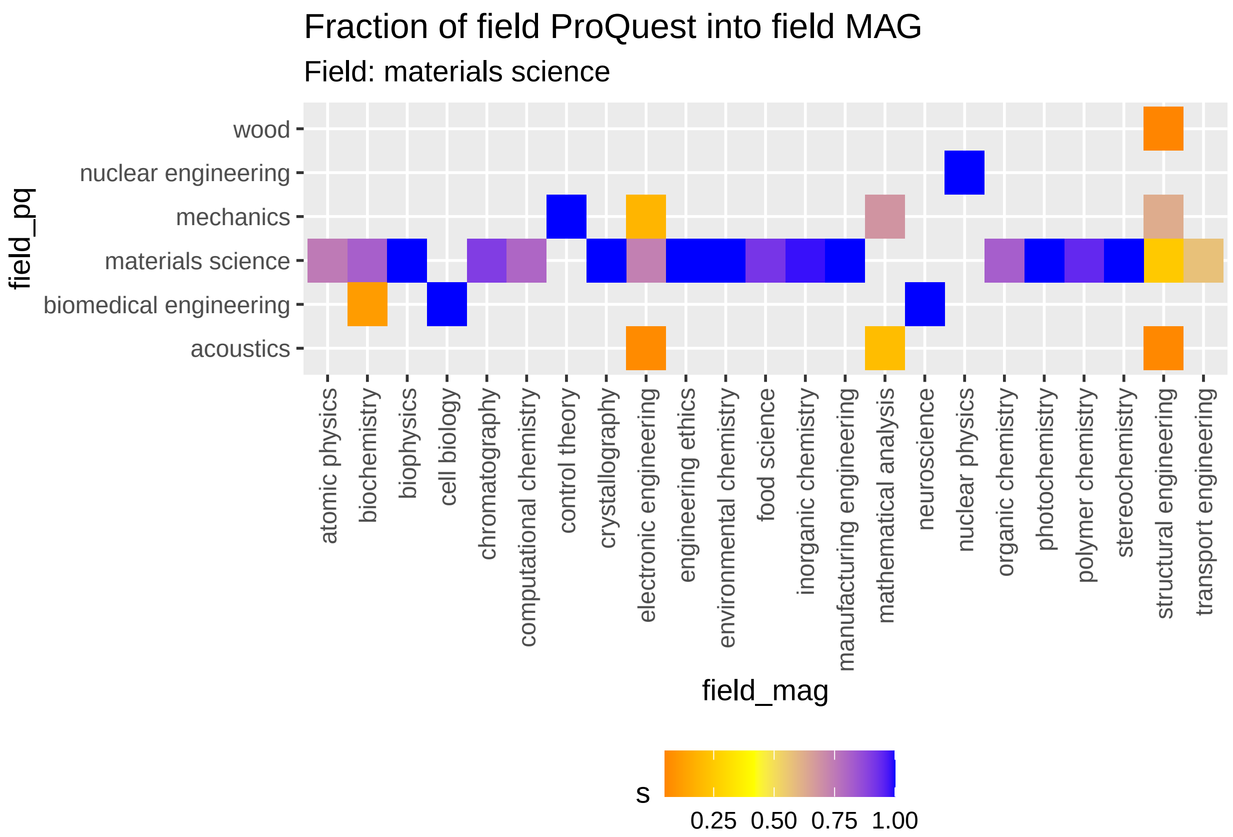
```
##
## [[10]]
```


Fraction of field ProQuest into field MAG
Field: geology
field_pq
physical geography
paleontology
mineralogy
geophysics
geology
geochemistry
field_mag
astrobiology
atmospheric sciences
climatology
geochemistry
geodesy
geomorphology
geophysics
geotechnical engineering
hydrology
inorganic chemistry
meteorology
oceanography
paleontology
seismology
s
0.2
0.4
0.6
0.8

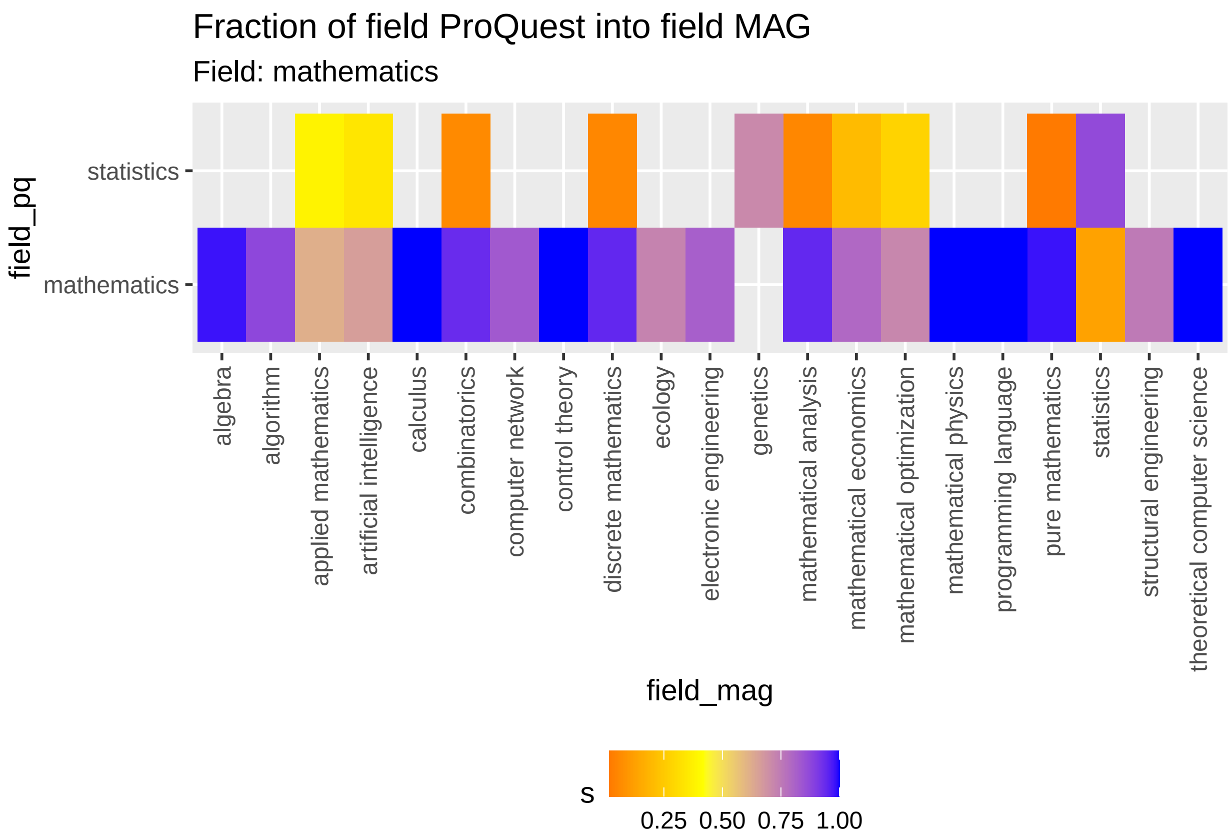
```
##
## [[11]]
```


Fraction of field ProQuest into field MAG
Field: history
field_pq
science history
religious history
middle eastern history
latin american history
history
european history
education history
archaeology
american history
african history
field_mag
archaeology
criminology
economic growth
economic history
economy
gender studies
law
library science
political economy
public administration
public relations
s
0.2
0.4
0.6

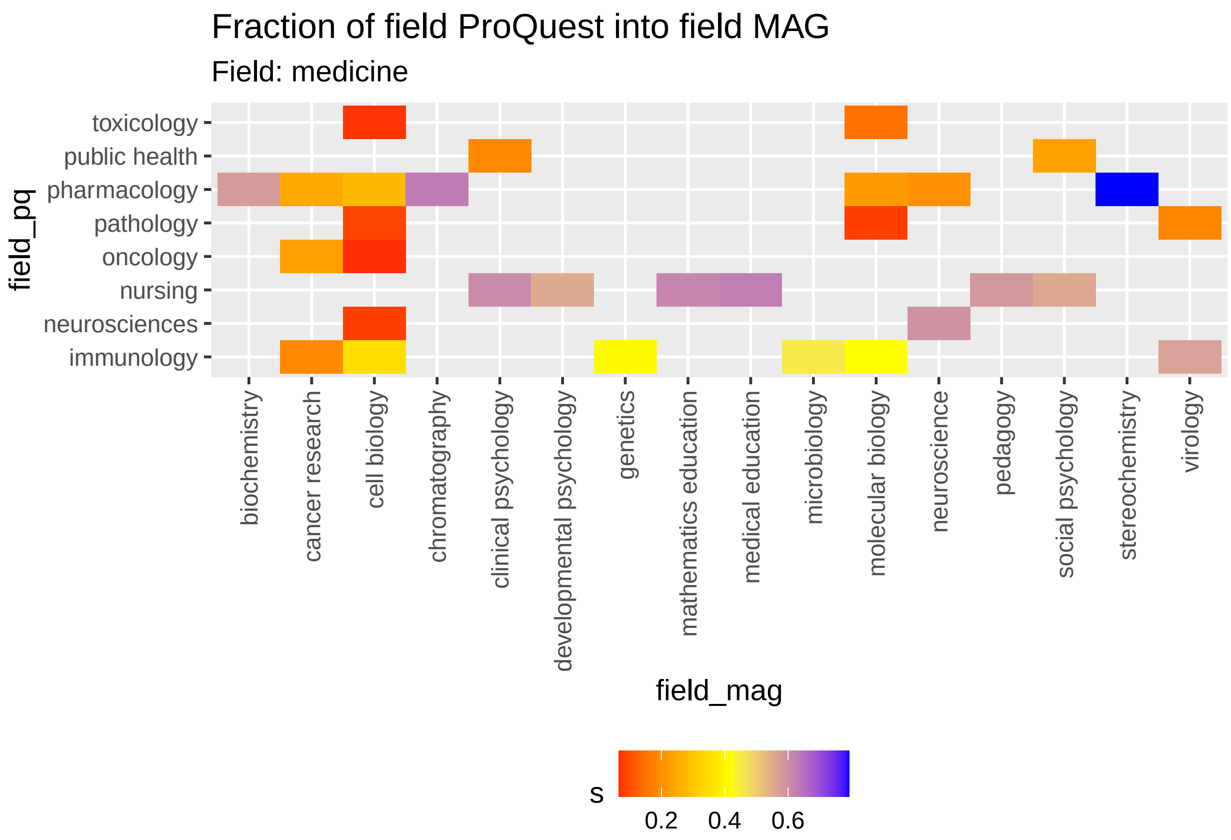
```
##
## [[12]]
```


Fraction of field ProQuest into field MAG
Field: materials science
field_pq
wood
nuclear engineering
mechanics
materials science
biomedical engineering
acoustics
atomic physics
biochemistry
biophysics
cell biology
chromatography
computational chemistry
control theory
crystallography
electronic engineering
engineering ethics
environmental chemistry
food science
inorganic chemistry
manufacturing engineering
mathematical analysis
neuroscience
nuclear physics
organic chemistry
photochemistry
polymer chemistry
stereochemistry
structural engineering
transport engineering
field_mag
s
0.25
0.50
0.75
1.00

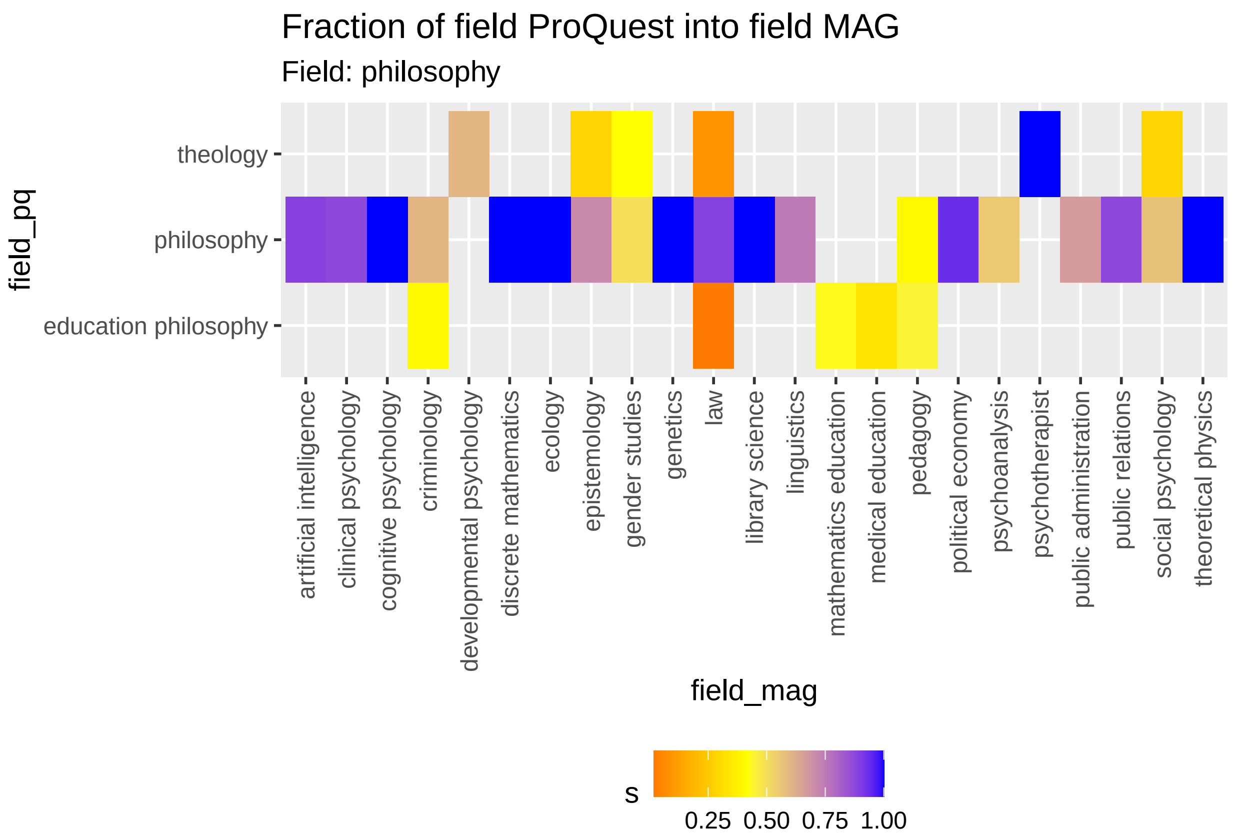
```
##
## [[13]]
```


Fraction of field ProQuest into field MAG
Field: mathematics
field_pq
statistics
mathematics
algebra
algorithm
applied mathematics
artificial intelligence
calculus
combinatorics
computer network
control theory
discrete mathematics
ecology
electronic engineering
genetics
mathematical analysis
mathematical economics
mathematical optimization
mathematical physics
programming language
pure mathematics
statistics
structural engineering
theoretical computer science
field_mag
s
0.25
0.50
0.75
1.00

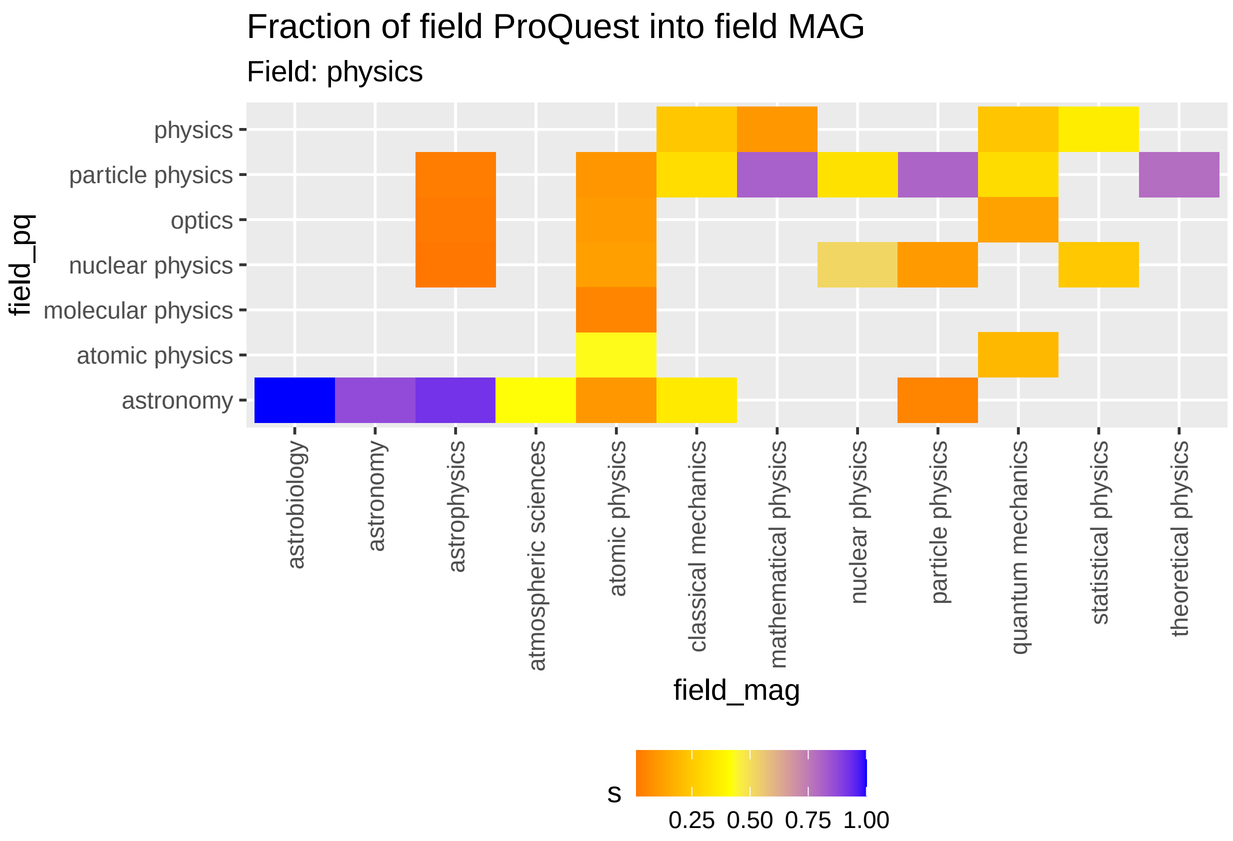
```
##
## [[14]]
```


Fraction of field ProQuest into field MAG
Field: medicine
field_pq
toxicology
public health
pharmacology
pathology
oncology
nursing
neurosciences
immunology
biochemistry
cancer research
cell biology
chromatography
clinical psychology
developmental psychology
genetics
mathematics education
medical education
microbiology
molecular biology
neuroscience
pedagogy
social psychology
stereochemistry
virology
field_mag
s
0.2
0.4
0.6

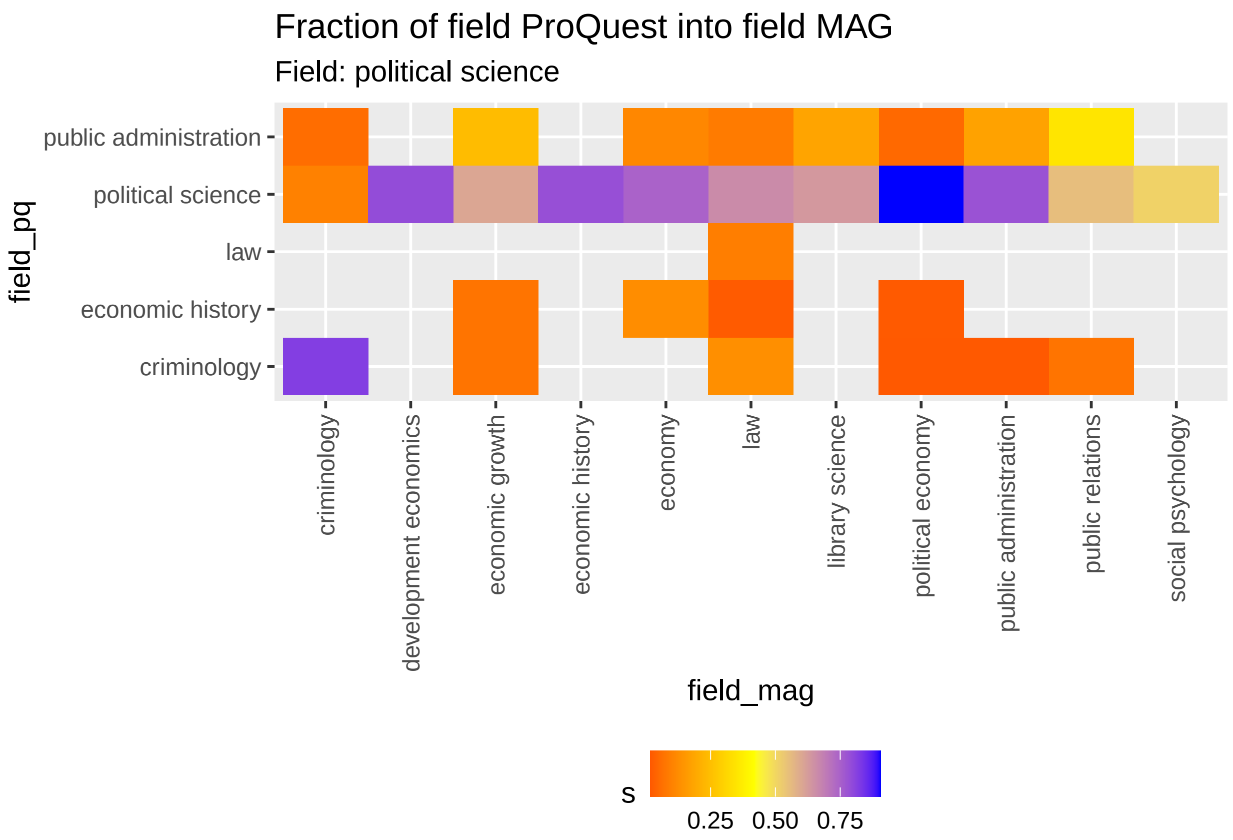
```
##
## [[15]]
```


Fraction of field ProQuest into field MAG
Field: philosophy
field_pq
theology
philosophy
education philosophy
artificial intelligence
clinical psychology
cognitive psychology
criminology
developmental psychology
discrete mathematics
ecology
epistemology
gender studies
genetics
law
library science
linguistics
mathematics education
medical education
pedagogy
political economy
psychoanalysis
psychotherapist
public administration
public relations
social psychology
theoretical physics
field_mag
s
0.25
0.50
0.75
1.00

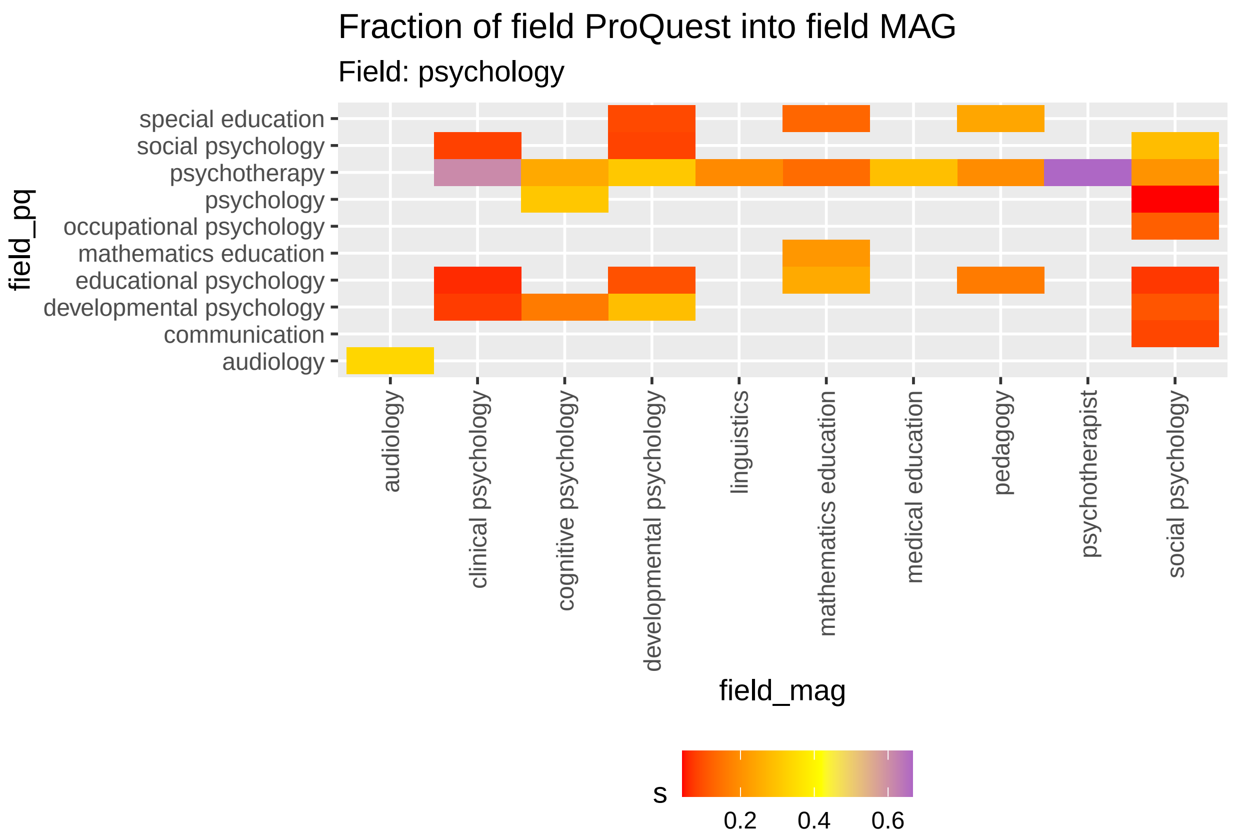
```
##
## [[16]]
```


Fraction of field ProQuest into field MAG
Field: physics
field_pq
physics
particle physics
optics
nuclear physics
molecular physics
atomic physics
astronomy
astrobiology
astronomy
astrophysics
atmospheric sciences
atomic physics
classical mechanics
mathematical physics
nuclear physics
particle physics
quantum mechanics
statistical physics
theoretical physics
field_mag
s
0.25
0.50
0.75
1.00

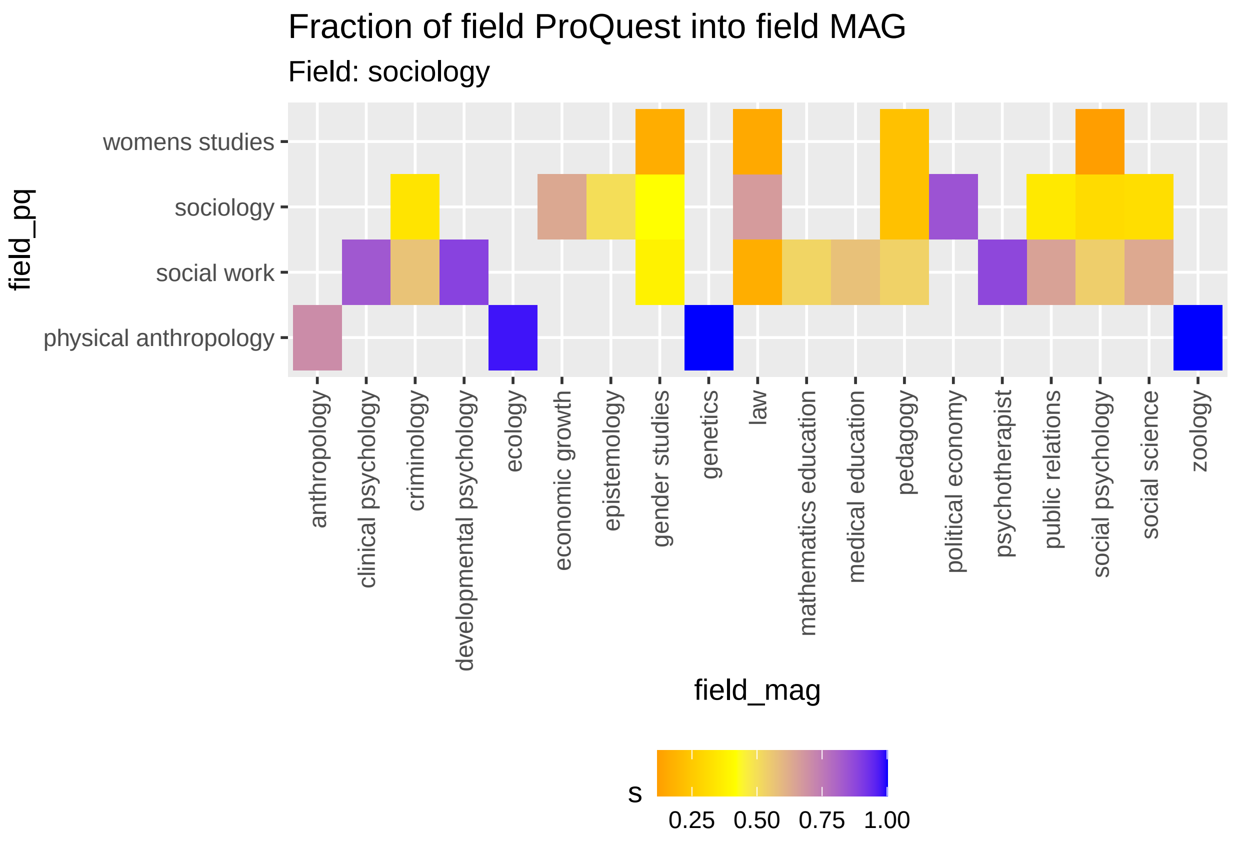
```
##
## [[17]]
```


Fraction of field ProQuest into field MAG
Field: political science
field_pq
public administration
political science
law
economic history
criminology
criminology
development economics
economic growth
economic history
economy
law
library science
political economy
public administration
public relations
social psychology
field_mag
s
0.25
0.50
0.75

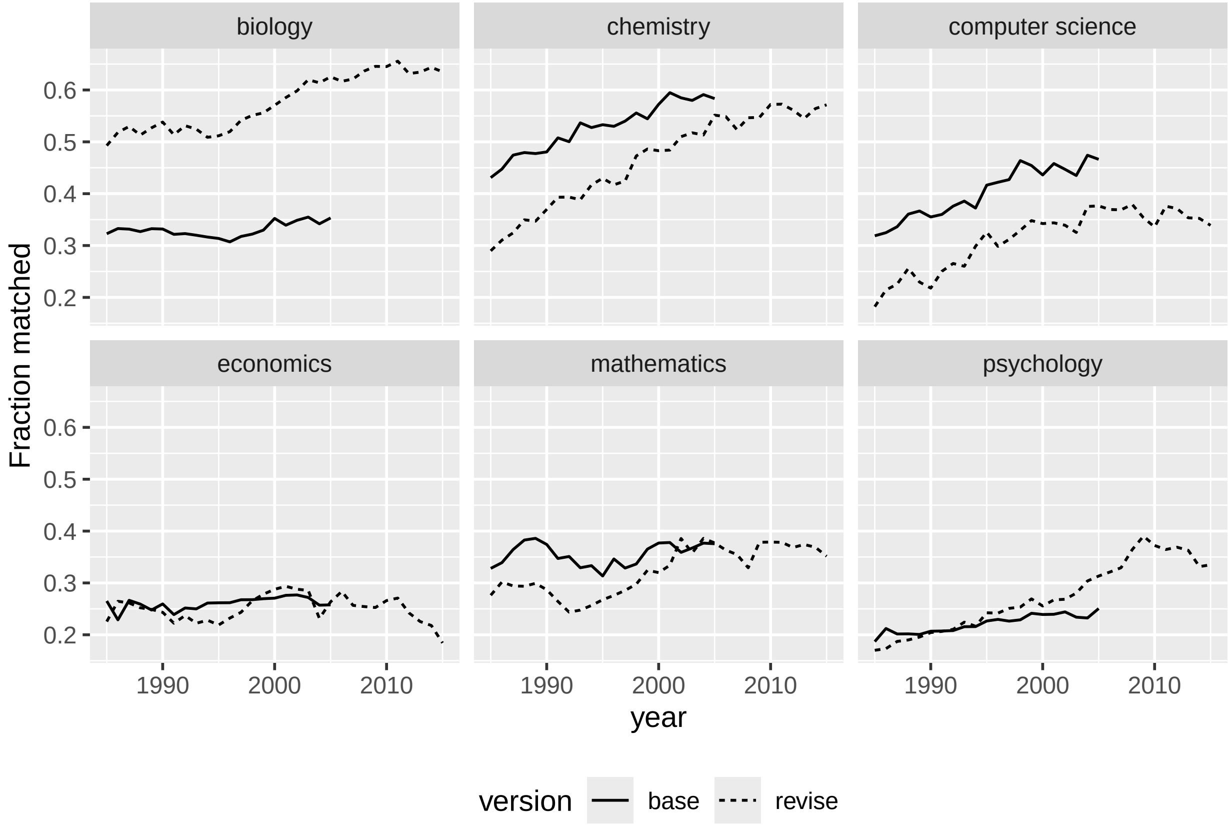
```
##
## [[18]]
```

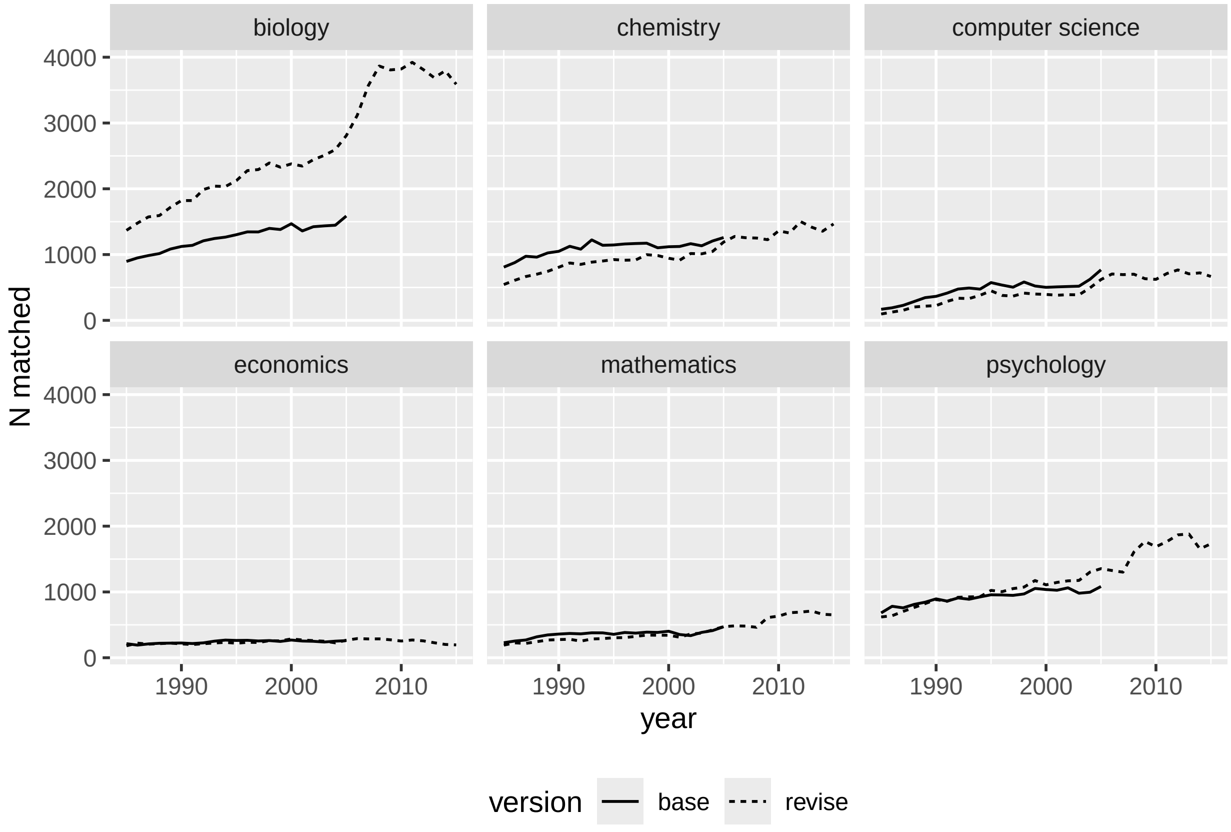

Fraction of field ProQuest into field MAG
Field: psychology
field_pq
special education
social psychology
psychotherapy
psychology
occupational psychology
mathematics education
educational psychology
developmental psychology
communication
audiology
field_mag
audiology
clinical psychology
cognitive psychology
developmental psychology
linguistics
mathematics education
medical education
pedagogy
psychotherapist
social psychology
s
0.2
0.4
0.6

```
##
## [[19]]
```


Fraction of field ProQuest into field MAG
Field: sociology
field_pq
womens studies
sociology
social work
physical anthropology
anthropology
clinical psychology
criminology
developmental psychology
ecology
economic growth
epistemology
gender studies
genetics
law
mathematics education
medical education
pedagogy
political economy
psychotherapist
public relations
social psychology
social science
zoology
field_mag
s
0.25
0.50
0.75
1.00

Fraction matched by year and field





Checking non-linked entities that should be a link

```
d_chem <- pq_authors %>%
   left_join(field_names_id %>%
                  rename(main_field = NormalizedName),
              by = c("mag_field0" = "FieldOfStudyId")) %>%
      mutate(link = ifelse(goid %in% d_links$revise$goid, "linked", "not linked")) %>%
  filter(main_field == "chemistry")

pq_unis <- tbl(con, "pq_authors") %>%
  left_join(tbl(con, "pq_unis") %>%
              select(university_id, normalizedname),
            by = "university_id") %>%
  select(goid, uni_name = "normalizedname") %>%
  collect()

d_chem <- d_chem %>%
  left_join(pq_unis, by = "goid")
```

```
d_chem %>%
  filter(year == 1995 & uni_name == "stanford university" & link == "not linked") %>% head(10)
```

```
## # A tibble: 10 x 11
##         goid  year firstname lastname  middlename fieldofstudy mag_field0
##      <int64> <int> <chr>     <chr>     <chr>      <chr>             <int>
##  1 304229925  1995 nancy     hansen    fisher     chemistry     185592680
##  2 304229722  1995 mark      pavlosky  alan       chemistry     185592680
##  3 304228620  1995 kristin   sannes    ann        chemistry     185592680
##  4 304218381  1995 glenn     jones     clark      chemistry     185592680
##  5 304201950  1995 david     offord    alan       chemistry     185592680
##  6 304238172  1995 robert    guettler  david      chemistry     185592680
##  7 304202002  1995 eric      remy      david      chemistry     185592680
##  8 304229882  1995 thomas    schoch    k          chemistry     185592680
##  9 304229838  1995 philip    merrill   bradley    chemistry     185592680
## 10 304218488  1995 claude    maechling ricketts   chemistry     185592680
## # i 4 more variables: university_id <int>, main_field <chr>, link <chr>,
## #   uni_name <chr>
```

```
#unique(d_chem$fieldofstudy)
## comparing to candidates:
# harvard:
# weldon in materials science
# beltrame in chemistry
# mit:
# lapointe is chemistry
# duff is chemistry
# stanford:
# shear in chemistry
# marcus is in biology
# hansen is in biology
# tokmakoff is in materials science

# update, chemistry check 8/11/22
# - tokmakoff still not linked; b/c of year first pub? -- yes, the linking score is 0.66...
# - nancy fisher hansen (2649181519) is not linked (unclear if she should be linked)
# - hopefully the keywords from topic models would help us here?
# - maybe david h offord (304201950) would also be linked with the keywords?
```

Chemistry: first affiliation of MAG authors should be the graduating institution. paper

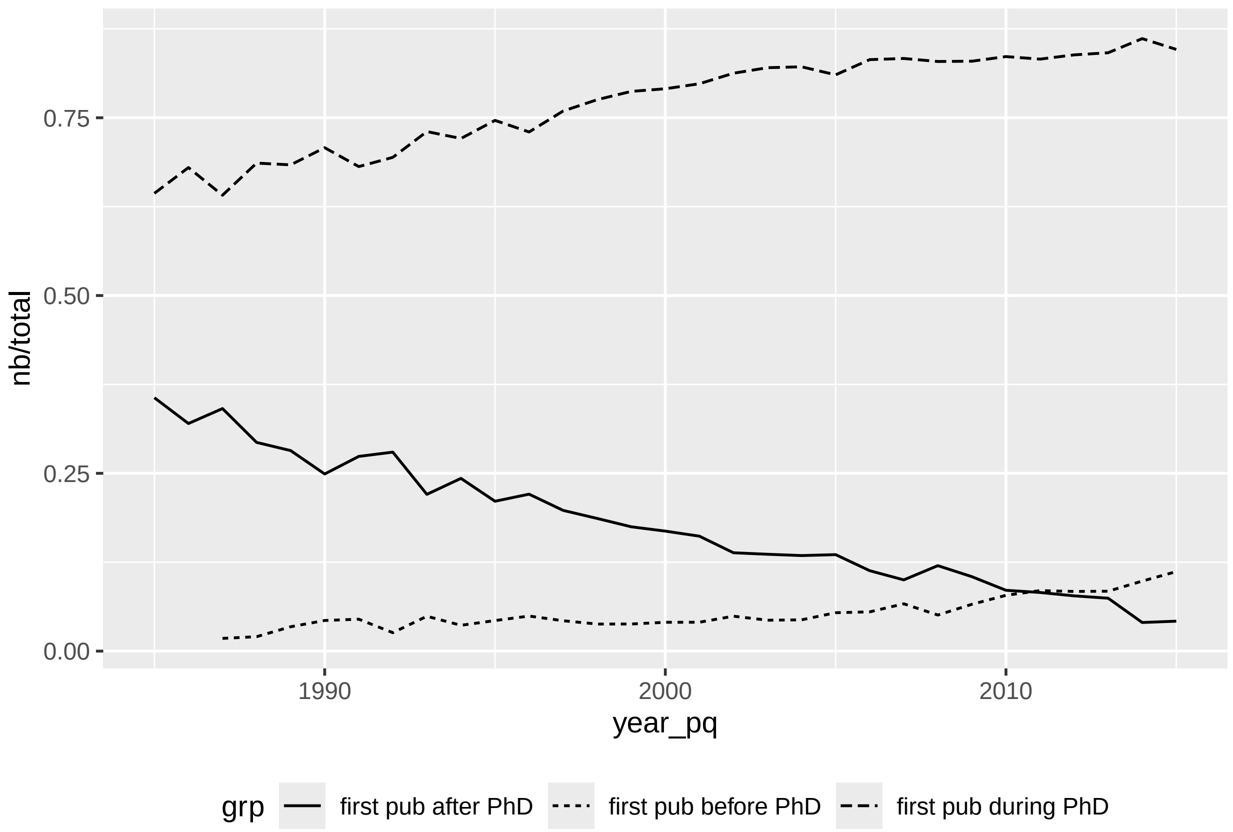
```
grads_chemistry <- d_links$revise |>
  filter(field0_mag == 185592680) |>
  group_by(AuthorId) |>
  filter(iteration_id == max(iteration_id)) |>
  ungroup() |>
  mutate(grp = case_when( # some people publish already way before the PhD
    year_mag > year_pq ~ "first pub after PhD",
    year_mag < year_pq - 6 ~ "first pub before PhD",
    TRUE ~ "first pub during PhD"
  )) |>
  select(AuthorId, goid, year_pq, grp)
```

```
head(grads_chemistry)
```

```
## # A tibble: 6 x 4
##     AuthorId      goid year_pq grp
##      <int64>   <int64>   <int> <chr>
## 1 2227604972 303417360    1986 first pub during PhD
## 2  641051114 303352848    1985 first pub during PhD
## 3 2143303641 881747820    2011 first pub after PhD
## 4 2168717013 304427153    1998 first pub during PhD
## 5 2504958925 305369745    2006 first pub during PhD
## 6 2225265093 304664910    2000 first pub after PhD
```

```
grads_chemistry |>
  group_by(grp, year_pq) |>
  summarise(nb = n()) |>
  ungroup() |>
  group_by(year_pq) |>
  mutate(total = sum(nb)) |>
  ggplot(aes(x = year_pq, y = nb/total)) +
  geom_line(aes(linetype = grp)) +
  theme(legend.position = "bottom")
```

```
## `summarise()` has grouped output by 'grp'. You can override using the `.groups`
## argument.
```



Gaule/Piacentini had 21154 graduates from 1999 to 2008; we have

```
grads_chemistry |>
  filter(year_pq >= 1999 & year_pq <= 2008) |>
  summarise(n())
```

```
## # A tibble: 1 x 1
##   `n()`
##   <int>
## 1 12992
```

- they had chemists and chemical engineers; we may miss the engineers in this sample.

```
grads_chemistry |>
  filter(year_pq >= 1990 & year_pq <= 2015) |>
  group_by(grp) |>
  summarise(nb = n()) |>
  ungroup() |>
  mutate(s = nb / sum(nb))
```

```
## # A tibble: 3 x 3
##   grp                     nb      s
##   <chr>                <int>  <dbl>
## 1 first pub after PhD   4977 0.147
## 2 first pub before PhD  1967 0.0579
## 3 first pub during PhD 27011 0.795
```

```
query_authors <- unique(grads_chemistry$AuthorId)
query_authors <- paste0(query_authors, collapse = ", ")
```

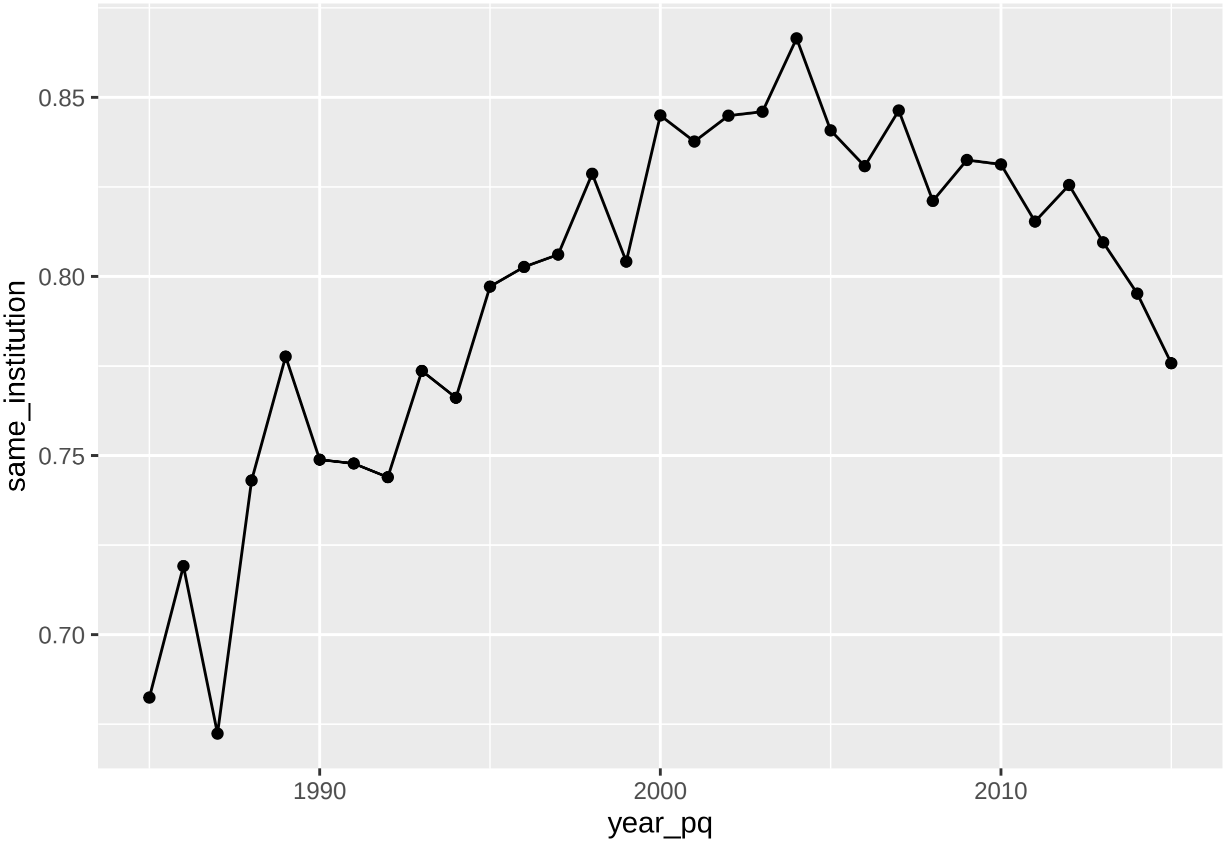
```r
q_authors_affil <- paste0(
  "SELECT AuthorId, AffiliationId, Year
  FROM AuthorAffiliation
  INNER JOIN (
    SELECT AuthorId, YearFirstPub
    FROM author_sample
  ) USING(AuthorId)
  WHERE AuthorId IN (", query_authors, ")
  AND Year <= YearFirstPub + 20"
)

authors_affil <- tbl(con, sql(q_authors_affil)) |>
  collect()

authors_first_affil <- authors_affil |>
  group_by(AuthorId) |>
  filter(Year == min(Year)) |>
  filter(!duplicated(AuthorId)) |>
  ungroup()
```

```r
links_to_cng <- tbl(con, "links_to_cng") |>
  collect()
```

Place of first publication

```r
place_first_pub <- grads_chemistry |>
  left_join(pq_authors |>
              select(goid, university_id),
            by = "goid") |>
  left_join(links_to_cng |>
              filter(from_dataset == "pq") |>
              select(from_id, unitid_graduate = unitid),
            by = c("university_id" = "from_id")) |>
  left_join(authors_first_affil |>
              select(AuthorId, AffiliationId),
            by = "AuthorId") |>
  left_join(links_to_cng |>
              filter(from_dataset == "mag") |>
              select(from_id, unitid_author = unitid),
            by = c("AffiliationId" = "from_id"))
```

```r
place_first_pub |>
  mutate(same_institution = ifelse(unitid_graduate == unitid_author, 1, 0)) |>
  group_by(year_pq) |>
  summarise(same_institution = mean(same_institution, na.rm = T),
            .groups = "drop") |>
  ggplot(aes(x = year_pq, y = same_institution)) +
  geom_line() +
  geom_point()
```



If publishing during PhD, does so at least once at the PhD university?

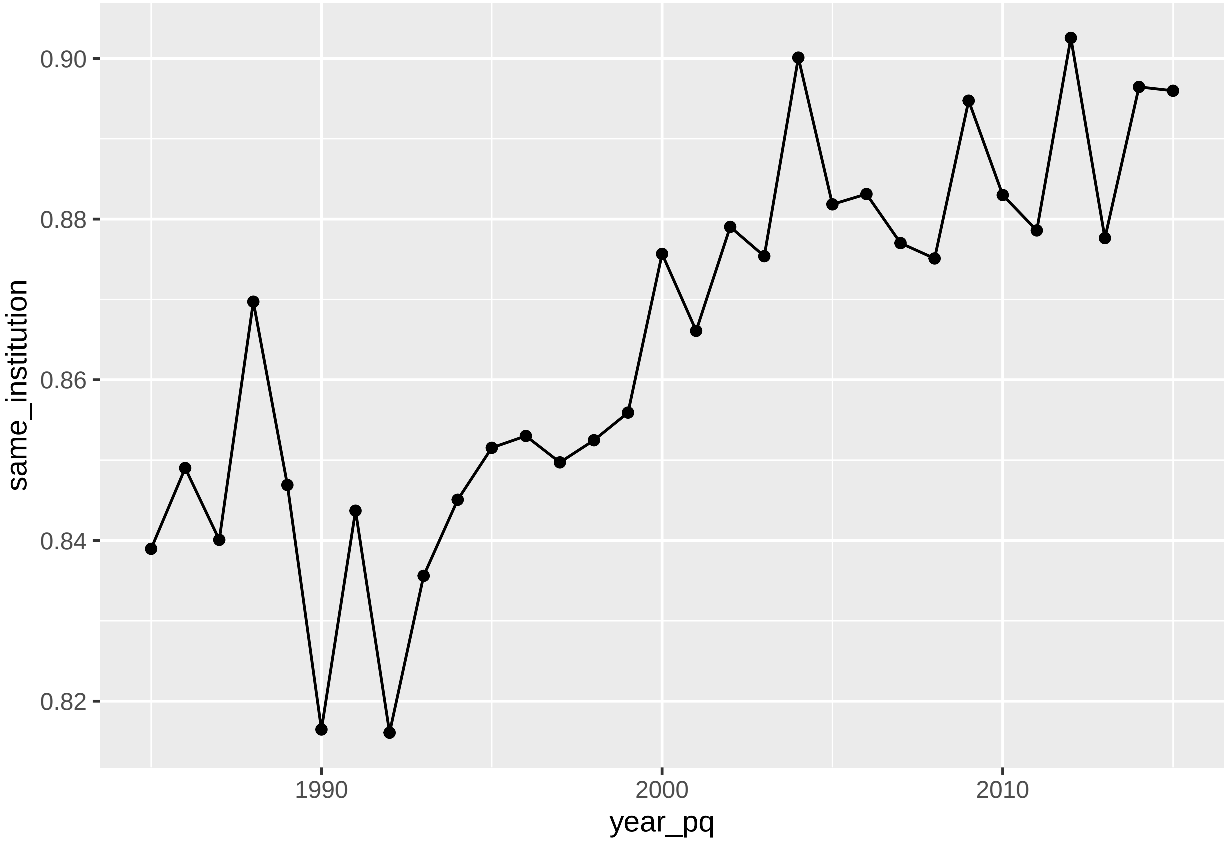
```r
publish_during_phd <- authors_affil |>
  left_join(grads_chemistry |>
              select(-grp),
            by = c("AuthorId")) |>
  filter(Year <= year_pq & Year >= year_pq - 6) |>
  left_join(links_to_cng |>
              filter(from_dataset == "mag") |>
              select(from_id, unitid_author = unitid),
            by = c("AffiliationId" = "from_id")) |>
  left_join(pq_authors |>
              select(goid, university_id),
            by = "goid") |>
  left_join(links_to_cng |>
              filter(from_dataset == "pq") |>
              select(from_id, unitid_graduate = unitid),
            by = c("university_id" = "from_id")) |>
  select(AuthorId, Year, year_pq, unitid_author, unitid_graduate, university_id) |>
  mutate(same_institution = ifelse(unitid_author == unitid_graduate, 1, 0),
         same_institution = ifelse(is.na(same_institution), 0, same_institution))
```

Fraction of students not publishing during PhD:

```r
1 - n_distinct(publish_during_phd$AuthorId) / n_distinct(grads_chemistry$AuthorId)
```

```
## [1] 0.2239471
```

```
# group by student: at least one pub with the PhD university?
publish_during_phd <- publish_during_phd |>
  group_by(AuthorId) |>
  filter(same_institution == max(same_institution)) |>
  filter(!duplicated(AuthorId))

publish_during_phd |>
  group_by(year_pq) |>
  summarise(same_institution = mean(same_institution, na.rm = T),
            .groups = "drop") |>
  ggplot(aes(x = year_pq, y = same_institution)) +
  geom_line() +
  geom_point()
```



```
summary(publish_during_phd)
```

```
##     AuthorId              Year         year_pq      unitid_author
##  Min.   :    797101   Min.   :1980   Min.   :1985   Min.   :100663
##  1st Qu.:2046494765   1st Qu.:1994   1st Qu.:1996   1st Qu.:144050
##  Median :2145361750   Median :2002   Median :2004   Median :174066
##  Mean   :2104704648   Mean   :2001   Mean   :2003   Mean   :181254
##  3rd Qu.:2435561831   3rd Qu.:2008   3rd Qu.:2010   3rd Qu.:212054
##  Max.   :3163604571   Max.   :2015   Max.   :2015   Max.   :495767
##                                                     NA's   :1793
##  unitid_graduate  university_id  same_institution
##  Min.   :100663   Min.   :   1   Min.   :0.0000
##  1st Qu.:144050   1st Qu.:  31   1st Qu.:1.0000
##  Median :174066   Median :  94   Median :1.0000
##  Mean   :180728   Mean   : 173   Mean   :0.8701
##  3rd Qu.:211440   3rd Qu.: 206   3rd Qu.:1.0000
##  Max.   :495767   Max.   :2849   Max.   :1.0000
##  NA's   :982
```

```
head(publish_during_phd |> filter(same_institution == 0))
```

```
## # A tibble: 6 x 7
## # Groups:   AuthorId [6]
##   AuthorId  Year year_pq unitid_author unitid_graduate university_id
##    <int64> <int>   <int>         <int>           <int>       <int64>
## 1  2387360  2004    2005        236948          131496           407
## 2  2683537  2000    2005        122597          141574           219
## 3  4924916  2001    2002        151111          243780            31
## 4  6283000  1990    1990            NA          131469           312
## 5  6395424  1999    1999        130943          176080           356
## 6  8227037  2002    2005            NA              NA           569
## # i 1 more variable: same_institution <dbl>
```

notes - some may publish after phd with the phd affiliation – not captured here - misses research institutes that are not in Carnegie, ie scripps research institute - all in all, this is a lower bound on the precision in the sample of people publishing during their PhD - the lower bound on precision for the sample of chemists can be calculated as follows - 19% publish after PhD; assume they are all false positives - of the remaining 81%, 87% publish at their graduating university - thus, our precision is at least 0.81 * 0.87 = 0.70 - this calculation is more difficult in fields where graduates publish more often after graduating