Online appendix to: Networks in the market for researchers

Flavio Hafner, Christoph Hedtrich

November 26, 2024

1 Record linkage

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

1.1 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.

1.2 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,[1] 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.

[1] We describe in section 1.5 the different distance functions for different data types.

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.

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 random sample of pairs that are labeled by humans. We discuss these concerns in detail below and address them empirically as best as possible.

1.3 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.

1.4 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.

1.5 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.

1.6 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 $\times$ year and lastname $\times$ 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.

1.7 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.

1.8 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.

1.9 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.[2] 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.

2 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

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

Field	Fraction of links found				Number of links
	Same entity	Only by 1	Only by 2	Different entity	
Art	0.48	0.52	0.00	0.00	1717
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
Medicine	0.44	0.22	0.31	0.03	8645
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	0.69	0.19	0.10	0.02	258017

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

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 section 3.

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.

	Fraction of links found				
Field	Same entity	Only by 1	Only by 2	Different entity	Number of links
Art	0.88	0.00	0.11	0.01	10628
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
Medicine	0.94	0.00	0.02	0.04	19264
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	0.83	0.02	0.12	0.03	670638

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 258'000 one-to-one links for graduates and 670'000 many-to-one links (from many dissertations to one MAG entity) for advisors.

3 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.

3.1 Precision and recall

In prediction tasks where a representative sample of correctly labeled pairs exists, 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. Precision is defined as the number of true links divided by the total number of links made. For chemistry, we provide a test based on correctly linked graduates: First, because most graduates already publish during their dissertation (Gaulé and Piacentini, 2013), we can calculate the fraction of our linked chemistry graduates that satisfy this condition, conservatively discarding any other links as false positives. Second, because we do not link on the name of the PhD university, we can compare the name of the alma mater to the affiliation name of the linked author entity in MAG before graduation. We find that 18% of linked graduates do not publish during their dissertation, and 4% of those that do, do not publish at their alma mater. Therefore, at least a fraction of $(1 - 0.18) \times (1 - 0.04) = 0.78$ linked records was plausibly linked correctly. This performance is 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.[3] 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 co-author connections across universities.

3.2 Human error

As the algorithm is trained through 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.[4]

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.

References

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."

[3]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.

[4]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).