Scientometrics (2018) 117:351–380
https://doi.org/10.1007/s11192-018-2840-5

The next generation (plus one): an analysis of doctoral students' academic fecundity based on a novel approach to advisor identification

Dominik P. Heinisch[1] · Guido Buenstorf[1,2,3]

Received: 16 February 2018 / Published online: 13 July 2018

Abstract
Scientific communities reproduce themselves by allowing senior scientists to educate young researchers, in particular through the training of doctoral students. This process of reproduction is imperfectly understood, in part because there are few large-scale datasets linking doctoral students to their advisors. We present a novel approach employing machine learning techniques to identify advisors among (frequent) co-authors in doctoral students' publications. This approach enabled us to construct an original dataset encompassing more than 20,000 doctoral student-advisor pairs in applied physics and electrical engineering from German universities, 1975–2005. We employ this dataset to analyze the "fecundity" of doctoral students, i.e. their probability to become advisors themselves.

Keywords Advisor identification · Fecundity · Ph.D. training · Advisor affects · Academic careers · Machine learning

JEL Classification PI23 · O30 · D83 · D85

Introduction: the production of scientists by means of scientists

Publications are the main output that scientists produce. In publications new results are described and disseminated to other researchers. Publications thus enable the cumulative nature of science, where each generation of scientists stands, in the metaphorical expression conventionally attributed to Isaac Newton, "on the shoulder of giants". However, it is implicit in Newton's metaphor that publications alone are not sufficient to keep science functioning. A second output is required: new generations of scientists willing and able to climb atop their predecessors' shoulders.

✉ Dominik P. Heinisch
heinisch@uni-kassel.de

[1] Institute of Economics and INCHER Kassel, University of Kassel, Kassel, Germany

[2] Institute of Innovation and Entrepreneurship, University of Gothenburg, Gothenburg, Sweden

[3] IWH Leibniz Institute of Economics Halle, Halle, Germany

Science is a self-reproducing system, and the training of doctoral students is a key step in the process of reproduction. New members of scientific communities are spawned within these communities. Specifically, they are trained by existing members of the community. This is most clearly observed in doctoral training: only established researchers—normally those who hold professorships—have the privilege to graduate doctoral students, i.e., admit new members to the community. Grounding on this incremental process, Sugimoto (2014) aptly describes studying academic genealogies as studying "intellectual heritage".

The relevance of producing doctoral students is recognized in the conventions by which reputation is allocated among researchers (e.g., in the notion that person X is "advisor Y's student", and also in disciplinary genealogies such as the Mathematics Genealogy Project or Neurotree).[1] The training of successful students is recognized as "the most lasting contribution" to science that any scientist can make. This is especially true for scientists who train those members of the next generation that subsequently train yet another generation (Marsh 2017). Andraos (2005) finds clear words when emphasizing the importance of training "fecund" students: "In fact, the most effective way for a scientist's work to live beyond their time is for them to populate the next generation of academics with people that they have mentored, otherwise their names and work will inevitably be driven to extinction". The influence of scientists in academic genealogies has been conceptualized into metrics similar to citation based measures (Rossi et al. 2017), and it is also a relevant parameter in many performance-based management systems set up by universities and national governments.

However, only a fraction of all doctoral students will eventually become advisors and produce doctoral students themselves, and it is not obvious, for example, whether successful self-reproduction is primarily due to having many offspring or more dependent on offspring quality (cf., e.g., Malmgren et al. 2010). To learn more about the self-reproduction of science, we therefore need to better understand the determinants of producing "fecund" students, i.e., those of today's doctoral students that tomorrow will train the next generation of students, and also how they relate to student, advisor and university characteristics.

In recent years, the interest in studying academic genealogies has increased, but evidence based on large scale genealogies is still scarce, in part because of limited data availability.[2] Advisor information is not part of the standard bibliometric data for doctoral dissertations, nor is it otherwise collected and published in systematic form. For some fields and disciplines, information about advisors and their students can be derived from publicly available academic genealogies, but these rely on information volunteered by members of the community. There is a lack of approaches allowing researchers to retrieve advisor information from existing datasets and thus extend and generalize our knowledge about the reproduction of science beyond the scope of available academic genealogies (Malmgren et al. 2010).

To help alleviate this problem we developed a novel approach to advisor identification based on co-publications and machine learning of matching algorithms. We consider the detailed description of this approach the main contribution of the present paper. Our approach disambiguates and matches authors of doctoral dissertations and publications,

[1] See: Jackson (2007) for a description of the Mathematics Genealogy project, David and Hayden (2012) for the Neurotree project, or for a comparable project in Germany: https://www.bibsonomy.org/persons (last access date: February 15, 2018).

[2] A recent overview of existing approaches is provided by Dores et al. (2016).

and identifies advisors in the set of co-authors with whom doctoral students published papers.

Using this novel approach, we constructed an original dataset containing more than 20,000 matched student-advisor pairs for German applied physics and electrical engineering in the period from 1975 to 2005. In the second part of the paper we employ this dataset to analyze factors influencing the probability that doctoral students are fecund, i.e. that they become doctoral advisors themselves. This issue has found little attention in the prior literature. The study most closely related to our analysis is the one of Malmgren et al. (2010) who investigate the fecundity of doctoral students in mathematics.

The remainder of this paper is structured as follows. In Section "Identifying doctoral advisors based on co-authored publications" we provide a detailed account of our approach to identify doctoral advisors based on dissertation and co-publication data. Section "Fecundity of doctoral students: theoretical considerations" then develops the theoretical background of the subsequent analysis of doctoral students' fecundity. It also formulates testable hypotheses. Section "Econometric analysis" presents the results of the empirical analysis, before we offer concluding remarks.

Identifying doctoral advisors based on co-authored publications

As noted in the Introduction, tracing advisor effects on doctoral students is complicated by the lack of advisor information in publicly available databases. In this section we outline a new large-scale approach to deal with this problem. The proposed approach, which underlies the empirical analysis presented below, exploits information about co-authored journal publications. We first present the general approach before discussing its implementation for the present paper. To keep the discussion tractable, only the most salient features of the approach are presented in the main body of the article, whereas technical details are provided in footnotes as well as in the Appendix "Assessment of data quality". A schematic representation of our approach is provided in Fig. 1.

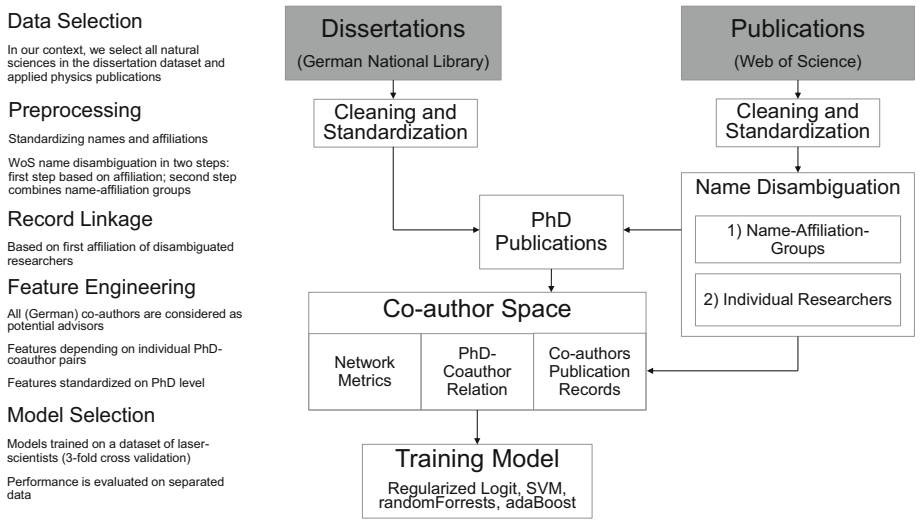

Fig. 1 Overview of the data processing and machine learning procedure

We start from the assumption that if doctoral students publish scientific articles based on their dissertation work they are likely to do so together with their advisor. By tracking co-authors of papers related to a given dissertation we should then be able to pick out the advisor with a high degree of accuracy. To this purpose we first need to identify doctoral students' papers as a subset of all scientific publications. We are able to solve this task by using additional information on completed doctoral dissertations that, as best we know, has not been used before in large-scale analyses. Based on knowing who is a doctoral student we can then identify their publications. Tracking the publication output of doctoral students we can retrieve a full list of all co-authors, which we expect to include the advisor. For this group of co-authors we can then calculate several characteristics that help reveal the advisor. By using a training dataset we are able to train a model that predicts the advisor with a known degree of accuracy.

Before providing details of our advisor identification approach, several characteristics of doctoral training in Germany, the empirical context of our data collection and empirical analysis, have to be noted. Among these is that traditionally the training of individual doctoral students has been concentrated in the hands of a single advisor (the "Doktorvater") who assumes the primary responsibility for the student. Doctoral advisors normally perform both mentoring (Sugimoto 2012) and examiner roles. Even though dissertations are graded by several examiners and defended before an examination committee, the other examiners and committee members do not normally engage in extensive mentoring. There are two constellations, however, in which not the formal advisor but another individual is the primary academic mentor of a doctoral student (and could in this sense be considered their "true" advisor). First, this may happen due to eligibility constraints. Only research universities have traditionally been allowed to examine doctoral students and grant doctoral degrees. Non-university public research organizations (such as the Max Planck Society or the Leibniz Association) and, until very recently, Universities of Applied Sciences are or were excluded from this right. It is quite common, though, that doctoral students work at public research organizations, and also in industrial R&D labs, and that their mentor is also employed there. In such cases, a university professor often becomes the student's official advisor and acts as the lead examiner.[3] Second, in the chair system that is still widespread in German universities, professors may delegate much of the day-to-day mentoring of their doctoral students to junior researchers employed at the professor's chair, while at the same time retaining their formal advisor and examiner status.

Dissertation data

Our point of departure is information about the universe of doctoral dissertations completed in Germany. Scholarly interest in dissertation data has been limited, especially when used as individual data and linked to other bibliographic information. However, there is a lot to learn from these data to understand the production of scientific knowledge (Morichika and Shibayama 2016). We take our information about doctoral dissertations from the online catalog of the German National Library (*Deutsche Nationalbibliothek* or DNB for short). Since 1969 German universities have been legally required to deposit copies of all completed doctoral dissertations at DNB. Universities, in turn, enforce mandatory deposit of doctoral dissertations vis-à-vis their doctoral students. To ensure compliance, the

[3] In practice these restrictions are somewhat less relevant because public or industrial researchers mentoring doctoral students may also hold positions at universities, e.g., honorary professorships, that allow them to examine doctoral students.

Table 1 Illustration of DNB data structure

	DNB_ID	Name	Surname	Dissertation title	Year	Affiliation
Raw	101	Wilhelm S.	Höhe	Studies based on H2O-LASER...	1996	Univserität Würzburg
	202	Karl	Saué	Physical experiments using a new laser...	91	Kassel Univ.
Clean	101	Wilhelm S	Hohe	studies based on h2o laser...	1996	Uni Wurzburg
	202	Karl	Saue	physical experiments using a new laser...	1991	Uni Kassel

The table describes the data originating from the German National Library (DNB) in a fictitious example before (raw) and after (clean) applying the standardization procedure

doctoral degree is granted only after the respective student has deposited the required number of copies. Due to this principle of mandatory deposal DNB holdings of doctoral dissertations, and consequently the information contained in its catalog, are virtually complete. Identifying doctoral dissertations in the catalog is unambiguous as they are indicated within the catalog's subsection of works from higher education institutions (*Hochschulschriften*), along with (non-standardized) information about the issuing university and the year of submission. After eliminating double entries (e.g., due to multiple versions—mostly print and online versions—or editions), a total of 894,086 unique authors of doctoral theses was identified for the time period from 1970 to 2010. An illustration of DNB data can be found in Table 1.

Publication data

Identifying doctoral advisors on the basis of co-authorships first requires us to identify publications authored by the doctoral students in our sample. We use the *Web of Science* (WoS) database to obtain these publications. An illustration of the WoS data structure is prestented in Table 2. Identifying doctoral students' publications in the WoS requires solving the namesake problem in the first place. The namesake problem (also known as ambiguity or homonym problem) arises when two or more publishing scholars share the same author name (cf., e.g., Smalheiser and Torvik (2009), for a detailed discussion of the namesake problem). Since WoS does not provide a unique identifier for each individual researcher, author identities have to be established by cleaning and standardizing the WoS data. Solving this problem is nontrivial for several reasons[4] and becomes especially challenging when very common names are considered (*ibid.*). Our disambiguation procedure uses a heuristic approach, and assessing its quality is complicated because we try to track publications records of doctoral graduates who frequently leave academia after just publishing a few papers. CVs including publications records are often not available for these individuals, and using the available ones would likely be biased. Accordingly, we do not have a sample to systematically evaluate the quality of disambiguation. However, the

[4] This task is best accomplished by using reliable automated procedures for two reasons: (a) the number of publications is too large for manual processing and (b) manual processed data seems to provide unreliable results, especially in the case of common names. Inaccuracies occur even if authors themselves are asked to select their own papers.

Table 2 Illustration of WoS data structure

	Paper_ID	Author	Title	Year	Affiliation
Raw	a011	W. S. Hoehe	Using laser to …	1995	Univseritaet Wuerzburg
	a012	Wilhelm Hohe	Improved LASER …	1996	U. Wurzburg
	a013	Wilhelm Stefan Höhe	A new laser for …	1998	Uni Bamberg
	b021	K. Saue	Analysing polymer …	2002	U. Kassel
	b022	Karl Saue	Using LASER to measure …	1990	Unikassel versität
	b023	Karl Saue	Experiments in …	1985	Uni Goettingen
Clean	a011	W S Hohe	using laser to …	1995	Uni Wurzburg
	a012	W Hohe	improved laser …	1996	Uni Wurzburg
	a013	W S Hohe	a new laser for …	1994	Uni Bamberg
	b021	K Saue	analysing polymer …	2002	Uni Kassel
	b022	K Saue	using laser to measure …	1990	Uni Kassel
	b023	Karl Saue	experiments in …	1985	Uni Gottingen

The table describes the data originating from the Web of Science (WoS) in a fictitious example before (raw) and after (clean) applying the standardization procedure

high quality of our disambiguation is indicated by a comparison of distributions of papers attributed to highly frequent and very rare names (see Appendix "Web of science author name disambiguation").

We decided to first identify individual researchers' publication oeuvres and then link them to the doctoral students dataset.[5] The applied procedure involves two steps (as comparable to, e.g., Wang et al. 2012; Strotmann et al. 2009; Ferreira et al. 2012): First, we search for clusters of papers sharing author names (surname and first name initial) and affiliation. While the retrieved name-affiliation groups might be sufficient to track doctoral student publications, researchers generally tend to change affiliations during their careers. Therefore we apply a second disambiguation step. By disambiguating name-affiliation groups sharing author name (but not affiliation), we identify all publications of individual researchers throughout their careers. This procedure has the advantage that, in the first step, same name-affiliations combinations are more likely to belong to an individual scientist, and, in the second step, once name-affiliation-groups are sorted into distinct individuals more information is available when comparing homonyms over different affiliations (e.g., self-citations are more easily identified if more papers are considered). The WoS author disambiguation procedure is illustrated in Table 3.

Before we divide authors by names and affiliations the WoS requires substantial cleaning and standardization. We first process author names (surname and first initial)[6] by correcting German Umlaute as well as removing punctuation and whitespaces [the latter is important for German data; cf. Schoen et al. (2014)], and assign affiliations to individual

[5] An alternative approach would be the use of an assignment procedure as recommended method (Ferreira et al. 2012; Reijnhoudt et al. 2014), where information about individual scientists is collected first and subsequently publications are assigned to the known individuals. Our decision against the use of assignment procedures is based on the need to identify all publications attributed to the co-authors as well.

[6] Only in recent years have full author names become available in the Web of Science publication database.

Table 3 Illustration of WoS disambiguation procedure

Paper_ID	Author	Affiliation	Author_affil_ID	Author_life_ID
a011	W Hohe	Uni Wurzburg	A01	A1
a012	W Hohe	Uni Wurzburg	A01	A1
a013	W Hohe	Uni Bamberg	A02	A1
b021	K Saue	Uni Kassel	B01	B1
b022	K Saue	Uni Kassel	B02	B2
b023	K Saue	Uni Gottingen	B03	B3

The table describes the WoS author disambiguation procedure in a fictitious example

authors wherever possible.[7] We identify all German universities by comparing affiliations with a list containing decomposed variations of university names. Within all publications corresponding to one name and affiliation we search for distinct clusters that represent one individual scientist active at this affiliation. Within groups, the cluster algorithm assigns authorship to one distinct researcher if at least one of the following criteria is fulfilled: the same department is mentioned,[8] there is at least one common co-author (identified by name), authors share a second name initial, self-citations between the considered articles can be found, authors use the same e-mail address,[9] or identical (author) keywords are used. In contrast, author references are identified as different individuals if there is a gap of 10 or more years between articles without any intermittent publications. Different full first names and different second name initials also divide publications into separate scientists.

Before applying the second step of the disambiguation procedure we exclude all subsets. Subsets emerge when a researcher is using several affiliations. Then it is not clear which affiliation is the main affiliation. We choose to decide for the affiliation most papers are assigned to, and all others are deleted from the dataset. The homonym problem to solve now gets at a sample of authors assigned to different affiliations. We still observe some overlaps of one to several papers on different disambiguated name-affiliations groups in the case of several affiliations being mentioned on the paper. This might be the case when researchers change affiliations, but still collaborate with previous colleagues. In the case of overlapping publications two name-affiliation groups are attributed to one single researcher. Non-overlapping distinct name-affiliations are then identified as the same (or different) individuals based on the following criteria: shared co-authors (based on disambiguated name-affiliation groups), self-citations between papers of one group to the other, and shared author keywords. Again, authors are identified as different individuals whenever they have different first names, second initials or publication lags of 10 or more

[7] Author affiliations can be attributed to a single author in certain cases, e.g. for corresponding authors [for a detailed discussion see Reijnhoudt et al. (2014)]. Similar to Reijnhoudt et al. (2014) we considered all affiliations listed in the paper as potential true affiliations of the authors whenever a clear assignment was not possible.

[8] The department is tracked from the remaining part of author affiliations after country and zip codes, cities and universities were extracted. Frequent occurring terms such as "dept." were eliminated as well. Because of the high variation in spellings the remaining department information was compared by computing a Jaccard-similarity measure.

[9] E-mail addresses are assigned to individual authors on the basis of highest Jaccard-similarities between all authors and e-mail addresses.

years. As mentioned above the quality of the applied disambiguation procedure is evaluated in Appendix "Web of science author name disambiguation".

Matching dissertations with WoS authors

To obtain publications related to doctoral dissertations the pre-processed DNB and WoS records have to be linked. The primary name matching procedure performs a simple name comparison of the authors' surnames and first name initials, which however leads to a high number of false positive matches [as described by D'Angelo et al. (2011)]. Therefore we additionally require the graduating university to be equal to the researcher's earliest affiliation found in the WoS data. To further reduce false positive matched pairs we only considered pairs to be true matches if they (a) either refer to a unique match (a dissertation author with a unique name in the whole DNB sample matches to only one of the identified authors in WoS); or (b) some similarity between paper titles and thesis title could be found. Similarity between titles is computed by using a longest common substring algorithm.[10] We decide to only include those matched pairs with at least one word (with five or more letters) in common. As a consequence of these requirements, our approach is rather conservative, and some WoS-dissertation pairs may not be retrieved because authors chose very different titles. Matched pairs with lags between publications exceeding seven years are excluded from the sample. Table 4 exemplifies the record linkage procedure. Matching quality is evaluated by comparing author first names whenever available. (Results are reported in Appendix "Publication dissertation record linkage").

Co-authors of doctoral papers and feature space construction

We expect doctoral students to publish results of their dissertation in co-authorship with their advisor. Consequently we should find the advisor among the student's co-authors (for a related approach cf. Wang et al. 2017). By using the matched papers of dissertation authors we can easily obtain a full list of co-authors. To identify the advisor among these co-authors we calculate several individual characteristics for each of the co-authors. These characteristics should describe the unique position and importance among all co-authors for the respective student. Co-authors who started publishing first after the first publication of a specific dissertation author are excluded from the dataset. Specifically, we use the following characteristics to identify advisors within the group of co-authors:

- Times published together (number of papers co-authored together):
  We expect doctoral students to publish larger shares of their dissertation-related papers in co-authorship with their advisors as opposed to other individuals.
- Publication lag (time lag between first paper of focal student and first paper of co-author):
  We expect advisors to be active in scientific publishing a certain time period before their students.

[10] Before we compare titles they are standardized by removing (German and English) stop-words, punctuations and numbers as well as whitespaces. In case several WoS authors match the same dissertation and vice versa, the ones with the highest score are taken, where scores are set in relation to the length of the dissertation title.

Table 4 Illustration of DNB-WoS record linkage

Author_ID	DNB_ID	Titles	First affiliation	Unique pair	Linked
A1	101	Similar	Same	Yes	True
B1	202	Differnt	Same	No	False
B2	202	Similar	Same	No	True
B3	202	Similar	Differnt	No	False

The table illustrates the record linkage of the DNB data with the WoS data by continuing the previous fictitious examples of Tables 1, 2 and 3

- Degree in co-author network:
  The network degree in the co-author network provides information on who is connected to how many of the co-authors. We expect advisors to be connected to more co-authors because of their central role in the research group.
- Betweenness centrality within the co-author network:
  We expect advisors to be in a highly central position, bridging between most or all of the student's co-authors.
- Burt constraint, based on citations within the co-author network:
  We expect that the person who receives most attention among the co-authors (receiving most of the citations) can be interpreted as the person with a high impact on the group's work.
- Number of publications and number of citations:
  We expect that co-authors with higher reputation/seniority are more likely to be the advisor.
- Author name position on paper:
  Being listed last or first among the co-authors captures information about the role of the specific author. We expect that group leaders are more likely to be named last on the paper.
- Number of dissertations a co-author is mentioned on:
  We expect advisors to contribute to many dissertations. Co-authors who frequently appear on students' papers are therefore more likely to be advisors than those who are found less frequently.
- Citations received by student papers (number of times previous work of the co-authors is cited by students' papers):
  We expect that doctoral students tend to follow the research agenda of their advisors, which is reflected by citations received by co-authors.
- Same university:
  Co-author is affiliated to the degree-granting university at the same time as the student.
- Co-author holds doctoral degree:
  Recently graduated students may not be indicated as holding a doctoral degree in our dataset, while already acting as a post-doc who might have a significant influence on students' work.

In Table 12 (Appendix "Statistical description of the features") we report general descriptive statistics for all of the above described features separately for true advisors and other co-authors. We further include several dissertation-related characteristics:

- Number of papers
- Number of co-authors
- Year of submission

Table 5 Illustration of co-authors and features

Author_ID	DNB_ID	Co-author	Times pub. together	Pub. lag	Degree	…	Advisor
A1	101	M Schmidt	3	− 16	6	…	1
A1	101	R Muller	1	− 2	2	…	0
A1	101	G Meier	2	+ 1	3	…	0
B2	202	F Mauer	1	− 3	1	…	0
B2	202	M Schmidt	2	−11	4	...	1

The table continues the fictive example of Tables 1, 2, 3 and 4 and exemplifies the identification of the doctoral advisor by using different features of the students' co-authors

Before applying machine learning algorithms, variables have to be standardized because we are interested in identifying one specific individual in the group of co-authors. Some groups have generally high values for some of the attributes (e.g., large numbers of publications) but not for others (e.g., citation rates). We use z-score standardization centered to the mean of all co-authors of the focal doctoral student to take this into account. Table 5 illustrates the applied procedure for doctoral advisor identification in the co-author space.

Machine learning algorithms

The final step of our identification approach for doctoral advisors is to use the characteristics of co-authors outlined above to find the doctoral advisor among them. To this end we split our sample in two parts. The first part is used for training algorithms and identifying best parameter settings. The other is used to evaluate the resulting models on an independent dataset. We train four different algorithms that are standard in the data mining literature: regularized logistic regression, support vector machine (SVM), random forests, and ada Boost.[11] All algorithms are available as R packages; for a detailed discussion see Bishop (2006).[12]

The regularized logistic regression model performs logistic regression including additional penalty terms to the optimized error function to avoid overfitting by reducing the model's complexity. However, the penalty parameter needs to be specified. We also need to define a threshold when the estimated probability is high enough to be classified in the "true" class. SVM uses a hyperplane to separate data by maximizing the distance to the vectors (data points) that are closest to the hyperplane. Those vectors are described as support vectors. To make the observed data linearly separable they are transferred into higher dimensional space. This is done by using a kernel, in our case a radial kernel. To avoid overfitting false classifications can be allowed, which however are penalized. Two parameters can be specified: a penalizing parameter specifying cost, and parameter gamma

[11] All four selected algorithms are comparable easy to implement and scalable. More advanced machine learning tools, as e.g. deep learning algorithms, become usually more computationally intense when applied to larger datasets. Beside this, the selected algorithms are implemented in most of the standard statistical programs.

[12] Using the programming language *R Version: 3.3.2* (R Core Team 2016) we apply the following R packages: *glmnet* for the regularized logistic regression (Friedman et al. 2010); for SVM we used *e1071* created by Meyer et al. (2015); for random forest we used the *randomForest* package of Liaw and Wiener (2002); and in the case of ada Boost we use *ada* (Culp et al. 2016).

specifying the radial basis function kernel. SVM has been used for disambiguation tasks e.g. by Wang et al. (2012). Random forest builds on decision trees. Multiple decision trees are constructed by choosing a random set of features. Decision trees simply split the dataset to achieve best separation of two classes. Combining several splits leads to decision regions. After training a specific number of trees, all trees are used together to give a majority vote on the class of an object. As parameters the number of randomly drawn features as well as the number of trees need to be specified. The ada Boost algorithm is a boosting technique specified for binary response that has also been used for author name disambiguation by Han et al. (2004). We employ it as a fourth method even though it is also based on decision trees. Different from random forests, boosting trains classifiers in sequence. Depending on the classification output to the input data weights are assigned giving higher importance to misclassified data points. Afterward also all classifiers give a majority vote. Iterations and a shrinkage parameter need to be specified as parameters.

The sample of doctoral students in applied physics and electrical engineering

Several modifications of the general approach discussed above were necessary to adapt it to the requirements of the present study. First and most importantly, we require a test sample allowing us to assess the reliability of our co-publication-based approach. To this purpose we utilize the sample of verified student-advisor matches developed by Buenstorf and Geissler (2014), which encompasses the universe of German laser-related doctoral dissertations completed since 1960 and is based on information obtained by university departments as well as the names of examiners provided in the dissertations. In addition to completeness, a further advantage of this dataset is its longitudinal character. However, for our purposes it has the disadvantage that advisors are often identified as the first examiner of the dissertation. The training dataset includes 6389 doctoral students with dissertations completed from 1960 to 2007. However sample size is reduced for two reasons. First we are stricter about excluding medical dissertation projects than the original authors, and second the doctoral student needs to have at least one publication. This reduces our sample to 3583 observations, of which 1,686 (47%) have publications identified and 1367 have their advisor among their co-authors. Only the latter are included in the training dataset.

Second, to ensure that our matching is not biased by field-specific differences in the role of student and advisor characteristics we focus on the broader fields of applied physics and electrical engineering into which most non-medical laser research falls as a subset. These fields are also characterized by a prevalence of publications in WoS-listed journals, a precondition for our approach to work. Specifically, we limit the sample to dissertations classified as science or engineering by the DNB.[13] This entails that we exclude all theses classified as medical dissertations, which account for about 50% of all doctoral theses in Germany. These exclusions reduce the number of authors of doctoral dissertations to 185,860, which also helps to limit the problem of false positive matches caused by homonyms, i.e. non-identical doctoral students and article authors sharing the same names.

Publication data for this set of authors are retrieved from the subset of WoS articles contained in the Science Citation Index dataset. Starting from the full set of all articles, proceedings and reviews with at least one author affiliation in Germany, we further restrict the dataset to articles published in relevant fields. These are identified as follows. Initially all articles containing “laser” in the title are selected. The set of relevant articles is then

[13] We used the classification scheme of the DNB corresponding to the Dewey-Decimal Classification (DDC) which is an international standard used by libraries for subject classifications.

extended using title terms that frequently co-occur with "laser". The journals in which these articles were published constitute our set of relevant journals.[14] For these fields a total of 334,945 articles is covered in the WoS.

The matching procedure of the dissertation data (DNB dataset) and the publication records (WoS datatset) resulted in 30.969 true positive matched dissertation-publication record pairs. We identified 25,856 linked records by finding similarities between the titles of the dissertation and the corresponding papers, and 22,777 were identified by using matched unique names on both datasets. Doctoral students in the dataset published on average 5.39 papers (median: 3) with 8.46 co-authors (median: 5). Because the sample size of the training data is limited, we apply a threefold cross validation within the training sample to find the best parameter setting (Witten and Frank 2005). In the cross-validation procedure we split into test and training set in 1:3 portions, rotating estimation and average results. From the cross-validation we take the best fitting parameter setting and use the algorithm on the evaluation set.[15] As already stated, the predictive power of the models is then evaluated by using a separated partition of the dataset. Evaluation results are presented in Table 6.

Random forests outperform the other algorithms in terms of precision and recall. Overall we obtain an 82% precision and recall rate, which is similar to matching results in other contexts and should provide sufficient statistical power to test the subsequent empirical analysis. We use the random forests model for predicting advisors for the whole sample. In doing so we use the whole test-training dataset and apply it to the full database. We also tested a second set demanding a minimum recall of 0.6 and tried to maximize precision. However, this did not yield a substantial improvement. Nevertheless, we tested all our models applying SVM for classification. For 3% of the doctoral students the algorithm was not able to identify one single advisor. In these cases several co-authors appeared to be very similar in the feature set. We decided to accept this decisions because doctoral students might truly have several advisors. In Appendix "The academic genealogy of german applied physics and electrical engineering" we describe some main features of the obtained genealogy and their network structure.

In the remainder of this paper, we employ the advisor-student dataset constructed with the help of the author identification method described in this section to analyze the probability that advisors have fecund doctoral students, i.e. students who become doctoral advisors themselves. The theoretical groundwork of this analysis is laid in the next section.

Fecundity of doctoral students: theoretical considerations

The socialization of young researchers

Doctoral training prepares students for an academic career by enabling them to independently advance the frontier of knowledge. Doctoral students need to acquire a broad range of knowledge and skills to perform successfully in academia. These range from mastering the intricate details of their field to insights into research strategies, skills in operating

[14] We use only WoS publications listed in the science citation index with at least one German affiliation. Then we select all Web of Science disciplines with a high coverage of relevant journals: "Physics, Applied"; "Optics"; "Physics, Atomic, Molecular and Chemical"; "Engineering, Electrical and Electronics"; "Chemistry, Physical"; "Physics, Condensed Matter"; "Materials Science, Multidisciplinary"; "Physics, Multidisciplinary". Papers with more than 25 authors are not considered.

[15] Details of the tested parameter settings are provided upon request.

Table 6 Training results of advisor classification (selection of best models)

Model	± 1 (Best parameter)			± 1 (Min recall 0.6)		
	Precision	Recall	F1	Precision	Recall	F1
Logistic	0.67	0.87	0.77	0.80	0.62	0.71
SVM	0.80	0.82	0.81	0.90	0.66	0.78
Random forests	0.84	0.80	0.82	0.84	0.80	0.82
AdaBoost	0.80	0.79	0.80	0.80	0.81	0.80

laboratory equipment, and also routines of professional etiquette. The various types of knowledge and skills can be learned in a variety of ways. The institutional setup known as Open Science (Dasgupta and David 1994) provides researchers with strong incentives to codify and freely communicate knowledge via publications. Significant parts of the required knowledge can therefore be accessed by reviewing the relevant literature. However, there is broad consensus that not all relevant knowledge can be acquired from the literature. For skills such as the handling of laboratory equipment learning from one's own experience is crucial, often based on a process of trial and error.

Besides access to codified knowledge and own experience, direct face-to-face interaction with others is crucial for becoming a successful researcher (e.g. Collins 1974; Stephan 2012). Face-to-face interaction allows doctoral students to access non-codified or "tacit" knowledge that cannot be found in the published scientific literature. It also enables "vicarious" learning from observing the behavior of others as well as the environmental reaction to this behavior (Bandura 1986).

In the socialization of doctoral students into their academic discipline, a major contribution is attributed to their advisor (Long and McGinnis 1985). Especially tacit knowledge cannot be acquired without mentoring, which includes, e.g., knowledge about scientific norms, but also higher-level expertise that shapes the way of thinking (de Mey 1982; Sugimoto 2014). This suggests an important role of the advisor in shaping the knowledge, skills, and attitudes of doctoral students (Buenstorf and Geissler 2014). Advisor attitudes and behaviors such as the willingness to engage in the commercialization of their research findings may be transferred from advisors to their students (Azoulay et al. 2017).

In light of these considerations, it can be expected that doctoral students learn from their advisor, and that more can be learned from better-performing advisors with higher research productivity. In turn, acquired knowledge and skills enhance a student's chances to remain in academia, i.e., to become an advisor him- or herself. However, only a fraction of all graduated doctoral students pursue academic careers, which holds for various time periods and countries (Stephan 2012; Waldinger 2016) and also for leading departments (Conley and Önder 2014). At the same time, while mentoring seems to affect student performance, it has been suggested that their career commitment is less strongly shaped by the advisor (Paglis et al. 2006). Doctoral students do not necessarily follow career paths of their advisors.

Advisors are not only an important sources of knowledge for their doctoral students, but also important as gatekeepers of the scientific community. The reputation that a given advisor enjoys in their community is an important precondition of their students' ability to signal their quality. Being advised by a highly recognized researcher and/or co-authoring

with them thus acts to certify student quality. Students of more highly reputed advisors may also be able to benefit from an extended "Matthew effect" (Merton 1968) resulting in their work getting more attention. To the extent that reputation builds upon academic achievement, this would provide further advantages to students of more productive advisors.

Advisors perform even more direct gate-keeping functions in referring students within the community (Baruffaldi et al. 2016) and by helping them secure academic jobs (a precondition for them to advise substantial numbers of students). In providing access to the larger scientific community, Long and McGinnis (1985) describe advisors as acting like "employment agents" for their students. How effectively a given advisor can perform these functions may depend less on their academic performance, and more on how well-connected they are in their community. Having a powerful position in the community network enables advisors to exert an influence on hiring decisions and generally to learn about opportunities to support their students' careers.

We expect both advisor performance and advisor influence to exert beneficial influences on their students' probability to become advisors themselves, and will try to disentangle them in the subsequent analysis by employing different empirical measures (citation counts as measures of performance vs. network-based variables as measures of influence). Specifically we predict the following effects on students' academic fecundity:

Hypothesis 1 *Doctoral students whose advisor has superior research output have a higher probability to become advisors themselves.*

Hypothesis 2 *Doctoral students whose advisor is better positioned in the network of the respective scientific community have a higher probability to become advisors themselves.*

Life cycle effects

As advisors get older and more experienced, there may be relevant changes in how they interact with and support their students. Advisors' age and career stage at the time they work with a particular student can thus be expected to matter for doctoral students' subsequent career outcomes. However, it is not obvious how the probability that students remain active in research and become advisors themselves changes over the career trajectory of their advisors.

On the one hand, we would generally expect that more senior advisors are better able to teach their students. We would also expect more senior advisors to be more established in their community. They should therefore be better positioned to leverage their own contacts to the benefit of their students, for example by arranging talks and visits to other research groups possessing complementary knowledge. In addition, doctoral advisors may switch positions during their career. Such moves would generally be expected from less to more highly reputed universities, suggesting that students advised at later career stages may enjoy a better working environment.

On the other hand, life-cycle models of individual research output (Levin and Stephan 1991) predict that as researchers get closer to their retirement age, they have weaker incentives to engage in research activities. Instead, they will often be more strongly inclined to engage in consulting work and/or decrease their own working hours. In addition, while being more experienced and possibly having a broader overview of their field, more senior

researchers may lose touch with the most recent developments in terms of theories and empirical methods. Their research foci may increasingly diverge from those that are currently "hot", which makes it harder for their students to find attention for their own work.

As scientific knowledge tends to decay relatively rapidly, more senior advisors may thus be less capable to teach their students. Furthermore, it is plausible that students of more senior researchers face stronger competition for the advisor's scarce time and attention. More senior advisors will often have a larger burden of administrative and managerial duties in the research group and the department. With increasing group size, they also have to divide their attention between a larger number of students. It moreover seems plausible that with an increasing "stock" of prior doctoral students, advisors' interest in securing the "survival" of their line of research, which requires students who become advisors of the new generation of students, is increasingly muted.

The above conjectures are related to what has been discussed as the "rising star hypothesis" which posits that ambitious, high-potential candidates are more likely to find a mentor who supports their further career (Singh et al. 2009). Malmgren et al. (2010) extend this hypothesis to account for effects of advisor career stage on students' career outcomes and suggest a homophilic matching process of promising students and young, ambitious advisors. Based on this conjecture, rising stars tend to select students with high potential, as mentoring these students is beneficial for their own further career. If the focus on high-potential students is subsequently copied by the students of "rising stars", then for these students a higher fecundity can be expected as well. Malmgren et al. (2010) provide empirical evidence supporting the "rising star hypothesis" for mathematicians. Among the students of otherwise comparable advisors, those trained in the first third of their advisor's career tend to have a higher fecundity. In general, Malmgren et al. (2010) find a negative association between the overall number of doctoral students and observed student fecundity.

Based on the prior evidence of mostly adverse effects of seniority on research performance, we expect the second set of arguments (suggesting a decreasing likelihood of student fecundity with increasing advisor seniority) to be more relevant than the first one (which suggest an increasing likelihood). We also expect that the role of advisors' accumulated stock of doctoral students can be separated from that of advisor career age. These considerations inform a second set of hypotheses:

Hypothesis 3 *Doctoral students whose advisor has a larger number of prior doctoral students have a reduced probability to become advisors themselves.*

Hypothesis 4 *Doctoral students whose advisor has reached a higher career age have a reduced probability to become advisors themselves.*

Peer effects and competition among doctoral students

Advisors are certainly not the only source of knowledge and skills that doctoral students draw upon. Relevant face-to-face interaction primarily takes place in the setting of the research group or laboratory, with fellow students and co-workers (Krabel 2012; Tartari et al. 2014; Hottenrott and Lawson 2017) being important partners from whom doctoral students can learn. The direct environment provides opportunities for informal mentoring, such as support and guidance of fellow students, as well as the sharing and discussion of ideas, before the official advisor is consulted (Sugimoto 2012). In the German university context, co-workers

in the research group have mostly been selected and/or trained by the student's own advisor. Accordingly, peer group effects exerted by fellow students and co-workers will generally not be independent from advisor effects. It nonetheless seems plausible that students benefit from being part of a strong cohort of researchers contemporaneously working in the same research group or laboratory. Other group members are sources of knowledge and social capital, and they may also provide role models affecting the career choices that a student makes. This suggests a positive peer group effect on individual students' performance and thus their ability and willingness to embark on an academic careers.

Hypothesis 5.a *Doctoral students have a higher probability to remain active in academia and to become advisors themselves if they are trained together with larger numbers of contemporary peers having the same advisor who subsequently become doctoral advisors.*

The above argument for positive peer effects does not take into account, however, that doctoral students in the German system need to go through additional career stages before assuming their first faculty positions. To a considerable extent these post-doctoral career stages are taken in the same research group or laboratory in which the individual received her doctoral training, which seems to be associated with superior career outcomes (Bäker 2015). This implies a competitive relationship among the student and her contemporary peers or "siblings", since often only a fraction of the graduated doctoral students can remain in the lab. Students may thus suffer from being trained together with exceptionally capable "siblings", which informs a competing hypothesis on peer impact:

Hypothesis 5.b *Doctoral students have a reduced probability to remain active in academia and to become advisors themselves if they are trained together with larger numbers of contemporary peers having the same advisor who subsequently become doctoral advisors.*

Table 7 summarizes all hypotheses with the predicted effect on students fecundity.

Econometric analysis

Based on the approach detailed in section "Identifying doctoral advisors based on co-authored publications" we were able to match a total of 25,735 student-advisor pairs in German applied physics and electrical engineering for the time period 1975 to 2005. We use this sample to test the above hypotheses about determinants of doctoral students' scientific fecundity (here defined by the student becoming a doctoral advisor him- or herself). Specifically, we estimate a set of logit models of the individual likelihood to be identified as an advisor of at least one dissertation in our dataset. This likelihood is obviously affected by the time when a given individual defended her own dissertation, which generates a truncation problem that might bias our results. To limit the potential impact of truncation, we restrict our sample in that we only consider individuals as potential advisors who completed their own dissertation no later than 1995. As a consequence the final sample is further reduced to 5373 doctoral students.[16]

[16] Note that due to this truncation problem, the dataset from Buenstorf and Geissler (2014), which we use as a training dataset in our approach to identify advisors of doctoral dissertations, cannot be used for the econometric analysis. Restricting this dataset to the years before 1996 would reduce the number of observations to 289, which provides insufficient statistical power to test our hypotheses.

Table 7 Summary of predictions

Hypothesis	Advisor characteristic	Predicted effect on student fecundity
1	Research output	+
2	Position in community network	+
3	Doctoral students previously advised	−
4	Career age	−
5.a	# of fecund contemporary peers	+
5.b	# of fecund contemporary peers	−

A brief description of all variables and general descriptive statistics are presented in Table 8. Pearson correlation coefficients of the variables in the final dataset are reported in Appendix "Additional descriptive statistics".

Doctoral students are not randomly matched to their advisors. In general, assortative matching of students and advisors can be expected (Azoulay et al. 2017). Students are often deliberately hired from within their future advisors' networks (Baruffaldi et al. 2016). We are limited in the extent to which we can control for the bias this introduces in the analysis of advisor effects on students. However, we conjecture that the inherent quality of students, as well as the quality of the match with the doctoral advisor, is reflected by the early success of the student. This can be observed in the number of publications a student produces before completing their doctoral degree. Indeed, as is indicated in Table 13, this variable is strongly correlated with advisor characteristics such as the advisor's citations, indicated by a correlation coefficient above 0.3. We further control for the number of co-authors a student accumulates in their early work, which can also be assumed to reflect the quality of students and of their matching with the advisor.

Given substantial correlation between some of the explanatory variables (some above 0.5, see Table 13), we first use them separately in individual models and then estimate a full model specification including all variables (Models 1 to 8 in Table 9). All models include a set of time dummies (5-year periods) for graduation years and advisors' first publications, as well as fixed effects for degree-granting universities. We then re-estimate the same models controlling for student quality (Table 10), and again including advisor fixed effects (Table 11, where the number of observations is reduced to 5140 because all advisors with only a single dissertation in the dataset, or with all-positive outcomes, are dropped from the estimation).

We proxy adviors' research productivity by the number of citations to their publications. A significantly positive coefficient for this variable is estimated in Model 1 (Table 9), but this finding is not robust to adding other explanatory variables or controlling for student quality. In the fixed effects specification, we even obtain a significantly negative coefficient for the number of advisor citations (Model 1 in Table 11). Accordingly, there is only limited support for Hypothesis 1.

The role of advisor positions in the scientific community is captured by three different network measures: their presence in the principal component of the publication network (*Adv. main-comp.*), their Bonacich centrality (*Adv. Bonacich*) as well as their degree (*Adv. degree*). Bonacich centrality is proposed by Ballester et al. (2006) as a suitable measure to identify key actors in networks. It goes back to Bonacich (1987) who used it to assess the power of central players, i.e., those who exert strong influence on surrounding players, by

Table 8 Summary statistics ($N = 5518$)

Variable	Description	Mean	SD	Min	Median	Max
Stud. fecund	Student is identified as an advisor of at least one dissertation in the complete genealogy	0.06	0.23	0	0	1
Stud. nbr. pub.	Number of student's publications associated with the student's dissertation	2.81	2.62	1	2	25
Stud. degree	Number of co-authors in the publication network at the time of submission	3.38	3.71	0	2	43
Adv. main-comp.	Advisor is member of the principal component of the publication network at the time of submission	0.82	0.39	0	1	1
Adv. Bonacich cent.	Number of paths to other advisors weighted by a decay factor that increases with path length[a]	11.18	16.15	0.00	3.19	84.62
Adv. degree	Number of co-authors in the publication network at the time of the dissertation	11.18	11.45	1	8	212
Adv. nbr. cit.	Sum of all citations received by the advisor till the time of the dissertation	224.37	305.15	0	117	2654
Adv. nbr. prev. stud.	Number of dissertations advised prior to that of the focal student	6.04	6.09	1	4	40
Adv. career age	Lag between the advisor's first publication and the time of submission	12.45	4.92	0	13	21
Fecund siblings	Number of doctoral students of same advisor who are fecund and in the students cohort	0.65	1.05	0	0	9

[a] In the publication network at the time of the dissertation. A more detailed description how the Bonacich centrality is calculated in our case is provided in the text

counting all paths to other nodes weighted by a decay factor increasing with path length. We use a variation of this measure by only considering other advisors as relevant.[17]

In Table 9, Models 2 to 4, we find support for a positive relationship of all three variables with the likelihood of students to become advisors themselves (i.e. their academic fecundity). In the full model, only advisor presence in the principal component remains statistically significant. Controlling for student quality (Table 10) moreover indicates the importance of selection. While we find that better students (in terms of more early publications; cf. Balsmeier and Pellens (2014)) are more likely to become advisors later on,[18] the coefficient of advisor membership in the principal component is reduced substantially, and both Bonacich centrality and degree lose all of their explanatory power. In the within-estimator framework of the fixed effects models of Table 11, the role of advisor network membership is further qualified, mostly reflecting imprecise measurement because the respective variables mostly vary in the cross section but less so over time. Accordingly, we only find limited support for the prediction of Hypothesis 2.

[17] We further introduce a cut-off value at path lengths above 5 for computational reasons. As decay factor we chose $2^{-(k-1)}$ depending on path length k.

[18] Coefficients obtained for the degree (number of co-authors) of students are generally negative and in some models marginally significant. This suggests that conditional on the number of early publications, having more co-authors is associated with a reduced likelihood to remain active in academia and become an advisor. In models without the publication variable, coefficients of student degree are positive.

Table 9 Logit estimations on determinants of individual scientific fecundity

	Dependent variable							
	Student fecund							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Adv. nbr. cit.	0.001***							0.0003
	(0.0001)							(0.0002)
Adv. main-comp.		0.764***						0.552**
		(0.212)						(0.217)
Adv. Bonacich cent.			0.009**					− 0.004
			(0.004)					(0.006)
Adv. degree				0.014***				0.006
				(0.004)				(0.006)
Adv. nbr. prev. stud.					− 0.015			− 0.058***
					(0.013)			(0.016)
Adv career age						− 0.038		− 0.054*
						(0.026)		(0.028)
Fecund sibling							0.375***	0.395***
							(0.047)	(0.057)
Constant	− 3.056***	− 2.534***	− 2.572***	− 2.603***	− 2.481***	− 2.329***	− 2.557***	− 2.776***
	(0.452)	(0.414)	(0.414)	(0.413)	(0.411)	(0.430)	(0.411)	(0.468)
Observations	5373	5373	5373	5373	5373	5373	5373	5373
Log likelihood	− 1, 136.543	− 1, 142.392	− 1140.218	− 1, 136.364	− 1, 143.823	− 1, 143.667	− 1, 115.821	− 1, 100.389
Akaike Inf. Crit.	2,401.086	2,412.783	2,408.437	2,400.728	2,415.645	2,415.334	2,359.642	2,340.778

$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Estimates from logit regression model, (robust) standard errors in parentheses.

All models include a set of time dummies (5-year periods) for graduation years and advisors' first publications, as well as fixed effects for degree-granting universities

Table 10 Logit estimations on determinants of individual scientific fecundity, including student variables

	Dependent variable							
	Student fecund							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Adv. nbr. cit.	0.0001							−0.00001
	(0.0002)							(0.0002)
Adv. main-comp.		0.430*						0.407*
		(0.223)						(0.223)
Adv. Bonacich cent.			−0.008					−0.012*
			(0.005)					(0.006)
Adv. degree				−0.00002				0.003
				(0.006)				(0.007)
Adv. nbr. prev. stud.					−0.019			−0.038**
					(0.013)			(0.017)
Adv. career age						−0.052*		−0.053*
						(0.028)		(0.029)
Fecund sibling							0.246***	0.315***
							(0.054)	(0.063)
Stud. nbr. pub.	0.238***	0.251***	0.244***	0.241***	0.243***	0.245***	0.229***	0.229***
	(0.021)	(0.022)	(0.021)	(0.021)	(0.021)	(0.021)	(0.021)	(0.022)
Stud. degree	−0.031	−0.025	−0.030	−0.031	−0.028	−0.031	−0.039*	−0.034*
	(0.020)	(0.020)	(0.020)	(0.020)	(0.020)	(0.020)	(0.021)	(0.020)
Constant	−3.608***	−3.323***	−3.317***	−3.323***	−3.302***	−3.087***	−3.297***	−3.333***
	(0.473)	(0.440)	(0.441)	(0.441)	(0.440)	(0.462)	(0.441)	(0.490)
Observations	5373	5373	5373	5373	5373	5373	5373	5373
Log likelihood	−1056.873	−1057.631	−1059.088	−1058.958	−1058.054	−1057.505	−1048.389	−1037.566

Table 10 continued

	Dependent variable							
	Student fecund							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Akaike inf. crit.	2245.746	2247.261	2250.175	2249.917	2248.107	2247.010	2228.777	2219.131

$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Estimates from logit regression model, (robust) standard errors in parentheses.

All models include a set of time dummies (5-year periods) for graduation years and advisors' first publications, as well as fixed effects for degree-granting universities

Table 11 Fixed-effect logit estimations on determinants of individual scientific fecundity, including student variables

	Dependent variable							
	Student fecund							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Adv. nbr. cit.	−0.001**							0.001
	(0.0005)							(0.001)
Adv. main-comp.		0.366						0.289
		(0.434)						(0.453)
Adv. Bonacich cent.			−0.031***					−0.018
			(0.009)					(0.012)
Adv. degree				−0.016				0.002
				(0.011)				(0.013)
Adv. nbr. prev. stud.					−0.058**			0.041
					(0.026)			(0.036)
Adv. career-age						−0.151***		−0.093
						(0.056)		(0.071)
Fecund sibling							−1.003***	−0.996***
							(0.117)	(0.123)
Stud. nbr. pub.	0.355***	0.367***	0.355***	0.360***	0.360***	0.359***	0.402***	0.403***
	(0.035)	(0.036)	(0.035)	(0.035)	(0.036)	(0.036)	(0.038)	(0.038)
Stud. degree	−0.044*	−0.039	−0.026	−0.036	−0.042*	−0.044*	−0.022	−0.026
	(0.023)	(0.024)	(0.027)	(0.024)	(0.024)	(0.024)	(0.026)	(0.030)
Constant	−19.952	−20.612	−20.194	−20.338	−20.548	−21.207	−21.194	−21.637
	(16,877.470)	(16,877.290)	(16,877.190)	(16,877.330)	(16,870.620)	(16,872.330)	(16,877.260)	(16,877.070)
Observations	5140	5140	5140	5140	5140	5140	5140	5140
Log likelihood	−668.809	−663.150	−667.981	−666.967	−666.543	−665.480	−626.166	−623.577

Table 11 continued

	Dependent variable							
	Student fecund							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Akaike inf. crit.	3427.619	3416.301	3425.962	3423.933	3423.086	3420.960	3342.331	3349.153

$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Estimates from fixed-effect logit regression model, standard errors in parentheses.

All models include a set of time dummies (5-year periods) for graduation years

Hypotheses 3 and 4 focus on life-cycle effects, i.e. the question whether more experienced (older) advisors are more likely to turn out fecund students than less experienced (younger) ones. We try to capture life-cycle effects by two variables: the number of dissertations advised prior to that of the focal student (*Adv. nbr. prev. stud.*), as well as advisor career age at the time of submission (time expired since their first publication; *Adv. career age*). Our findings point to negative effects of seniority on student fecundity, as the coefficients obtained for both variables are mostly negative and often significantly different from zero. In particular, we find strong negative associations between the seniority measures and students' likelihood to become advisors in the fixed-effects framework of Table 11 (Models 5 and 6), which controls for (time-invariant) advisor characteristics. Further support for this interpretation is provided by the significantly negative coefficient of the number of previous citations in the fixed effects model (Model 1 in Table 11).

Finally we look at the peer effects that were the subject of Hypotheses 5a and 5b (*Fecund siblings*). A complex pattern emerges. On the one hand, we find robust evidence suggesting a positive peer impact in Models 7 and 8 of Tables 9 and 10. This is consistent with benefits from working in groups where other students successfully embark on academic careers. On the other hand, in line with our above conjecture that peer quality is not independent of advisor quality, the positive impact of peers appears to be mostly cross-sectional. In contrast, in the fixed-effects models of Table 11, we obtain an even more pronounced negative association between the fecundity of contemporaneous "siblings" and the focal student's likelihood to become an advisor herself. This indicates that the competitive effect predicted by Hypothesis 5b dominates.

Concluding remarks

In this paper we presented a novel approach to identify the advisors of doctoral students based on co-publications and machine learning of matching algorithms. This approach can help alleviate the problem that large-scale datasets linking doctoral students to their advisors are lacking for many disciplines and fields of research. The limited availability of suitable data currently requires researchers interested in student-advisor relationships to rely on academic genealogies from specific contexts that are not necessarily representative of academic research more generally.

The proposed approach enabled us to construct and analyze a large-scale dataset of German doctoral graduates in applied physics and electrical engineering from 1975 to 2005. Using this dataset, we studied factors associated with the subsequent likelihood of these individuals to be "fecund" as academics, i.e. to become doctoral advisors themselves. This issue has found little attention in the prior literature even though it is directly relevant to the self-reproduction of science.

We obtained substantial evidence indicating that advisor seniority is associated with a decreased likelihood of students to become advisors themselves. This finding, which was robust to controlling for student quality and advisor heterogeneity, suggests that working with good advisors early in the advisor's career is helpful for the academic career of doctoral students. One interpretation is that younger advisors are more in touch with current developments in their fields, and that this is more important than their academic reputation or their position in the community. In line with this interpretation, our analysis suggests a limited importance of citations to advisors' prior work or of their network position. Our results may also reflect early matching and selection of high potential researchers, consistent with the conjecture that "rising stars" among advisors match with

particularly promising students (Malmgren et al. 2010). In addition, our results indicate that competition for academic career opportunities among the students co-advised by an advisor at the same time plays a relevant role. The dynamics that our results imply are consistent with dynastic relationships in science. However, in contrast to what is argued by Horta et al. (2010), they seem to be more suggestive of meritocratic processes than of navel gazing.

Both the methodological approach presented above to match students and advisors using publication data and the specific results on academic "fecundity" are not without limitations. First and perhaps most importantly, we can only match individuals for whom we can obtain co-authored publications in openly accessible databases. This excludes fields and disciplines in which the respective type of publication is uncommon, which holds for large parts of the social sciences and humanities. Second, our approach identifies those individuals as advisors who leave their mark on students' research output. While we think that these individuals are indeed the key persons in socializing the next generations of scientists, they may differ from those who are officially listed as advisors, e.g., university professors "filling in" as examiners for researchers lacking the formal qualification to do so. Third, we have so far only considered student-advisor pairs where both parties completed their dissertations at a German university. In an increasingly global system, this is an increasingly limiting restriction. It is likewise not clear to what extent the above findings on the determinants of student "fecundity" are generalizable to other fields and countries, and how strongly the production of fecund students is associated with conventional measures of researcher productivity. These and other limitations of our analysis, including the imperfect empirical measurement of theoretical concepts, should be addressed in future research. Doing so will require substantial further effort in data collection. Finally, while we tried to control for the quality of students and the student-advisor match, a causal interpretation of our results would require randomized matching which is not what we have in our data.

Acknowledgements We would like to thank two anonymous reviewers of the ISSI 2017 conference, as well as two reviewers of this journal, for their helpful comments. This work was funded by the German Federal Ministry of Education and Research (BMBF) in its program "Forschung zu den Karrierebedingungen und Karriereentwicklungen des Wissenschaftlichen Nachwuchses (FoWiN)" under Grant 16FWN001.

Appendix 1: Assessment of data quality

Web of science author name disambiguation

To test the assignment quality of the WoS author disambiguation procedure we compare the number of publications attributed to very frequent versus rare names. To get accurate numbers of the name frequencies we use the DNB database on doctoral students. Because foreign names are overrepresented in unique name combinations we consider only doctoral students with German nationality in the two name groups. Doctoral students from abroad might suffer from systematically lower publication records, since they are more likely to leave Germany after graduation. We select the 100 most frequent name combinations (surname and first name initial) in the whole DNB dataset (which covers about one million records) and compare these against names which occur only once. The distribution of publication oeuvre sizes should be equal if our disambiguation procedure works accurately. Figure 2 compares the two distributions, which indeed appear to be equal. We cannot find

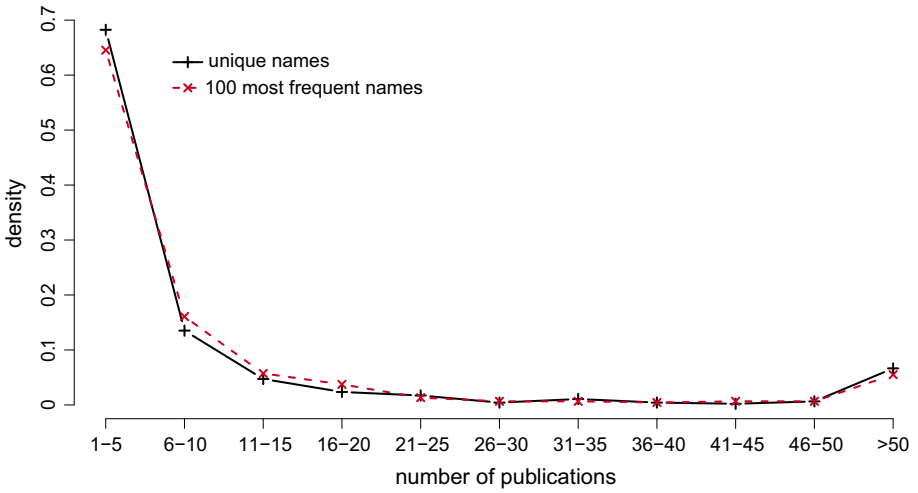

Fig. 2 Density of publications assigned to 100 most frequent (German) names and unique (German) names

a significant difference between the mean number of publications authored by researchers with one of the 100 most frequent names and those of authors with unique names (means: 12.37, 13.36; *p*-value: 0.72). The median equals 3 for both groups.

Publication dissertation record linkage

To check the quality of our matching of dissertation data with WoS scientists we employ first names, which we could not use in large scale. We find full first names in both sources for 6466 pairs. Comparing all first names manually, we found 257 pairs with different name. This corresponds to an error rate of 3.97% (or 96% correctly matched pairs).

Statistical description of the features

See Table 12

Table 12 Descriptive statistics of the features in the training data

Feature	True advisor	Mean	Median	Min	Max
Times published together	1	5.137	3	1	112
	0	2.021	1	1	45
Publication lag	1	13.34	13	− 8	34
	0	0.594	0	− 24	33
Degree in co-author network	1	10.08	6	0	185
	0	9.127	7	0	192
Betweenness centrality	1	108.616	5.667	0	14,466.893
	0	12.39	0	0	3817.51
Burt constraint	1	0.4109	0.4082	0	1.5515
	0	0.293	0.216	0	1.637
Number of publications	1	53.61	38	0	649
	0	11.07	1	0	698
Number of citations	1	799.6	420	0	12,284
	0	146.5	5	0	18,678
Number of dissertations	1	18.08	13	0	166
	0	4.797	1	0	173
Citations received by student	1	5.679	3	0	114
	0	1.594	1	0	94
Same university	1	1	1	1	1
	0	0.917	1	0	1
Co-author holds doctoral degree	1	0.05552	0	0	1
	0	0.335	0	0	1

The table reports descriptive statistics on the distribution of the features used to predict the advisor among doctoral recipients' co-authors. The data is split into two samples: advisors ($N = 1369$) and other co-authors ($N = 18{,}665$), while the advisor subset is indicated by true advisor equal to one

Appendix 2: The academic genealogy of German applied physics and electrical engineering

The genealogies obtained capture 38,037 members (students and advisors). We do not obtain one single genealogy connecting all members, but find a multiple of non-connected genealogies scattered into 5936 independent sub-graphs. This reflects different origins of advisors aggregated in our data, including advisors who obtained their own doctoral degree outside the German university system. About half of the independent genealogies (3077) only capture one single advisor-student link. Within the independent genealogies we mostly find genealogies that only reach over one generation ($N = 5391$). Reaching over two generations we count 507 genealogies, 33 reach over three generations, and three genealogies reach over four generations. The advisor with most offspring has 70 students, while there are 2818 advisors with only one student. The largest connected sub-graph contains 3263 members (see Fig. 3b). This genealogy covers four generations of doctorate students which are attributed to only one advisor (see Fig. 3a). In the sub-graph, we find 18 advisors with offspring in the third generation, and 90 advisors with offspring in the second

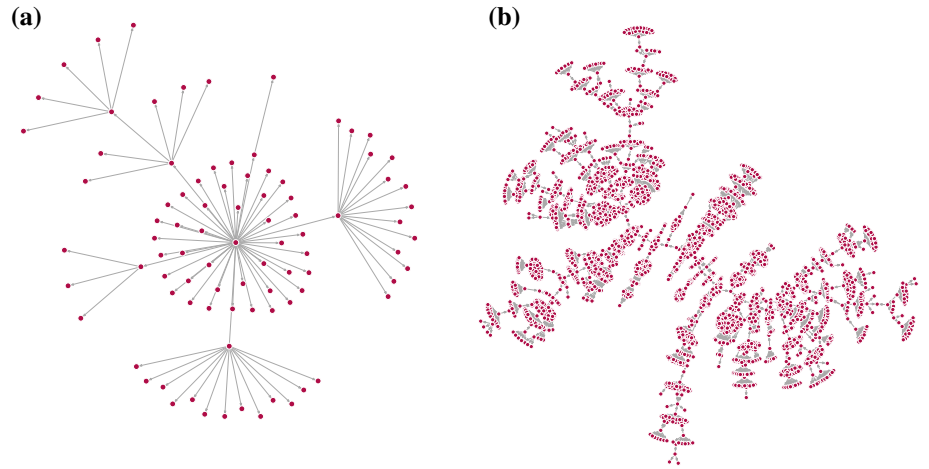

Fig. 3 Exemplary visualization of the obtained genealogies **a** Four generation genealogy, **b** largest connected sub-genealogy

generation. However, the size of the largest connected genealogy is partially due to students with several advisors who connect different sub-graphs.

Appendix 3: Additional descriptive statistics

See Table 13

Table 13 Correlations

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
(1) Stud. fecund	1	0.21	0.08	0.05	0.02	0.04	0.07	− 0.02	− 0.02	0.12
(2) Stud. nbr. pub.		1	0.52	0.2	0.33	0.34	0.34	0.09	0.17	0.28
(3) Stud. degree			1	0.16	0.34	0.57	0.23	0.13	0.16	0.23
(4) Adv. main-comp.				1	0.33	0.26	0.27	0.19	0.3	0.24
(5) Adv. Bonacich cent.					1	0.61	0.6	0.36	0.43	0.35
(6) Adv. degree						1	0.5	0.42	0.32	0.42
(7) Adv. nbr. cit.							1	0.37	0.28	0.5
(8) Adv. nbr. prev. stud.								1	0.47	0.39
(9) Adv. career age									1	0.2
(10) Fecund siblings										1

References

Andraos, J. (2005). Scientific genealogies of physical and mechanistic organic chemists. *Canadian Journal of Chemistry*, *83*, 1400–1414.

Azoulay, P., Liu, C. C., & Stuart, T. E. (2017). Social influence given (partially) deliberate matching: Career imprints in the creation of academic entrepreneurs. *American Journal of Sociology*, *122*(4), 1223–71.

Bäker, A. (2015). Non-tenured post-doctoral researchers' job mobility and research output: An analysis of the role of research discipline, department size, and coauthors. *Research Policy*, *44*(3), 634–650.
Ballester, C., Calvo-Armengol, A., & Zenou, Y. (2006). Who's who in networks. Wanted: The key player. *Econometrica*, *74*(5), 1403–1417.
Balsmeier, B., & Pellens, M. (2014). Who makes, who breaks: Which scientists stay in academe? *Economics Letters*, *122*(2), 229–232.
Bandura, A. (1986). The explanatory and predictive scope of self-efficacy theory. *Journal of Social and Clinical Psychology*, *4*(3), 359–373.
Baruffaldi, S., Visentin, F., & Conti, A. (2016). The productivity of science and engineering PhD students hired from supervisors' networks. *Research Policy*, *45*(4), 785–796.
Bishop, C. M. (2006). *Pattern recognition and machine learning*. New York, NY: Springer.
Bonacich, P. (1987). Power and centrality: A family of measures. *American Journal of Sociology*, *92*(5), 1170–1182.
Buenstorf, G., & Geissler, M. (2014). Tracing role model learning in the evolution of German laser research. *Jahrbücher für Nationalökonomie und Statistik*, *234*(2+3), 158–184.
Collins, H. M. (1974). The TEA set: Tacit knowledge and scientific networks. *Science Studies*, *4*(2), 165–185.
Conley, J. P., & Önder, A. S. (2014). The research productivity of new PhDs in economics: The surprisingly high non-success of the successful. *Journal of Economic Perspectives*, *28*(3), 205–216.
Culp, M., Johnson, K., & Michailidis, G. (2006). ada: An R package for stochastic boosting. *Journal of Statistical Software*. https://doi.org/10.18637/jss.v017.i02.
D'Angelo, C. A., Giuffrida, C., & Abramo, G. (2011). A heuristic approach to author name disambiguation in bibliometrics databases for large-scale research assessments. *Journal of the American Society for Information Science and Technology*, *62*(2), 257–269.
Dasgupta, P., & David, P. A. (1994). Toward a new economics of science. *Research Policy*, *23*(5), 487–521.
David, S. V., & Hayden, B. Y. (2012). Neurotree: A collaborative, graphical database of the academic genealogy of neuroscience. *PLoS One*, *7*(10), e46608.
de Mey, M. (1982). *The cognitive paradigm*. Dordrecht: D. Reidel Publishing Company.
Dores, W., Benevenuto, F., & Laender, A.H. (2016). Extracting academic genealogy trees from the networked digital library of theses and dissertations. In Proceedings of the 16th ACM/IEEE-CS on joint conference on digital libraries—JCDL '16 (pp. 163–166).
Ferreira, A. A., Gonçalves, M. A., & Laender, A. H. (2012). A brief survey of automatic methods for author name disambiguation. *SIGMOD Record*, *41*(2), 15–26.
Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1), 1–22.
Han, H., Giles, L., Zha, H., Li, C., & Tsioutsiouliklis, K. (2004). Two supervised learning approaches for name disambiguation in author citations. In Proceedings of the 2004 joint ACM/IEEE conference on IEEE. (pp. 296–305).
Horta, H., Veloso, F. M., & Grediaga, R. (2010). Navel gazing: Academic inbreeding and scientific productivity. *Management Science*, *56*(3), 414–429.
Hottenrott, H., & Lawson, C. (2017). Flying the nest: How the home department shapes researchers' career paths. *Studies in Higher Education*, *42*(6), 1091–1109.
Jackson, A. (2007). A labor of love: the mathematics genealogy project. *Notices of the American Mathematical Society*, *54*(8), 1002–1003.
Krabel, S. (2012). Scientists' valuation of open science and commercialization: The influence of peers and organizational context. In G. Buenstorf (Ed.), *Evolution, organization and economic behavior* (pp. 75–102). Cheltenham: Edward Elgar.
Levin, S. G., & Stephan, P. E. (1991). Research productivity over the life cycle: Evidence for academic scientists. *The American Economic Review*, *81*(1), 114–132.
Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News*, *2*(3), 18–22.
Long, J. S., & McGinnis, R. (1985). The effects of the mentor on the academic career. *Scientometrics*, *7*(3–6), 255–280.
Malmgren, R. D., Ottino, J. M., & Amaral, L. A. N. (2010). The role of mentorship in protégé performance. *Nature*, *465*(June), 622–627.
Marsh, E. J. (2017). Family matters: Measuring impact through one's academic descendants. *Perspectives on Psychological Science*, *12*(6), 1130–1132.
Merton, R. K. (1968). The Matthew effect in science. *Science*, *159*(3810), 56–63.
Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., & Leisch, F. (2015). *e1071: Misc functions of the department of statistics, probability theory group (Formerly: E1071)*. Wien: TU Wien.

Morichika, N., & Shibayama, S. (2016). Use of dissertation data in science policy research. *Scientometrics*, *108*(1), 221–241.

Paglis, L. L., Green, S. G., & Bauer, T. N. (2006). Does adviser mentoring add value? A longitudinal study of mentoring and doctoral student outcomes. *Research in Higher Education*, *47*(4), 451–476.

R Core Team. (2016). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna.

Reijnhoudt, L., Costas, R., Noyons, E., Börner, K., & Scharnhorst, A. (2014). Seed + expand : A general methodology for detecting publication oeuvres of individual researchers. *Scientometrics*, *101*(2), 1403–1417.

Rossi, L., Freire, I. L., & Mena-chalco, J. P. (2017). Genealogical index: A metric to analyze advisor—advisee relationships. *Journal of Informetrics*, *11*(2), 564–582.

Schoen, A., Heinisch, D., & Buenstorf, G. (2014). Playing the name game to identify academic patents in Germany. *Scientometrics*, *101*(1), 527–545.

Singh, R., Ragins, B. R., & Tharenou, P. (2009). What matters most? The relative role of mentoring and career capital in career success. *Journal of Vocational Behavior*, *75*(1), 56–67.

Smalheiser, N. R., & Torvik, V. I. (2009). Author name disambiguation. *Annual Review of Information Science and Technology*, *43*(1), 1–43.

Stephan, P. E. (2012). *How economics shapes science*. Cambridge, MA: Harvard University Press.

Strotmann, A., Zhao, D., & Bubela, T. (2009). Author name disambiguation for collaboration network analysis and visualization. *Proceedings of the American Society for Information Science and Technology*, *46*(1), 1–20.

Sugimoto, C. R. (2012). Are you my mentor? Identifying mentors and their roles in LIS doctoral education. *Journal of Education for Library and Information Science*, *53*(1), 2–19.

Sugimoto, C. R. (2014). Academic genealogy. In B. Cronin & C. R. Sugimoto (Eds.), *Beyond Bibliometrics: Harnessing multidimensional indicators of scholarly impact* (pp. 365–382). Cambridge, MA: MIT Press.

Tartari, V., Perkmann, M., & Salter, A. (2014). In good company: The influence of peers on industry engagement by academic scientists. *Research Policy*, *43*(7), 1189–1203.

Waldinger, F. (2016). Bombs, brains, and science: The role of human and physical capital for the production of scientific knowledge. *The Review of Economics and Statistics*, *98*(5), 811–831.

Wang, J., Berzins, K., Hicks, D., Melkers, J., Xiao, F., & Pinheiro, D. (2012). A boosted-trees method for name disambiguation. *Scientometrics*, *93*(2), 391–411.

Wang, W., Liu, J., Xia, F., King, I., & Tong, H. (2017). Shifu: Deep learning based advisor-advisee relationship mining in scholarly big data. In *Proceedings of the 26th international conference on world wide web companion* (pp. 303–310).

Witten, I. H., & Frank, E. (2005). *Data mining: Practical machine learning tools and techniques*. San Francisco, CA: Morgan Kaufmann Publishers.