# Sample size for linking: some computations

Flavio & Christoph

02 September, 2024

## Contents

Note: I have not run this script but only copied from another. Perhaps something does not work!

# How to extend the coverage of interdisciplinary people?

## What do we do now?

- for graduates: we take the field level 1 of the person at the beginning of the career.
- the person is sampled for linking when this field level 1 is a subfield of biology, in *our* correspondence between parents and children (which we built because the correspondence from children to parents is not always unique). This table `crosswalk_fields` is generated in `main/prep_mag/paper_fields.py`.
- this sampling strategy misses people whose first field does not map into biology—again according to our correspondence.

## What should we be do instead?

- When linking biology graduates, sample from MAG so that it also covers fields in chemistry where biologists often/sometimes/tend to work.
- "Find all fields level 1 (and lower?) that are similar to biology, but whose assigned parent is *not* biology"

## How can we do it?

### Approach 1: extend the parent definition, using the mag table

- include all entities from mag that have a first field whose parent field can be in biology
- but would it be sufficient to cover the relevant cases?

- the examples from graduates suggest not
- can we extend? to more levels?

### Approach 2: use empirical distributions of papers by field level 1

1. take all papers with main paper in biology (assigned by us: table `papermainfieldsofstudy`).
2. count number of papers by fieldofstudyid level 1.
3. take top X fields of study level 1 whose parent is NOT biology (table `fieldofstudychildren`)
4. include all researchers whose first field falls into these fields when linking biology.

### Notes

- How different are the two approaches? → How does MAG build its correspondence?
- How can we check whether this "works" as expected?

## Solution: children and grand children of the proquest dissertation field

1. Get all fields that the researcher ever publishes in their career (from paperfieldsofstudy)
2. Get children and grandchildren of biology
3. Get authors who publish at least once in one of the fields from 2

## Investigate the field0 of authors

```r
author_field0 <- tbl(con, "author_field0") %>%
  # focus on the sample we are interestd in
  inner_join(tbl(con, "author_info_linking") %>%
               filter(!is.na(main_us_institutions_career)) %>%
               select(AuthorId),
             by = "AuthorId"
             )
author_field0 <- collect(author_field0)
```

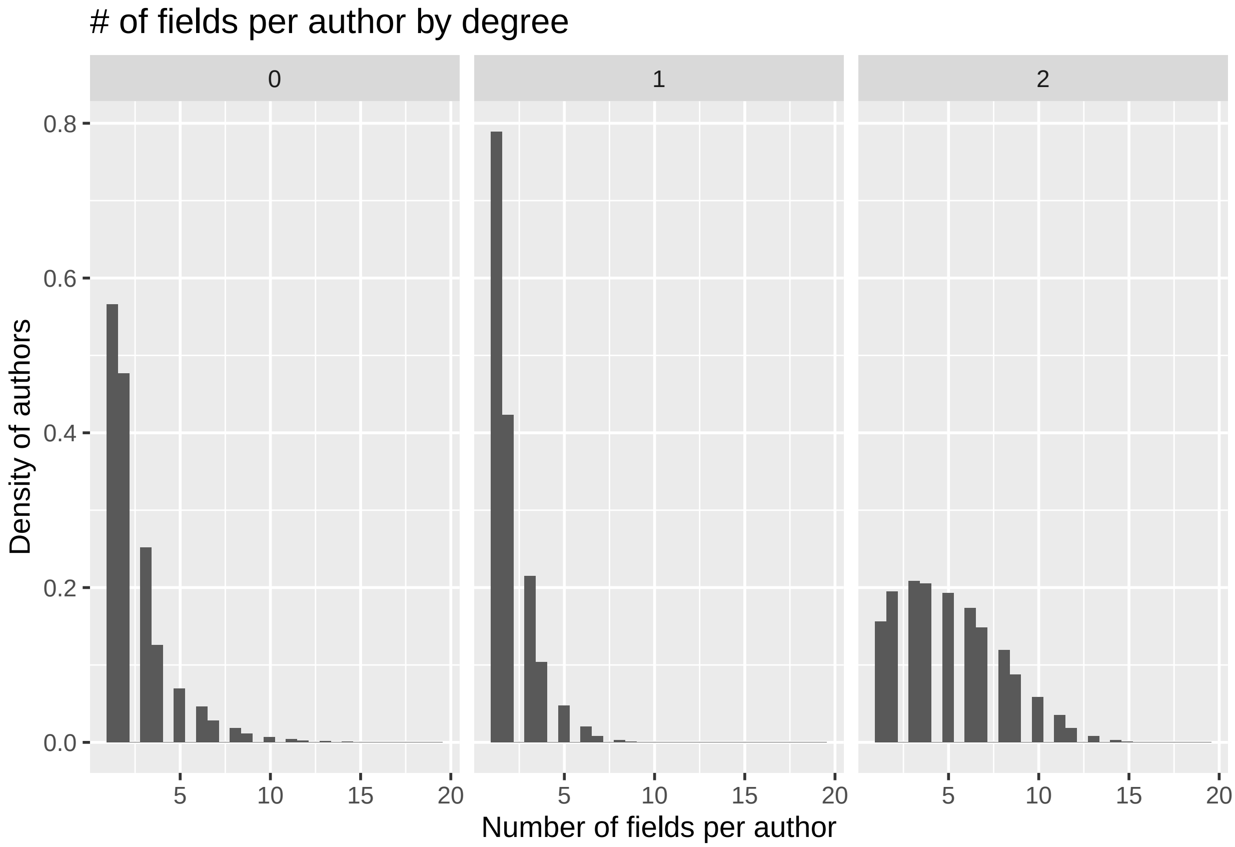
```r
sample_size <- 100000

# sample the authors
sampled_authors <- unique(author_field0$AuthorId) %>%
  data.frame()
names(sampled_authors) <- "AuthorId"
sampled_authors <- sampled_authors %>%
  slice_sample(n = sample_size)
```

```r
xlab <- "Number of fields per author"
ylab <- "Density of authors"

field_count <- author_field0 %>%
  filter(AuthorId %in% sampled_authors$AuthorId) %>%
  group_by(AuthorId, Degree) %>%
  summarise(nb = n(),
            .groups = "drop")
```

```
field_count %>%
  ggplot(aes(x = nb)) +
  geom_histogram(aes(y = ..density..)) +
  facet_wrap(~factor(Degree)) +
  labs(x = xlab, y = ylab, title = "# of fields per author by degree")
```

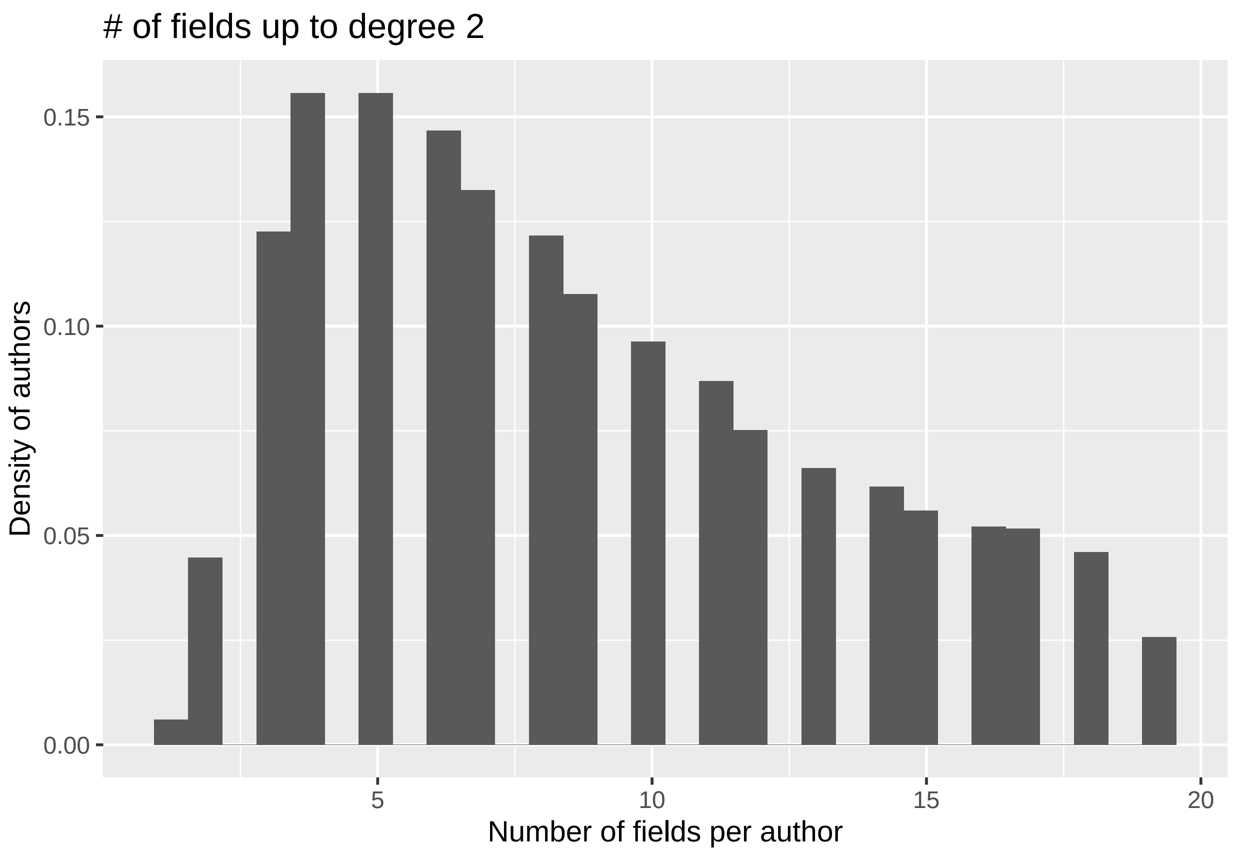
```
## Warning: The dot-dot notation (`..density..`) was deprecated in ggplot2 3.4.0.
## i Please use `after_stat(density)` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

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



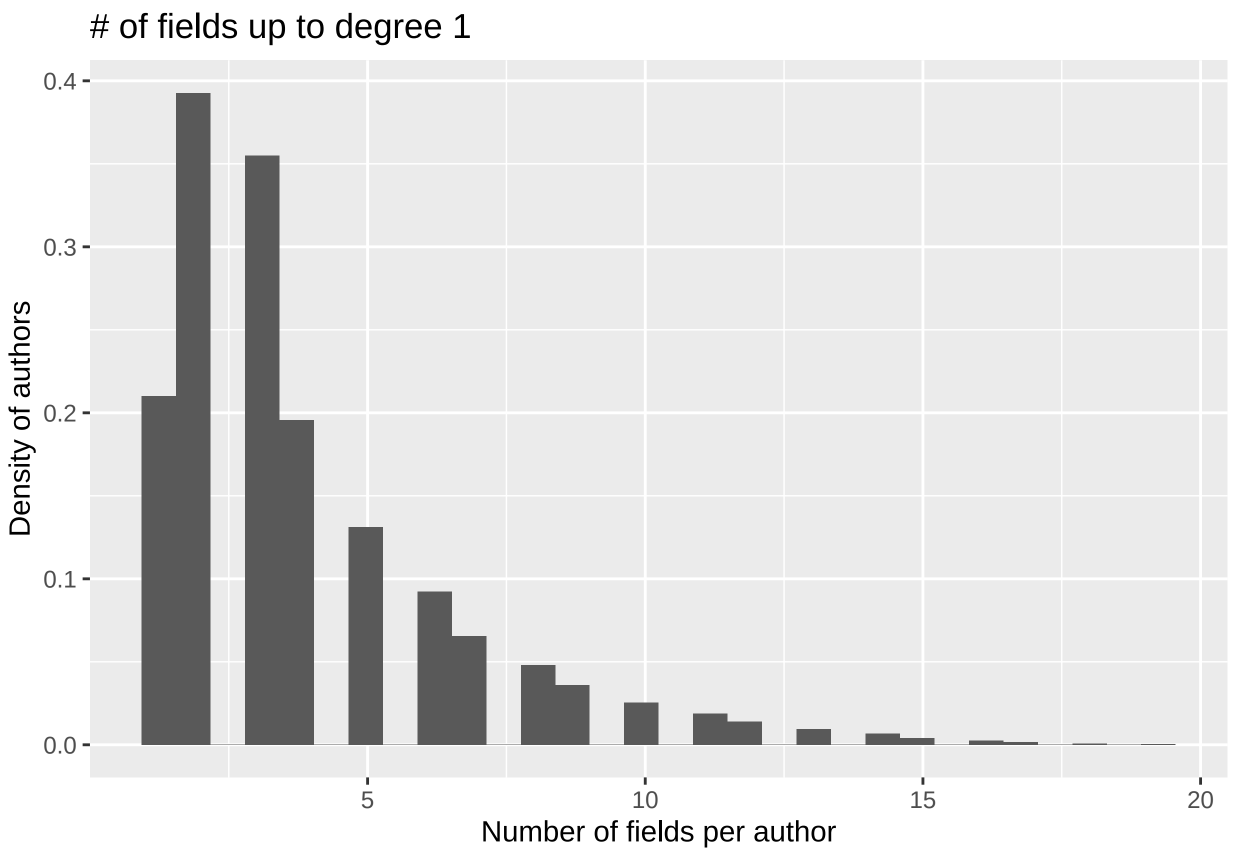
```
author_field0 %>%
  filter(AuthorId %in% sampled_authors$AuthorId) %>%
  group_by(AuthorId) %>%
  summarise(nb = n_distinct(FieldOfStudyId_lvl0),
            .groups = "drop")  %>%
  ggplot(aes(x = nb)) +
  geom_histogram(aes(y = ..density..)) +
  labs(x = xlab, y = ylab, title = "# of fields up to degree 2")
```

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



```
author_field0 %>%
  filter(AuthorId %in% sampled_authors$AuthorId) %>%
  filter(Degree <= 1) %>%
  group_by(AuthorId) %>%
  summarise(nb = n_distinct(FieldOfStudyId_lvl0),
            .groups = "drop")  %>%
  ggplot(aes(x = nb)) +
  geom_histogram(aes(y = ..density..)) +
  labs(x = xlab, y = ylab, title = "# of fields up to degree 1")
```

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



```
author_field0 %>%
  filter(AuthorId %in% sampled_authors$AuthorId) %>%
  group_by(AuthorId) %>%
  summarise(nb = n_distinct(FieldOfStudyId_lvl0),
            .groups = "drop") %>%
  summary()
```

```
##     AuthorId                 nb
##  Min.   :     65055   Min.   : 1.000
##  1st Qu.:2124167748   1st Qu.: 5.000
##  Median :2425936830   Median : 8.000
##  Mean   :2358362743   Mean   : 8.708
##  3rd Qu.:2683914536   3rd Qu.:12.000
##  Max.   :3163940281   Max.   :19.000
```

## Notes

- Performance with the entities not linked
  - Graduates above
    - hansen has chemistry only in degree 2!
      - this would not change even if we only look at papers in the early stage of the career
    - tokmakoff has chemistry in degree 0
    - weldon has chemistry in degree 0
    - marcus has chemistry in degree 0
    - This still does not address the missing links from actual
      - thus, up to field 1 should work (maybe even field 0); would still not capture all; we still haven't address the missing links of actual chemists

    - from advisor linking
        - in biology and sociology, we would capture sample all of the advisors in the respective field of the dissertation
        - and most of them with only degree 0
- Many people have many fields!
    - can we do better than this?
- Next steps
    - implement and try it out for chemistry graduates with degree <= 1?
    - think about other improvements for the currently unlinked?

## Compare number of MAG authors loaded for linking with old and new approach

```
count_sample <- function(field, con, max_degree = 1) {

  cat(field, "\n")

  q_base <- paste0("
  select count(*) as n_authors
  from (
    select authorid
    from author_sample
    inner 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 g
        inner join (
          select fieldofstudyid
          from fieldsofstudy
          where normalizedname = '", field, "'
        ) f on (g.ParentFieldOfStudyId = f.FieldOfstudyId)
        where parentlevel = 0
      ) as e on (e.childfieldofstudyid = c.fieldofstudyid)
      where fieldclass = 'first'
    ) as e using(authorid)
    inner join (
      select authorid
      from author_info_linking
      where main_us_institutions_career is not null
    ) using (authorid)
    where yearfirstpub > 1980
  )
  ")

  q_new <- paste0(
    "select count(*) as n_authors
    from (
      select authorid
      from author_sample
      inner join (
        select authorid
        from author_info_linking
        where main_us_institutions_career is not null
      ) using (authorid)
      inner join (
        select authorid
        from author_field0
        inner join (
          select fieldofstudyid
          from fieldsofstudy
          where normalizedname = '", field, "'
        ) on (fieldofstudyid_lvl0 = fieldofstudyid)
        where degree <= ", max_degree, "
      ) using(authorid)
      where yearfirstpub > 1980
    )
    "
  )

  d_base <- tbl(con, sql(q_base)) %>%
    collect()
  d_new <- tbl(con, sql(q_new)) %>%
    collect()

  d <- tibble(
    n_base = d_base$n_authors,
    n_new = d_new$n_authors
    ) %>%
    mutate(field = field,
           max_degree = max_degree)

  return(d)

}
```

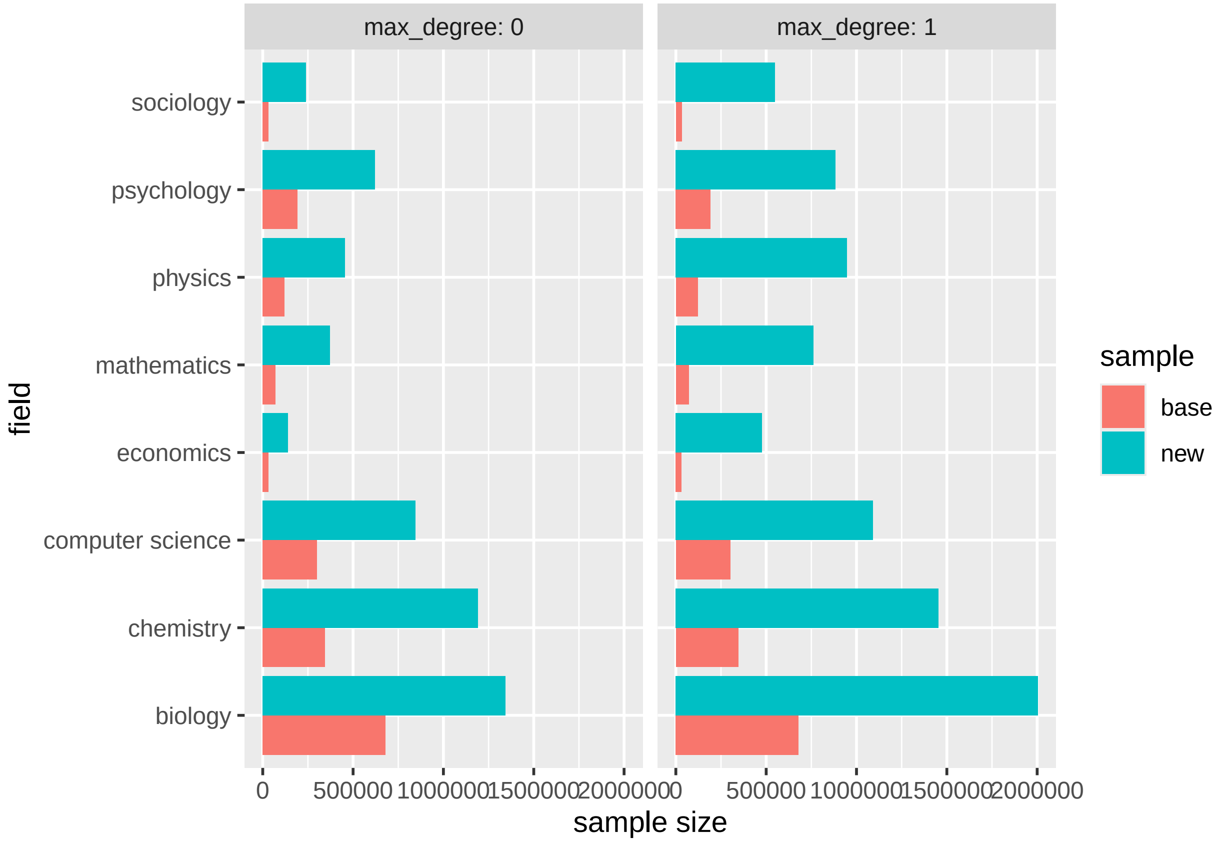
```
fields <- c("biology", "chemistry", "computer science",
            "mathematics", "psychology", "sociology",
            "psychology", "physics", "economics")

sizes_degree1 <- map(
  .x = fields,
  .f = ~count_sample(field = .x, con = con, max_degree = 1)
)
```

```
## biology
## chemistry
## computer science
## mathematics
## psychology
## sociology
## psychology
## physics
## economics
```

```r
sizes_degree0 <-  map(
  .x = fields,
  .f = ~count_sample(field = .x, con = con, max_degree = 0)
)
```

```
## biology
## chemistry
## computer science
## mathematics
## psychology
## sociology
## psychology
## physics
## economics
```

```r
bind_rows(
  sizes_degree0 %>% bind_rows(),
  sizes_degree1 %>% bind_rows() %>% mutate(max_degree = 1)
) %>%
  pivot_longer(cols = starts_with("n_"),
               names_to = "sample",
               values_to = "nb") %>%
  mutate(sample = gsub("n_", "", sample),
         max_degree = paste0("max_degree: ", max_degree)) %>%
  ggplot(aes(x = field, y = nb)) +
  geom_bar(stat = "identity",
           aes(fill = sample),
           position = position_dodge()) +
  labs(y = "sample size") +
  facet_wrap(~max_degree) +
  coord_flip()
```


max_degree: 0
max_degree: 1
sociology
psychology
physics
mathematics
economics
computer science
chemistry
biology
field
0
500000
1000000
1500000
2000000
sample size
sample
base
new