Online appendix to: Networks in the market for researchers

Flavio Hafner, Christoph Hedtrich

December 6, 2023

Contents

1 Data appendix

The broad steps in our data pipeline are shown in figure 1. The following subsections explain the details.

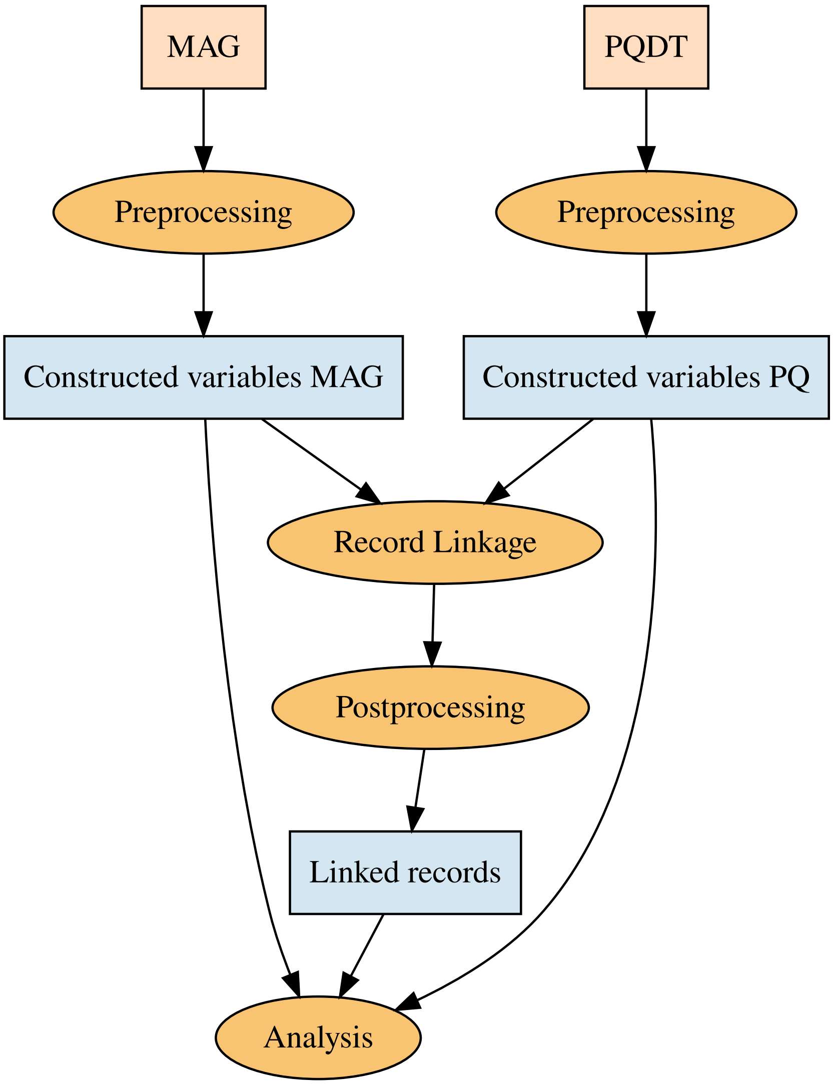

Figure 1: The end-to-end data pipeline.

The figure shows a schema of the data pipeline. Oval nodes are data handling steps; rectangular nodes are data (single or multiple relational tables). See text for details.

1.1 Data sources

1.1.1 Microsoft Academic Graph (MAG)

The data are collected from the internet by Microsoft's search engine Bing and from RSS Feeds of publishers (Sinha et al, 2015). Six entities are then extracted from the data: Author, Institution,[1] Paper, Field of Study, Venue (e.g., "American Economic Review") and Event (e.g.,

[1] Institutions can universities, firms or international organizations. Since we focus on universities, we will refer to this entity as "university".

a specific issue of the AER). For some entities (Author, Affiliation), the database reports both the original string and a disambiguated string together with an entity ID generated by MAG's algorithms, described in more detail by Sinha et al. (2015). The Affiliation entity is extracted from the metadata of published papers, and it is possible that an author has multiple affiliations for the same paper.

To create the Field of Study entity, they use additional proprietary in-house data and online data. They find the procedure to be highly accurate. There are overall five levels of Field of Study, and the labels represent the semantics of a paper. At the highest level 0, the labels roughly correspond to major fields of study such as chemistry, biology and mathematics. A paper can have several such Field of Study from any level. The Field of Study is what we use to calculate the similarity in research concepts as discussed in the main text.

Compared to other databases often used in bibliometric analysis such as Scopus or Web of Science, several studies find MAG to be comparable in coverage. Visser, van Eck and Waltman (2021) benchmark several databases (MAG, Crossref, Dimensions, and a restricted version of Web of Science) against Scopus for the years 2008 to 2017. They report that MAG has the highest coverage (81 percent). They also report that MAG covers many documents that are not in Scopus, and in a random sample of these, that many of them are of a scientific nature. Martín-Martín et al. (2020) report that Microsoft Academic has the second-highest coverage of citations among a set of services including Google Scholar, Scopus and Web of Science. Microsoft Academic finds 60% of all citations, with good coverage for many fields except Physics and some Humanities categories. Hug, Ochsner and Brändle (2017) find that a citation analysis with MAG gives the same results as with Scopus. Although they find some errors in metadata on papers such as missing authors and wrong year of publication, they believe that MAG has the potential for "full-fledged bibliometric analysis". Hug and Brändle (2017) benchmark the quality and coverage of MAG with Scopus and Web of Science for the research output of a specific university. They find that high rank-rank correlations between the databases in citation counts are high and MAG has a good coverage of venue entities. Harzing and Alakangas (2017) compare citation counts for a set of academics in different disciplines (Engineering, Social Sciences, Humanities, Life Sciences, Sciences) and find that MAG, behind Google Scholar, performs at least as good or better than WoS and Scopus across discplines.

1.1.2 Proquest Dissertation & Theses (PQDT)

PQDT is a repository of PhD theses, provided by Clarivate Analytics. It is to our knowledge the most complete repository of PhD theses in the United States. We use the metadata of this database: the name, graduating university, reported advisors of the student, the title and the keywords of the dissertation.

1.2 Preprocessing

MAG Based on the MAG database, we build several intermediate tables for each author. We extract the major fields of study they publish in and the start and end of their publishing career. We extract the unique affiliation-year combinations of their published papers. From the first ten papers published within the first five years of an author's publication career, we extract:

the unique keywords, and the year-title combinations of the publications. The keywords are the MAG fields of study labels at level 1. The year-title combinations are stored as a list of tuples: : `[(int, string), (int, string), ...]` , where `int` and `string` are integer and string data types.

We extract the unique combination of year and US university of the author's publications across the author's career. We store them also as list of tuples as previously described. Because an author can publish papers with different affiliations in the same year, we use two types of universities: `main` and `all`. `main` refers to the one affiliation in a year at which the author publishes most of their papers.

MAG is built from the raw data of each paper, and thus requires entity disambiguation. The disambiguation in MAG is conservative in a sense that high precision is traded off for lower recall. For instance, this creates a lot of entities that are duplicates of other author entities. These duplicates, however, tend to be associated with only one or a few papers, and have shorter publication careers than the "ground truth" author entities. For this reason, we follow Huang et al. (2020) and restrict the author sample to entities that publish at least two articles during their whole career and, on average, at most 20 papers per year. We call this the main author sample.

ProQuest Starting with the individual dissertation files, we extract tables for advisors, authors, tagged fields of study,[2] and predicted fields of study. The predicted fields of study are predicted from the abstract of each thesis with the language model from MAG. We keep the first 10 predicted fields as long as their score is above 0.4 (a heuristic threshold, below which we found the predictions to be noisy). We export these tables from the Clarivate server and process them further.

We use the dissertation identifier from PQDT as an identifier for graduates. As advisors are not uniquely identified, we create identifiers called `relationship_id`. They uniquely identify the graduate and the position in which the advisor appears in the metadata. We uniquely identify universities based on their unique name. We correspond the fields of study reported in ProQuest to the major fields in MAG by hand, based on the MAG data and based on the classification by Organisation for Economic Co-operation and Development (2007).

Crosswalking university entities In several steps of our pipeline, it is important to distinguish universities from each other, but reported names differ between MAG and ProQuest. We create a crosswalk of universities from the set of US universities on the Carnegie classification (American Council on Education, 2021): We correspond the university identifiers from MAG and from PQDT to the Carnegie list. We do so with the name and where possible the zip code and city of the university; for this task we use the table of zip codes from Missouri Census Data Center (2020). We link all R1 universities, most of R2 and R3 universities and some other universities. Where necessary, we then use the university name from the Carnegie classification, to which we refer as the crosswalked university name.

[2] These tags are readily available, but are different from the labels used in MAG.

1.3 Record linkage

We look for graduates and advisors in ProQuest that publish in MAG with the following steps.

1.3.1 Data preparation

We normalize the author and university names in both data sets. We extract the first, last and middle author name: the first name is the first string of the name until the first white space. The last name is defined equivalently, starting from the end of the name. The middle name is any string occurring in between.

1.3.2 Learning algorithm: dedupe

We use the open-source library `dedupe` (Gregg and Eder, 2022) to train our record linking algorithm. `dedupe` learns with two interdependent steps: learning blocking rules and learning distance functions.

Learning blocking rules with predicate functions The first step is to use blocking which assigns entities into blocks of maximum similarity. `dedupe` learns a blocking rule for all the specified features (except the custom comparators). A blocking rule consists of predicate functions such as "the first three characters" of a feature (the first name, for instance). `Dedupe` also tries out cross-field blocks such as "same city" and "same zip code". The algorithm selects the smallest number of blocking rules that cover all labeled pairs but minimizes the number of pairs to be compared. A learned blocking rule can also mean that the feature is not used for blocking.

Learning distance functions The second step is to calculate the similarity within the blocks. The algorithm compares all pairs with each other and predicts,[3] with logistic regression, how likely it is that a pair of records refer to the same entity. Then, it groups links together by hierarchical clustering (centroid linkage clustering) within blocks. The algorithm uses the predicted values from the logistic regression as the distance measure between entities.

Active learning For training the model, `dedupe` uses active learning. Active learning selects pairs for labeling which are expected to increase accuracy of the model the most when added to the training set. In particular, the algorithm selects random pairs for labeling from a pool of pairs where the current learned blocking rule clusters a pair together, but the current learned classifier does not group the pair as a match. The reverse is also possible. After labeling one pair, the disagreement set is updated and a new pair for labeling selected. The algorithm works with as few as 10 pairs labeled as a match and 10 pairs labeled as a non-match, but we usually aim for a larger training set. Moreover, in the labeling one can also skip pairs when one is not sure whether they are a match.

[3] We describe in section 1.3.5 the different distance functions for different data types.

Pros and cons The advantages of `dedupe` are its flexibility and efficiency: The blocking rules are learned from the data, and the pairs that improve prediction the most are labeled by humans. This is also a possible disadvantage because we lack a ground truth. We discuss these concerns in detail below and address them empirically as best as possible.

1.3.3 Sampling records for linking: graduates

From the sample of graduates in PQDT, we search for links according to the following criteria:

- they graduate between 1990 and 2015
- their first name has at least two characters
- the university is in the United States

From the main author sample in MAG, we search for links among authors in the main authors table according to the following criteria:

- their first name has at least two characters
- they start publishing between 1985 and 2020
- they publish at least once with an affiliation in the United States

We do the linking in batches—separately for each major field of study. For example, for the field "mathematics", we load all graduates from ProQuest whose dissertation is classified in this field. From MAG, we load all authors that ever publish at least one paper that is tagged with "mathematics". This approach should also cover a large set of interdisciplinary researchers, because an author in MAG appears in the linking for multiple fields. For instance, we consider physicists who published a paper in mathematics once to be possible links for mathematics dissertations. This approach also could create duplicated links when a record is linked in two different fields. We take care of such cases in the postprocessing step.

1.3.4 Sampling records for linking: advisors

From the universe of advisors in PQDT, we search for links in a sample defined according to the following criteria:

- the graduate finished the PhD between 1990 and 2015
- the university is in the United States

From the main author sample in MAG, we search for links in a sample defined according to the following criteria:

- their year of first publication is before 2020 and their year of last publication is after 1985
- who publish at least once with an affiliation in the United States

We link for each major field of study in the same way as for graduates.

1.3.5 Comparing records

One can specify the model features that the algorithm uses to predict whether a pair refers to the same entity or not. Possible feature types are strings, number, categorical variables, date and time, whether a field has a missing value, as well as any interaction between the existing variables. When comparing strings, `dedupe` learns a distance function using affine gap distance functions. This gives more flexibility to the specific learning task at hand. When comparing numbers, `dedupe` calculates the difference between the logarithms of the two. When comparing sets, `dedupe` explores different measures of set similarity such as “one common element”, “two common elements”, “first common element” and term frequency-inverse document frequency.

Custom comparators Dedupe also allows custom comparator functions. We construct the following.

- The `year_title_comparator` returns the maximum similarity of the titles in the year-title tuples between two records (the information on the year is ignored). The similarity is calculated as the Term frequency-inverse document frequency (Tfidf) on the titles, after stemming them with the Snowball stemmer for English.
- The function `compare_range_from_tuple` compares whether the number in record x is in the range between y_1 and y_2 in record 2.
- The functions `compare_startrange_from_tuple` and `compare_endrange_from_tuple` return 0 when the singleton year of the first tuple lies within the year range defined in the second tuple, and the difference in years to the start (end) of the second tuple otherwise.
- The `set_of_tuples_distance` comparator is a family of functions that compare either the first, the second or both entries of all elements in a list of tuples. For string types it uses the Jaro-Winkler similarity; for numeric types it uses the difference in logs. When comparing both entries in a tuple it returns the product of the similarities of the first and second entries, respectively. When we use the `set_of_tuples_distance` comparator we use all three family members as separate features for the same list of tuples.

1.3.6 Linking PhD Graduates from PQDT to MAG Authors

We search for one-to-one links between graduates in PQDT and authors in MAG using the following features.

- A string comparator for first name, last name and middle name.
- An indicator for whether the first and lastname match exactly.
- A set comparator for the dissertation keywords and the keywords in the publications at the start of the publishing career in MAG. The keywords are the fields of study entity from the MAG semantic language model. We use the fields at level 1 and aggregate fields at level 2 to 5 to their most likely parent at level 1. We only consider fields where the algorithm is confident enough about the field (a score of at least 0.4).

- A number comparator for the year, which is the year of graduation in PQDT and the year of first publication in MAG. We also interact the number comparator with an indicator whether the number is negative; this allows for varying slopes for positive and negative differences.
- An interaction firstname × year and lastname × year.
- A `year_title_comparator` that compares the year and thesis title in PQDT and the set of paper titles and publication years in MAG at the start of the career. We also interact this variable with the similarity of the first and lastname as well as the similarity of the year.

We do not use the graduating university as a feature because graduates may only publish their first paper after their PhD with an affiliation different from their PhD university. Because our sample selection for linking described previously could link the same person multiple times (in different fields of study), we only use links where the entity in MAG is linked only once to ProQuest and vice versa.

1.3.7 Linking PhD Advisors from PQDT to MAG Authors

Because an author in MAG can be advisor for multiple theses, we search for many-to-one links between advisors in PQDT and authors in MAG. We use the following features.

- A string comparator for first name, last name and middle name.
- An indicator for whether the first and lastname match exactly.
- For the student's graduating year in PQDT and the start and end year of the advisor's publishing career, a `compare_range_from_tuple`, a `compare_startrange_from_tuple` and a `compare_endrange_from_tuple` comparator.
- A `set_of_tuples_distance` comparator for the tuples `(year, university)` of the students' graduating year and crosswalked university name in PQDT and the advisor's publication year and crosswalked university name in MAG, respectively. In the latter, we use both the `main` and `all` universities; they are highly correlated but the cross-validation in the algorithm will select the relevant one depending on the context.
- We also interact the year and `main` university similarities with the similarities of the authors' first and last names.

1.3.8 Dedupe parameters

`dedupe` requires a few user-supplied parameters. First, we use a sample size for training of 50'000 for graduates and 100'000 for advisors. Second, we set the algorithm to propose as possible links to be labelled a blocked pair 2/3 of the time, and a random pair 1/3 of the time. The blocked pairs are taken from the current blocking rules. Third, we set the recall to 0.9, which means that the blocking rule needs to include at least 90 percent of pairs labelled as true links in the same block. Finally, we consider links with a score of at least 0.7.

1.3.9 Training

Using the above setup, each of the the two authors created one training sample per field of study. We provided 40 to 60 labelled pairs and followed the following protocol. In order to label a proposed link as a true link, the following needs to hold

1. For students
   (a) Similar names.
   (b) Overlap in paper titles or overlap in keywords.
   (c) Graduating year and year of first publication not more than 10 years apart.
2. For advisors
   (a) Similar names.
   (b) The student does not graduate before the year of the first publication of the advisor.
   (c) The advisor is at a university with a similar name as the student in a window around the student's graduation.[4] Proposed links that are more than 10 years apart are labelled as "no", even if the previous conditions are true.

Each of us then labelled data for each of the research fields. Within the above boundaries, we used our own judgement to actively label proposed links.

We then trained models with the training data from both labellers, and predicted respective links.

1.4 Postprocessing of the predicted links

In this step, we combine the predicted links from the two models into a single prediction. The step is based on a comparison as illustrated in table 1 for graduates and table 2. For each field, we make a full join of the predicted links of the two models. Starting from the sample of graduates/advisors in ProQuest, this then allows us to classify the predicted links into four groups: those where the predicted MAG entity is the same (column "Same entity"), where only one of the models found a link ("Only by 1", "Only by 2"), and where both models found a link but to different entities ("Different entity"). The table reports these fractions for each field of study as well as a weighted average across all fields. We will return to these numbers in subsection 1.5.

We combine the links as follows. First, using the comparison previously explained, we only keep predicted links where the predictions from the two models agree. This means that, for both graduate and advisor entities, we accept links of three kinds:

1. The two models link the same MAG entity identifier to the ProQuest entity.
2. Only one of the models predicts a link, but the entity name in MAG is very similar to the entity name in ProQuest—a Jaro-Winkler similarity of 0.9 or more.

[4] This often resulted in requiring exactly matching university names, although in some special cases (such as the University of California system) it did not.

	Fraction of links found				
Field	Same entity	Only by 1	Only by 2	Different entity	Number of links
Biology	0.67	0.29	0.03	0.01	51420
Business	0.67	0.15	0.13	0.05	12147
Chemistry	0.86	0.12	0.01	0.01	24490
Computer Science	0.88	0.04	0.07	0.02	19933
Economics	0.76	0.16	0.07	0.01	7680
Engineering	0.59	0.24	0.14	0.04	34597
Environmental Science	0.79	0.03	0.17	0.01	5091
Geography	0.66	0.29	0.04	0.01	4097
Geology	0.68	0.26	0.05	0.01	5260
History	0.88	0.07	0.05	0.01	5574
Materials Science	0.46	0.26	0.22	0.05	9771
Mathematics	0.63	0.32	0.04	0.01	12239
Philosophy	0.85	0.06	0.08	0.01	2721
Physics	0.62	0.04	0.31	0.03	7459
Political Science	0.85	0.08	0.06	0.01	7107
Psychology	0.89	0.06	0.04	0.01	33327
Sociology	0.71	0.05	0.23	0.01	4742
Total	0.74	0.16	0.08	0.02	247655

Note: The table summarises the links found from from ProQuest graduates to MAG authors. Graduates are defined as the authors of the dissertations in ProQuest. First, the columns headed by "Fraction of links found" compare the identified links across two different labellers as described in the text. The columns show the fraction of links found for two training sets constructed by two different labellers. "Same entity" are graduates for which the models trained on the different training sets find the same MAG identifier. "Only by 1" and "Only by 2" are graduates for which only the model trained on either of the training sets found a link to MAG at all. "Different entity" are graduates for which both models find links to MAG, but to different identifiers. Second, the last column reports the total number of links found for each field, after all postprocessing (see text for details). Third, the last row reports the total across fields. The fractions are weighted by the number of graduates in the respective fields in ProQuest 1990–2015.

Table 1: Linking the graduates

	Fraction of links found				
Field	Same entity	Only by 1	Only by 2	Different entity	Number of links
Biology	0.78	0.00	0.22	0.00	101852
Business	0.69	0.01	0.30	0.00	30963
Chemistry	0.93	0.00	0.02	0.05	48670
Computer Science	0.89	0.08	0.02	0.01	39618
Economics	0.96	0.00	0.02	0.02	27266
Engineering	0.81	0.00	0.19	0.00	99106
Environmental Science	0.73	0.23	0.03	0.01	11055
Geography	0.62	0.32	0.02	0.03	12229
Geology	0.84	0.00	0.14	0.01	11478
History	0.85	0.01	0.12	0.02	30228
Materials Science	0.74	0.01	0.13	0.12	23998
Mathematics	0.78	0.15	0.01	0.06	32938
Philosophy	0.93	0.00	0.06	0.00	12153
Physics	0.51	0.15	0.03	0.31	19825
Political Science	0.87	0.00	0.08	0.04	24309
Psychology	0.91	0.00	0.08	0.00	94144
Sociology	0.46	0.01	0.51	0.02	20914
Total	0.81	0.03	0.13	0.03	640746

Note: The table summarises the links found from from ProQuest advisors to MAG authors. An advisor is one relationship id as described in the text. First, the columns headed by “Fraction of links found” compare the identified links across two different labellers as described in the text. The columns show the fraction of links found for two training sets constructed by two different labellers. “Same entity” are advisors for which the models trained on the different training sets find the same MAG identifier. “Only by 1” and “Only by 2” are advisors for which only the model trained on either of the training sets found a link to MAG at all. “Different entity” are advisors for which both models find links to MAG, but to different identifiers. Second, the last column reports the total number of links found for each field, after all postprocessing (see text for details). Third, the last row reports the total across fields. The fractions are weighted by the number of graduates in the respective fields in ProQuest 1990–2015.

Table 2: Linking the advisors

3. The two models assign a different MAG entity to the same ProQuest entity. Here, we only keep cases where two conditions hold: First, the names of the two entities in MAG are almost identical (a Jaro-Winkler similarity of 0.99 or more). Second, the entity of one of the predicted links publishes at more than five times as many papers over the whole career as the entity of the other predicted link. This case covers duplicated entities in MAG, where a few papers are not merged to the main entity of an author.

Second, we deal with duplicates that arise from linking different fields separately. For graduates, we require that both the MAG and the ProQuest entity are uniquely linked; in other words, if a chemistry graduate in ProQuest is linked to different MAG entities in the models for chemistry and biology, we drop them. For advisors, we require that the ProQuest entity is uniquely linked to MAG, but not vice versa—since advisors can supervise multiple theses.

The last row in tables 1 and 2 report the number of links that result at the end of this postprocessing step. In total, we find around 250'000 one-to-one links for graduates and 640'000 many-to-one links (from many dissertations to one MAG entity) for advisors.

1.5 Empirical performance of the linking algorithm

We discuss the performance of the record linkage in terms of precision and recall given the "truth" and human error.

1.5.1 Precision and recall

In prediction tasks where the ground truth is representative of the population, one can check the performance of the prediction by comparing the predicted labels to the true labels. Because of active learning, the labelled records are not randomly selected and therefore not suitable to calculate recall and precision. Nevertheless, we suggest a lower bound for the precision in the linking of graduates and an approximation to recall for advisors.

First, we provide a check on the precision of our links for graduates in chemistry where many graduates publish during their dissertation (Gaulé and Piacentini, 2013). Because we do not use the name of the PhD university in the predictive model, we can use this feature to validate our links. First, we find 18 percent of our linked PhD graduates do not publish a paper during their PhD (in the six years before graduating). Second, among graduates that do publish during their PhD, 96 percent do so with an affiliation in MAG that corresponds to their PhD university. Assuming that among our links, only the ones that publish during the PhD at their alma mater true positives, we can now suggest a lower bound on the precision of our procedure as $0.96 \times (1 - 0.18) = 0.78$—a performance in the upper tercile of the methods discussed in Bailey et al. (2020, Table 1) for historical US census data.

Second, because advisors tend to be established researchers and publish regularly we should find a high fraction of advisors in the MAG data. In most fields and years, we find about 75 percent or more advisors in the MAG data.[5] This not only indicates that our linking strategy has a high recall for advisors, but it also provides reassuring evidence that the affiliation information in MAG is accurate, since it is a requirement for identifying the links.

[5] The exceptions are philosophy, where we find between 50 and 75 percent of advisors in each year, and business where the fraction of advisors linked declines from 0.75 to 0.6 over the sample period.

1.5.2 Human error

Since the ground truth is defined by active labelling by humans, it is possible that errors in labelling propagate to a biased linked sample. By aggregating out errors from individual labellers, the postprocessing of links reduces such concerns. Moreover, the detailed data in tables 1 and 2 show that the models, even without such aggregation, make very similar predictions.[6]

First, in 74 percent of linked graduates and in 81 percent of linked advisors the models agree. Across fields, it varies more, but in most cases the agreement is at least 60 percent. Second, the predicted entity differs only 2 to 3 percent of the ProQuest entities. Across fields, this number is always below five person for graduates. It is similar for advisors with the exception of Materials Science and Physics. Third, the columns “Only 1” and “Only 2” also show that one of the two models is more conservative than the other, but again the postprocessing only keeps such links if the names of the two linked entities are very similar.

1.6 Constructing the analysis sample

We look at graduates between 1990 and 2014 whom we can link to MAG and for whom we can link at least one advisor to MAG.[7] We build the following variables.

First university after graduation This is a university that is not the PhD university where the graduate publishes their first paper after graduating. Following Kramarz and Skans (2014) we allow for a gap of up to seven years between the graduation and the first publication.

Research output after graduation We summarise the publications and 10-year forward citations of all papers the graduate publishes in the first 7 years after graduating.

Connections to universities through co-authors For each graduate–university pair, we determine whether they are connected at the time of graduation through their own, or their advisors co-author network.. There are two types of connections. The first is a direct co-author connection that arises when the graduate publishes a paper with a co-author from the university before their graduation. The second is a co-author connection through the advisor before the graduation of the PhD student. In both cases we require that the co-author publishes a paper at the connected university in the last five years before the graduates’ dissertation.

1.7 Field composition of sample

In Figure 2 we show the field composition of the sample. The figure shows the distribution of graduates in all major fields of study, except medicine and art. The fields of study are based on the highest level of the Microsoft Academic Graph (MAG) classification.

[6] These similarities do not stem from overlap in the training sets: We verified in one field (graduates chemistry) that the training sets of the two labellers do not overlap (for one labeller, two out of 50 pairs are also in the training set of the other labeller).

[7] This means we implicitly condition also on the set of universities that we can crosswalk to the Carnegie list.

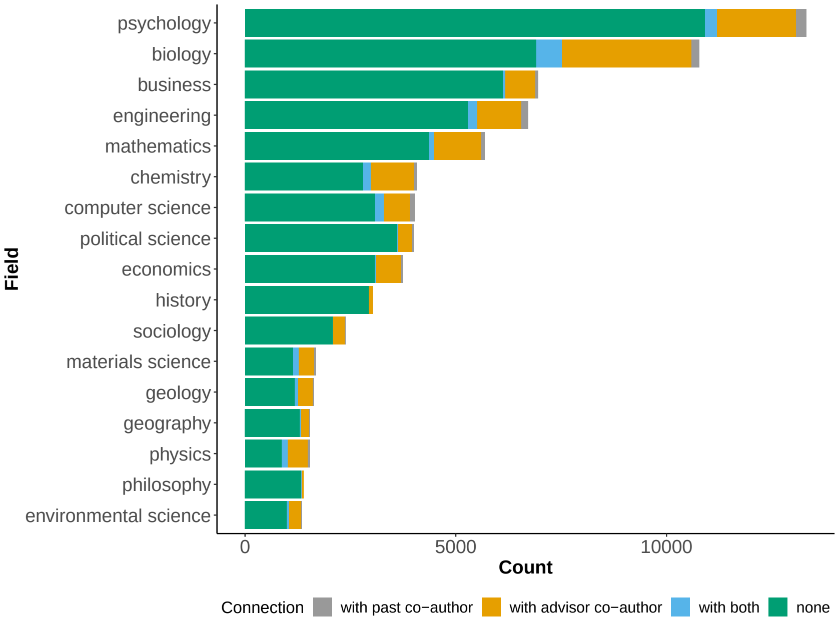

Figure 2: Field composition of PhD graduates

1.8 Observed signals of pre-hiring productivity

We construct position of the graduate's pre-graduation and the advisor's publications in their respective citation distributions as follows.

For graduates, we compute their expected citations of the pre-graduation publications. We predict the number of citations with the average citations of papers published in the same year and venue (conference or journal).[8] Then, we compute the student's position in the distribution, relative to other students in the same research field and within a five-year interval.

For advisors, we measure their position in the citation distribution in the last 10 years, within the respective research field, and a five-year interval before graduation.

2 Additional Results

2.1 Post-PhD Outcomes

In the main text we presented results for a set of post PhD outcomes. Here we present the results without the controls for pre-graduation outcomes of the PhD student and their advisor. These results are provided to show the role of the controls in explaining the differences between connected and non-connected hires. Pre-graduation productivity predictors can explain a substantial part of differences in outcomes between connected and not-connected hires. This is true for the number of papers, and co-author outcomes. However, it is not true for the outcomes

[8] We consider this measure more precise than the acutal citations received by the time of graduation.

Same Affililation PhD + 6 years and *Any Output PhD + 6 years*.

Panel A: Comparison of post-PhD outcomes with Class Fixed Effect

Dependent Variables:	N Cites PhD graduate	N papers	Co-authors	Same Affil	Any output	N Cites of
			First Affil	PhD+6yrs	PhD+6yrs	First Affil
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Variables						
Advisor connection	0.262	0.113	0.228	-0.142	0.001	0.597
	(0.023)	(0.011)	(0.022)	(0.013)	(0.004)	(0.100)
PhD's connection	0.215	0.198	-0.022	-0.597	-0.005	0.011
	(0.029)	(0.013)	(0.022)	(0.029)	(0.006)	(0.025)
Fixed-effects						
PhD Class	Yes	Yes	Yes	Yes	Yes	Yes
Fit statistics						
Pseudo R^2	0.44	0.24	0.34	0.04	0.008	0.76
Observations	73,775	73,885	69,566	71,704	73,427	73,705

Panel B: Comparison of post-PhD outcomes with Destination Fixed Effect

Dependent Variables:	N Cites PhD graduate	N papers	Co-authors	Same Affil	Any output
			First Affil	PhD+6yrs	PhD+6yrs
Model:	(1)	(2)	(3)	(4)	(5)
Variables					
Advisor connection	0.129	0.051	0.049	-0.056	-0.002
	(0.020)	(0.010)	(0.016)	(0.012)	(0.004)
PhD's connection	0.191	0.178	-0.024	-0.564	-0.005
	(0.026)	(0.012)	(0.018)	(0.030)	(0.006)
Fixed-effects					
Field×5 Year Window	Yes	Yes	Yes	Yes	Yes
Hiring University Id×Field	Yes	Yes	Yes	Yes	Yes
Subfield (MAG lvl 1)	Yes	Yes	Yes	Yes	Yes
Fit statistics					
Pseudo R^2	0.41	0.24	0.34	0.04	0.007
Observations	73,672	73,885	66,819	72,008	73,273

Panel C: Comparison of post-PhD outcomes with Class and Destination Fixed Effect

Dependent Variables:	N Cites PhD graduate	N papers	Co-authors	Same Affil	Any output
			First Affil	PhD+6yrs	PhD+6yrs
Model:	(1)	(2)	(3)	(4)	(5)
Variables					
Advisor connection	0.091	0.048	0.053	-0.058	-0.003
	(0.021)	(0.011)	(0.018)	(0.012)	(0.004)
PhD's connection	0.174	0.169	-0.040	-0.575	-0.003
	(0.028)	(0.013)	(0.022)	(0.030)	(0.006)
Fixed-effects					
PhD Class	Yes	Yes	Yes	Yes	Yes
Hiring University Id×Field	Yes	Yes	Yes	Yes	Yes
Subfield (MAG lvl 1)	Yes	Yes	Yes	Yes	Yes
Fit statistics					
Pseudo R^2	0.54	0.32	0.41	0.07	0.014
Observations	73,596	73,885	64,634	70,109	72,893

Notes: Unit of observation is a PhD graduate. See main text for the poisson regression specification. Observations with zero outcomes that are perfectly predicted by fixed effects are dropped. Clustered (PhD university-field-5 year window + Hiring university) standard-errors in parentheses. N Cites PhD graduate measures citations received on articles published in the first 6 years post PhD graduation. N papers is the number of articles published in the same period. Co-authors First affil measures the number of new co-authors at the first post-PhD affiliation. Same Affil PhD+6yrs and Any Output PhD+6yrs indicate whether the PhD graduate, 6 years after the PhD graduation or later, is still affiliated with their first post-PhD affiliation and whether they publish any papers at that point.

Table 3: Post-PhD outcomes of connected vs. not-connected hires

2.1.1 The role of the advisor

In the paper, we compare graduates within a PhD class or within the same hiring university. To further sharpen the results we now establish whether the positive selection of connected hires is driven by differences between or within advisors. This also addresses the concern that estimated network effects are driven by more connected advisors having more productive students.

Figure 3 presents the results from variations of the post-PhD regressions. The dots are point estimates; the error bars 95% confidence intervals. In grey, we reproduce the results from Table 3 with controls for PhD graduate productivity during the PhD and advisor cites.

We first compare these baseline estimates to results—shown in yellow—from a model that replaces the advisor cites with advisor fixed effects. The point estimates are very similar between the models. While the standard errors increase, the results remain significant at the five percent confidence level. Thus for the outcomes we consider, the advisor fixed effects account for similar variation in the outcome as the advisor cites—and that our results are not spuriously driven by more connected advisors having more productive students. Instead, the results show that the same advisor—and associated network—place graduates that are ex-post more productive to universities with high output.

Now we compare these latter results to results where we include again fixed effects for the advisor, but we drop the controls for pre-graduation productivity of the PhD graduate. These results are shown in blue; they are very similar to the results from the previous specification. Thus, after accounting for the identity of the advisor, there is limited additional information in the pre-graduation productivity of the PhD graduate in explaining the connected hire gap in post-PhD outcomes.

2.2 Robustness checks

2.2.1 Placebo analysis

Our main results highlight the role of the PhD advisor's collaboration network for the placement of PhD students. In this section we present a placebo analysis by repeating the analysis with a "placebo advisor". To do so, we draw another advisor from the same PhD class and repeat the estimation of equation (1) in the main text with the placebo advisor's connections. If the personal connection of the advisor with the student is important, then the estimated effect should be substantially lower for the placebo advisor. Table 4 shows the results. Placebo connections predict a small decrease in the probability to match, conditional on controls in the main specification in column (6). The values are small in absolute value compared to the main effect. Further, contrasting the coefficient on the placebo connection in colum (1) with column (2-6) highlights that the controls we use capture common matching determinants at the

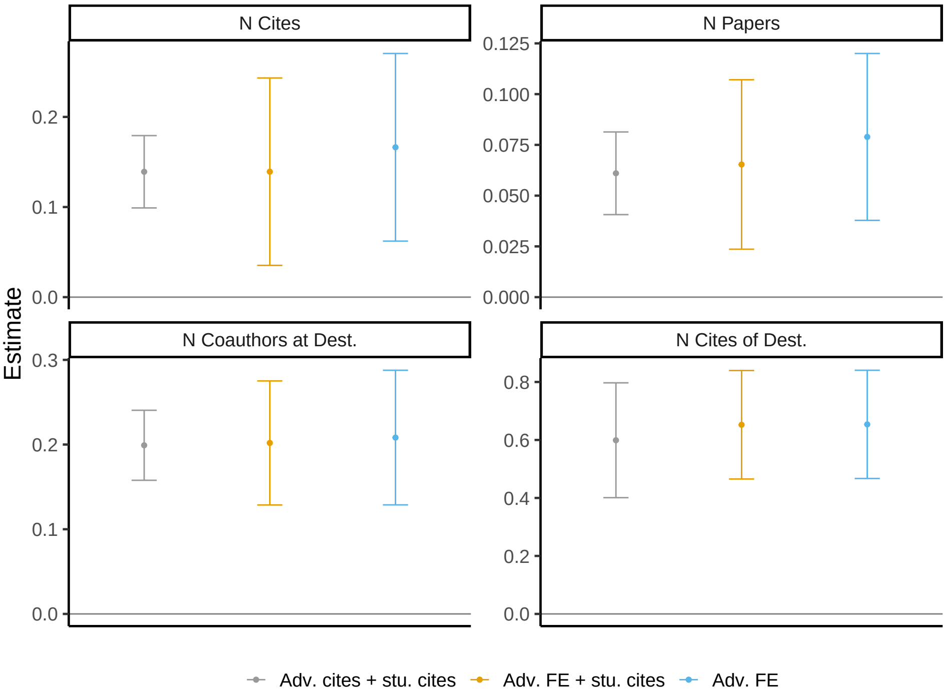

Figure 3: Advisor effect and Post-PhD outcomes

Estimates and associated 95 % confidence intervals of γ for the advisor's network. Different colors refer to different specifications of equation (2) in the paper: *Adv. cites + stu. cites* includes controls for the advisor's and the student's pre-PhD citations; *Adv. FE + stu. cites* includes advisor fixed effects and student pre-PhD citations; and *Adv. FE* includes advisor fixed effects only. Standard errors are clustered at the level of (First Affiliation-Field-5 Year Window).

class-destination level, that otherwise would confound the effect of the advisor's network.

Dependent Variable:	Match formed					
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Variables						
Constant	0.414					
	(0.017)					
Advisor connection	1.13	0.595	0.594	0.599	0.645	0.599
	(0.096)	(0.037)	(0.036)	(0.035)	(0.038)	(0.035)
"Placebo Advisor" connection	0.529	-0.049	-0.049	-0.049	-0.055	-0.054
	(0.072)	(0.014)	(0.014)	(0.014)	(0.015)	(0.015)
PhD's connection	2.85	3.31	3.30	3.30	3.37	3.31
	(0.172)	(0.152)	(0.149)	(0.149)	(0.155)	(0.153)
Fixed-effects						
PhD Class×Potential Hiring University ID		Yes	Yes	Yes	Yes	Yes
Pre-Graduation Productivity×Field×5 Year Window			Yes	Yes		
Advisor Citation Decile×Field×5 Year Window				Yes		
Student Id					Yes	Yes
Varying Slopes						
Max similarity to faculty members ×Field						Yes
Avg. similarity to faculty members ×Field						Yes
Fit statistics						
Observations	5,396,046	5,396,046	5,396,046	5,396,046	5,396,046	5,396,046

Notes: Placebo connections are constructed by randomly assigning a PhD graduate another "placebo" advisor from the same PhD class. The placebo advisor's connections are then used as an additional type of network in equation (1) in the main text.

Table 4: Matching - Placebo connection

2.2.2 Connections - Decay with time since last collaboration

In our main results we considered only connections of the advisor in the last 5 years up to the graduation year of the PhD student. Here we compare results as a function of the most recent year of collaboration between the advisor and the co-author at another institution. We expect connections further distant in the past to be weaker predictors of placement. Table 5 shows the results with the most recent collaboration year binned into 3 year windows. The effect is strongest for connections in the last 3 years, cut into almost half for connections 3 to 6 years in the past and down to about one third for connections 6 to 9 years in the past. We see the year of most recent collaboration as a proxy for the strength of the connection between the advisor and the co-author at another institution, as it relates to how recent contact was. In line with that interpretation, the estimated effect of a network connection decays with time since last collaboration.

Dependent Variable:	Match formed					
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Variables						
Constant	0.434					
	(0.017)					
Advisor connection ∈ [0,3) years ago	1.29	0.669	0.668	0.677	0.735	0.683
	(0.118)	(0.044)	(0.043)	(0.042)	(0.045)	(0.042)
Advisor connection ∈ [3,6) years ago	0.718	0.361	0.360	0.366	0.414	0.378
	(0.061)	(0.035)	(0.035)	(0.036)	(0.037)	(0.036)
Advisor connection ∈ [6,9) years ago	0.707	0.238	0.237	0.244	0.291	0.261
	(0.086)	(0.048)	(0.048)	(0.048)	(0.049)	(0.048)
Advisor connection ∈ [9,100) years ago	0.534	0.104	0.103	0.109	0.145	0.120
	(0.057)	(0.021)	(0.021)	(0.021)	(0.022)	(0.021)
PhD's connection	2.83	3.29	3.28	3.28	3.35	3.29
	(0.171)	(0.150)	(0.147)	(0.147)	(0.154)	(0.151)
Fixed-effects						
PhD Class×Potential Hiring University ID		Yes	Yes	Yes	Yes	Yes
Pre-Graduation Productivity×Field×5 Year Window			Yes	Yes		
Advisor Citation Decile×Field×5 Year Window				Yes		
Student Id					Yes	Yes
Additional controls with varying slopes						
Max similarity to faculty members×Field						Yes
Avg. similarity to faculty members×Field						Yes
Observations	5,396,046	5,396,046	5,396,046	5,396,046	5,396,046	5,396,046

Notes: The table shows the results of reestimating equation (1) in the main text while allowing for heterogeneity in the effect of the advisor's network by the most recent year of collaboration between the advisor and the co-author at another institution. The most recent year of collaboration is binned into 3 year windows.

Table 5: Matching - Heterogenous Effects by most recent year of collaboration

2.2.3 Time trends in post-PhD outcomes of connected and non-connected graduates

In table 6, we report the results from estimating equation (4) from the main text: We test whether the differences between connected and non-connected graduates has changed over time.

We report the results that include fixed effects for both the destination and the class. For all the outcomes, we find no significant changes in the gap between graduates placed through the network and graduates not placed through the network.

Dependent Variables:	N Cites PhD graduate	N papers	Co-authors First Affil	Same Affil PhD+6yrs	Any output PhD+6yrs
Model:	(1)	(2)	(3)	(4)	(5)
Variables					
Advisor connection $\times(t-1990)$	0.002	0.0006	0.0007	0.001	0.0008
	(0.003)	(0.001)	(0.002)	(0.002)	(0.0006)
PhD’s connection $\times(t-1990)$	0.009	0.001	0.002	0.005	0.003
	(0.005)	(0.002)	(0.004)	(0.004)	(0.0009)
Fixed-effects					
PhD Class	Yes	Yes	Yes	Yes	Yes
Pre Graduation Productivity×Field	Yes	Yes	Yes	Yes	Yes
Advisor Citation Decile×Field	Yes	Yes	Yes	Yes	Yes
Hiring University Id×Field	Yes	Yes	Yes	Yes	Yes
Subfield (MAG lvl 1)	Yes	Yes	Yes	Yes	Yes
Degree Year ×Field	Yes	Yes	Yes	Yes	Yes
Advisor connection×Field	Yes	Yes	Yes	Yes	Yes
PhD’s connection×Field	Yes	Yes	Yes	Yes	Yes
Fit statistics					
Observations	73,596	73,885	64,625	70,109	72,893
Pseudo R^2	0.62	0.37	0.43	0.07	0.015

Table 6: Post-PhD Outcomes: Network effect over time

References

American Council on Education. 2021. "Carnegie Classification of Institutions of Higher Education. `https://carnegieclassifications.acenet.edu/wp-content/uploads/2023/02/CCIHE2021-PublicData.xlsx` (accessed June 2022)."

Bailey, Martha J, Connor Cole, Morgan Henderson, and Catherine Massey. 2020. "How well do automated linking methods perform? Lessons from US historical data." *Journal of Economic Literature*, 58(4): 997–1044.

Gaulé, Patrick, and Mario Piacentini. 2013. "Chinese graduate students and US scientific productivity." *Review of Economics and Statistics*, 95(2): 698–701.

Gregg, Forest, and Derek Eder. 2022. "Dedupe, version 2.0.11. URL: https://github.com/dedupeio/dedupe."

Harzing, Anne-Wil, and Satu Alakangas. 2017. "Microsoft Academic: Is the phoenix getting wings?" *Scientometrics*, 110(1): 371–383.

Huang, Junming, Alexander J Gates, Roberta Sinatra, and Albert-László Barabási. 2020. "Historical comparison of gender inequality in scientific careers across countries and disciplines." *Proceedings of the National Academy of Sciences*, 117(9): 4609–4616.

Hug, Sven E., and Martin P. Brändle. 2017. "The coverage of Microsoft Academic: Analyzing the publication output of a university." *Scientometrics*, 113(3): 1551–1571.

Hug, Sven E., Michael Ochsner, and Martin P. Brändle. 2017. "Citation analysis with Microsoft Academic." *Scientometrics*, 111(1): 371–378.

Kramarz, Francis, and Oskar Nordström Skans. 2014. "When strong ties are strong: Networks and youth labour market entry." *Review of Economic Studies*, 81(3): 1164–1200.

Martín-Martín, Alberto, Mike Thelwall, Enrique Orduna-Malea, and Emilio Delgado López-Cózar. 2020. "Google Scholar, Microsoft Academic, Scopus, Dimensions, Web of Science, and OpenCitations' COCI: A multidisciplinary comparison of coverage via citations." *arXiv preprint arXiv:2004.14329.*

Missouri Census Data Center. 2020. "ZIP Code Lookup, complete list in excel file. `https://mcdc.missouri.edu/applications/zipcodes/ZIP_codes_2020.xls` (accessed June 2022)."

Organisation for Economic Co-operation and Development. 2007. "Revised Field of Science and Technology (FOS) Classification in the Frascati Manual. `https://www.oecd.org/science/inno/38235147.pdf` (accessed June 2022)."

Sinha, Arnab, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-June (Paul) Hsu, and Kuansan Wang. 2015. "An Overview of Microsoft Academic Service (MAS) and Applications. `https://zenodo.org/record/2628216` (accessed 2021)." 243–246.

Visser, Martijn, Nees Jan van Eck, and Ludo Waltman. 2021. “Large-scale comparison of bibliographic data sources: Scopus, Web of Science, Dimensions, Crossref, and Microsoft Academic.” *Quantitative Science Studies*, 2(1): 20–41.