

Comparing links in chemistry

Flavio & Christoph

13 September, 2023

Contents

Chemistry: first affiliation of MAG authors should be the graduating institution. paper	3
Place of first publication	5
If publishing during PhD, does so at least once at the PhD university?	6

This document compares the final links we obtain for chemistry.

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

```
con <- DBI::dbConnect(RSQLite::SQLite(), db_file)
cat("The database connection is: \n")
```

The database connection is:

```
src_dbi(con)
```

```
## src:  sqlite 3.40.0 [/mnt/ssd/AcademicGraph/AcademicGraph.sqlite]
## tbls: affil_fields_temp, affiliation_fields, affiliation_outcomes,
## Affiliations, author_citations, author_coauthor, author_collab,
## author_field0, author_fields, author_fields_detailed, author_gender,
## author_info_linking, author_output, author_panel, author_performance,
## author_sample, author_selfcites, AuthorAffiliation, AuthorNameSplits,
## Authors, cng_distances, cng_institutions, cohort_career,
## cohort_career_decomp, conference_citations, crosswalk_fields, current_links,
## current_links_advisors, estimation_sample_art, estimation_sample_biology,
```

```
## estimation_sample_business, estimation_sample_chemistry,
## estimation_sample_computer_science, estimation_sample_economics,
## estimation_sample_engineering, estimation_sample_environmental_science,
## estimation_sample_geography, estimation_sample_geology,
## estimation_sample_history, estimation_sample_materials_science,
## estimation_sample_mathematics, estimation_sample_medicine,
## estimation_sample_philosophy, estimation_sample_physics,
## estimation_sample_political_science, estimation_sample_psychology,
## estimation_sample_sociology, FieldOfStudyChildren, FieldsOfStudy, FirstNames,
## FirstNamesGender, graduates_closest_collaborators,
## graduates_similarity_to_institutions, graduates_similarity_to_self,
## journal_citations, linked_ids, linked_ids_advisors, linked_ids_grants,
## linking_info, linking_info_advisors, linking_info_grants, links_new,
## links_old, links_to_cng, NSF_Appropriation, nsf_fields0_collapsed,
## nsf_fields1_collapsed, NSF_FieldsOfStudy, NSF_FoaInformation, NSF_Fund,
## NSF_Institution, NSF_Investigator, NSF_MAIN, NSF_Performance_Institution,
## NSF_ProgramElement, NSF_ProgramReference, nsffos, paper_citations,
## paper_outcomes, PaperAuthorAffiliations, PaperAuthorUnique,
## PaperFieldsOfStudy, PaperMainFieldsOfStudy, PaperReferences, Papers,
## pq_advisors, pq_authors, pq_fields, pq_fields_mag, pq_info_linking,
## pq_magfos, pq_unis, quantiles_papercites, sqlite_stat1, UnclearNamesGender
```

```
field_names_id <- tbl(con, sql(paste0(
  "SELECT FieldOfStudyId, NormalizedName
  FROM FieldsOfStudy
  WHERE Level = 0
        AND NormalizedName IN (",
    paste0(paste0("'", select_fields, "'"), collapse = ", "),
    ")")
)))
field_names_id <- collect(field_names_id)
```

```
query_mag <- paste0(
  "SELECT AuthorId
        , year
        , fieldofstudy
        , mag_field0
  FROM (
    SELECT a.AuthorId
          , a.YearFirstPub AS year
          , e.NormalizedName AS fieldofstudy
          , e.ParentFieldOfStudyId as mag_field0
    FROM author_sample AS A
    INNER JOIN (
      SELECT AuthorId, NormalizedName, ParentFieldOfStudyId
      FROM author_fields c
      INNER JOIN (
        SELECT FieldOfStudyId, NormalizedName
        FROM FieldsOfStudy
      ) AS d USING(FieldOfStudyId)
      INNER JOIN (
        SELECT ParentFieldOfStudyId
          , ChildFieldOfStudyId
```

```

        , ParentFieldOfStudyId
      FROM crosswalk_fields
      WHERE ParentLevel = 0
        AND ParentFieldOfStudyId IN (",
          paste0(field_names_id$FieldOfStudyId, collapse = ", "),
        ")
    ) AS e ON (e.ChildFieldOfStudyId = c.FieldOfStudyId)
    WHERE FieldClass = 'first'
  ) AS e USING(AuthorId)
)
WHERE year >= 1980 and year <= 2022
")

```

```

linked_ids <- tbl(con, "current_links")
linking_info <- tbl(con, "linking_info") %>%
  filter(mergemode == "1:1" & fielfofstudy_str == "False")
pq_authors <- get_proquest(conn = con, from = "graduates", start_year = 1990, end_year = 2015)
mag_authors <- tbl(con, sql(query_mag))

```

Combine data pq and MAG

```

d_linked <- linked_ids |>
  select(AuthorId, goid) |>
  left_join(pq_authors |>
    select(-gender),
    by = "goid") |>
  left_join(mag_authors |>
    select(AuthorId, year_firstpub = year),
    by = "AuthorId")

```

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

```

d_main <- d_linked |>
  filter(fieldname0_mag == "chemistry") |>
  mutate(grp = case_when( # some people publish already way before the PhD
    year_firstpub > degree_year ~ "first pub after PhD",
    year_firstpub < degree_year - 6 ~ "first pub before PhD",
    TRUE ~ "first pub during PhD"
  )) |>
  select(AuthorId, goid, degree_year, grp)

```

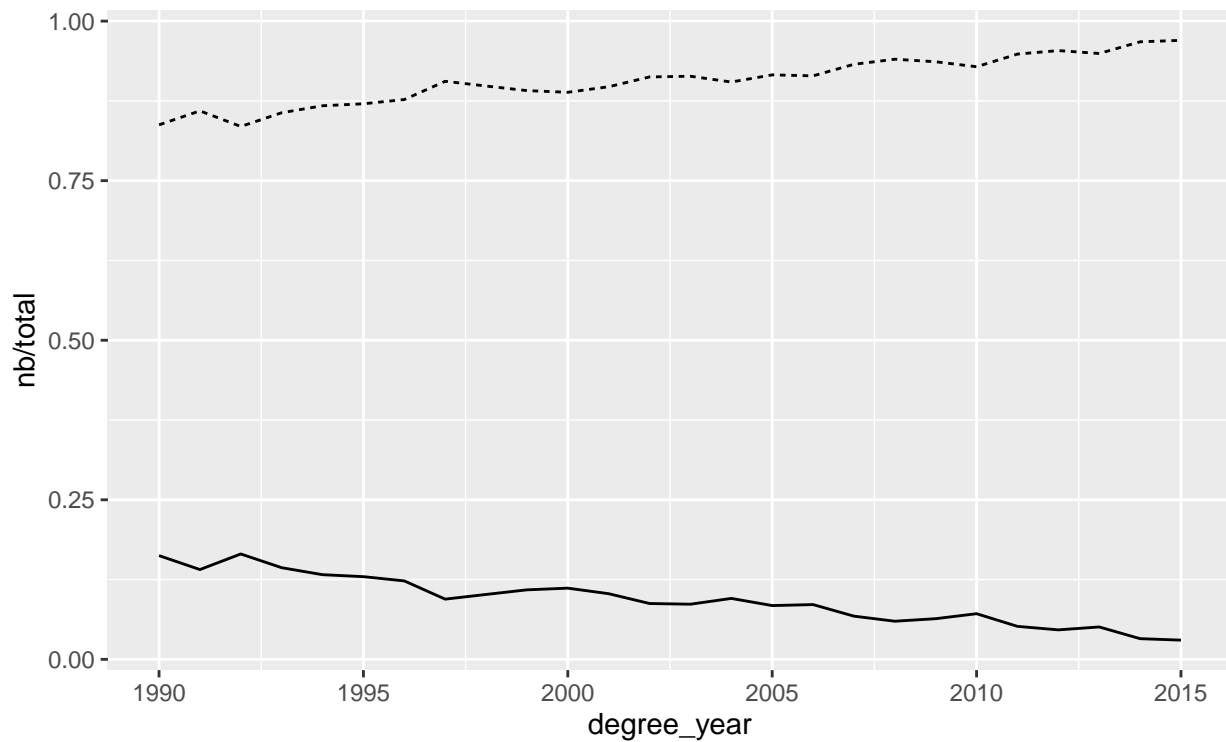
```

d_main |>
  group_by(grp, degree_year) |>
  summarise(nb = n()) |>
  ungroup() |>
  group_by(degree_year) |>
  mutate(total = sum(nb)) |>
  ggplot(aes(x = degree_year, 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.
```



grp — first pub after PhD ---- first pub during PhD

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

```
d_main |>
  filter(degree_year >= 1999 & degree_year <= 2008) |>
  summarise(n())
```

```
## # A tibble: 1 x 1
##   `n()`
##   <int>
## 1  9407
```

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

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

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

```

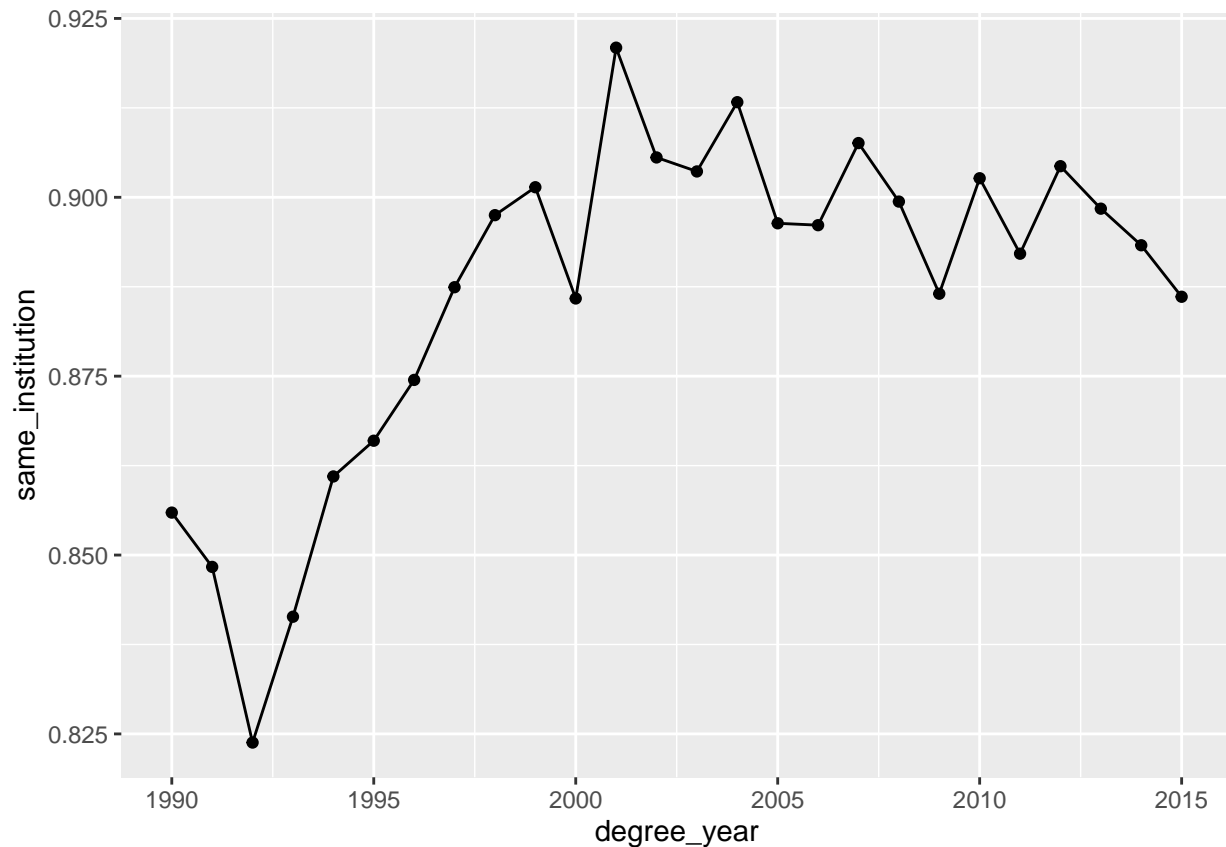
Place of first publication

```

place_first_pub <- d_main |>
  left_join(pq_authors |>
    select(goid, university_id),
    by = "goid") |>
  inner_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") |>
  inner_join(links_to_cng |>
    filter(from_dataset == "mag") |>
    select(from_id, unitid_author = unitid),
    by = c("AffiliationId" = "from_id"))

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

```



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

```
publish_during_phd <- authors_affil |>
  left_join(d_main |>
    select(-grp),
    by = c("AuthorId")) |>
  filter(Year <= degree_year & Year >= degree_year - 6) |>
  inner_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") |>
  inner_join(links_to_cng |>
    filter(from_dataset == "pq") |>
    select(from_id, unitid_graduate = unitid),
    by = c("university_id" = "from_id")) |>
  select(AuthorId, Year, degree_year, 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:

```
1 - n_distinct(publish_during_phd$AuthorId) / n_distinct(d_main$AuthorId)
```

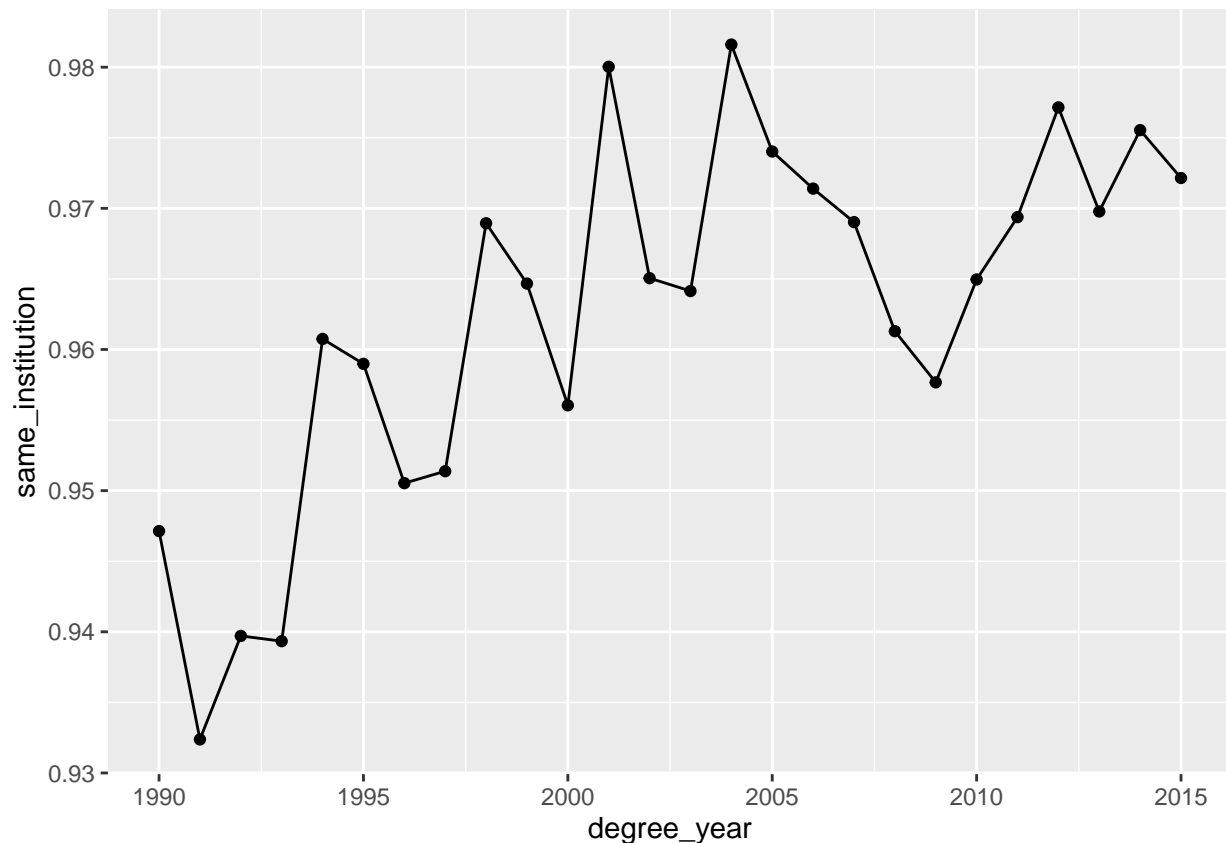
```
## [1] 0.1776644
```

```

# 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(degree_year) |>
  summarise(same_institution = mean(same_institution, na.rm = T),
            .groups = "drop") |>
  ggplot(aes(x = degree_year, y = same_institution)) +
  geom_line() +
  geom_point()

```



```
summary(publish_during_phd)
```

```

##      AuthorId      Year      degree_year      unitid_author
##  Min.   : 590343  Min.   :1985  Min.   :1990  Min.   :100663
## 1st Qu.:2020547864 1st Qu.:1997 1st Qu.:1999 1st Qu.:141574
## Median :2130305377 Median :2004 Median :2006 Median :174066
## Mean   :2039660888 Mean   :2003 Mean   :2005 Mean   :181846
## 3rd Qu.:2318164168 3rd Qu.:2009 3rd Qu.:2011 3rd Qu.:212106
## Max.   :3163059217 Max.   :2015 Max.   :2015 Max.   :495767
## unitid_graduate university_id same_institution
##  Min.   :100663  Min.   : 1  Min.   :0.0000
## 1st Qu.:141574  1st Qu.: 29 1st Qu.:1.0000
## Median :174066  Median : 88  Median :1.0000

```

##	Mean	:181789	Mean	: 139	Mean	:0.9648
##	3rd Qu.	:212106	3rd Qu.	: 184	3rd Qu.	:1.0000
##	Max.	:495767	Max.	:2589	Max.	:1.0000

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
 - chemical engineers
- all in all, this is a lower bound on the precision in the sample of chemists publishing during their PhD
- the lower bound on precision for the sample of chemists can be calculated as follows
 - 18% publish after PhD; assume they are all false positives
 - of the remaining 82%, 96% publish at their graduating university
 - thus, our precision is at least $0.82 * 0.96 = 0.78$
- this calculation is more difficult in fields where graduates publish more often after graduating