

Predicting Academic Success Using Academic Genealogical Data, a Data Science Approach

Joost E.S. van Weert
STUDENT NUMBER: 2043002

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN DATA SCIENCE SOCIETY
DEPARTMENT OF COGNITIVE SCIENCE & ARTIFICIAL INTELLIGENCE
SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
TILBURG UNIVERSITY

Thesis committee:
dr. Michal Klincewicz
dr. Henry Brighton

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science & Artificial Intelligence
Tilburg, The Netherlands
December 2020

Preface

I would like to thank NeuroTree for granting me access to their data. This brought the basis for this research which I value greatly. Furthermore, I would like to thank my supervisor, Michal Klincewicz for supporting me greatly throughout this project, answering my questions and providing me with tips and useful information. Finally, I would like to thank the creators of the scholarly module, without which the h-index would not be included in this research.

Predicting Academic Success Using Academic Genealogical Data, a Data Science Approach

Joost E.S. van Weert

Many academic genealogy studies include the backtracking of ideologies or scientific concepts. In the age of data however, many more possibilities become apparent for academic genealogical data (information concerning an academics place in the scientific world). In this paper, one central question is examined:

Can academic success be predicted?

To study this question, data from NeuroTree, an academic genealogy focused on neuroscientist, is used. In addition, the h-index is added from google scholar to serve as a measure of success in combination with fecundity. Fecundity is a metric based on successful mentor student relations.

This study uses two main approaches to answer the research question, prediction via regression and via classification. Regression allows variables to be accurately estimated. But due to the messy nature of the data, classification is used in combination with a up and down sampling strategy to balance the dataset for more robust variables.

In conclusion this study found insufficient evidence using regression models for predictive values in academic genealogies for academic success (highest scores, Explained variance 0.085 and MAE 8.66). However, one interesting finding within this study is that there is no evidence supporting a positive relationship between the two success metrics fecundity and h-index.

1. Introduction

Academic success often determines an academic's value in his field and overall professional life. However, we do not have a good and reliable measure of academic success. Given this, research concerning academic success exists across disciplines and multiple distinct academic literatures about it never intersect. Some studies suggest that a successful mentorship relation is itself academic success ([Marsh 2017](#); [Heinisch and Buenstorf 2018](#); [Malmgren, Ottino, and Amaral 2010](#)) while others use funding as academic success ([Rezek, McDonald, and Kallmes 2011](#); [Bol, de Vaan, and van de Rijt 2018](#)), other studies claim an academic's citation or publication count as success ([Li et al. 2019](#); [Fortunato et al. 2018](#); [Acuna, Allesina, and Kording 2012](#)), which involves complicated metrics and questionable metrics, such as the h-index, as operationalisations of academic success ([Schubert and Schubert 2019](#); [Lazaridis 2010](#); [Hirsch 2005](#)). All of these approaches use metrics that focus on a different part of an academics professional life and can conclude different results and thus assign disparate values to individual academics.

Furthermore, whenever a particular metric of success is defined, other questions arise, for example: Who in a heterogenous group of academics is successful? ([Vanclay 2008](#); [Cronin and Meho 2006](#)); Are there correlators with this particular metric? Does

the metric correlate with other metrics of success? (Li et al. 2019; Liénard et al. 2018; Heinisch and Buenstorf 2018; Ayaz, Masood, and Islam 2018), etc. In short, there is an abundance of interest and research concerning academic success, but little focus. This means that many of the conclusions from this research contradict one another. For example, there are studies which conclude that a top performing academic mentor increases an individual's academic success (Li et al. 2019; Liénard et al. 2018), while evidence that shows the exact opposite trend can also be found (Malmgren, Ottino, and Amaral 2010; Heinisch and Buenstorf 2018), or yet other studies which suggest that mentorship does not matter at all (Paglis, Green, and Bauer 2006). In summary, there is a literature about the research question at the center of this thesis, but many individual research gaps all of which demand attention (details about the literature can be found in Section 2). Given the state of the field and the potential impact that it could have on an individual academic's life (through evaluations, institutional review, or policy) more research concerning this subject is needed in general, but also specifically at the level of the most basic questions about what academic success is, how it can be measured, and what the value of mentorship may be. This thesis approaches these questions carefully, but with an eye to this more general problem of a lack of stable definitions, operationalisations, and contradictory results (see Section 1.1 for details about the specific goals of this paper).

We start with the most basic data about academic student-mentor relationships, which can be found in an academic genealogy characterized as a family tree of academics, their communities and their work. Genealogies are a common way to ascertain information hidden in large quantities of sparsely connected data points. Several academic fields have published some form of mentorship history (Kelley and Sussman 2007; Jackson 2007; Chang 2003). This paper focuses on the genealogy and surrounding data from NeuroTree (<https://neurotree.org> 2020). Research concerning genealogies involve relatively basic concepts such as backtracking ideologies (Kelley and Sussman 2007) to more complex subjects like mentorship-derived k-means clustering to distinguish subgroups within a research area (David and Hayden 2020).

Academic fecundity and the h-index are relatively basic measures of individual academic performance. Within this paper, these metrics are used to find possible correlations and characteristics of successful academics, such as the number of students or an academic's place in the scientific community. Academic fecundity is based on the number of successful student relations of an academic (David and Hayden 2020) (formula shown in Section 2.2). It is "one of the most lasting and important contributions a scientist can make", adding to the importance of student guidance (Marsh 2017). The h-index, on the other hand, is a measure of the amount of publications and citations of an individual paper (Hirsch 2005). It is one of the most common metrics to determine the impact of an academic, being used in several academic rankings (Cronin and Meho 2006; Vanclay 2008; Lazaridis 2010), and multiple academic success analyses (Ayaz, Masood, and Islam 2018; Acuna, Allesina, and Kording 2012; Jensen, Rouquier, and Croissant 2009). One study that examined the h-index found more than 6110 citations of the paper introducing the h-index and more than 87 published different variants of the metric (Schubert and Schubert 2019). In summary, academic fecundity and the h-index are proven metrics of academic success, so both will be used with the addition of aforementioned genealogical data for possible correlations. That said, they are far from perfect in giving us a definition of what individual academic success actually is.

1.1 Research questions

The goal of this study is to add to the existing literature concerning academic success overall and possible correlations with academic genealogical data. To improve the current literature, this paper focuses on filling the existing gap concerning academic success, its metrics and possible correlators. This aim can be translated into the following general and guiding research question:

RQ: Can academic success be predicted?

To answer this research question in an organised manner, a number of hypotheses will be tested. These are as follows:

- H1 Academic genealogical data* correlates positively to the success of an academic.
- H2 There is a positive relationship between the number of direct PhD students and subsequent PhD students of students, of an academic and his success.
- H3 Different metrics of academic success have a positive relation towards each other.

* : Data concerning an academic and his place in the scientific community.

These hypotheses are derived from current literature concerning the subject of academic success and details regarding an academics place in their scientific community. In short, the current literature describes conflicting results concerning correlations towards academic success (Li  nard et al. 2018; Malmgren, Ottino, and Amaral 2010; Paglis, Green, and Bauer 2006). This argues for more research on the respective subject and thus a foundation for Hypothesis H1. To go into detail, a more specified hypothesis is formulated as Hypothesis H2. This individualizes the students of an academic which emphasizes the prestige and necessity concerning qualitative mentorship as mentioned in the literature (Marsh 2017; Seibert, Hall, and Kram 1995; Hunt and Michael 1983). Finally, Hypothesis H3 is a consequence of the current dilemma concerning different metrics of academic success and their conflicting results (Chariker et al. 2017; Heinisch and Buenstorf 2018; Waltman and Van Eck 2012). A more detailed argumentation and association with the literature for each of these hypotheses is included in Section 2.

Due to the nature of the data and initial results of the analyses, an opportunity arose for further interpretation and exploration of the data and possible trends or groups concerning academics and their genealogical information. To exploit this opportunity and improve on the conclusions based on testing the hypotheses, a number of unsupervised clustering tasks (CT's) are created. These tasks are detailed below, further details and argumentation is shown in Section 3.2.

- CT1 Clustering based on student count per academic.
- CT2 Clustering based on different metrics of academic success.
- CT3 Clustering based on formal academic relations of an academic.

2. Related Work

This section will focus on previous studies related to the subject of this paper. This will involve studies concerning academic genealogies in general (Section 2.1), fecundity and its function (Section 2.2), other metrics of academic success (Section 2.3), and previous research concerning academic genealogies and academic success (Section 2.4).

2.1 Academic genealogies

The interest in genealogies has only increased in recent years (Heinisch and Buenstorf 2018; Li et al. 2019). Especially because of the growth of more sophisticated databases such as NeuroTree (David and Hayden 2012), and the Mathematical genealogy (Jackson 2007) and the growing possibilities using Data Science methodologies. The most common interest for academic genealogies is using the data to trace ideologies through descendants and finding the root (Kelley and Sussman 2007). There are nonetheless, many more possibilities for the information that genealogies unveil.

An academic genealogy does not differ greatly from a biological family tree. An academic is linked via lines to his or her mentor (parent), linking to the preceding (family of) academics. In addition, they are linked to their own students (children) and thereby the following students of those students (grandchildren), etc. This creates a clear visualisation of the structure that is the academic world in person.

Academic genealogies include information that is hidden in individual papers and other collaborations. This leads to interesting metrics for an academic and his or her field. To quickly touch on such metrics, these include, academic fertility, cousins, descendants and generations (Rossi et al. 2018). Moreover, such relations cause an opportunity for clustering. In (David and Hayden 2012), 60 groups were identified within the field of neuroscience using k-means clustering, identifying and visualising clusters with close relations. Moreover, comparable genealogies are used to determine academic success (Malmgren, Ottino, and Amaral 2010; Heinisch and Buenstorf 2018), and compare multiple success factors with awards won by an academic (Marsh 2017).

Moreover, recent research shows a growing interest in academic genealogies and the possibilities they offer, especially concerning academic success (Malmgren, Ottino, and Amaral 2010; Rossi et al. 2018; Marsh 2017; Heinisch and Buenstorf 2018) (Li et al. 2019) (more detail concerning this subject in Section 2.4). Adding to these recent papers and the growing interest, a relatively new study concluded that female mentors reduce female student success and the gain of new female mentors (AlShebli, Makovi, and Rahwan 2020). Moreover, this paper was later (19-11-2020) retracted by Nature (which published the paper) after serious complaints concerning the methods of the study, indicating more significance for further research concerning the subject at hand.

2.2 Fecundity

By using academic genealogical data, fecundity can be calculated. This is a metric to assess the successful student relations of an academic. This is measured in the amount of students an academic directly mentors, adding the mentorship students of those students, and so on. This can be described in the formula, calculating the fecundity sum (Sugimoto 2014):

$$FecunditySum = n_1 + n_2 + \dots + n_m$$

However, this formula counts the students of students indefinitely and equally across generations. Therefore, this formula will favor historic academics and create a bias causing it to be problematic as a measure of success. Thus, a normalization factor can be introduced for a more balanced comparison (David and Hayden 2012). The same study concluded an initial normalization factor of one, followed by a multiplication of 0.5 with each generation step suffices to mitigate the initial bias. The formula of this met-

ric is shown below. From this point in the paper, fecundity refers to the FecundityScore as stated in the formula. In both formulas, the notations shown in Table 6 apply.

$$FecundityScore = \gamma n_1 + \gamma n_2 + \dots + \gamma n_m$$

Table 1
FecunditySum and FecundityScore formula notation explanation

Notation	Meaning
n_1	Direct mentored students
n_2	Students of n_1
n_m	The number of mentored students stepping down through m successive generations
γ	Normalization factor which scales with successive generations

When examining the FecundityScore function as a metric of academic success, one might conclude an aspect that the function ignores. In theory, academic X can have one student that becomes a high performing mentor, thereby creating a large amount of students himself. In this case, academic X would receive credit for his second generation of students (and the students of those students, etc.), since they are included in the formula. However, academic X might have contributed little to none to ascertain this credit, making academic X a lucky and not a good mentor per se. This can also be the other way around, where academic Y would have a lot of direct students but little to no students of students. This could be seen as a lazy mentor, since he has a lot of students but might have not actually guided or counseled them. To determine the practical side of this theory, a clustering task is formed based on the sum of students per generation (Section 3.2.3).

Within the academic community, the training of new students is of vital importance to the upkeep and advance within a research field. The training of successful students is perceived as one of the most important and lasting contributions (Andraos 2005; Marsh 2017). This conclusion is the cause for the specification of Hypothesis H1 to allow the current study to explicitly address students of an academic and its relation to academic success as stated in Hypothesis H2. Since mentoring is an essential part of the aforementioned training (Campbell and Campbell 1997), this would logically extrapolate to fecundity as a fitting metric of success. Moreover, fecundity has repeatedly been stated as a measure of success in analytical studies (Heinisch and Buenstorf 2018; Malmgren, Ottino, and Amaral 2010), grounding fecundity even further as a widely used metric of success.

2.3 Measures of academic success

Firstly, as stated in Section 2.2, fecundity is a commonly used metric of success. This would coincide with mentorship being a form of prestige besides other benefits for the mentor himself (Hunt and Michael 1983) and (Seibert, Hall, and Kram 1995). Adding

to these benefits, one qualitative study found that mentors perceive building close professional relationships, satisfaction in attributing to personal growth, and the passing of knowledge as reasons to mentor (Allen, Poteet, and Burroughs 1997). Another paper adds that mentorship is a form of upwards striving from the mentor, within a community (Allen et al. 1997). These conclusions emphasise fecundity, not only as a metric for how many students an academic has, but also as a metric of performance, learning on the job and willingness to provide guidance. Adding its frequent occurrence as a metric of success (Heinisch and Buenstorf 2018; Malmgren, Ottino, and Amaral 2010), argues the use of fecundity as a metric of academic success and will therefore be used to (partly) test Hypotheses H1–H3.

Adding to the importance of mentorship as a role in academic success, the sheer output of PhD students has also been used, for example, to determine success of academic faculties (Campbell and Campbell 1997). Some other relatively simplistic measures of success are the summation of citations or publications of an academic (Li et al. 2019; Fortunato et al. 2018; Acuna, Allesina, and Kording 2012). Citation count has, however, received critique for its robustness and vulnerability to extraneous variables (Li et al. 2019). Both citation and publication count serving as success metrics were also recognized to distribute unfair credit over coauthored papers (Hirsch 2007).

Apart from mentorship and producing new academics, success could also be seen as the amount of funding that is trusted upon an academic. One paper concluded that professors of radiology with an h-index under 10 (low success) were "significantly less likely to receive [...] funding" (Rezek, McDonald, and Kallmes 2011), odds ratio 0.07; $p = .0321$. However, this study found no further significant relationships between the h-indices over 10 (more successful academics) and other funding indicators. Another study found an upwards spiral where winners of fundings would earn more funding later in their academic career (Bol, de Vaan, and van de Rijt 2018). Although this does not state a clear relationship or provide evidence for funding as a metric of success, one could argue that funding would be directed to more productive or successful academics. Funding does, however, include considerable extraneous variables including precise allocation of credit(s), field dependencies, and social variables allowing questionable results (Sugimoto and Larivière 2018).

There are more approaches of assessing academic success, in fact there are several metrics with different formulas. One of the most generally and widely used metrics of academic success, is the h-index (Hirsch 2005; Braun, Glänzel, and Schubert 2006; Hirsch 2007; Schubert and Schubert 2019). An example of the extensive use of the h-index is its use in academical rankings which encompasses a wide variant of papers including informational sciences (Cronin and Meho 2006), forestry (Vanclay 2008), and university departments (Lazaridis 2010). The h-index focuses on the papers published by an author and their impact, measured in cite count. For example, an h-index of 20 indicates at least 20 papers that were individually cited at least 20 times. This creates a more robust combination of citations and publications of an academic. To put the h-index in perspective, Einstein, Darwin and Feynman had a respective h-index of 96, 63 and 53. Furthermore, an academic with an h-index of 12 could qualify him for tenure at a major university (Acuna, Allesina, and Kording 2012). While this metric is widely used, there has also been proportional critique. The single use of the h-index as a metric for academic success appeared to be inadequate and inaccurate in a multitude of studies (Costas and Bordons 2007; Waltman and Van Eck 2012; Cerchiello and Giudici 2014). In an attempt to mitigate the difficulties surrounding the h-index, researchers added to this metric of success. One paper, for example, looks at an addition of a time element where only the last x amount of years are counted (Egghe 2010). This is only the tip

of the figurative iceberg, since another paper found more than 87 different published adaptations of the h-index, along with 6110 citings of Hirsch's paper (which introduces the h-index) (Schubert and Schubert 2019). Although not perfect, the wide application and the extensive use of the h-index, validates the use of this variable as a metric of success and therefore, will be used to (partly) answer Hypotheses H1–H3.

2.4 Academic genealogy and success analyses

As previously stated, fecundity and mentor analysis based on genealogies is not a new concept. Several fields of science have some published form of mentorship history, including computer science (Chang 2003), mathematics (Jackson 2007), and primatology (Kelley and Sussman 2007). By analysing such genealogies, some conclusions have been drawn.

Focusing on academic success studies with genealogical data, interesting relationships between academics' positions in the scientific community and metrics of their success were found. One conclusion entails that if an academic is in the top 10% 'core' of his 'coauthorship network', his paper will be in the top 10% most cited papers after five years of publication (Sarigöl et al. 2014). The same paper stated that these results might, however, be a result of a confounding variable, influencing both metrics. This could, for example, be the 'Matthew effect' or 'preferential attachment' which is commonly referred to when researching academic genealogies and academic growth. It states that the rich (a lot of connections) get richer (more connections and citations) (Perc 2014; Jeong, Nédá, and Barabási 2003; Newman 2001). Expanding on this concept, one recent paper found a positive relation between having a top scientist (being in the top 5% of cited authors in that field of research in the year of mentoring) as a mentor and the students' prestige measured in published papers in a set of academic journals (Li et al. 2019). This coincides with other studies claiming a positive relationship between mentor and student performance (Liénard et al. 2018; Crosta and Packman 2005; Andraos 2005) and (Chariker et al. 2017). However, as stated in papers by (Paglis, Green, and Bauer 2006) and (Li et al. 2019), these results do not account for the selection process that mentors have for their students.

One paper by (Malmgren, Ottino, and Amaral 2010) that did take this selection process into account, focused on a mathematics genealogy. This study found that mentors with small (less than 3 students) fecundity (success) train students that have a 37% larger fecundity, that mentors train students with 31% smaller fecundity in the last third of their career and train students with 29% larger fecundity in the first third of their career. Another study created a database using machine learning, including academics in applied physics and electrical engineering from German universities. This study found that mentors with a large amount of previous students and mentors later in their career, both produce students with a lower probability to become advisors ($p < 0.01$) (Heinisch and Buenstorf 2018). This complies with the findings from the previously mentioned paper, adding to the contradiction towards the papers mentioned before. Furthermore, this study only found little evidence to confirm their hypotheses concerning the positive correlation between both research output and connections of the mentor, and the probability of his students becoming an advisor (thus becoming a successful mentor, student relationship). Moreover, a longitudinal study who followed PhD students found no significant contribution of mentorship towards student productivity or commitment to a research career (Paglis, Green, and Bauer 2006), adding to the clashes of findings in the literature. In total, this field of research is recognised as under-researched (Heinisch and Buenstorf 2018) and multiple studies seem to have contradicting observations. The find-

ings stated within this section could mean significant changes in how a good academic is determined and how good mentorship is viewed. Adding the statement that interest towards these studies has grown (Li et al. 2019), formulates a decent argument for this study to focus on examining possible correlators to success, Hypothesis H1, relations between metrics of success Hypothesis H3, and in total the main Research Question QR.

The combination of papers as depicted in this section, raise additional questions concerning the different metrics of academic success and their relations. A number of studies researched these relations. One study included 3085 neuroscientists to predict the h-index achieved over time and yielded $R^2=0.92$ ($R^2=1$ being a perfect predictor) for predicting one year in the future, $R^2=0.67$ for five years, and $R^2=0.48$ for ten years, all with cross-validation. The variables that were included are (with direction of correlation), number of articles written (positive), current h-index (positive), years since publishing first article (negative), number of distinct journals published in (positive), and number of articles in prespecified respectable scientific journals (positive) (Acuna, Allesina, and Kording 2012). In all models, the h-index has a significant (if not overshadowing) importance as predictor. Which coincides with another paper (Ayaz, Masood, and Islam 2018) and a study by the person that introduced the h-index. That study concluded that a researcher with a high h-index 12 years after his first publication is highly likely to have a high h-index after 24 years with a predictive power of $r=0.91$ (Hirsch 2007). Adding to the predictive analyses concerning the h-index, one paper examined bibliometric indicators to predict promotion for all academic fields of study. It concluded no single indicator as best predictor over all disciplines, the h-index however, provided the "least bad" results with 48% accuracy compared to a 30% chance baseline (Jensen, Rouquier, and Croissant 2009). In total, as shown by papers included in this section, the main result that keeps appearing is that h-index predicts its future self. Other relationships for metrics of academic success are still unclear. One study concluded that fecundity is "strongly correlated" to publication count, but argues for "substantial extra effort" on the subject (Malmgren, Ottino, and Amaral 2010). This enforces the argument for this paper to focus on analyzing correlations between the h-index and fecundity (Hypothesis H3), and the h-index and genealogical data (Hypothesis H1).

3. Experimental Setup

3.1 Data

The data used in this study is obtained from NeuroTree (<https://neurotree.org> 2020), which is an academic genealogy which mainly focuses on neuroscientists. This includes a data dump from their database as of 17-06-2020. Within this dump, three .tsv files are included. These files encompass the details of each individual (people table), all connections between the individuals (connect table), and all the locations, defined by the university of the individual (locations table). This includes all data NeuroTree uses for their website. Moreover, data from Google Scholar (<https://scholar.google.com/> 2020) is used to add a more credible and consistent source for the h-index. This data is attained via a self made program that that allows communication with Google Scholar, this procedure is described in Section 3.1.2. More details about the NeuroTree data will be discussed in the following section.

3.1.1 Raw Data. The three datasets 'people', 'connect', and 'locations' each contain a different detailed part of the information within Neurotree. The structures of these sets

are as follows. The 'people' dataset contains 16 columns and 758599 rows. Each row represents one person, identified by 'pid' (person id), which is their unique value. The columns, or variables, are explained in Appendix A. The "connect" dataset contains 11 columns and 1501062 rows. Each row represents one connection between a mentor and student, each connection is identified by the unique value 'cid' (connection id). More details concerning this dataset are displayed in Appendix B. Finally, the "locations" dataset contains 11 columns and 31941 rows. Each row represents one location identified with the unique 'locid' (location id) value. The columns contain the variables shown in Appendix C. The different tables are all connected as is visualised in Appendix D.

This study focuses on the individuals and their success. Therefore, the 'people' set is used as a basis and the other tables are converted via the key values to fit into the 'people' set. Furthermore, since fecundity is not included in the data dump, the 'people' and 'connect' datasets are used to calculate this metric. To calculate fecundity without a bias towards historic researchers, the fecundity score is used as is stated in Section 2.2. In practice, the unique value 'pid' is used to search academics in the 'connect' dataset in the mentor column ('pid2') when the 'relation' column states a mentor relation type. The outputted student(s) ('pid1') are used to calculate the students in the next generation, and so on. This process creates the subsequent student sums per generation, which are also separately included into the data ('fec1', 'fec2', 'fec3', 'fec4', 'fec5') as can be seen in Appendix E. For each generation the normalization factor is multiplied by 0.5 before summation as is recommended in the literature (David and Hayden 2012). Due to computational limitations, this process is repeated till the 5th generation of students. In addition, fecundity related variables are also calculated and implemented into the dataset, including the summation of different relation types (see Appendix B) per academic. Using the data as stipulated within this section, one dataframe including all information is created containing variables as stated in Appendix E.

3.1.2 Procedure H-index Data Collection. The h-index is included in the NeuroTree data dump. However, the h-index variable consists for 97.06% of missing data. Therefore, Google Scholar (<https://scholar.google.com/> 2020) is used as a separate source for the h-index. Since no proper Google Scholar API exists, a custom program is created to ascertain the h-index on a large scale.

Before data can be collected, a variable that can serve as identifier needs to be established. The unique academic identifier ORCID (<https://orcid.org/> 2020) is included in the NeuroTree data dump. This variable, however, contains 99.81% missing data and therefore has little potential as a basis for further data gathering. Without the ORCID there is no consistently included unique identifier for the academics as to be able to search them on Google Scholar. Therefore, the data is accumulated by inputting the full name of an academic into the custom program.

After inputting the full name of an academic, the program activates the scholarly module (<https://pypi.org/project/scholarly/> 2020), allowing a connection with Google Scholar. The custom program then imports all data from the specified academic. While importing the data, a custom search log is created to track the searches and record metadata concerning the imports. By using this search log, unique results are filtered and non-unique instances are deleted, counteracting the non-unique nature of the search input variable (full name of an academic). From the unique instances, the h-index is gathered via a text search, which are added to the full dataset as stated in Section 3.1.1.

The custom program is not recognized by Google Scholar as a proper API. This causes the program to receive CAPTCHA errors, which aim to prevent such large scale

requests that intent to disturb Google Scholar's functionality. These errors eventually forced the search to stop. To resolve this issue, a workaround is included in the program. This recognizes the occurrence of such errors via the custom search log and adapts the request rate accordingly, which allowed the implementation of the h-index into the data to continue.

3.1.3 Data cleaning. The full dataset, as stated in Section 3.1.1 is used as the raw data for this study. This data is largely collected from NeuroTree (<https://neurotree.org> 2020). NeuroTree is a crowd sourced website, meaning "any internet user can add information about researchers and the connections between them" (David and Hayden 2012). Since there are no limitations to the input fields, variables can take an almost limitless amount of different values (example in Appendix F), and there are large portions not filled in. Therefore, all missing values are first summed and divided by the amount of instances (rows) to get insights into the percentage of missing data, these results are shown in Appendix G. As is visualised, the columns 'award', 'hindex', 'orcid_id', 's2id', and 'homepage' all have missing value percentages higher than 90%. These columns are therefore dropped from the dataset. As previously stated in Section 3.1.2, the h-index is later replaced by data retrieved from google scholar via a self made program.

From the remainder of the data, there are five categorical variables which must first be interpreted for overarching definitions and then clustered into a limited number of bins. This concerns the following variables, 'degrees', 'location', 'country', 'area', and 'majorarea'. The data distributions before binning are shown in Appendix H. This figure shows that there is a large amount of unique values that occur variety of times. In total, the variables 'degrees', 'location', 'country', 'area', and 'majorarea' have 2495, 23808, 121, 167786, and 1719 unique values respectively.

The 'degrees' column is defined by the degree(s) an academic has, it contains the abbreviation of the diploma(s). This variable mainly contains PhD values, therefore this is contained in its own bin. Furthermore, due to frequent occurrences in the data, doctorates in psychology and doctorates in education are also filtered in their own bins. Furthermore, due to estimated relevance for analyses, degrees higher than a PhD, were assigned into a specific bin. Finally, all other specific doctorates are assigned into the 'misc_doc' bin. The resulting bins are shown in Appendix I.

The 'location' column contains all universities, as stated within NeuroTree. This variable contains the specific name of a university containing, 23808 unique values (23808). Since these are so specifically named, the upper 90% of occurrences are inspected to ascertain a lower number of unique values. The result is 1470 unique values which could not be logically binned, pressuring this variable to be dropped from the dataframe.

The 'country' column identifies the country of the university of an academic. This variable has 65.11% missing data. Because the missing data is a significant amount of the overall data, this variable is dropped from the dataframe.

The 'area' column involves the research area(s) of an academic, as is reported by NeuroTree. Within this variable an academic can enter a multitude of areas of research. Therefore, the binning structure accounts for academics to be sorted into multiple bins. The bins that are used within this study express the following research areas electronics, mathematics, engineering, computer science, psychology, biology and business. All bins are selected to encompass as much instances as possible, without losing implication.

The 'majorarea' column describes an academics place between genealogical trees. This stems from NeuroTree and their aim to cluster academics (David and Hayden 2020). The unique values mainly stem from the combination of multiple groups. Consequently,

the binning structure aims to even out its different samples without losing information and accounts for academics to be sorted into multiple bins. The bins that are used for this variable are as follows, neuroscience, physics, academical tree research, biology and chemistry, sociology and education, economy and business, and a misc bin covering all other options (the distribution is shown in Appendix I).

The steps as discussed previously in this section, totals the data cleaning for this study. The resulting distributions of the categorical variables are shown in Appendix L and of the continuous variables are shown in Appendix M. The resulting dataset is referred to as the 'full dataset'. This includes the fecundity and h-index variables ('fec_tot', 'hindex') and all other variables, as previously stated in this section which entail:

- The academic degrees of an individual, binned
- Research area of an academic, binned
- Area within scientific community, binned
- Sum of direct students of the academic
- Sum of students of first generation of students
- Sum of students of second generation of students
- Sum of students of third generation of students
- Sum of students of fourth generation of students
- Sum of undergraduate mentorship relations (mentorship role)
- Sum of grad student mentorship relations (mentorship role)
- Sum of Postdoc student mentorship relations (mentorship role)
- Sum of research scientist mentorship relations (mentorship role)
- Sum of collaboration relations (not a mentorship role)

These listed variables form all genealogical variables that are used in this study, this is referred to as 'All genealogical data'. This forms the basis to answer Hypotheses H1-H3, with different metrics of success (fecundity and h-index). To properly answer these questions, however, different variables and instances need to be treated, included or excluded. These specific approaches are formed as stated in the following section (Section 3.1.4).

3.1.4 Preprocessing. Starting with Hypothesis H1 with the h-index variable as a metric of success. The data for this hypothesis requires all genealogical data and the h-index variable. The h-index is extracted from Google Scholar (as mentioned in Section 3.1.2). Due to the lack of a unique identifier, some instances were forced to be deleted. This caused the h-index data to have a different shape than the data from NeuroTree. Thus the corresponding instances in the 'All genealogical data' dataset are deleted to fit on all instances where h-index contains a value as imputed from Google Scholar. The result of this adaptation is shown in Appendix O. The results based on this data will also serve as an answer for Hypothesis H2 (with h-index as a measure of success), since the sum of students per generation is included in this data.

Continuing with Hypothesis H1 with the fecundity variable as a metric of success. The task of answering this hypothesis involves correlating the full dataset with the target value fecundity. To avoid the loss of any information, only the student sums per generation ('fec1', 'fec2', 'fec3', 'fec4', 'fec5') are excluded from all genealogical data, since these are included in the formula to create fecundity. Therefore, the distributions remain unchanged (see Appendices M and L).

To enable the answering of Hypothesis H2, using fecundity as a metric of success, one aspect of the data must be taken into account. When an academic has no students, he can not have students of further generations and thus has a fecundity of zero (Section 2.2. Since the data for this task consists entirely of student sums per generation (fec1 - fec5 in Appendix E) and fecundity, all included variables would default to zero. Thus, an argument can be made to exclude all instances where fecundity is zero. Therefore the results concerning this task will be reported both with instances where fecundity is zero, and without these instances (see Appendices J K).

Hypothesis H3 requires both the success metric variables, fecundity and h-index. As stated in Section 3.1.2, the h-index variable is retrieved from google scholar via a self made program. This does not encompass all instances involved in the full data as described in this section. Therefore the instances are adapted to properly fit the h-index data. This results in the data distributions as shown in Appendix N.

In the raw dataset fecundity is relatively skewed towards the zero, as is visualized in Appendix O. A too large inbalance might prove problematic for the results of analyses. To counteract this inbalance, up and down-sampling could be a viable option. This entails analyzing the distribution of the data and correcting its skewness by excluding instances with frequently occurring values and extrapolating the instances with less common values. To enable proper up and down-sampling, the data must be adapted. Since this method works relatively poorly on continuous data, the fecundity variable will be binned, creating a categorical variable instead.

Therefore, the tasks including fecundity will also be analyzed with categorical prediction models. To enable classification, the fecundity variable is split into seven different classes, with the following boundaries, 0-2, 2-4, 4-8, 8-16, 16-32, 32-64 and 64 and up. These bins are chosen according to the distribution of the fecundity data. To adequately perform this categorical prediction task, categorical models will be included to analyze the data as previously specified and an up and down-sampled version of it. Through up and down-sampling, these models might have increased performance over the regression models.

After the data is adapted for the specific tasks, all datasets for regression and classification follow another shared pipeline. This includes the normalization of each of its input variables, using sklearn. The function uses a simple formula as shown below and explained in Table 2. Normalization is a common technique that allows unbalanced variables to perform better when using predictive models (Pedregosa et al. 2011).

$$z = (x - u)/s$$

Table 2
Standardize formula notation explanation

Notation	Meaning
z	Standardized feature
x	Input sample
u	Mean of sample
s	Standard deviation of training samples

The clustering tasks as specified in Section 1.1, can be operationalized at this point in the paper. The specification and the data used concerning each of the clustering tasks are specified as follows:

- CT1 Clustering based on sum of students per generation. This uses the same data as specified for answering Hypothesis H2 with fecundity
- CT1 Clustering based on both metrics of academic success. This uses the same data as previously specified to answer Hypothesis H3
- CT1 Clustering based on sum of academic relationship per relationship type. This uses the 'All genealogical data' dataset, filtered to only contain the variables concerning academic relationship types ('relation_0', 'relation_1', 'relation_2', 'relation_3', 'relation_4', in Appendix E).

For each task, the respective data is collected before normalization as previously described for the regression and classification tasks. The selected clustering models (Section 3.2.3) use distance in multidimensional space to split the clusters. Therefore, outliers and different variable ranges could decrease performance of the models. To limit these factors, the following measures are performed.

Each dataset is first run through an outlier detection analysis. Since no theoretical distribution such as Gaussian applies, the outliers are interpreted per variable by examining what logical values are in the distribution of the selected variable. After checking the distributions via histograms and other meta data information (for example, from scatter plots), values who are three standard deviations greater than the mean are deleted to handle extreme values within variables. Although it is recognized that no Gaussian distribution is present, the resulting data has less extreme values without losing sensible data according to the respective distributions. This data is normalized using a MinMax scaler. This forces all variables in a range between zero and one (while not assuming a specific distribution), thereby equalizing the distance scales between the variables and thus allowing multidimensional space to be more equally divided. Due to computational limitations, a sample of 10,000 instances is randomly selected from all datasets. To achieve comparable results, a random state is selected, causing the same sample to be selected every time.

3.2 Methods & Models

As is mentioned in Section 3.1.3, most of the tasks defined in this paper are approached by both regression models and classification models. Within this section, the methods and models for both types are described in a corresponding section (Sections 3.2.1 and 3.2.2). Before this split is made, a number of shared methods and models are stated.

After standardizing the data, the test set is extracted from the original dataset. This will be used to test the performance of the predictive models. This is followed by a Kfold train validation split. This splits the data into 'K' amount of smaller sets called folds. A model can then be trained using 'K-1' of the folds as training data, followed by a validation on the remaining fold. This process is reiterated 'K' amount of times, each time holding out a different part for validation. All performance metrics are consequently averaged. This process is included for all regression and classification tasks. In total, this method allows the data to be used more efficiently and the validation metrics to become more robust (Pedregosa et al. 2011). Due to a large dataset in some instances, limited processing power and comparability between tasks, a Kfold with five

splits is selected for all regression and classification tasks. To add to the comparability, a set random state is used to retrieve repeatable and comparable results.

Furthermore, during this study, the importance of clustering arose. The nature of the data and initial results of the prediction models left an opportunity for further interpretation and exploration of the data and possible trends or groups concerning the academics and their information. Therefore, clustering models are also used within this study, the respective methods and models are described in Section 3.2.3. By creating specific clustering tasks, a deeper understanding of the genealogical data and the metrics of academic success might be gained. These clustering tasks are depicted in Section 1.1 and specified according to the current data in Section 3.1.4 as CT1, CT2, and CT3. Since these tasks are a consequence of early testing and model performance, a concise overview of the argumentation for each task is depicted below:

- CT1 For a deeper understanding concerning the formation of fecundity (Section 2.2) and initial analytical results (Section 4.1.3)
- CT2 To create further insights into the distribution of both metrics of success and possible relationships (results leading to this argumentation are detailed in Section 4.1.4)
- CT3 To extend on possible relations and groups between the sum of different academical relations per relationship type and metrics of academic success (more detailed explanation in Section 4.1.1)

Finally, the libraries shown in Table 3 are used throughout the study. More precisely, all models described within this section are implemented with sklearn and all the unspecified hyperparameters are implemented according to the default as stated sklearn (Pedregosa et al. 2011). Furthermore, the included neural networks are created using tensorflow.

Table 3

Used libraries, their version and their purpose within this study

Library/language	Version	Purpose
python	3.8.6	Programming language
pandas	1.1.3	Data handling
numpy	1.18.5	Data handling
sklearn	0.23.2	Analysis and Preprocessing
tensorflow	2.3.1	Creating neural networks
matplotlib	3.3.2	Visualisation
seaborn	0.11.0	Visualisation
scholarly	v1.0b1	Google Scholar data import

3.2.1 Regression models and parameters. For the continuous regression tasks, a multitude of options for models are available. Considering that this research area is relatively under-researched, easily interpretable models are overall preferred. Moreover, the data that is used in this study has the following key characteristics that need to be accounted for in model selection:

- Unbalanced variables
- Binned categorical values and continuous variables
- No Gaussian or other theoretical distribution
- Relatively large dataset in both rows and columns

These characteristics can be transformed to the following requirements for models:

- Versatility
- Does not assume any prespecified distribution
- Not compute heavy
- Works well with multidimensionality

For model selection, the sklearn website (Pedregosa et al. 2011) is consulted. Bayesian Regression, and Nearest Neighbor Regression were considered as alternatives, but failed to meet the requirements as previously stated. A Support Vector Regression was used in initial testing, but was excluded from the selection due to the large amount of required computations. In Table 4 an overview of the models that are used, is shown with reasoning for each of them.

Table 4
Model options for regression tasks

Model name	Upsides	Downsides
SGDRegressor	Efficiency, Versatile	Sensitive to feature scaling
Lasso	Tendency for more Zero coefficients	Data handling
Decision forest	More robust version of Decision tree	Compute heavy

After defining the options for regression models, the SGDRegressor, Decision forest and Lasso are picked as best suited models for this study due to their upsides and compliance with the aforementioned model requirements. Moreover, the Lasso model is included because it tends towards many zero valued coefficients, therefore effectively reducing the number of variables included in the model and highlighting the most important ones. Since this study aims to find relationships between many variables, this model is well suited and could conclude high contributing variables.

Adding to the use of the Lasso model, the regression task concerning Hypothesis H2 with fecundity as a metric of success, is only approached with this model. As is defined in Section 3.1.4, only the sums of students per generation is used as input for this task. As stated in Section 2.2, these variables are all included in the function to calculate fecundity. Therefore, logic would dictate that that these are correlated and thus prediction models will not retrieve useful information. The Lasso model, however, could still imply useful information since it shows the influence of the best performing input variables.

The SGDRegressor is implemented using the ‘squared-loss’ loss function, ‘l2’ penalty and an alpha of 0.0001. The hyperparameters for the Decision forest include `n_estimators = 100` (number of trees), MSE as the criterion and a maximum depth of 2. Finally, the Lasso hyperparameters are kept at default.

All regression tasks are analyzed by running these models except the task concerning Hypothesis H2 with fecundity as a metric of success. For this task, only the

Lasso regression will be used because the interpretation of its coefficients define which variable(s) are the best correlators.

Due to initial results of these models, a consequential neural network (NN) is added to the regression models used in this study. Although a neural network is not easy to interpret, early results show a possible need for a more complex model. The results accumulated by a neural network will not show a clear correlation between specific variables. The results could, however, show a more complex relationship concerning the included variables.

The neural networks that are used within this study are created using the tensorflow module. Tuning the models followed a random search approach for the optimal hyperparameters. This includes picking a random value from a list of pre-selected hyperparameter settings which are shown in Table 3.2.1. The epochs, number of hidden layers, and number of neurons per hidden layer were all capped due to initial results showing a tendency for overfitting when increasing complexity. The optimizers and loss functions are selected based on the same criteria as the other regression models. Initially, dropout, batch normalization and Principal Component Analysis (where multiple variables were inputted) were tested but were later excluded from the model due to a drop in model performance. Early stopping based on validation MAE (Mean Absolute Error) was included, but had no impact on the performed number of epochs. The hyperparameter settings of the best performing neural networks are shown in Appendix P.

Table 5
Features used for random search

Hyperparameter		Values
Number of epochs		2, 4, 8, 16
Hidden layer activation	Elu, relu, sigmoid, softplus, linear	
Optimizer	Adagrad, adam, SGD, rmsprop	
Loss function	mae, mse, CosineSimilarity	
Number of hidden layers		2, 3, 4, 8, 14
Hidden layer neuron count		16, 32, 64

To determine the success of the regression models, the following performance metrics are included. Mean Squared Error (MSE), Mean Absolute Error (MAE), Explained Variance (Expl. Var), and R^2 score (R^2). Since the MSE punishes outliers heavier than comparable metrics, it will be useful to evaluate the error of a model. However, as stated in Section 3.1.4 and Appendix M, the data used in this study is relatively unbalanced. The MAE will be the main performance metric to calculate the error, since outliers could cause the MSE to overreact. To estimate a proper fit of a model, Explained Variance and R^2 are used. Due to a lack of research concerning the subject of this paper, no baseline can be formed. Therefore, interpretation of the fit of the models is based on the ability to explain more than a mean estimation for each instance. The formulas for all performance metrics are shown below.

$$MSE = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |Y_i - \hat{Y}|$$

$$Expl.Var = 1 - \frac{Var(y - \hat{y})}{Var(y)}$$

$$R^2 = 1 - \frac{\sum (Y_i - \hat{Y})^2}{\sum (Y - \bar{Y}_i)^2}$$

Table 6

FecunditySum and FecundityScore formula notation explanation

Notation	Meaning
n	sample size
Y	True value
\hat{Y}	predicted value
\bar{Y}	Mean value

3.2.2 Classification models and parameters. As argued in Section 3.1.4, all tasks concerning fecundity as a metric of success are additionally approached via classification techniques. As stated in Section 3.2.1, when testing Hypothesis H2 with fecundity as metric of success, prediction will not yield useful information. This is due to the correlation of the included input and output variables because of the nature of the formula for fecundity (Section 2.2).

The classification tasks in this paper try to correctly predict the binned version of fecundity (as stated in Section 3.1.3). Due to the multiple output variables (separate fecundity bins), these tasks entail a multidimensional output (the fecundity bins). Furthermore, the bins created from the fecundity variable are generated while accounting for the ordered nature, allowing one class to be defined as ‘larger’ or ‘higher’ than its predecessor.

The models that are used for the classification are selected on the same bases as the models for the continuous tasks since the characteristics of the data remain unchanged. Therefore, the SGDClassifier is used, due to its versatile nature and a good fit for larger datasets. Furthermore, Random Forest Classification and Support Vector Classification are selected since both are versatile and have advantages in high dimensional spaces. Moreover, all classifiers can be applied to multilabel classification which is a requisite, as previously stated.

The performance metrics that are selected include Accuracy, precision, recall, and f1-score (formulas shown below). Since each metrics covers a different component of the performance, all are included to allow a detailed interpretation of the results. The f1-score, however, returns a better estimation for uneven class distributions and is

therefore the main performance metrics for the classification models. All these metrics are implemented according to the multidimensional nature of the tasks by allowing the metrics to calculate a 'micro' average over the different classes. This causes the metrics to calculate the metrics globally and count the total true positives, true negatives, false negatives and false positives. To better interpret the performance of the models, a random chance baseline is set (accuracy=0.143).

$$Accuracy = \frac{TruePositives + TrueNegatives}{AllSamples}$$

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives}$$

$$Recall = \frac{TruePositives}{TruePositives + FalseNegatives}$$

$$F_1score = \frac{Precision * Recall}{Precision + Recall}$$

3.2.3 Clustering models and parameters. The clustering tasks included in this paper contain no 'true' clusters. Therefore, the clusters need to be approximated via clustering models. Such a task is identified as an unsupervised learning task. In terms of model selection, this means that the models need to define the optimal number of clusters themselves. Adding the same model requirements as stated in Section 3.2.1, sets the total clustering model requirements for this study. When consulting the sklearn website (Pedregosa et al. 2011), a few models fit the requirements. These are Kmeans, Mean shift (Mshift), and Affinity propagation. Although Affinity propagation was initially selected, early testing concluded the model being to compute heavy, which resulted in the model being excluded from the study. The other models are explained in detail in this section.

With Kmeans, the number of clusters needs to be identified. By adding another method, the 'elbow method', the number of optimal clusters can be approximated. This uses the Kmeans model and the within cluster sum of squares to calculate the best performing number of clusters. It is called the 'elbow method' because one needs to determine where the respective plot goes from a seemingly exponential function to a linear one, creating a kink or elbow in the plot (an example is visible in Section ??). By adding this method, Kmeans clustering suffices to the set model requirements, being a flexible model without distribution assumptions or a need for much computing power. One downside of Kmeans, is the initialization trap, which specifies possible complications due to a wrong starting point. Therefore, the 'Kmeans++' value is inputted for the initialization method hyperparameter which counteracts the initialization trap.

The Mshift model can determine the amount of clusters. However, it does need a set bandwidth of the clusters, which is hard to define since no information regarding the clusters is known beforehand. By using the 'estimate_bandwidth' setting for this hyperparameter, a separate function will calculate this unknown variable, circumventing the need for a set bandwidth (Pedregosa et al. 2011). Adding the advantages of this model

concerning its performance with uneven cluster sizes and non-flat geometry, completes the argumentation for this model.

The performance metrics for the clustering models are selected on the same basis as the clustering models. These requirements led to the following selection, Silhouette Coefficient, Calinski-Harabasz Index, and Davies-Bouldin Index. The first two both represent within cluster distance versus between cluster distance. The Silhouette Coefficient provides better comparison between different datasets since the output range is set between -1 for incorrect and +1 for highly dense clustering. The Calinski-Harabasz Index is mainly used during model and hyperparameter selection due to the little computational load. Higher values for this index indicate more optimal clustering. The Davies-Bouldin Index signifies the average similarity between clusters by comparing the distance between clusters and size of the clusters. Lower values for this index indicated better performance.

4. Results

The results formed in this study can be subset into multiple subsections. This paper first divides on type of prediction task, including regression, and classification. Secondly a split is made between the different hypotheses as stated in Section 1.1. Finally, Hypotheses H1 and H2 are answered using both metrics of academic success, fecundity and h-index.

4.1 Regression

Within this section, results concerning regression tasks are addressed. The results are accomplished and determined by the regression models and performance metrics as stated in Section 3.2.1.

4.1.1 All genealogy data versus fecundity (as academic success) H1. As shown in Table 7, the Explained Variance and R^2 show no significant predictive power for any of the included models. Concluding, the regression models based on the 'All genealogical data' dataset show no clear correlations between the variables in this dataset and the fecundity variable (best performing model, Random Forest Regression, $R^2=0.079$).

Furthermore, by fitting the Lasso model on the data, a select number of input variables are shown. This result shows the variables that account for the most contribution towards a higher explained variance. The non-zero coefficients are shown in Figure 1. By examining this graph, one can see that only the sum of the different types of relations are included as non-zero coefficients. Moreover, relation_4 appears the most influential, despite being the only relation type that does not depict a mentorship role. These non-zero coefficients argue for a deeper understanding of these variables, thus arguing for the creation of the Clustering Task CT3, answered in Section 4.3.3.

4.1.2 All genealogical data Versus H-index (as academic success), H1. By observing Table 8, the Explained Variance and R^2 show near zero results. This shows little to no predictive power when calculating a best fit line based on the data. The best performing model is Random Forest Regression ($R^2=0.085$). Adding to these results, Figure 2 shows the non-zero coefficients as determined by the Lasso model.

4.1.3 Sums of students per generation versus fecundity (as academic success), H2. As is stated in Section 3.1.4, this task is performed on one dataset where instances

Table 7
Model metrics for All genealogical data Versus Fecundity

Metric	SGD	RandForReg	Lass	NeuralNet
Test MAE	14.77	8.664	12.826	5.36
Test MSE	21196.32	17390.08	18336.37	21087.48
Test Expl. Var	-0.122	0.079	0.03	0.000
Test R ²	-0.123	0.079	0.028	-0.001

Figure 1
Non-zero Lasso coefficients for All genealogical data versus Fecundity

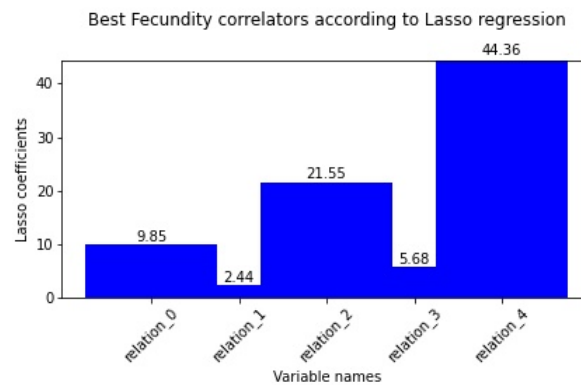


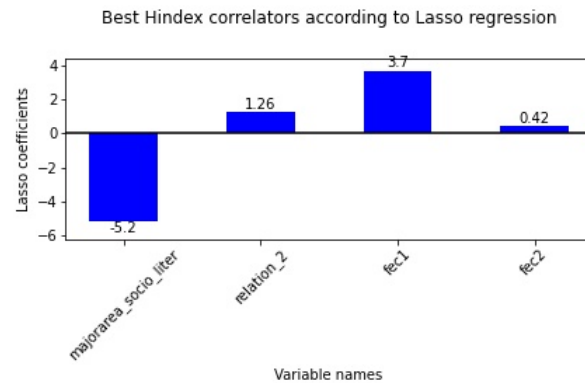
Table 8
Model metrics for All genealogical Data Versus H-index

Metrics	SGDReg	RandForReg	Lass	NeuralNet
Test MAE	20.166	20.199	20.611	32.915
Test MSE	740.306	737.323	751.908	1856.544
Test Expl. Var	0.081	0.085	0.067	0.002
Test R ²	0.081	0.085	0.067	-1.395

containing fecundity zero values are included and one where they are excluded. The sums of students (fecundity sum) per generation are the sole input for the fecundity variable (See FecundityScore formula in Section 2.2). Therefore, these variables have a logical correlation, which the Lasso performance confirms for both datasets ($R^2=1.000$), as visible in Table 9. From the non-zero Lasso coefficients as shown in Figure 4.1.3, it appears that the coefficients for the sums of students per generation increase with each generation. To gain more insights into these results and the relation between the sums of students per generation and fecundity, Classification Task CT1 is created. The results of this task are shown in Section 4.3.1.

Figure 2

Non-zero Lasso coefficients for All genealogical data versus H-index

**Table 9**

Lasso coefficients of 'Students of students' per generation for Fecundity. Without fecundity zero values (left) and with fecundity zero values (right)

Metric	Without zeros	With zeros
MAE	0.543	0.397
MSE	1.682	2.686
Var	1.000	1.000
R ²	1.000	1.000

4.1.4 H-index versus Fecundity, H3. The performance metrics visualized in Table 10, show both an Explained Variance and an R^2 of near zero or less with all models. The best performing model is SGDRegression, slightly outperforming the other models on Explained variance (Explained variance= 0.002, $R^2=0.000$) To visualize the result, a scatter plot is included into this section (Figure 7. This shows the data as distributed across both variables with the best performing regression model. Due to a lack of significant predictive power of any of the models, Clustering Task CT2 is established which might gain further insights, the respective results are in Section 4.3.2.

Table 10

Model metrics for H-index Versus Fecundity

Metrics	SGDReg	RandForReg	Lass	NeuralNet
Test MAE	22.998	23.007	22.991	20.704
Test MSE	936.452	936.81	938.186	891.169
Test Expl. Var	0.002	0.001	0.000	-0.002
Test R ²	0.000	0.000	0.000	-0.122

Figure 3
Lasso coefficients of Students of students per generation for Fecundity. With fecundity zero values (left) and without fecundity zero values (right)

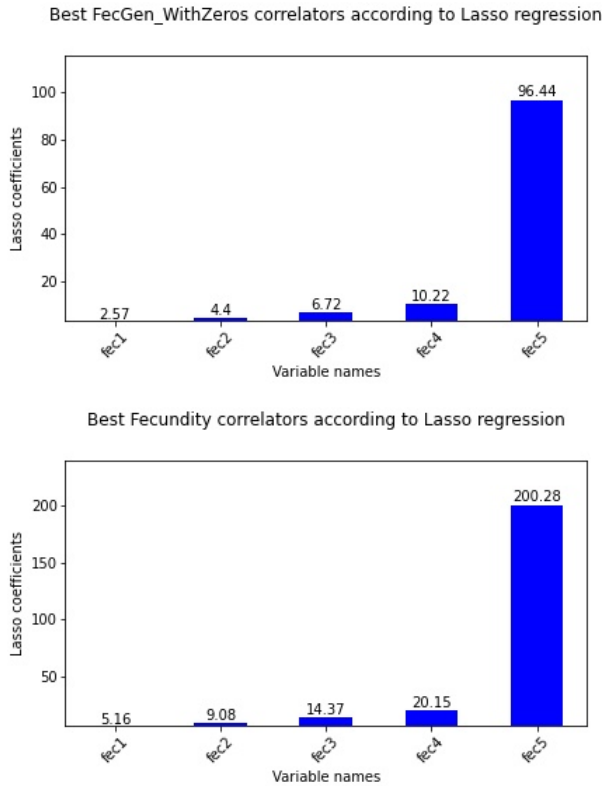
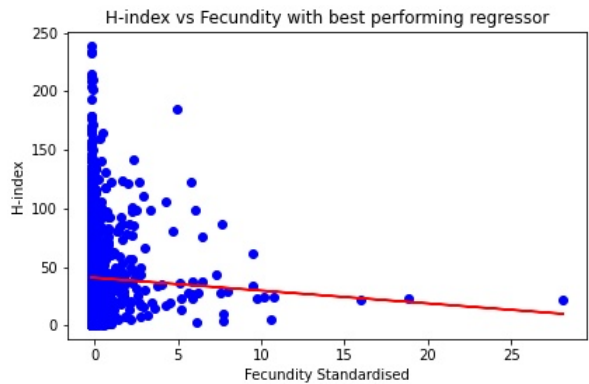


Figure 4
H-index versus Fecundity with SGDRegression as best regressor fit for the data



4.2 Classification

This section describes the results concerning the classification tasks of this paper. To achieve these results and be able to adequately measure the successes, the models, model metrics, and random chance baseline (accuracy=0.143) mentioned in Section 3.2.2 are used. Finally, all tasks are performed for both the original data and an up and down-sampled version of this data (Section 3.1.4, both datasets comply to the descriptions in Section 3.1.3).

4.2.1 All genealogy data versus fecundity (as academic success) H1. The performance metrics of the classification models trained on the original data are shown in Table 11. The best performing model for this data is the SGDClassifier (f1-score=0.900). The performance metrics of the classification models trained on the up and down-sampled data are shown in Table 12. The best performing model for this dataset is the Random Forest Classifier (f1-score). Not that all included models on both datasets perform better than the random chance baseline (accuracy=0.143).

Table 11

Classification model metrics for All genealogical data Versus Fecundity

Metrics	SGDClass	RandForClass	SupVecClass
Test recall	0.900	0.873	0.078
Test precision	0.900	0.873	0.078
Test accuracy	0.900	0.873	0.078
Test f1-score	0.900	0.873	0.078

Table 12

Classification model metrics for All genealogical data Versus Fecundity, with up and down-sampling for balance

Metrics	SGDClass	RandForClass	SupVecClass
Test recall	0.425	0.470	0.156
Test precision	0.425	0.470	0.156
Test accuracy	0.425	0.470	0.156
Test f1-score	0.425	0.470	0.156

4.2.2 H-index versus Fecundity, H3. The classification task regarding Hypothesis H3 concerns both the academic success metrics. The target value is fecundity, which is predicted by the input value h-index. In Table 13 and Table 14, the model metrics for the classification models are shown. Interpreting these results leads to the conclusion that Random Forest Classification is the best performing model for the original dataset (f1-score=0.203). Considering the up and down-sampled dataset, the SGDClassifier is the best performing model (f1-score=0.160). Both models outperform the random chance baseline of 0.143.

Table 13
Classification model metrics for H-index versus Fecundity

Metrics	SGDClass	RandForClass	SupVecClass
Test recall	0.179	0.203	0.141
Test precision	0.179	0.203	0.141
Test accuracy	0.179	0.203	0.141
Test f1-score	0.179	0.203	0.141

Table 14
Classification model metrics for H-index versus Fecundity, with up and down-sampling for balance

Metrics	SGDClass	RandForClass	SupVecClass
Test recall	0.160	0.136	0.098
Test precision	0.160	0.136	0.098
Test accuracy	0.160	0.136	0.098
Test f1-score	0.160	0.136	0.098

4.3 Clustering

This section goes into detail about the results concerning the clustering tasks within this study. Although these results are not directly linked to a hypothesis, they provide necessary background information and extra insights to back the results as aforementioned in this Section (Section 4). The formatting of this section follows the Clustering Tasks CT1-CT3 as depicted in Section 1.1.

4.3.1 Sums of students per generation, CT1. Beginning with the Kmeans model, as depicted in Appendix Q, the optimal amount of clusters identified by the elbow method is three. Using the Kmeans model, results in the scatter plots per cluster as illustrated in Appendix R. The performance metrics shown in Table 15, portray a positive view on the outputted clusters and their ability to distinguish groups using the current data.

When applying Mshift clustering on the current data, 155 clusters are defined by the model. The first cluster, however, is the only one with a sample size larger than 23.

Table 15

Performance metrics of Kmeans clustering based on sums of students per generation

Metric	Metric values
Silhouette score	0.672
Calinski harabasz score	6232.254
Davies Bouldin score	1.055

Table 16

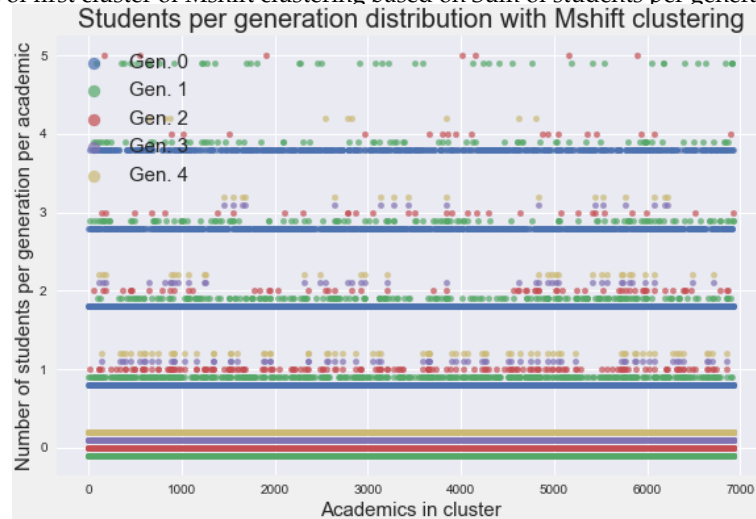
Performance metrics of Mshift clustering based on sums of students per generation

Metric	Metric values
Silhouette score	0.532
Calinski harabasz score	118.025
Davies Bouldin score	0.456

Therefore, the first cluster is selected for further interpretation. A scatter plot of this cluster is shown in Figure 5. The performance metrics regarding Mshift based on the current data are shown in Table 16

Figure 5

Scatter plot of first cluster of Mshift clustering based on Sum of students per generation



4.3.2 Metrics of academic success, CT2. Starting with the Kmeans model. By using the elbow method to determine the amount of clusters, an argument can be made for both two and four clusters (Visualised in Appendix Q). When looking at the performance metrics in Table 17, the selection for two clusters becomes apparent. When performing Kmean clustering with two clusters, a split appears based solely on the h-index.

When looking at the Mshift model, however, a different split can be recognized. This model shows a split on both axis, creating groups in the lower, middle and upper

Table 17

Performance metrics of Kmeans clustering based on h-index and fecundity

Metrics	2 Clusters	4 Clusters
Silhouette score	0.589	0.505
Calinski harabasz score	16.182	16062.926
Davies Bouldin score	0.613	0.636

Figure 6

Scatter plot for Kmeans clustering based on h-index and fecundity



ranges of each axis, adding one group located in the middle of both axis (cluster 3). Although the performance metrics of this clustering technique perform worse than the Kmeans, the clusters show logical groups.

Table 18

Performance metrics of Mshift clustering for h-index and fecundity

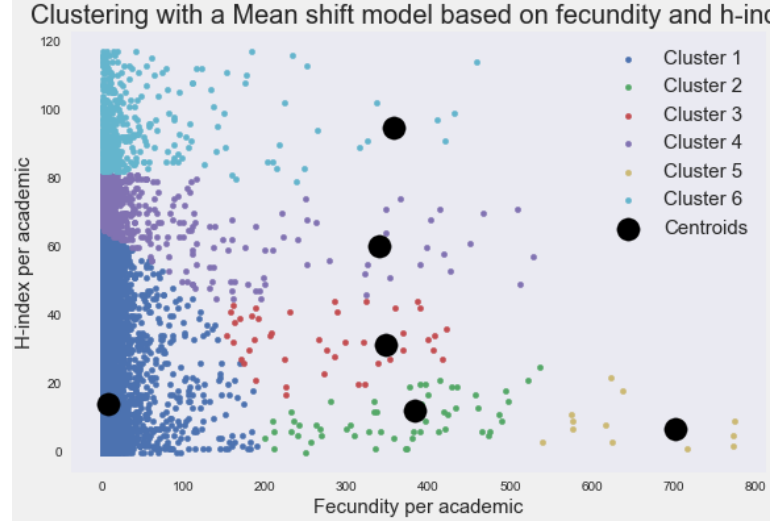
Metrics	Metric results
Silhouette score	0.441
Calinski harabasz score	2.184
Davies Bouldin score	0.921

4.3.3 Sum of academic relationship per relationship type, CT3. When running the elbow method to determine the amount of clusters for the Kmeans model, a clear preference for two clusters is shown (Appendix Q). The resulting Kmeans model returns two clusters as is visualized in Appendix S. The corresponding performance metrics are shown in Table 19.

Allowing the Mshift model to cluster the current data, results in four different clusters. These clusters are shown in Appendix T. This figure shows the count of a

Figure 7

Scatter plot for Mshift clustering based on h-index and fecundity

**Table 19**

Performance metrics of Kmeans clustering for sum of academic relationship per relationship type

Metric	Metric values
Silhouette score	0.681
Calinski harabasz score	3283.96
Davies Bouldin score	1.107

Table 20

Performance metrics of Mshift clustering for sum of academic relationship per relationship type

Metric	Metric values
Silhouette score	0.672
Calinski harabasz score	6232.254
Davies Bouldin score	1.055

certain relationship on the y-axis and the type of relationship on the x-axis (which follows the same numeric interpretation as the relation_0 through relation_1 as depicted in Appendix E). By examining the plot, one can see a limit of five on the range of the relation count (y-axis). The Mshift model appears to leave out all other values. The corresponding performance metrics are shown in Table 20.

5. Discussion

The goal of this paper is to answer the research question "*Can academic success be predicted?*". And thereby fill a current gap in the body of literature. To answer this question,

several hypotheses are formed, creating the structure of this paper. To answer the main research question, first the hypotheses must be reexamined.

Starting with Hypothesis H1, "*Academic genealogical data correlates positively to the success of an academic.*". To answer this statement, genealogical data extracted from NeuroTree is, after some preprocessing steps, used as training data for both regression and categorical prediction models. The result of the best performing model for regression prediction is Random Forest Regression (R^2 0.079 and MAE 8.66), which does not provide clear evidence of a correlation between the genealogical data and fecundity. The best performing regressor for h-index was also Random Forest Regression (Explained Variance 0.085 and MAE 20.06), which also provides no clear evidence for a positive correlation. The results of the best models for the respective classification tasks are, with original data SGDClassifier (f1-score 0.900), and with up and down-sampled data for balancing, Random Forest Classifier (f1-score 0.470) for predicting fecundity. Both perform way over their random chance baseline of 0.143, showing there is a relationship between academic genealogical data and the success of an academic, measured in fecundity. However, this shows no clear positive or negative relation. To answer the same question for the h-index, the best performing regression model is Random Forest Regression (R^2 0.085 and MAE 20.20), which also provides no evidence of a significant positive correlation. These findings coincide with other studies who aimed to predict a future h-index of an academic. Although these papers report high predictive power, a significant part of the explained variance seems to be attributable to current or historic values of the h-index. This argues for the predictive power of the h-index towards its future self but provides little argument for the predictive power of other genealogical data (Acuna, Allesina, and Kording 2012; Ayaz, Masood, and Islam 2018; Jensen, Rouquier, and Croissant 2009). Summarizing the result concerning Hypothesis H1, this study finds no sufficient support to confirm this hypothesis.

One interesting discovery during the research to test Hypothesis H1, is the interpretation of the non-zero lasso coefficients, presented in Section 4.1.1. when looking at the relation between all genealogical data and fecundity. These show that relation type 4 (relation_4) is the most influential variable from the 'All genealogical data' dataset for predicting fecundity according to the Lasso model. One would expect some positive relationship with the other relation types, since those are mentorship relations and are thus indirectly included into the formula calculating fecundity. However, relation type 4 is a collaborative relation, and not a mentorship relation. In an attempt to get a deeper understanding of the variables containing the sums of the different types of relations, clustering models are applied (results in Section 4.3.3). The created clusters, however, show a tendency to split based on all relation types without indicating informative results. To better understand the relation between collaborative academic relationships and academic fecundity, more research is needed.

Continuing with Hypothesis H2, "*There is a positive relationship between the number of direct PhD students and subsequent PhD students of students, of an academic and his success.*". As earlier stated, this hypothesis is a more detailed and specific version of Hypothesis H1. This choice is made because the literature seems to agree that producing new academics is an important part of being a good and successful academic yourself. To calculate a proper model for the fecundity variable, two sets were created, one with, and one without instances containing zero values for fecundity. This is due to the fact that the data is skewed towards the zero, moreover, if fecundity is zero, the amount of students is zero as well. The deletion of these instances aimed to improved results. When training the Lasso regression model on the sum of students per generation and allowing it to predict fecundity, resulted in an MAE of 0.543 and an R^2 of 1.000. Since the

fecundity is calculated using only the sum of students per generation, such a correlation is not surprising. For a possible correlation with the other academic success metric, h-index, one can look at the results depicted in Section 4.1.3, where the relation between h-index and all genealogical data is shown. Since the sum of students per generation is included in this input data, this will show similar to higher results. The Random Forest Regression (R^2 0.085 and MAE 20.199) does not provide clear evidence for a significant positive relationship. This result indicates that, although mentoring of qualitative new academics is considered of vital importance to the academic community and scientific advancements (Marsh 2017; Andraos 2005; Allen et al. 1997; Hunt and Michael 1983), it is not included in the h-index and thus a widely used metric of success. In total, the results do not provide adequate support for Hypothesis H2.

However, when interpreting the Lasso non-zero coefficients, as depicted in Section 4.1.3, some interesting details appear. These results show that the Lasso regression increases the estimated influence with each generation of students. In combination with the theoretical possibility of 'lucky' and 'lazy' mentors as stated in Section 2.2, the argumentation for a clustering task was formed. This task includes clustering based on the sums of students per generation to form a deeper understanding of the aforementioned theory in practice and non-zero Lasso results. As visualized in Appendix R, the second cluster (Silhouette score=0.672) resulting from the Kmeans model (in Section 4.3.1) seems to identify mentors with relatively high direct students and relatively low students from further generations (Gen. 1 in the plot). This cluster could identify the aforementioned 'lazy' mentors. However, a cause for the distribution of sums of students cannot be determined. Still, this clustering output raises further questions regarding the cause of such groups and how this group compares to the total data and other groups. To answer such questions, further research is necessary.

Finally, Hypothesis H3 *"Different metrics of academic success have a positive relation towards each other."* Since there are different values of success for the same academic, this is an interesting hypothesis. Do the metrics measure different parts of an academic's performance, or do they coincide? By combining data from NeuroTree and Google Scholar, the relationship between h-index and fecundity is analyzed. When looking at the best performing regression model, Random Forest Regression (R^2 0.000 and MAE 23.01), no clear evidence is shown for a positive (or any) relationship. When examining the classification predictors for the same task, the highest scoring model is Random Forest Classification (f1-score 0.203), using the original dataset and SGDClassification (f1-score 0.160) using the up and down-sampled dataset. Both show a slightly higher performance than the random change baseline of 0.143. This indicates little predictive power between the two metrics of success. In total the results of this paper show no clear evidence for a positive (or any) correlation between fecundity and the h-index. To go into more detail concerning these variables, two clustering models were used in an attempt to identify possible groups (results in Section 4.3.3. The Kmeans model (Silhouette score=0.532) shows only a split based on the h-index without implications concerning fecundity. The Mshift model (Silhouette score=0.441), shows a lower performance, but a logical grouping. It appears group academics in the middle of both variable ranges (cluster 3). All other groups seem to be clustered based on the performance based on either one of the success metrics (cluster 1, 4 and 6 based on h-index and 2 and 5 on fecundity). The splits based on the single variables, add to the question concerning the existence of any relation or overlap between the two metrics of success.

These results fit in the conflicting results in the literature concerning academic success and possible correlations (Li  nard et al. 2018; Li et al. 2019; Malmgren, Ottino, and Amaral 2010; Heinisch and Buenstorf 2018). The current conflicts might be attributable

to the fact that different metrics cover considerable different sections of an academic's performance. As is stated in several studies (Cerchiello and Giudici 2014; Waltman and Van Eck 2012; Costas and Bordons 2007), the use of a single value to distinguish the success of an academic might be inaccurate and inadequate.

The lack of correlation and overlap between two widely used metrics of academic success could have a significant societal impact. Both metrics are, as stated in Section 2, widely used to assess the performance of an academic. By adhering to one of the stated metrics of success, academic success might be consistently misjudged. A misjudgement of an individuals success directly links to real life consequences ranging from less attention towards an academic's research output up to the rejection for certain jobs. When interpreting the results in this paper, this misevaluation could be due to an estimation error which is inherited by the very metric(s) meant to fairly and consistently judge performance.

In summary, this study shows little evidence for Hypothesis H1. There seems to be some relation when accounting for the best performing categorical model predicting fecundity (f1-score 0.900). A specific relation can, however, not be deducted from these results, imploring more research on the subject. Furthermore, Hypothesis H2 is, obviously, true for fecundity due to its nature, but shows no clear positive (or any) correlation with the h-index. Lastly, there is little evidence supporting Hypothesis H3. The best classification models perform only slightly better than the random chance baseline and the best regression model resulted in near zero R^2 . In total, these results answer the Research Question RQ based on the current data. Only the classification results seem to show some predictive power of the genealogical data towards fecundity, but no clear predictors or relations are identified. Moreover, this study found little evidence supporting any predictive power of the genealogical data towards the h-index. Finally, by examining the increase in interest towards studies researching academic success (AlShebli, Makovi, and Rahwan 2020; Li et al. 2019; Heinisch and Buenstorf 2018), contradicting results in such studies (Li et al. 2019; Liénard et al. 2018; Malmgren, Ottino, and Amaral 2010; Paglis, Green, and Bauer 2006; Waltman and Van Eck 2012), and the results in this paper, a pressing, more fundamental question arises 'what is true academic success, and how should it be measured?'

As a final note, this paragraph will emphasize some difficulties concerning this study. First and foremost, the crowd sourced nature of the data from NeuroTree caused several obstacles. These include the numerous missing and invalid values. This caused a multitude of assumptions, including assuming overarching definitions for categorical data, and missed opportunities, such as the inclusion of an academics country or university. Adding to difficulties concerning the data, due to the absence of a unique identifier which could function as such outside the NeuroTree dataset caused an extremely sub-optimal approach for data enrichment. The h-index importation from Google Scholar, included a significant number of double values, which were forced to be dropped. This might have caused an unknown bias in the data. Furthermore, without access to a server or other tools, computational limitations caused considerable complications in the model selection, analysis and h-index importation.

References

- Acuna, Daniel E, Stefano Allesina, and Konrad P Kording. 2012. Predicting scientific success. *Nature*, 489(7415):201–202.
- Allen, Tammy D, Mark L Poteet, and Susan M Burroughs. 1997. The mentor's perspective: A qualitative inquiry and future research agenda. *Journal of vocational behavior*, 51(1):70–89.
- Allen, Tammy D, Mark L Poteet, Joyce EA Russell, and Gregory H Dobbins. 1997. A field study of factors related to supervisors' willingness to mentor others. *Journal of Vocational Behavior*, 50(1):1–22.
- AlShebli, Bedoor, Kinga Makovi, and Talal Rahwan. 2020. The association between early career informal mentorship in academic collaborations and junior author performance. *Nature communications*, 11(1):1–8.
- Andraos, John. 2005. Scientific genealogies of physical and mechanistic organic chemists. *Canadian journal of chemistry*, 83(9):1400–1414.
- Ayaz, Samreen, Nayyer Masood, and Muhammad Arshad Islam. 2018. Predicting scientific impact based on h-index. *Scientometrics*, 114(3):993–1010.
- Bol, Thijs, Mathijs de Vaan, and Arnout van de Rijt. 2018. The matthew effect in science funding. *Proceedings of the National Academy of Sciences*, 115(19):4887–4890.
- Braun, Tibor, Wolfgang Glänzel, and András Schubert. 2006. A hirsch-type index for journals. *Scientometrics*, 69(1):169–173.
- Campbell, Toni A and David E Campbell. 1997. Faculty/student mentor program: Effects on academic performance and retention. *Research in higher education*, 38(6):727–742.
- Cerchiello, Paola and Paolo Giudici. 2014. On a statistical h index. *Scientometrics*, 99(2):299–312.
- Chang, Sooyoung. 2003. Academic genealogy of american physicists. *AAPPS Bulletin*, 13(6):6–41.
- Chariker, Julia H, Yihang Zhang, John R Pani, and Eric C Rouchka. 2017. Identification of successful mentoring communities using network-based analysis of mentor–mentee relationships across nobel laureates. *Scientometrics*, 111(3):1733–1749.
- Costas, Rodrigo and María Bordons. 2007. The h-index: Advantages, limitations and its relation with other bibliometric indicators at the micro level. *Journal of informetrics*, 1(3):193–203.
- Cronin, Blaise and Lokman Meho. 2006. Using the h-index to rank influential information scientists. *Journal of the American Society for Information Science and technology*, 57(9):1275–1278.
- Crosta, Peter M and Iris G Packman. 2005. Faculty productivity in supervising doctoral students' dissertations at cornell university. *Economics of Education Review*, 24(1):55–65.
- David, Stephen and Ben Hayden. 2020. Analysis.
- David, Stephen V and Benjamin Y Hayden. 2012. Neurotree: A collaborative, graphical database of the academic genealogy of neuroscience. *PloS one*, 7(10):e46608.
- Egghe, Leo. 2010. The hirsch index and related impact measures. *Annual review of information science and technology*, 44(1):65–114.
- Fortunato, Santo, Carl T Bergstrom, Katy Börner, James A Evans, Dirk Helbing, Staša Milojević, Alexander M Petersen, Filippo Radicchi, Roberta Sinatra, Brian Uzzi, et al. 2018. Science of science. *Science*, 359(6379).
- Heinisch, Dominik P and Guido Buenstorf. 2018. The next generation (plus one): an analysis of doctoral students' academic fecundity based on a novel approach to advisor identification. *Scientometrics*, 117(1):351–380.
- Hirsch, Jorge E. 2005. An index to quantify an individual's scientific research output. *Proceedings of the National academy of Sciences*, 102(46):16569–16572.
- Hirsch, Jorge E. 2007. Does the h index have predictive power? *Proceedings of the National Academy of Sciences*, 104(49):19193–19198.
- <https://orcid.org/>. 2020.
- <https://neurotree.org/>. 2020.
- <https://pypi.org/project/scholarly/>. 2020.
- <https://scholar.google.com/>. 2020.
- Hunt, David Marshall and Carol Michael. 1983. Mentorship: A career training and development tool. *Academy of management Review*, 8(3):475–485.
- Jackson, Allyn. 2007. A labor of love: the mathematics genealogy project. *Notices of the AMS*, 54(8):1002–1003.
- Jensen, Pablo, Jean-Baptiste Rouquier, and Yves Croissant. 2009. Testing bibliometric indicators by their prediction of scientists promotions. *Scientometrics*, 78(3):467–479.
- Jeong, Hawoong, Zoltan Néda, and Albert-László Barabási. 2003. Measuring preferential attachment in evolving networks. *EPL (Europhysics Letters)*, 61(4):567.

- Kelley, Elizabeth A and Robert W Sussman. 2007. An academic genealogy on the history of american field primatologists. *American Journal of Physical Anthropology: The Official Publication of the American Association of Physical Anthropologists*, 132(3):406–425.
- Lazaridis, Themis. 2010. Ranking university departments using the mean h-index. *Scientometrics*, 82(2):211–216.
- Li, Weihua, Tomaso Aste, Fabio Caccioli, and Giacomo Livan. 2019. Early coauthorship with top scientists predicts success in academic careers. *Nature communications*, 10(1):1–9.
- Liénard, Jean F, Titipat Achakulvisut, Daniel E Acuna, and Stephen V David. 2018. Intellectual synthesis in mentorship determines success in academic careers. *Nature communications*, 9(1):1–13.
- Malmgren, R Dean, Julio M Ottino, and Luís A Nunes Amaral. 2010. The role of mentorship in protégé performance. *Nature*, 465(7298):622–626.
- Marsh, Elizabeth J. 2017. Family matters: Measuring impact through one's academic descendants. *Perspectives on Psychological Science*, 12(6):1130–1132.
- Newman, Mark EJ. 2001. Clustering and preferential attachment in growing networks. *Physical review E*, 64(2):025102.
- Paglis, Laura L, Stephen G Green, and Talya N Bauer. 2006. Does adviser mentoring add value? a longitudinal study of mentoring and doctoral student outcomes. *Research in Higher Education*, 47(4):451–476.
- Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.
- Perc, Matjaž. 2014. The matthew effect in empirical data. *Journal of The Royal Society Interface*, 11(98):20140378.
- Rezek, Issa, Robert J McDonald, and David F Kallmes. 2011. Is the h-index predictive of greater nih funding success among academic radiologists? *Academic radiology*, 18(11):1337–1340.
- Rossi, Luciano, Rafael JP Damaceno, Igor L Freire, Etelvino JH Bechara, and Jesús P Mena-Chalco. 2018. Topological metrics in academic genealogy graphs. *Journal of Informetrics*, 12(4):1042–1058.
- Sarigöl, Emre, René Pfitzner, Ingo Scholtes, Antonios Garas, and Frank Schweitzer. 2014. Predicting scientific success based on coauthorship networks. *EPJ Data Science*, 3(1):9.
- Schubert, András and Gábor Schubert. 2019. All along the h-index-related literature: A guided tour. In *Springer Handbook of Science and Technology Indicators*. Springer, pages 301–334.
- Seibert, Kent W, Douglas T Hall, and Kathy E Kram. 1995. Strengthening the weak link in strategic executive development: Integrating individual development and global business strategy. *Human Resource Management*, 34(4):549–567.
- Sugimoto, Cassidy R. 2014. *Academic Genealogy In Cronin, Blaise Sugimoto. Beyond bibliometrics: Harnessing multidimensional indicators of scholarly impact (pp. 365-380)*. MIT Press.
- Sugimoto, Cassidy R and Vincent Larivière. 2018. *Measuring research: What everyone needs to know*. Oxford University Press.
- Vanclay, Jerome K. 2008. Ranking forestry journals using the h-index. *Journal of informetrics*, 2(4):326–334.
- Waltman, Ludo and Nees Jan Van Eck. 2012. The inconsistency of the h-index. *Journal of the American Society for Information Science and Technology*, 63(2):406–415.

Appendix A: People table, variable explanation

Table 1
People tsv file description

Notation	Meaning
pid	Unique key of an individual
degrees	Degrees of an individual
location	Most recent institution
locid	Location unique key
area	Research areas of the academic
majorarea	Indication of location between genealogical trees
award	Awards won by the individual
hindex	h-index of an academic
orcid_id	Individual unique key within orcid
s2id	Individual unique key within s2
homepage	The url of an academic
addedby	Person that added the information
dateadded	Date the information is added

Appendix B: Connect table, variable explanation

Table 1
Connect tsv file description

Notation	Meaning
cid	Connection unique id
pid1	Trainee unique key
pid2	Mentor unique key
relation	Type of relation relation_0: Undergraduate relation_1: Grad student relation_2: Postdoc relation_3: Research scientist relation_4: Collaboration—not a mentor relationship
locid	Location unique key
location	Most recent insitution
startdate	Start date of relation
stopdate	Stop date of relation
dateadded	Date the information is added

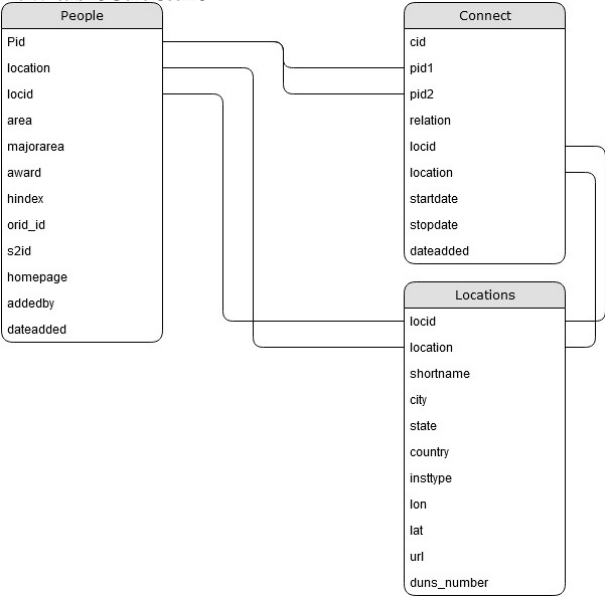
Appendix C: Raw dataset, combination of

Table 1
Locations tsv file description

Notation	Meaning
locid	Location unique key
location	Location name
shortname	Location name in short
city	City name of the location
state	State name of location
country	Country name of location
insttype	Unknown and will not be used
lon	Longitude value of location
lat	Latitude value of location
url	Website of the location
duns_number	specific duns number of the location

Appendix D: Data structure NeuroTree data dump, visualized

Figure 1
Raw table structure



Appendix E: Dataframe created by combining datasets from NeuroTree**Table 1**

Column description of total data

Column	Meaning
pid	Unique key of an individual
degrees	Degrees of an individual
location	Most recent institution
locid	Location unique key
area	Research areas of the academic
majorarea	Indication of location between genealogical trees
award	Awards won by the individual
hindex	h-index of an academic
orcid_id	Individual unique key within orcid
s2id	Individual unique key within s2
homepage	The url of an academic
addedby	Person that added the information
dateadded	Date the information is added
country	Country name of location
fec1	Fecundity sum over first generation of students
fec2	Fecundity sum over second generation of students
fec3	Fecundity sum over third generation of students
fec4	Fecundity sum over fourth generation of students
fec5	Fecundity sum over fifth generation of students
fec_tot	Fecundity score over first five generations of students
fec_parent	Fecundity score of the mentor
students	pid's of direct students
relation_0	Accumulation of relations "0", undergraduate (mentor function)
relation_1	Accumulation of relations "1", grad student (mentor function)
relation_2	Accumulation of relations "2", postdoc (mentor function)
relation_3	Accumulation of relations "3", research scientist (mentor function)
relation_4	Accumulation of relations "4", collaboration (not a mentor function)

Appendix F: Example for freedom of input NeuroTree entry

Figure 1

Example of freedom of input NeuroTree

NEW ENTRY

Name:

Affiliation(s): Please include current and previous non-mentored positions. Click "Add another" to create additional entries.

1. Years: /

OTHER DATA (OPTIONAL)

Research Areas:

Homepage:

Photo URL:


Pub-med search:

Google Search:

Orcid ID ([info](#)):

BIOGRAPHY

I am sorry to bother you, but no information on this entry can be used. Feel free to delete this



(We especially appreciate biographical information for individuals and historical figures who do not otherwise have a web presence.)

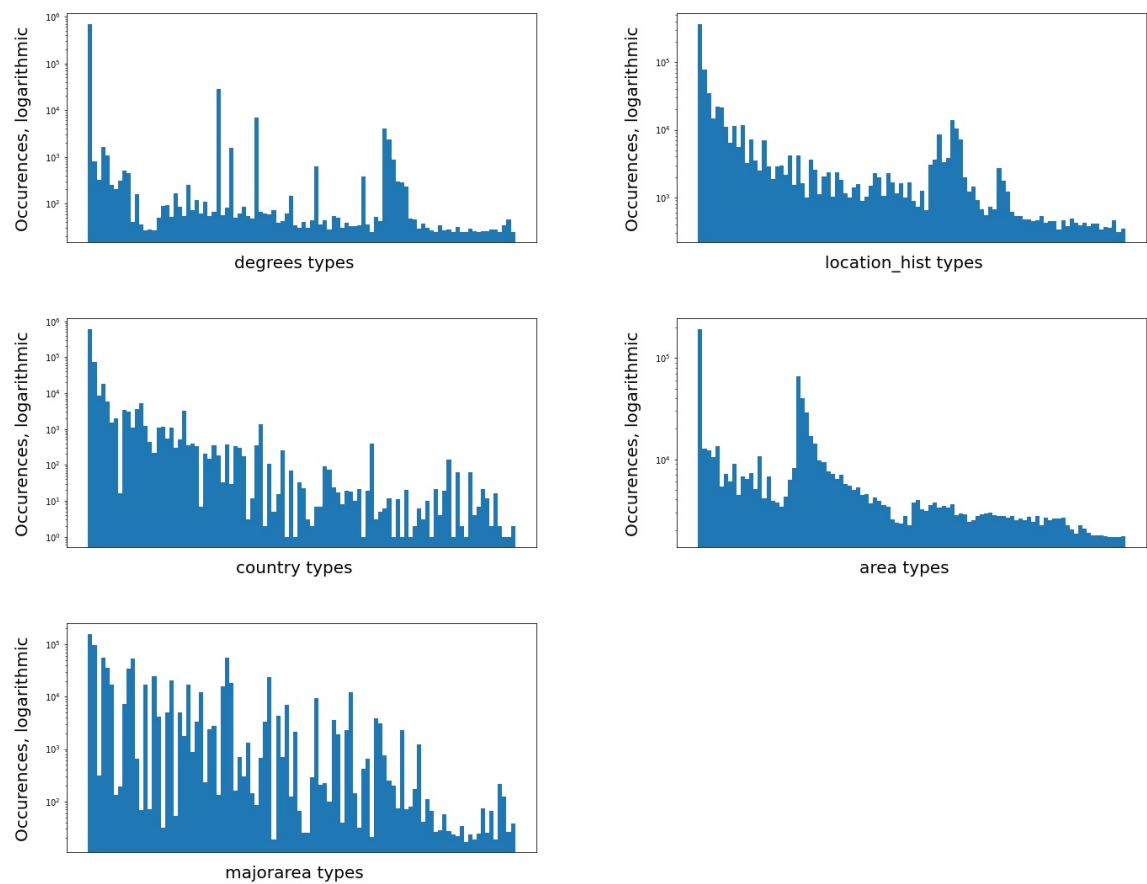
([Cancel](#))

Appendix G: Missing data percentages in raw combined dataset**Table 1**
Missing data per variable

Column	Percentage missing data
location	1.41%
degrees	0.00%
area	21.11%
majorarea	0.00%
award	99.90%
hindex	97.06%
orcid_id	99.82%
s2id	97.06%
homepage	93.72%
addedby	0.00%
dateadded	0.00%
country	65.11%
fec1	0.00%
fec2	0.00%
fec3	0.00%
fec4	0.00%
fec5	0.00%
fec_tot	0.00%
fec_parent	0.00%
students	0.00%
relation ₀	0.00%
relation ₁	0.00%
relation ₂	0.00%
relation ₃	0.00%
relation ₄	0.00%

Appendix H: Raw categorical data distribution

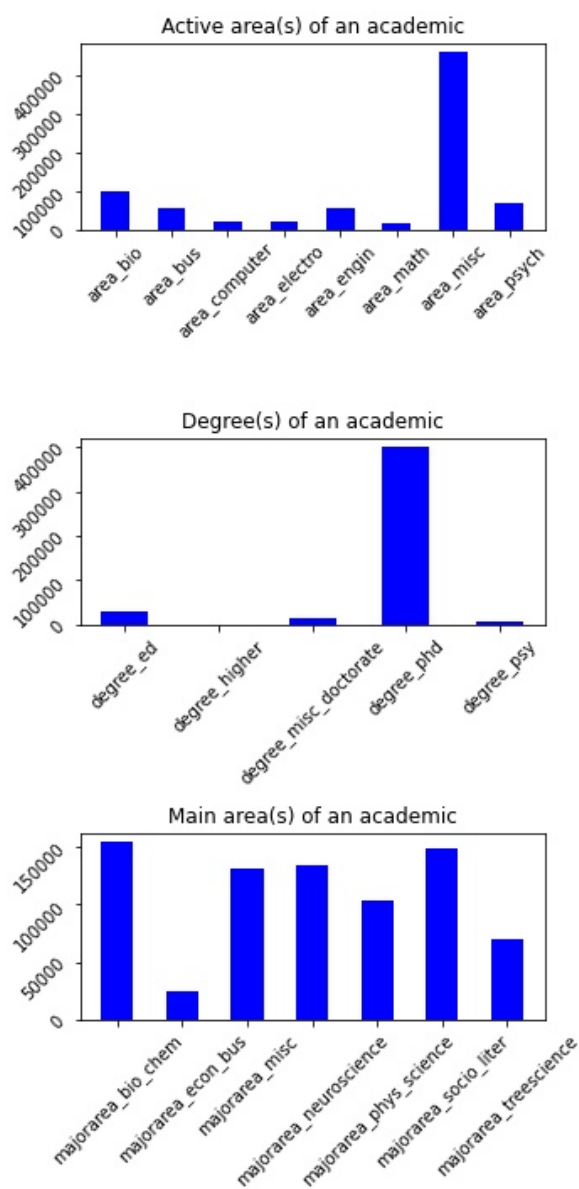
Figure 1
Histograms of raw categorical data



Appendix I: Distributions of categorical values, binned

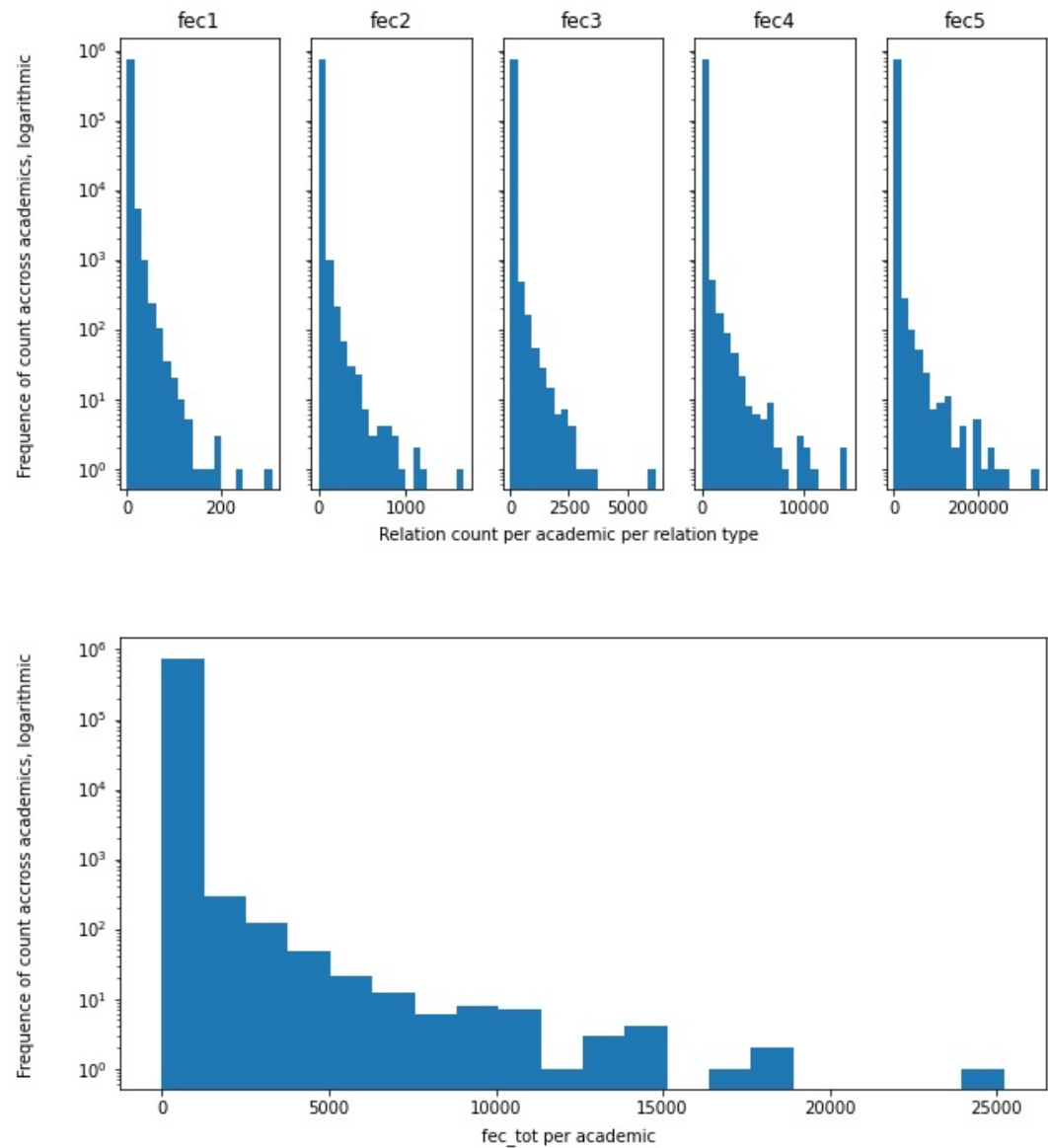
Figure 1

Value counts of categorical variables in bins



Appendix J: Distribution of student sum per generation and fecundity

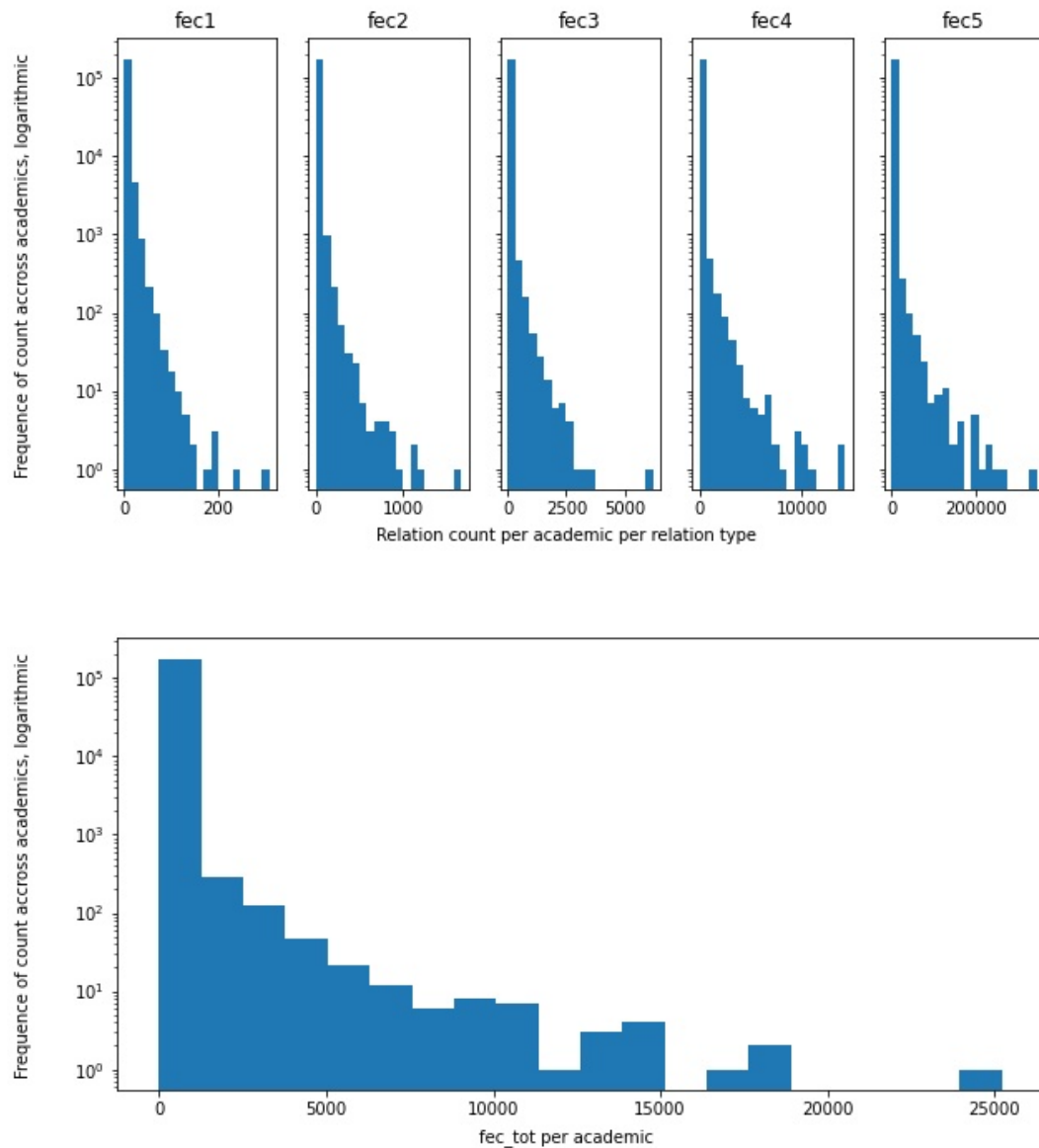
Figure 1
Histograms of sum of students per generation and fecundity



Appendix K: Distribution of student sum per generation and fecundity without fecundity zero values

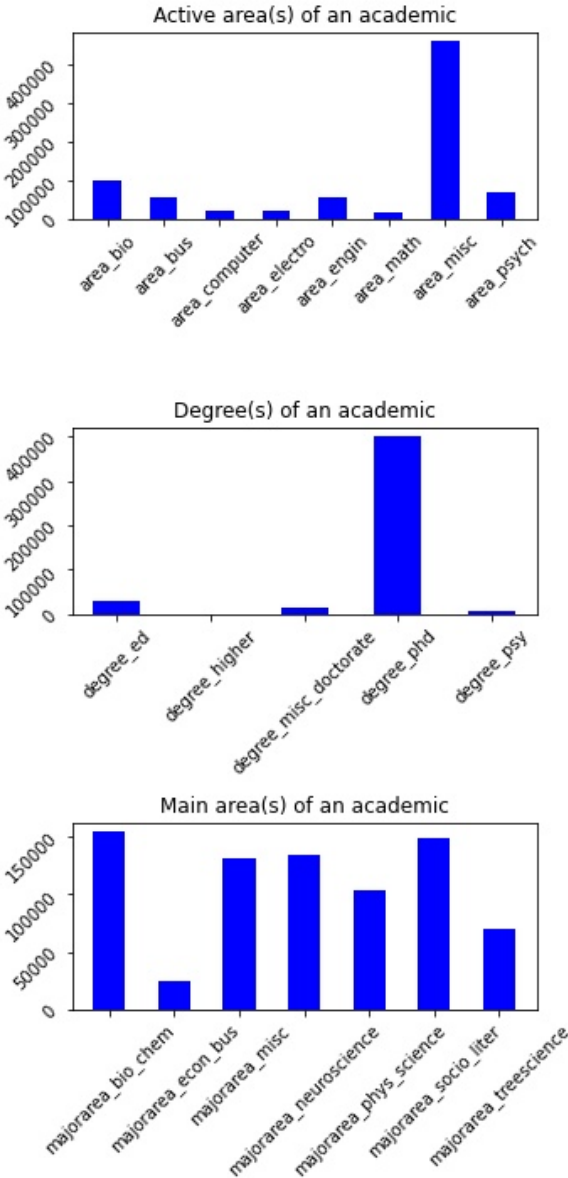
Figure 1

Histograms of student sum per generation (fec1-fec5) and fecundity



Appendix L: Raw binned variables distributions

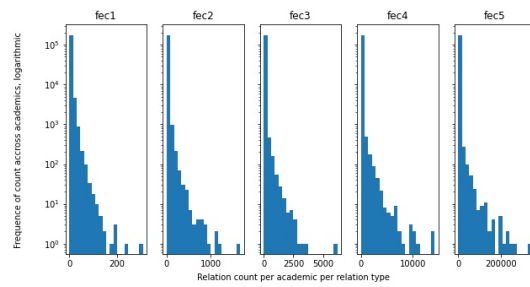
All binned categorical variables shown as histograms.



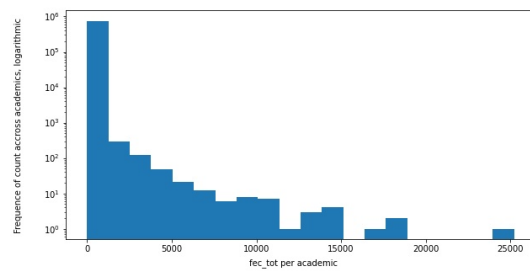
Appendix M: Raw continuous variables distributions

All continuous variables shown as histograms. Note that the y axis is logarithmic.

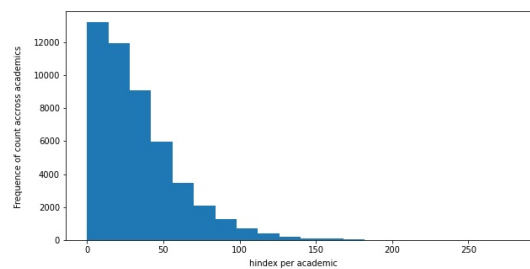
Sum of students per generation



Fecundity per academic



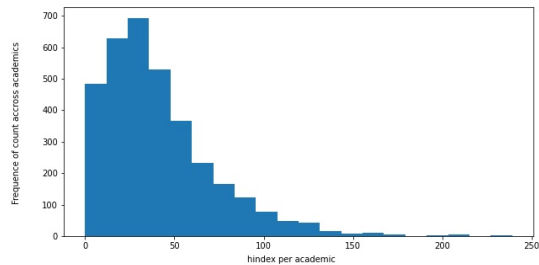
H-index per academic



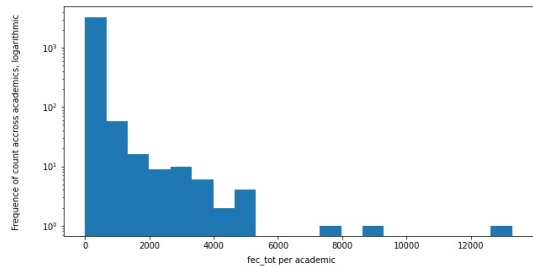
Appendix N: Fecundity and H-index distribution for analyses concerning their relationship, Hypothesis H3

Both variables shown as histogram, note that the fecundity histogram has a logarithmic y axis.

H-index histogram



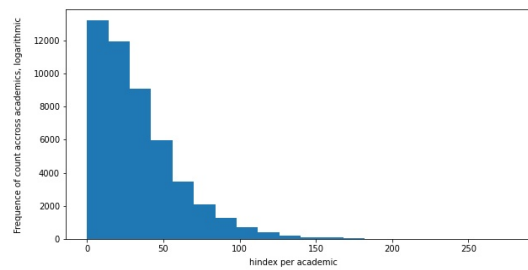
Fecundity histogram



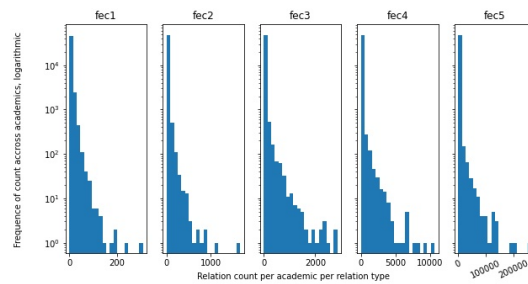
Appendix O: H-index and all genealogical data distributions

All variables displayed as histogram.

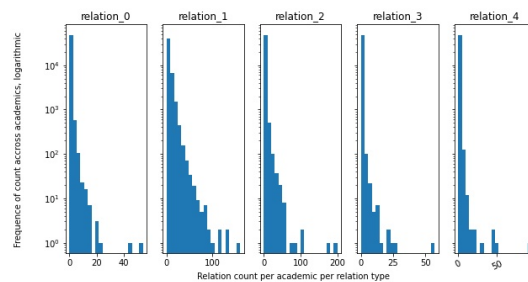
H-index histogram



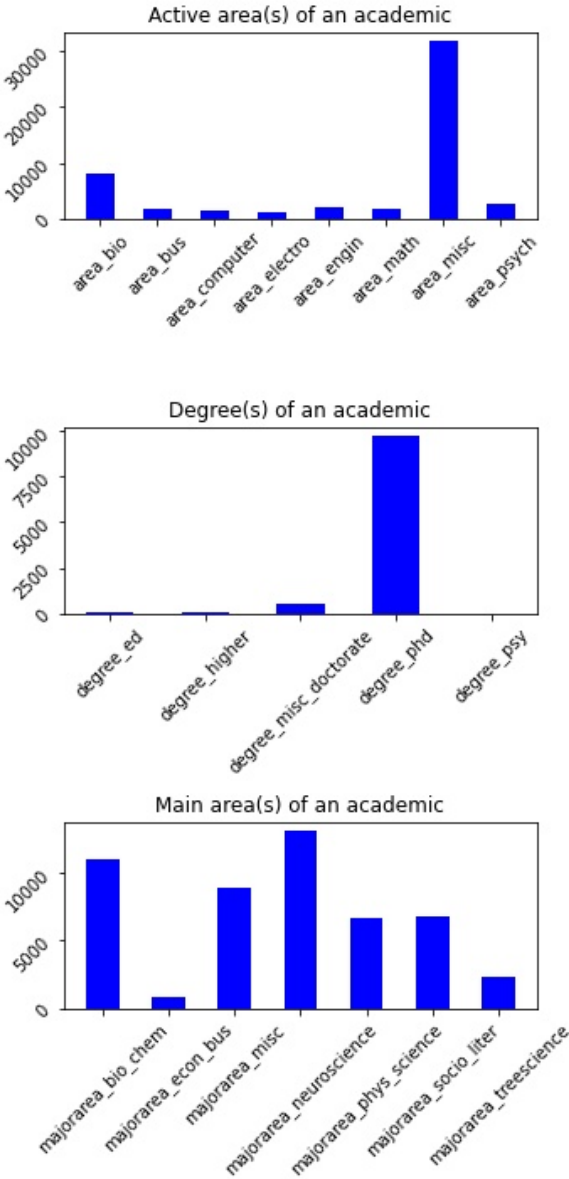
Sum of students per generation (y axis is algorithmic)



Sum of types of relations (y axis is algorithmic)



Sum of categorical values



Appendix P: Best performing Consequential Neural Network hyperparameters per regression task

Table 1

Best performing Consequential Neural Network hyperparameters All genealogical data versus Fecundity

Hyperparameter	Setting
Epochs	8
Activation	relu
Optimizer	SGD
Loss	CosineSimilarity
Final_acti	linear
Layers	14
Perceptrons	32

Table 2

Best performing Consequential Neural Network hyperparameters All genealogical data versus H-index

Hyperparameter	Setting
Epochs	8
Activation	relu
Optimizer	SGD
Loss	CosineSimilarity
Final_acti	linear
Layers	4
Perceptrons	32

Table 3

Best performing Consequential Neural Network hyperparameters Fecundity versus H-index

Hyperparameter	Setting
Epochs	16
Activation	softplus
Optimizer	Adam
Loss	mae
Final_acti	linear
Layers	4
Perceptrons	32

Appendix Q: Elbow method used for all Kmeans clustering tasks

Figure 1
Elbow method for Kmeans clustering based on sums of students per generation

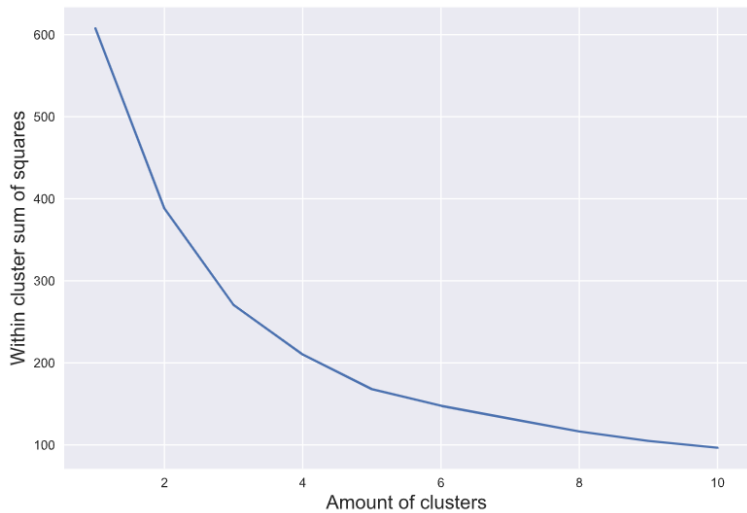


Figure 2
Elbow method for Kmeans clustering based on h-index and fecundity

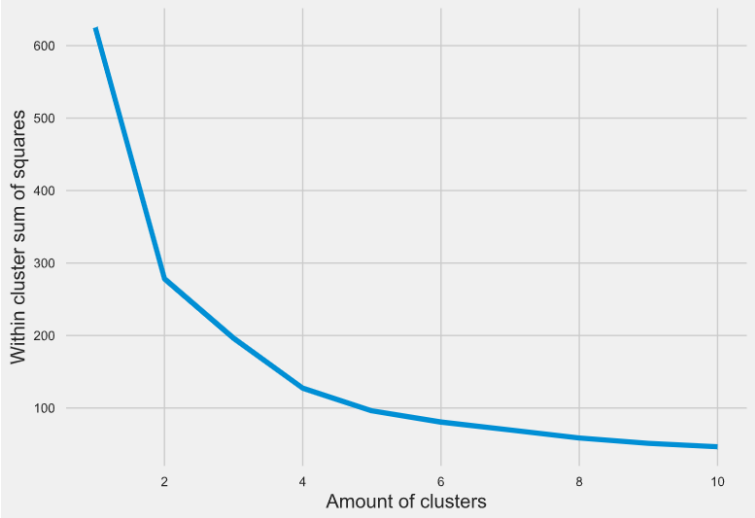
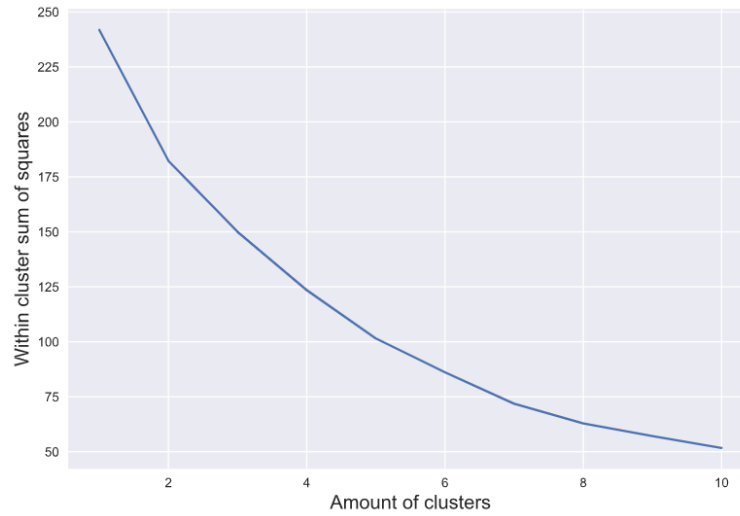


Figure 3

Elbow method for Kmeans clustering based on h-index and fecundity



Appendix R: Scatter plots for Kmeans clustering based on Sums of students per generation

Figure 1
Scatter plot Kmeans cluster 1

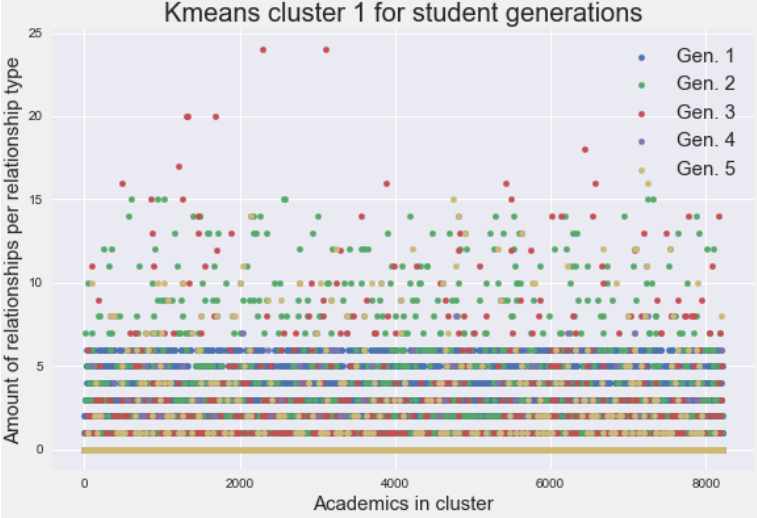


Figure 2
Scatter plot Kmeans cluster 2

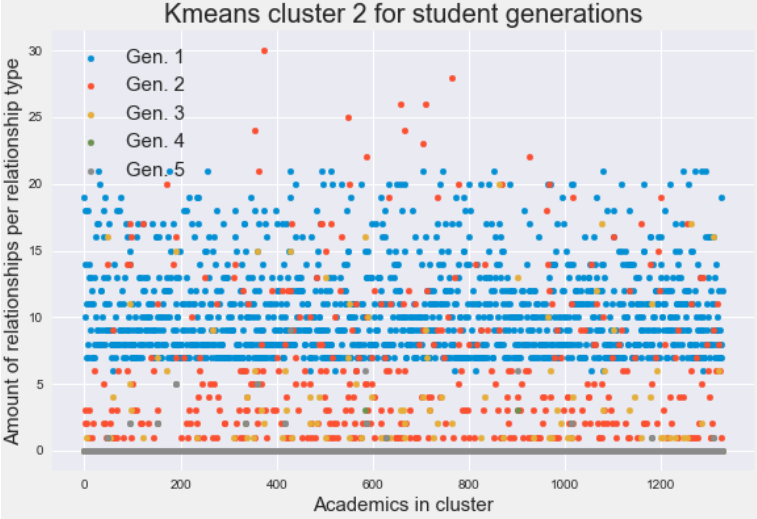
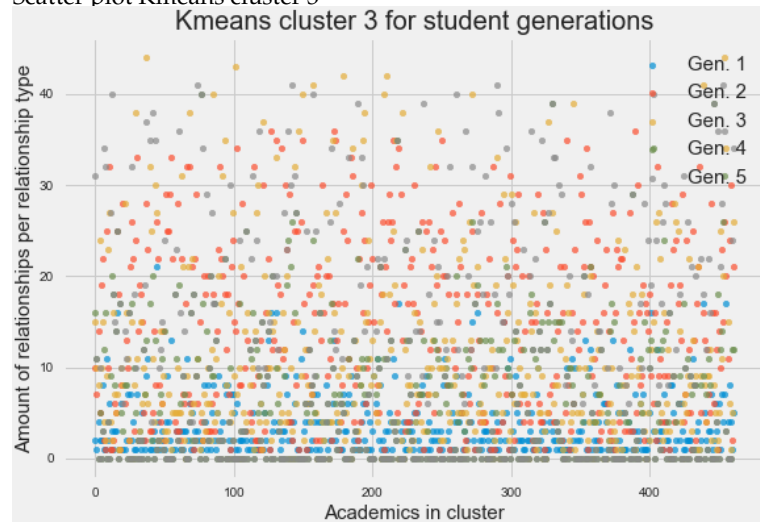


Figure 3

Scatter plot Kmeans cluster 3



Appendix S: Plotted clusters of Kmeans clustering based on sum of academic relationship per relationship type

Figure 1
Scatter plot Kmeans cluster 1
Kmeans cluster 1 for types of academic relation

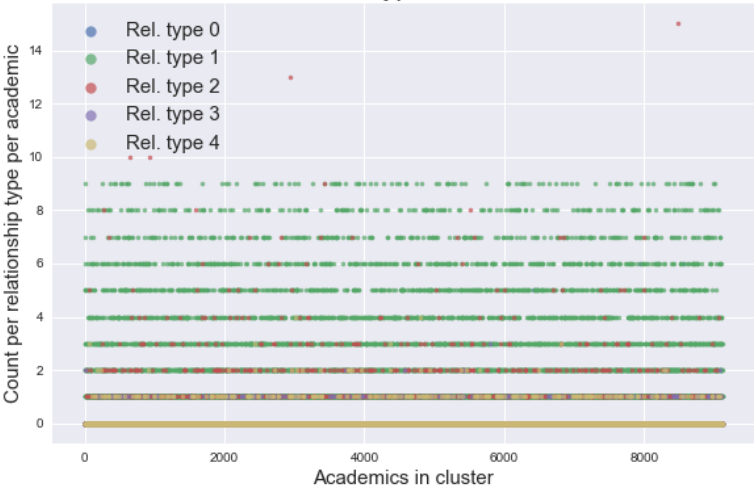
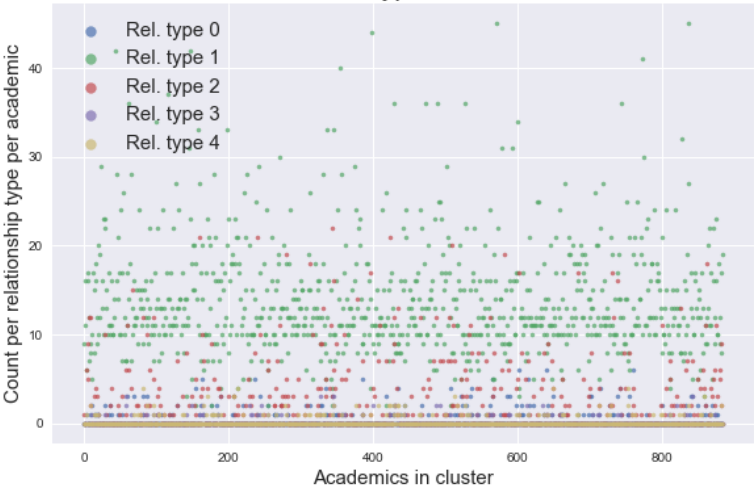


Figure 2
Scatter plot Kmeans cluster 2
Kmeans cluster 2 for types of academic relation



Appendix T: First cluster of Mshift clustering based on sum of academic relationship per relationship type**Figure 1**

Scatter plot for Mshift clustering based on sum of academic relationship per relationship type

