ACADEMIC SUPPORT NETWORK REFLECTS DOCTORAL EXPERIENCE AND PRODUCTIVITY

by
ÖZGÜR CAN SEÇKIN

Submitted to the Graduate School of Engineering and Natural Sciences
in partial fulfilment of
the requirements for the degree of Master of Sciences

Sabancı University
July 2022

ACADEMIC SUPPORT NETWORK REFLECTS DOCTORAL EXPERIENCE AND PRODUCTIVITY

Approved by:

Asst. Prof. Onur Varol .....
(Thesis Supervisor)

Prof. Ayşe Berrin Yanıkoğlu .....

Assoc. Prof. Hamid Akın Ünver .....

Date of Approval: July 6, 2022

ABSTRACT

ACADEMIC SUPPORT NETWORK REFLECTS DOCTORAL EXPERIENCE AND PRODUCTIVITY

ÖZGÜR CAN SEÇKIN

Data Science M.Sc. THESIS, July 2022

Thesis Supervisor: ASST. PROF. ONUR VAROL

Keywords: science of science, network science, text mining, natural language processing, machine learning

Current practices of quantifying academic performance by productivity raise serious concerns about the psychological well-being of graduate students. These efforts often neglect the influence of researchers' environment. Acknowledgments subsections in dissertations shed light on this environment by providing an opportunity for students to thank the people who supported them. We analysed 26,236 acknowledgments to create an "academic support network" that reveals five distinct communities supporting students along the way: Academic, Administration, Family, Friends & Colleagues, and Spiritual. We show that female students mention fewer people from each of these communities, with the exception of their families, and that their productivity is slightly lower than that of males when considering the number of publications alone. This is critically important because it means that studying the doctoral process may help us better understand the adverse conditions women face early in their academic careers. Our results also suggest that the total number of people mentioned in the acknowledgements allows disciplines to be categorised as either individual science or team science as their magnitudes change. We show that male students who mention more people from their academic community are associated with higher levels of productivity. University rankings are also found to be positively correlated with productivity and the size of academic support networks. However, neither university rankings nor students' productivity levels correlate with the sentiments students express in their acknowledgements. Our results point to the importance of academic support networks by explaining how they differ and how they influence productivity.

ÖZET

AKADEMİK DESTEK AĞI DOKTORA TECRÜBESİNİ VE VERİMLİLİĞİ YANSITMAKTADIR

ÖZGÜR CAN SEÇKİN

VERİ BİLİMİ YÜKSEK LİSANS TEZİ, TEMMUZ 2022

Tez Danışmanı: ASST. PROF. ONUR VAROL

Anahtar Kelimeler: ağ bilimi, metin madenciliği, doğal dil işleme, makine öğrenmesi

Akademik performansı üretkenlikle ölçmeye yönelik mevcut çalışmalar, doktora öğrencilerinin psikolojik refahlarına dair ciddi endişeler uyandırmaktadır. Bu çalışmalar genellikle araştırmacıların çevresinin etkisini ihmal etmektedir. Tez alt bölümlerinden biri olan Teşekkür, öğrencilerin kendilerini destekleyen kişilere teşekkür etmelerine olanak sağlayarak bu çevreye ışık tutmaktadır. Bu çalışmada, öğrencileri doktora sürecinde destekleyen beş farklı topluluğu ortaya çıkaran bir "akademik destek ağı" oluşturmak için 26.236 tezin Teşekkür bölümü analiz edilmiştir: Akademik, Yönetim, Aile, Arkadaşlar & Meslektaşlar ve Manevi. Kadın öğrencilerin aileleri dışında bu toplulukların her birinden daha az kişiye teşekkür ettikleri ve yayın sayılarına bakıldığında verimliliklerinin erkeklere göre biraz daha düşük olduğu görülmektedir. Bu kritik öneme sahiptir çünkü doktora sürecini incelemenin, kadınların akademik kariyerlerinin başlarında karşılaştıkları olumsuz koşulları daha iyi anlamamıza yardımcı olabileceği anlamına gelmektedir. Ayrıca, teşekkür edilen kişi sayıları disiplinler arasında değişmekte ve bu disiplinlerin "bireysel bilim" ya da "takım bilimi" olarak kategorize edilmesini sağlamaktadır. Bununla birlikte akademik topluluklarından daha fazla kişiye atıfta bulunan erkek öğrenciler daha yüksek verimlilik seviyeleri ile ilişkilendirilebilmektedir. Üniversite sıralamalarının ise üretkenlik ve akademik destek ağlarının boyutu ile pozitif ilişkili olduğu bulunmuştur. Ancak, ne üniversite sıralamaları ne de öğrencilerin üretkenlik düzeyleri, öğrencilerin teşekkürlerinde ifade ettikleri duyguların pozitifliği ile ilişkilendirilememektedir. Sonuçlarımız, akademik destek ağlarının nasıl farklılık gösterdiklerini ve üretkenliği nasıl etkilediklerini açıklayarak çevresel faktörlerin önemine işaret etmektedir.

ACKNOWLEDGEMENTS

Last two years were truly nourishing and this is thanks to the people who supported me throughout my journey. First of all, I want to acknowledge Dr. Onur Varol for his guidance, support and encouragement. I am indebted to him for his immense contributions to my academic and personal development. I may never express all of the appreciation for his advice and training as my advisor. If there were more people like him, academia would be a more preferred and more productive career path for so many people.

I want to thank to my fellow colleagues from VRL Lab and my other friends as well. So many people contributed to my projects intentionally or unintentionally by their valuable comments and proofreading my papers. They were also there for me when I needed emotional support and a helping hand. It is great to know that there are people who will help you when you stumble.

Lastly, acknowledgements to my wonderful family. My mother, who always supported and loved me unconditionally. Whatever I did would not have been possible without her. I would also like to express my gratitude to my extended family. They helped and embraced me on this pathway to become a scholar and for this, I am forever grateful.

Dedication page

I dedicate this thesis to my beloved family.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

In recent years, well-being and mental health concerns for PhD students have been increasing. According to a survey conducted in 2019 by Nature on 6,300 PhD students, 36% responded that they sought help for anxiety or depression caused by their studies (Woolston (2019)). Another devastating fact is that doctoral students are 2.43 times more likely to have a common psychiatric disorder than the rest of the highly educated population (Levecque, Anseel, De Beuckelaer, Van der Heyden & Gisle (2017)). It is therefore important to look through the journey of doctoral students not only through the lens of academic "success measures" such as publication numbers, citation counts, fellowships received *etc.* but also at their overall well-being and the quality of the environment that supports them in fulfilling their potential.

Although obtaining a doctoral degree is often viewed as an isolated process, it is a collaborative endeavor in which family, friends, colleagues, advisors, faculty, and administrative staff are directly or indirectly involved in and can influence the well-being of the students. At the end of the journey, students can show their gratitude by mentioning these names in their work through "acknowledgements" section of their dissertations. Acknowledgements, even though existed before, could not be found explicitly in the academic work before 1940s, and did not become a common subsection until 1960s (Bazerman & others (1988)). Hyland (2003) named acknowledgements as "Cindirella" genre because of its suffering from an undeserved neglect. Through time, these sections got longer and their use have become more prevalent (Cronin, McKenzie & Stiffler (1992)), making them more "insightful" in terms of understanding how and with whom doctoral students complete their journeys. Since there is almost no guideline or style guide to receive help when writing this section (Hyland (2004)), students have more freedom, compared to the other parts of their dissertations. Acknowledgements also serve purposes other than expressing gratitude, such as exhibiting associations with respected academics to display a special connection to which the author has been admitted (Scrivener (2009)). Thus, introducing their strategic decision in their professional identities by illustrating the

author in a positive aspect and governing their connections with the disciplinary community (Ben-Ari (1987)).

Acknowledgements contain such profound details of their authors' academic journey; however, research efforts to study how they vary concerning disciplinary and demographic differences have remained limited. Mantai and Dowling examined the type of social support that are provided for PhD students using 79 acknowledgements gathered from Australian universities (Mantai & Dowling (2015)). Hyland (2004) examined 240 acknowledgements of MA and PhD dissertations to characterize their narrative structure.

Using acknowledgement sections to delve into hidden networks outlined by the gratitude and appreciation expressed by students helps drawing conclusions that cannot be obtained from measures of academic success alone. For this task, we examined 26,236 PhD dissertations, obtained from *ProQuest Open Access Dissertations & Theses* database (PQDT-Open hereafter), 99% of which are from the United States in the last 20 years. In this thesis, we introduce a quantitative approach to examine the acknowledgements and how they are related to productivity levels of PhD students. It is noteworthy to mention that even though there are research on acknowledgements that examine their structure, to our knowledge, none of them uses such a comprehensive dataset or linking the productivity of PhD students with acknowledgements' characteristics. We aimed to shed light on the doctoral process by examining who is acknowledged, and how they are recognised from the perspective of students, using the tools of network science and natural language processing that enable research on large-scale data. Employing a community detection technique, we uncovered five support provider groups acknowledged by PhD students: Academic, Family, Administration, Friends & Colleagues and Spiritual. We revealed gender based and disciplinary differences when acknowledging support provider communities in terms of number of people mentioned and sentiment scores. We also investigated the factors derived from academic support networks influencing productivity levels. Lastly, we point out to linguistic differences between those who are located in the extreme cases of productivity and of sentiment.

This thesis is structured as follows: Chapter 2 provides background information by referring to the studies in the literature related to the subject. Chapter 3 describes the methodology of how we extracted information and how we processed it. Chapter 4 introduces our results and findings for descriptive and predictive analyses. Chapter 5 concludes the thesis, proposes ideas for future work, and explains the limitations.

2. LITERATURE REVIEW

In recent years, researchers started to search the ways in which the academic productivity can be predicted and explained. Many different data sources and methodologies were utilized for this task. Throughout this chapter, dissertation acknowledgements will be explained in further detail, the factors that have an influence on PhD students' productivity levels will be discussed and how it is possible to benefit the most from acknowledgement texts will be studied.

2.1 Dissertation Acknowledgements

Acknowledging social or professional environment and dedicating a work to the others in academia is a usual practice. It is widely practiced in PhD and master's dissertations under acknowledgements subsection as well. Acknowledgements, in general, are providing authors a place to express and develop social ties (Yang (2012a)) and to show gratitude to those who made contributions (Al-Ali (2010)). It is also an opportunity to represent the writers' own character and cultural norms, where thanking strategies differ from one culture to another (Cheng & others (2012)). Since we are more interested in the structural and quantifiable characteristics of acknowledgements, we focused on the works that utilized computational methods rather than qualitatively investigating them.

Hyland (2004) studied the acknowledgements subsection of 240 master's and doctoral dissertations authored by students at five Hong Kong universities in six different academic areas. They found that while 80% of masters theses had an acknowledgement subsection, this number was 98% in PhD dissertations. They argued that the difference is partly due to the fact that PhD students being already in the academic environment and knowing the common practices. On the other hand, master's students can be considered as part-time researchers who are less interested in tra-

ditions of academia and that some of which are expecting to work in another field. For example, only 30% of the Master of Business Administration (MBA) students had an acknowledgements subsection and they stated that these subsections should be brief and formal. They also investigated the structure of dissertation acknowledgements and showed that, in general, they have a "reflecting move" consisting of self-reflective commentaries on research experience, followed by a "thanking move" and an "announcing move" including the statements explaining responsibilities and inspirations at the end. They showed that in general, "thanking move" section, in which authors starts by presenting the participants, continues by thanking them for academic assistance (i.e. intellectual support, ideas), then for resources (i.e. technical, financial support) and lastly for moral support (i.e. friendship, patience).

Yang (2012b) investigated the dissertation acknowledgements of 120 PhD students to see if scientific disciplines had an impact on the acknowledgements in terms of their general framework and the language used to change thanking behaviors. They used Hyland's coding system for detecting the moves in acknowledgement structure and found similar results to those of Hyland's in terms of the "moves" included. They showed that between soft science (applied linguistics, business studies, and public administration) and hard science (medical science, electronic engineering, and biology) PhD students, there were modest differences in how they write. Also, both Hyland's and Yang's works demonstrated that, even though they were taken from different countries with people from different backgrounds, academic and emotional support are the main aspects that were appreciated by the authors.

Another quantitative approach is examining the stance expressions in dissertation acknowledgements (Chan (2015)) namely, how the authors express their point of view, attitude, or evaluation of the associated subject. In this study, they investigated 256 doctoral dissertations authored by Hong Kong Chinese students at three Hong Kong universities. They compared soft and hard sciences in terms of the frequency of modals, adverbs and complement constructions. According to the results, prediction/volition modals were used 21% more in hard sciences while ability/permission/possibility and necessity/obligation were shown to be utilized more frequently in the soft disciplines. Therefore, showing that there exists various differences in linguistic preferences among disciplines.

2.2 Productivity Among PhD Students

Measuring academic productivity is an important subject especially of science of science, which sheds light on the circumstances underlying creativity and the source of scientific breakthrough, with the aim of providing tools and policies that have the potential to accelerate science (Fortunato, Bergstrom, Börner, Evans, Helbing, Milojević, Petersen, Radicchi, Sinatra, Uzzi & others (2018)). Efforts to assess academic productivity generally focuses on some quantitative measurements such as citation counts, number of publications (Carpenter, Cone & Sarli (2014)), h-index (Hirsch (2005)) or depends on the various tools of bibliometrics. However, some of these measurements are criticized since they do not reflect editorship duties, mentorship efforts, and textbook authorship (Sarli & Carpenter (2014)). Moreover, evaluating the research quality of universities depend on main global indices such as Science Citation Index, Web of Science, or Scopus which make up only a small portion of journals that are endorsing the publications in English (Altbach (2015)). Although there are imperfections, the importance of evaluating the research performance should not be neglected since it determines the decisions on rewards, tenure, promotion decisions and staff recruitment (Blackburn & Lawrence (1995); Bland, Center, Finstad, Risbey & Staples (2006); Costas, Van Leeuwen & Bordons (2010)). Therefore, it is also crucial to examine which factors have an influence on academic performance.

In this paper, we are focusing on the PhD students' productivity levels, hence our aim is to give more details about how and by whom they are influenced from. Levecque et al. (2017) worked on 3,659 PhD students, who were drawn from a cross-sectional survey organized in 2013 to show what are the important aspects of these students that affect their mental health status. They showed that work-family conflict has been the most significant predictor of psychological distress and the probability of having a common mental disease. It was followed by job demands, family-work conflict, job control, and inspiring leadership style - which were all significant variables in the multivariate analysis. Meanwhile these last results do not associate productivity with mental health status, it is shown by Hysenbegasi, Hass & Rowland (2005) on undergraduate students that depression is linked with a 0.49 point lower GPA, indicating that it might be the case for PhD students as well.

Besides these factors, it is possible to think that advisor and university quality can influence the productivity since it plays a crucial role in research (Pearson & Brew (2002)). Working on recent graduated PhDs in economics, García-Suaza, Otero

& Winkelmann (2020) demonstrated that after controlling for additional variables such as gender and field of specialization, both advising quality and graduation institution rank are positively associated to academic production. In the meantime, although graduates from the top 25 universities seem to be more productive in terms of quality-adjusted production - where the number of publications are adjusted with the quality of the journal, students from non-top institutions can bridge this gap if they are mentored by the most productive professors in their programs. Another study examined the Swedish doctoral graduates in science, engineering, mathematics and medicine to better understand the linkage between the productivity and research environment of students (Broström (2019)). Evidence showed that students who were taught in groups with a smaller number of PhD students were more productive in the first five years of their academical trajectories. In addition, it is found that PhD students who were working in a funded group in which the research topic was constrained by the funder influence were more likely to remain in academia. It is also presumable that one's perception towards writing can be crucial for a researcher's productivity. Using a cross-sectional data collected from Spain and UK postdoctoral researchers, Castelló, McAlpine & Pyhältö (2017) studied the how their view the writing process influences productivity levels. They demonstrated that those who think writing as an activity that allows generation of knowledge were associated with higher productivity. On the other hand, those who showed high levels of perfectionism and reported that they were experiencing blocks during the writing process were linked with lower levels of productivity.

In the light of these findings, it is possible to state that while academic productivity should be encouraged and acknowledged, it should also be remembered that there are various aspects influencing it. Therefore, it is crucial to develop a deeper understanding of these aspects to help individuals improve themselves mentally and academically.

2.3 Word Embeddings and Community Detection

Word embeddings can be used to explain semantic and syntactic meanings of words in a machine readable format and to find the similarities between words. The standard method to represent words as bag-of-words vectors or distinct symbols is generally not enough to complete this task. A frequently practiced method for finding such representations relies on the distributional hypothesis of Harris (1954), claiming that words utilized in similar contexts have similar meanings. Word2Vec method is a good example since it leverages this idea to find word representations using a simple neural network. Given a window size of n words around a center word, the model predicts either the word in the middle by considering the neighboring words in the window or predicts the neighboring words by considering the word in the middle (Mikolov, Chen, Corrado & Dean (2013)). Doc2Vec model uses the same architecture with an additional document matrix involved in the neural network, which enables to discriminate one document to another or to find the similarities among documents (Le & Mikolov (2014)).

However, these are only a few examples from the methodologies to represent text data and many work demonstrated that these text representations can be successfully employed to detect communities or classify documents. Rashed, Kutlu, Darwish, Elsayed & Bayrak (2020) used Google's Convolutional Neural Network (CNN) based multilingual universal word encoder on Twitter data and expressed Twitter users in an embedding space for a stance detection model. Another work on social media data proposed an algorithm that builds word networks based on word embeddings to determine the discussion topics by employing community detection algorithms Mu, Lim, Liu, Karunasekera, Falzon & Harwood (2022). Therefore, it can be argued that document and word embeddings can help us discriminate different communities or topics since they are able to keep the contextual information.

3. METHODOLOGY AND DATASET

This section presents methodological steps followed in this work. Fig.3.1 demonstrates the overall framework of the study.

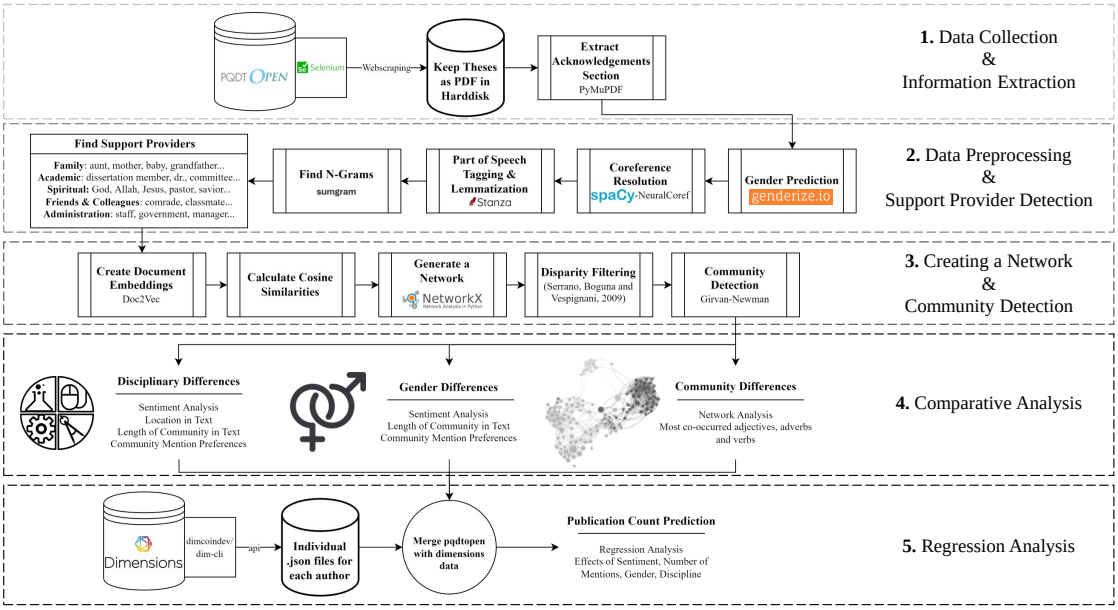

Figure 3.1 **The Complete Flowchart of the Study**

3.1 Data Collection & Information Extraction

To run a large-scale analysis of dissertation acknowledgements, we first needed a comprehensive and reliable database that includes the dissertations, as well as their metadata such as publication year, institution, thesis subject, and content of the study. Secondly, we had to extract the acknowledgement subsection of these dissertations and convert them into a machine readable format.

3.1.1 Dissertation data scraping

We retrieved data from `pqdtopen.proquest.com`, also called PQDT-Open, which offers open access to dissertations from institutions all around the world, most frequently located in the United States. Since PQDT-Open provides a publicly accessible data that allows our work to be reproducible, we chose to collect the dissertations by scraping the data directly from the website using Selenium library offered in Python.

We have gathered documents for 47,000 researchers, 26,264 of which are doctoral dissertations, written between 2000 and 2020. This collection of dissertation data also included metadata on dissertation abstract, title, author name, university, year of publication, page number, advisor name, department, subjects, and keywords.

3.1.2 Extracting acknowledgement text

To extract the acknowledgement subsection from the dissertations, we parsed raw data obtained in Portable Document Format (PDF). We used PyMuPDF library to extract textual information for each page, and then, we utilized a rule-based approach to identify pages that are likely to belong to acknowledgement subsection – e.g.,accepting the pages with the first word being "Acknowledgements," or ignoring pages that contain text such as "Table of Contents," "List of Figures," or "Appendix." When we encountered a page that starts with "Acknowledgements," we took the whole text from that page. If text in a page exceeds 1100 characters - which was the observed mean character limit for one page in our case, our system checked the next page if there is any sign of that page not belonging to an acknowledgement section. The system looked for at most four pages using the same heuristic explained above.

3.1.3 Obtaining number of publications

To obtain the number of publications of PhD students, we used the Dimensions Application Programming Interface (API) (Hook, Porter & Herzog (2018)) and matched the PhD students' records on Dimensions with the ones in our data. For

this task, we looked up the names, surnames and graduated university information of students, then we counted the total number of papers that they have published four years before and after their graduation. Out of 26,264 PhD students, we were able to find 3,989 individuals who published at least one paper in that time interval. We used this data as the dependent variable in our regression analysis to reveal how the characteristics of acknowledgements (sentiment scores, number of people mentioned etc.), gender and discipline category are related with the publication counts.

3.2 Data Preprocessing

Our analysis relied on individuals and institutional offices supporting a doctoral student. We analyzed acknowledgement texts to determine these entities and extract information on how they are referred to and how they are being acknowledged. We then constructed an "academic support network" based on their contextual similarities.

3.2.1 Text Preprocessing and Support Provider Detection

In the acknowledgement section, support providers are introduced by their names and affiliation, but they can also be referenced in a series of consecutive sentences by personal pronouns. We use Coreference Resolution, to identify pairs of references that are associated with the same entity. Coreference resolution approach helped us to link the support providers with all other related words and was the first step in the text preprocessing.

Next, we lemmatized and applied POS tagging to the text by using Stanza library to identify different functional words (Qi, Zhang, Zhang, Bolton & Manning (2020)). Since some POS types such as punctuation, numerals, and conjunctions are not relevant to our analysis, we built our models by using only the words that are tagged as adjective, adverb, verb or noun.

In the last step, we employed Sumgram library to identify n-grams that are used regularly in documents, therefore helping us to find some of the word groups that might indicate a support provider, such as "thesis advisor" or "sounding board"

(Nwala, Weigle & Nelson (2018)).

By using the corpus that we created after preprocessing the data, we employed both manual and data-driven approaches to find the support providing entities within text. Firstly, we created a corpus of word and n-grams that have been found in acknowledgements, we went over the word or word groups that have more than 50 occurrences by manual inspection to clarify whether these tokens belong to a support provider such as mother, girlfriend, father, colleague, thesis advisor, God etc.

To complement with the previous approach, we used *Word2Vec* and *Sumgram* to identify the remaining support providers. Once we created a list of support providers from manual inspection, we examined 10 most similar words for each support provider and included them to the collection if needed. This hybrid process helped us to minimize any bias when determining the words and phrases that are being examined in this study. Among 155 words support providers identified in this analysis, we removed the ones that occur less than 50 times in acknowledgements, leaving us 144 support providers which can be further examined under Table A.1. While the number 50 is selected arbitrarily, it is important to mention that 50 dissertations make up only 0.02% of our dataset, which indicates that we are including the ones that are available enough to make inferences on. Also note that since we employed computational methods to represent these words in vector format, we would obtain poorer results as we decrease the number of minimum dissertations since these methods are data-driven.

3.2.2 Creating a Network & Community Detection

We created document embeddings for support providers by using Doc2Vec model trained on the dissertation acknowledgement corpus explained above (with only adjectives, adverbs, verbs and nouns). Each sentence that a support provider mentioned is tagged with the related support provider and these tagged document vectors are given as inputs to the model. We used the default parameters in the model, which provides 100 dimensional vectors for each support provider. These vectors are then used as feature vectors for support providers.

After learning the vector representations for support providers, we created a network representation where nodes correspond to support providers and their sizes are determined by their number of occurrences. The edge weight between each pair of nodes is defined as the associated cosine similarity between vectors representations

learned from Doc2Vec model (Le & Mikolov (2014)).

Support providers network will be fully connected since edge weights are computed using cosine similarities between dense vector representations. We filtered edges using disparity filtering (Serrano, Boguná & Vespignani (2009)) to focus only on the statistically significant associations. Resulting network gave us the sparsest network possible with 144 nodes and 759 edges instead of 10,296. Subsequently, to detect the communities in the network, we used Girvan-Newman algorithm (Girvan & Newman (2002)) and observed the formation of the 5 communities as given under Table A.1.

3.2.3 Methodological Toolkit

In this section, we will briefly explain the text analysis, natural language processing and clustering tools that we have used throughout the paper.

3.2.3.1 Tokenization, Part of Speech Tagging (POS) and Lemmatization

Tokenization is the process of identifying meaningful units of a sentence, such as words, numbers, or punctuation. This gives us the capability of labelling the words in a text to a specific part of speech, regarding their definition and context (e.g., verb, noun, preposition) which is called part of speech (POS) tagging. At last, after detecting each word in the sentence, we can take off their inflectional suffixes, and then only consider the dictionary form of the word, which is called the lemmatization. E.g., running, ran and run will be all transformed into "run." This process helps us to reduce the number of words and to detect the entities that are associated with the support providers while keeping the integrity of the meaning in texts. For these processes, we used the pretrained, neural network based natural language processing package Stanza (Qi et al. (2020)), offering a state-of-the-art performance on various NLP tasks.

3.2.3.2 Sumgram

Adjacent sequence of N number of words is called an N-gram. E.g., “nice picture” is a 2-gram (or bigram), and “beautiful Saturday morning” is called a 3-gram.

While N-grams can be useful for several different tasks, our goal is to catch the words that together represent meaningful noun phrases.

Sumgram library generates n-grams by combining other several lower-order n-grams (e.g., by conjoining “thesis committee” and “committee member” we get “thesis committee member”). It leverages Stanza’s POS tagging module and takes into account the inter document term frequencies for n-grams. The algorithm tries to find and replace the base n-gram fragment with its multi-word proper noun (MWPN) (e.g., "emergency management" is the first n-gram that is found by the algorithm, "federal emergency management agency") by comparing the frequency of fragment child n-gram with the term frequency of MWPN.

3.2.3.3 Coreference Resolution

Finding and pairing all textual references that are associated with the same entity is called coreference resolution. Since we seek to pair all the support providing entities with the other words that are used in the same sentence, changing the personal nouns into entities themselves helps to find all the pairs, which in turn allows us to make use of the entire acknowledgement. In traditional models, coreference resolution is done by manually designed scoring algorithms (Bagga & Baldwin (1998)), which may cause a variation in performance across different languages and datasets (Clark & Manning (2016b)).

For this task, we used neuralcoref module in Python (huggingface (2021)). The module firstly leverages a rule-based mentions-detection module of spaCy Library which distinguishes tokens and then detects named entities and part-of-speech tags (POS tag) to uncover the possible coreference mentions in a text. Secondly, it uses a feed-forward neural network to compute a coreference score for each of these possible mentions. This feed-forward neural network model is pre-trained on the manually labeled OntoNotes Corpus dataset (Weischedel, Hovy, Marcus, Palmer, Belvin, Pradhan, Ramshaw & Xue (2011)), and uses a two-fold scoring system where one neural network tries to find the antecedent of a given entity while the other one is forced to understand whether there is an antecedent at all or not (i.e.,

whether the given entity is mentioned for the first time or not) (Clark & Manning (2016a)). An example of how the model links an entity with a personal pronoun is given in Fig. 3.2.

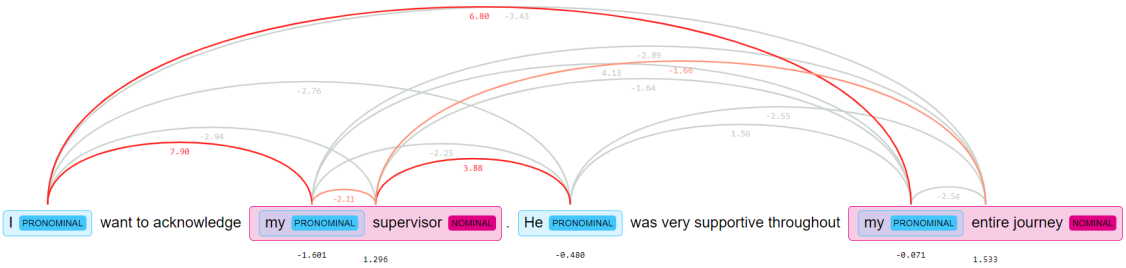

Figure 3.2 **Coreference Resolution** An example of how the model links entities with pronouns.

3.2.3.4 Word & Document Embeddings

Word embeddings are vector representations of words which are capable of keeping their semantic and syntactic properties. The idea here is that words that occur in the same contexts tend to have similar meanings (Harris (1954)). Therefore, these vectors allow us to find the similarities in a numerical format between words. Word2Vec model uses a shallow feed forward neural network with only one hidden layer and an output layer with Softmax activation function. While Word2Vec has two architectures as continuous bag-of-words (CBoW) and Skip-gram, we used the first one since it provided better results for our case. In this approach, during the training phase, the vector representations of surrounding words are used to predict the word in the center.

We used Word2Vec model to check the similar words that we labeled as support providers so that we can make sure that we are including every support provider in the corpus.

Since our aim is to represent each support provider by using each word (verb, adverb, adjective and noun) in the same sentence as features (not only the n number of words that surround the support provider entity), we employed the Doc2Vec model for revealing the latent representations of support providers as document vectors. Doc2Vec model is very similar to the Word2Vec model and uses nearly the same architecture. It uses an extra document matrix that is learned throughout the backpropagation process.

We tried three different approaches to challenge the robustness of our results and decided to continue with Doc2Vec because of the following three reasons:

(1) It is proven to be capable of capturing the context of documents (in our case, sentences that include support providers) and shown to be more successful than traditional methods such as bag-of-words or term-frequency inverse-document-frequency (TF-IDF).

(2) It is much more memory efficient than TF-IDF and bag-of-words method. While all the support providers are represented as 100-dimensional vectors with the Doc2Vec method, TF-IDF uses the whole words in the corpus – even though we remove the words that have occurred once or twice in the dataset, there are 15 thousand dimensions left for each support provider.

(3) Since Doc2Vec is not pre-trained and learns the word representations directly from the dataset, its outputs are context-dependent and well-crafted for our task. Therefore, we did not use any pre-trained, neural network based models such as GloVe Pennington, Socher & Manning (2014).

3.2.3.5 Sentiment Analysis

We used a pretrained BERT model (Devlin, Chang, Lee & Toutanova (2019)) for sentiment analysis toolkit "bert-base-multilingual-uncased-sentiment" for this task.[1]

When running analyses on acknowledgements, we considered each sentence separately, which gave us the opportunity to link the support providers with their associated sentiment scores. After forming the communities, we aggregated the sentiment scores of the support providers. Moreover, to see whether our model's results hold if we were to use other algorithms based on different architectures, we compared the BERT scores with those of Vader's (Hutto & Gilbert (2014)). We observed that the models provide similar results where the scores for Spiritual, Family and Friends & Colleagues communities are higher than those of Academic and Administration communities. Therefore, we showed that while PhD students talk positively about each of their support provider communities on average, the positiveness alters among communities; regardless of the model that is being used to assess the sentiment scores. We choose to use the BERT model since it is the state-of-the-art architecture and is proven to deliver better results than older models.

[1] `https://huggingface.co/nlptown/bert-base-multilingual-uncased-sentiment`

3.2.3.6 Disparity Filtering

Disparity filtering is a network reduction technique, which can be regarded as the creation of a network with fewer edges, allowing us to uncover the relevant features of the original network.

We employ the disparity filtering algorithm introduced Serrano et al. (2009) to extract the reduced representation of our network while keeping the key connections we want to highlight. Disparity filtering is based on the p-value statistical significance test of the null model. Given $s_i = \sum_j w_{ij}$ where w_{ij} is the weight of the link between nodes i and j:

$p_{ij} = w_{ij}/s_i$

Disparity filtering uses normalized weights p_{ij} for each node to avoid under representing small-scale interactions along with a k parameter that represents degree of nodes to decide whether an edge is statistically significant or not. The edges with $\alpha_{ij} \leq \alpha$ reject the null hypothesis and can be considered as significant linkages between node pairs:

$a_{ij} = 1 - (k-1)\int_0^{p_{ij}} (1-x)^{k-2} dx < \alpha$

Since our vertices stand for the support providing entities and edges are computed by cosine similarity between each node pair (which are represented as 100x1 vectors), all the pairs are densely connected to each other, i.e. there is no such pair that are orthogonal to each other. Therefore, we remove 92.7% of the edges using the algorithm as shown in Fig.3.3. We choose $\alpha = 0.255$ since it provided the network with all the nodes with the minimum number of edges.

3.2.3.7 Community Detection

To uncover the communities that PhD students are mentioning in the acknowledgement section, we used the doc2vec network that is reduced by disparity filtering as the data. For this task, we used the Girvan-Newman Algorithm (Girvan & Newman (2002)).

Girvan-Newman Algorithm leverages the edge betweenness measure, which is computer by counting the number of shortest paths passing through an edge. It uses an iterative approach by following the steps below:

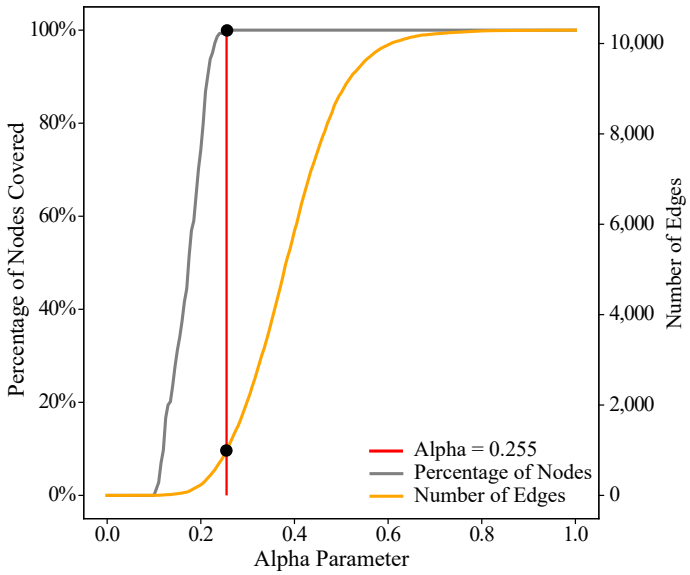

Figure 3.3 **Determining Alpha Value for Disparity Filtering** Percentage of Nodes Covered and Number of Edges for Associated Alpha Threshold Values in Disparity Filtering

(1) Compute edge betweenness for each edge in the network.

(2) Eliminate the edge with highest edge betweenness.

(3) Calculate the edge betweenness for the remaining each edges.

(4) Repeat steps 2 and 3 until no edge remains.

The idea is to remove the edges that link communities together, which intuitively have the highest edge betweenness. Therefore, since communities will be separated from one another, the underlying community structure of the network will be uncovered.

However, to check the validity of the communities, we performed a similar analysis using hierarchical clustering and obtained the results as shown in the Fig 3.4, which provides 5 clusters that are significantly overlapping with the results obtained using the Girvan-Newman Algorithm.

3.2.3.8 Bootstrapping Method

We employed bootstrap sampling approach to estimate mean and confidence interval of sentiment scores over a population. We repeated this procedure 5000 times for each estimation. We have randomly drawn 5% of the population with replacement

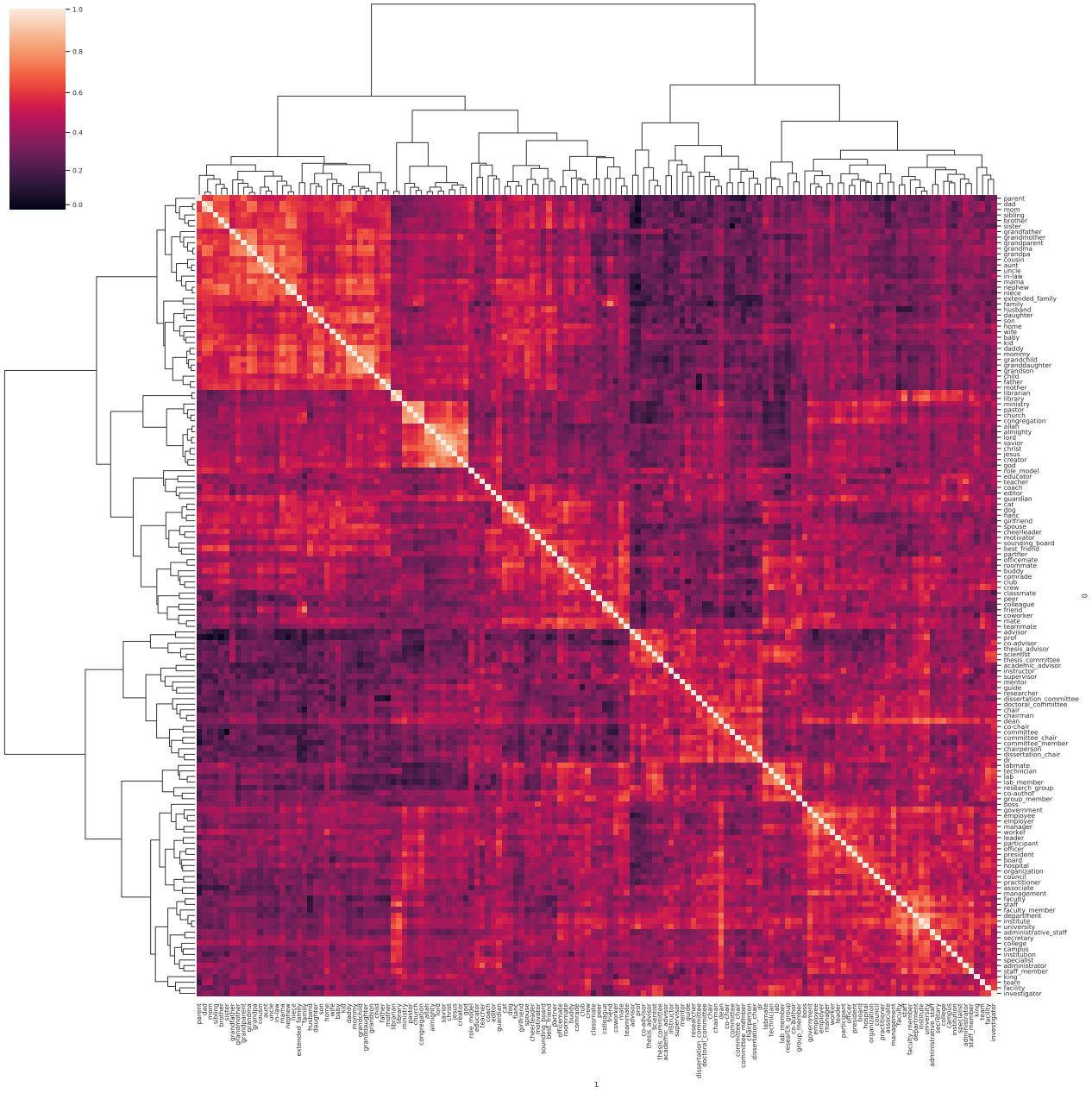

Figure 3.4 **Detecting Communities with Hierarchical Clustering** Support providers vector similarities using hierarchical clustering with complete linkage and dendograms computed following agglomerative approach.

and calculated the mean of the sample. Using these mean values, we ran two-tailed T-tests to see whether difference between two means is significant or not ($p \leq 0.05$, $p \leq 0.01$, $p \leq 0.001$ and thresholds are denoted by "*", "**" and "***", respectively).

3.3 Gender Inference & Gender Based Differences

Since information about authors' gender is not provided in the metadata provided by PQDT-Open, we inferred gender information using first names of the authors. For this task, we used the online service named `genderize.io`. It relies on a database that has over 110 million entries from 242 countries to examine whether a name is more frequently used amongst females or males.

We used this data to determine whether there are differences between females and males in terms of acknowledging the support providers as percentage of mentions, number of people acknowledged, and sentiment reflected towards a particular support provider group. We checked whether these results hold for disciplines as well. Our results are robust with some minor differences from one discipline to another (see Figures A.1, A.2, A.3, A.4 and A.5).

3.4 Determining Discipline Categories

Although there is no categorization agreed in the literature and guideline or consensus on how research fields should be classified, it could be done by following previous research efforts. Hence, to have a clearer view of disciplinary differences, we divided the subjects into 5 categories as: Biology & Health Sciences, Life & Earth Sciences, Mathematics & Computer Sciences, Physics & Engineering and Social Sciences & Humanities (Lamers, Boyack, Larivière, Sugimoto, van Eck, Waltman & Murray (2021)). To assign each subject into one category, we separately labeled each subject with a discipline and reached to an 85% agreement and 0.75 Cohen's Kappa score for inter-annotator reliability (Cohen (1960)) (see SI - List of Discipline Categories for the detailed list of categories).

3.5 Regression Analysis

A regression analysis is used to estimate the effect of independent variables on the target variable. We used generalized linear model (GLM) with Inverse Gaussian distribution since linearity and normality assumptions do not hold in our case. The histogram that shows the publication counts and Inverse Gaussian distribution can be seen on Fig.3.5(a). A QQ-plot is also given on Fig.3.5(b) to motivate our preference using this model.

After selecting the regression model, to detect multicollinearity and select the variables that are going to be used in regression analysis, we checked the variation inflation factor (VIF), which is calculated as $1/(1-R_j^2)$ in which R_j^2 denotes the coefficient of determination derived by regressing the j^{th} predictor on the remaining predictors. We removed those which had higher than 10 VIF score. The remaining covariates were included as given in the Fig.3.5.

Note that the number of publications of PhD students were counted for four years before and after their graduation. To ensure that these results are robust to the time interval, we ran the regression analysis for the publication counts of three, five and six years before and after the graduation. Our results showed that the results demonstrated on Table 4.1 do not differ significantly.

3.6 University Rankings

We analyzed rankings of university affiliations of doctoral students with their performance and academic support networks. However, since nomenclatures for same universities may change from one database to another, we had to match ranking lists and university names in our PhD acknowledgement database manually before conducting a research on this part. We investigated three widely used world ranking lists for universities: Center for World University Rankings (CWUR), Times Higher Education (THE), and Quacquarelli Symonds (QS), which included 2000, 1663 and 1000 universities, respectively. As can be seen on Fig.3.6(a-c), pairwise correlation between these three lists can be considered as at least moderate with a minimum value of 62% for CWUR-QS pair. Therefore, we have concluded that they are consis-

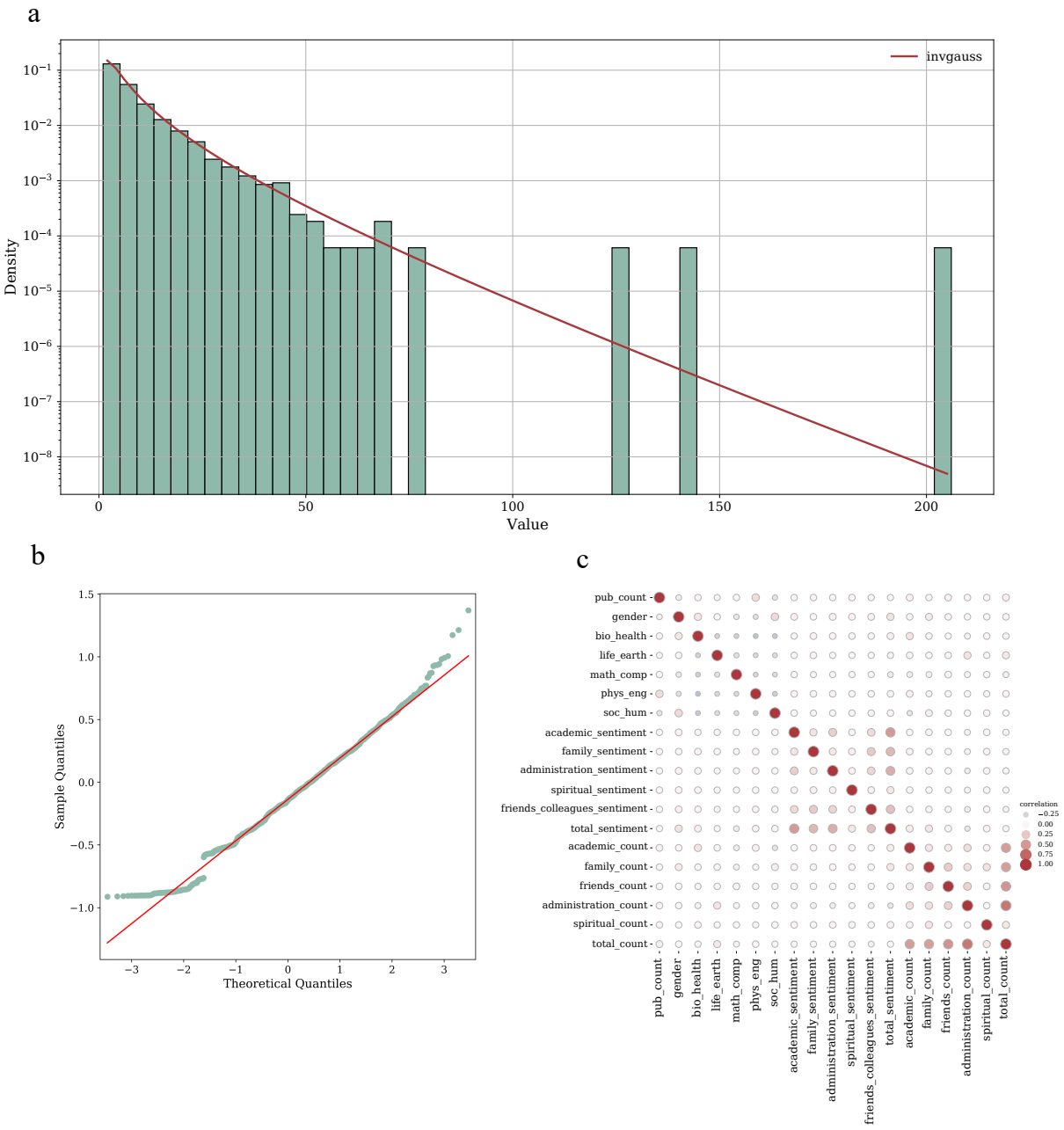

Figure 3.5 **Distribution Identification and Independent Variable Selection** Ten different distributions were fitted on the Publication Count histogram and the best distribution was plotted (a). Before checking the variation inflation factor (VIF), the correlations between independent variables were visually represented (b). While the points fall along a line in the middle of the graph, they curve off in the extremities (c).

tent enough to make an analysis on. Hence, we checked whether they are correlated with the number of publications or not. Fig.3.6(d-f) shows that universities with higher rankings have on average PhD students with higher productivity levels.

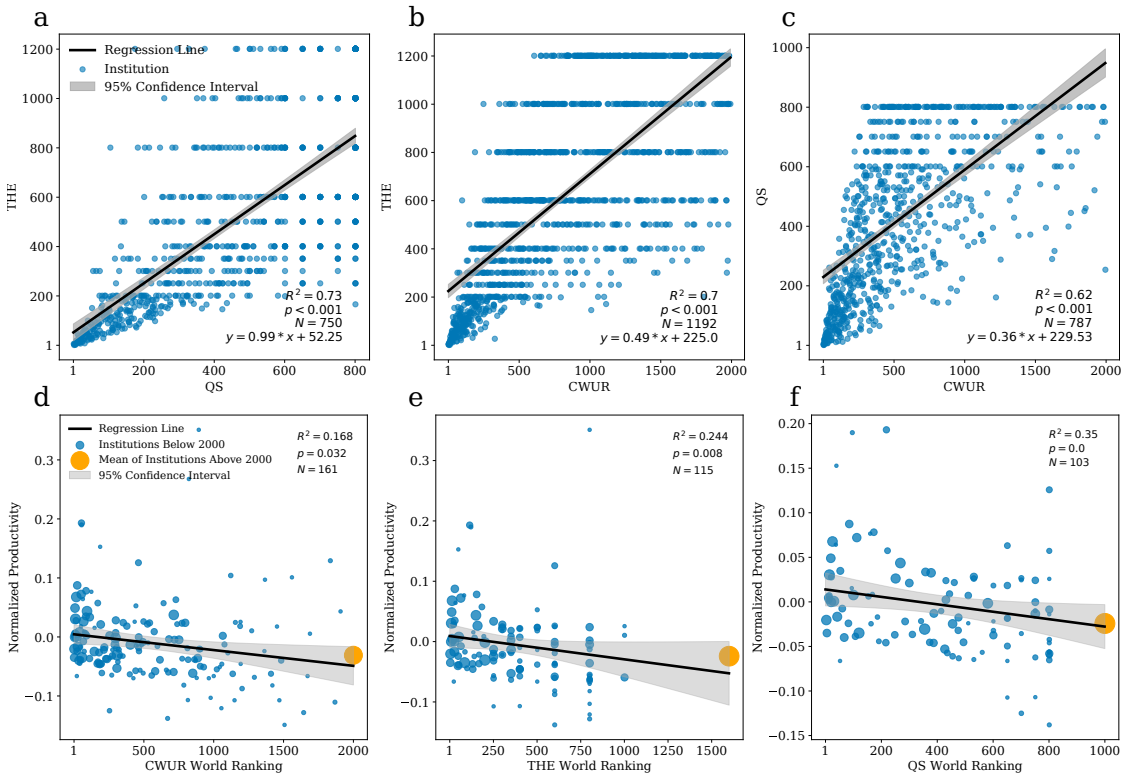

Figure 3.6 **Comparison of Different University Ranking Lists and Correlation with Productivity Levels** Relationship between QS and THE rankings (a), CWUR and THE rankings (b), CWUR and QS rankings (c). Relationship between normalized productivity and CWUR (d), THE (e), QS (f). N represents the number of institutions that match for both lists while the equations are associated with the given regression lines in figures. R squared values denote Spearman's rank correlation.

Our results have shown that number of academic mentions and total mentions have a positive correlation with university rankings for each of the ranking lists as well. And, when we normalized the number of mentions regarding students' gender and discipline, the analyses yielded the same results as shown in Fig.A.6, Fig.A.7 and Fig.A.8.

We also examined whether there is a significant relationship between university rankings and sentiment scores. The results have shown that none of the sentiment scores can be explained by university rankings (see Fig.A.9, A.10, A.11).

4. RESULTS

4.1 Characterization of a Support Network

The acknowledgements section of dissertations contains statements about individuals or institutions who have provided emotional, economic, and administrative support to students on their journey towards attaining their degree. To systematically identify acknowledged individuals and institutions, we used a data-driven approach supported by manual inspection to identify distinct types of support providing entities in the acknowledgements.

To build the academic support network, we extracted different individual roles and institution types as nodes from each text and computed contextual similarities learned from the text as edge weights (Fig.4.1(a)). Our entity extraction approach identified 144 support providers that were mentioned in at least 50 dissertations. We used a deep learning approach, called Doc2Vec (Le & Mikolov (2014)), to learn embeddings for each support provider within the context they were used in the dissertation corpus. Using these embeddings, we calculated pairwise similarities and used them as edge weights for each node pair. We then used disparity filtering and retained only the statistically significant edges (see SI Appendix, Section 2), providing us the network capturing significant relations between these entities. Next, we employed Girvan-Newman (Girvan & Newman (2002)) algorithm for community detection to identify groups of support providers.

The network representation of all of the support providers is given in Fig.4.1(a). Community detection analysis identified 5 distinct communities in this network and each of them is illustrated by a different color: Spiritual (purple), Academic (yellow), Administration (gray), Family (blue) and Friends & Colleagues (green). These communities are consistent with those identified with other clustering approaches

such as hierarchical clustering (see SI Appendix, Section 2). Node sizes were determined by the occurrence of support providers and the edges were weighted with cosine similarity of embeddings between node pairs.

Connectivity among these communities reveals separation between social and professional networks. Friends & Colleagues are located among Family, Academic and Administration communities. Some dissertations also refer to spiritual entities and community consisting of these entities is loosely connected with the rest of the network and has few links with the family community. Factors influencing community relations can be explained by comparing the words that are used to acknowledge these communities. By analyzing bipartite connections of support providers and prominent words, as seen in Fig.4.1(b), we present the most frequent 20 words used for support providers in these communities. While four of the most widely used words for acknowledging Spiritual support providers are not linked with the other communities; words like thank, acknowledge, and grateful used approximately at the same rate for each group.

To further support rank order in acknowledgements, we checked the locations of the support providers in the text and observed that different communities can be distinguished by analyzing the locations in which they are frequently mentioned (Fig.4.1(c)). While academic support providers are most frequently mentioned at the beginning of the acknowledgements, students tend to start talking about their families towards the end. We also observed that Friends & Colleagues and Administration are generally mentioned in the middle of the text. In contrast, Spiritual entities are mentioned either at the beginning or at the end.

Although acknowledgements are expected to have an overall positive sentiment, certain entities receive more formal tone. To highlight these subtle differences, we explored the interplay between sentiment scores and how many times they are mentioned for each category separately as shown in Fig.4.1(d). The sentiment analysis results have shown that Spiritual, Family and Friends & Colleagues communities are being acknowledged roughly at the same level with average sentiment scores 4.33, 4.33, and 4.31 defined in $[0,5]$ range. Academic and Administration communities have lower scores with on average 4.24 and 4.17, respectively. Similarly, we analyzed the number of people mentioned from these categories. Not surprisingly, PhD students tend to acknowledge the Academic community the most (9.05 people per acknowledgement), it is followed by Administration (6.76), Family (5.44), Friends & Colleagues (4.16), and Spiritual (1.97). Families, friends, and spiritual figures generally do not involve in research as workforce; however, they provide emotional and financial support to make life easier for doctoral students and are a crucial part

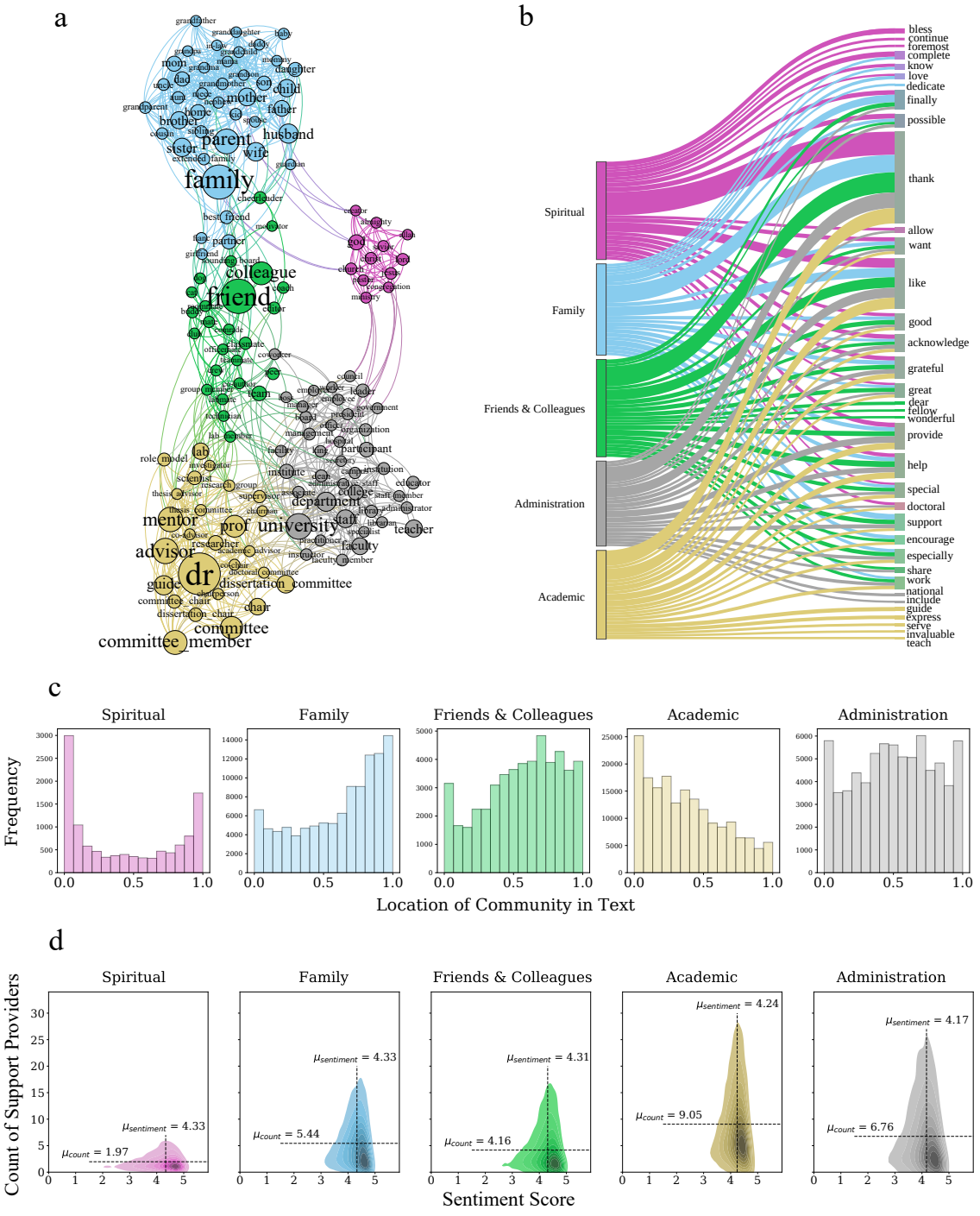

Figure 4.1 **Analyzing support providers in acknowledgements.** Different support providing entities identified in dissertation documents represented as nodes and their contextual similarities learned from document embeddings used as edge weights (a). Community detection revealed 5 distinct groups: Family, Friends & Colleagues, Academic, Administration and Spiritual. These groups are acknowledged using specific words and bi-partite relation points group specific properties (b). Location of mention in the acknowledgement text indicates norms among scholars to highlight distinct groups (c). Support providers also differ in terms of their occurrence in acknowledgements and the corresponding sentiment they are referred (d).

of the support network deserving an appropriate mention.

4.2 Gender based differences

PQDT-Open provides metadata on universities, authors, and committee but lacks details about demographics of the authors such as gender of the author. We inferred this information using the names with a widely used public API and examined the differences between genders in terms of their academic support networks (see see SI Appendix, Section 3). Previous work studied how female and male students acknowledge support providers both in quantitative and qualitative terms. Alotaibi (2018), using Metadiscourse, studied 120 dissertation acknowledgments written by Saudi students studying in the U.S. and revealed that while all male and female students acknowledge their academic environment, there exist differences when thanking God, resources and moral support. It is also shown that women in academia have less access to powerful social networks and inter-personal bounds that provide resources and create other advantages, which limits their opportunities to achieve their goals (Casad, Franks, Garasky, Kittleman, Roesler, Hall & Petzel (2020); Collins & Steffen-Fluhr (2019)).

Fig. 4.2(a) shows the ratio of students who mention the respective support provider community at least one time in acknowledgements. We observed that female students are slightly more likely to thank each community at least once except for the administration. The largest gap is observed in Friends & Colleagues category where the difference is 5.6% between males and females. This is followed by a roughly 4% difference for the family members. These results differed when we examined the number of mentions instead of percentage of mentioners; number of people acknowledged from each category is higher in male students except for the family members. In fact, the highest difference is observed in Academic, Administration, and Friends & Colleagues groups, which may imply that women have limited access to their academic advisors, administrative staff, and peers.

Besides the occurrence rates, we analyzed the sentiment of the language used for different support providers. Females are inclined to express more positive sentiment towards the ones that help them through their journey. Meanwhile, the gender gap between Friends & Colleagues community seem to be highest; the difference is narrower for the Spiritual characters, but still significant despite the large variance

of the distribution.

Looking from an overall perspective, it is readily apparent that females tend to thank their families both qualitatively and quantitatively more compared to males. This is in line with the existing work on dissertation acknowledgements showing that while both men and women appreciate social support evenly, they highlight different aspects of it; men value companionship and collegiality, women note emotional support (Mantai & Dowling (2015)). Taken together, this may be an indicator of the level of importance of families for females and lack of professional support from the other communities during their doctoral journey.

4.3 Disciplinary Differences

Each academic discipline has different research practices and collaborations. These differences are also reflected by mentorship styles and academic environment. To analyze these disciplinary differences, we manually categorized the subjects given in the PQDT-Open metadata into five main category of academic disciplines (Biology & Health Sciences, Life & Earth Sciences, Mathematics & Computer Science, Physics & Engineering Sciences, Social Sciences & Humanities) (see see SI Appendix, Section 4). Fig.4.3(a) shows the subject co-occurrence network and categories labeled as different disciplines.

We argued that while dissertations on certain subjects may be considered as individual work and require less academic collaboration and administrative support, other subjects might require cooperation, teamwork, and access to resources and field work. We calculated the number of support providers mentioned for each subject and presented in Fig.4.3(b). While Social Sciences & Humanities students mention the least number of people with 23.14 on average, this number is the highest with 37.87 for Life & Earth Sciences students. Number of support providers mentioned for each discipline aligns with academic norms of individual and team science as shown in the literature (Fortunato et al. (2018)). Here, we measured not only size of academic groups, but also other support provider categories.

Moreover, it is also reasonable to presume that different disciplines might have different preferences in terms of acknowledging the support provider categories. While the results show that there is a small gap between occurrence rates, this ratio varies the most for Administration and Spiritual communities (see Fig.4.3(c)). Mathemat-

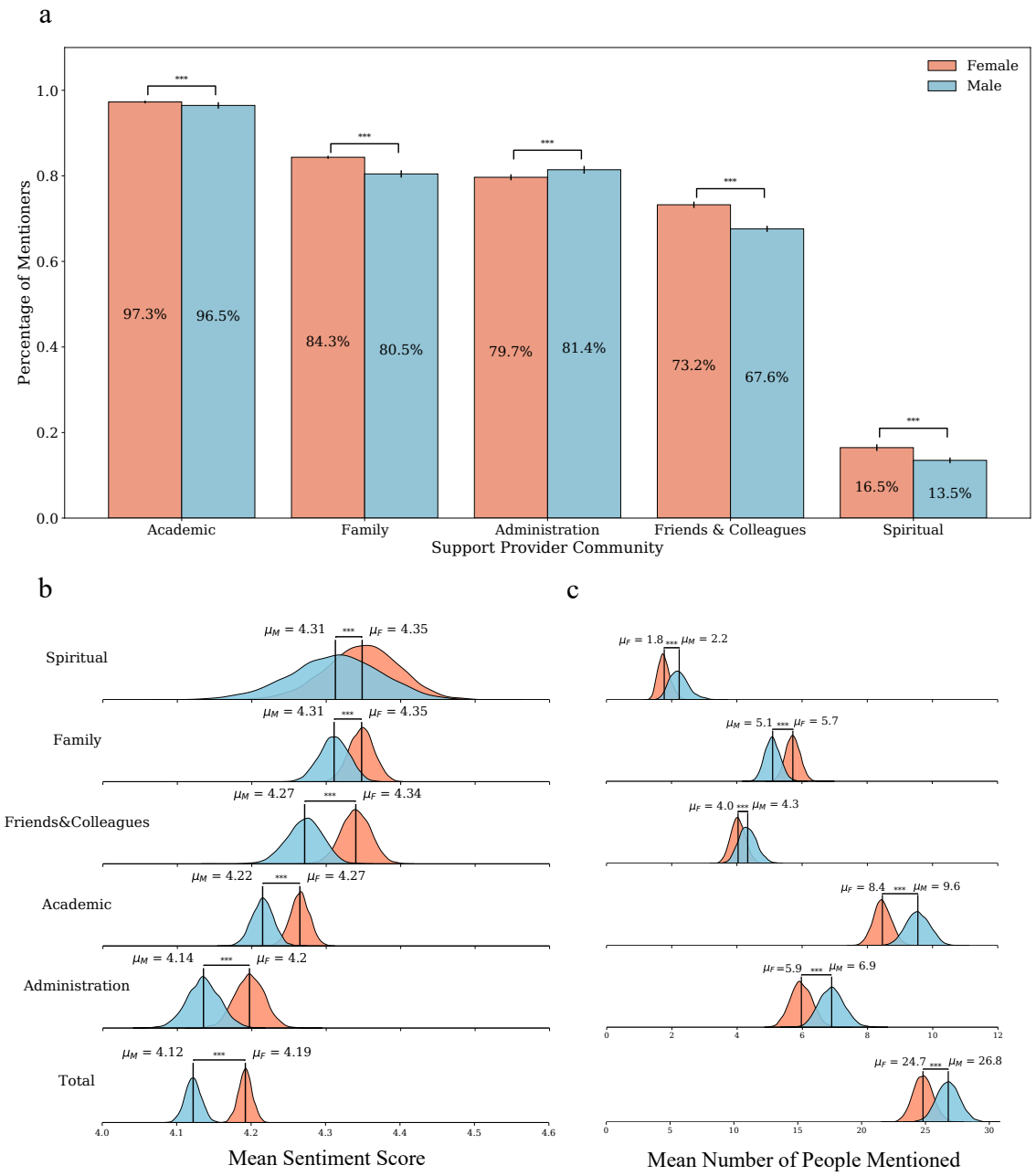

Figure 4.2 **Gender differences in support provider communities.** Ratio of students referring to different categories of support providers varies across genders (a). Sentiment scores differ when mentioning the support providers (b). Mean number of people mentioned alter across genders for different support provider categories (c). Individual groups were compared using the two-sided t-test. ***, $p \leq 0.001$; **, $p \leq 0.01$; *, $p \leq 0.05$

ics & Computer Science students seem to mention their families, friends, colleagues, and the administration less than other disciplines. Additionally, almost one fifth of Social Sciences & Humanities students acknowledge spiritual characters, which may be explained by dissertation studies in religion and relevant fields (e.g. "Biblical Studies", "Islamic Studies") covered under this discipline. We also investigated the gender ratios in these disciplines and, consistent with the past work (Huang, Gates,

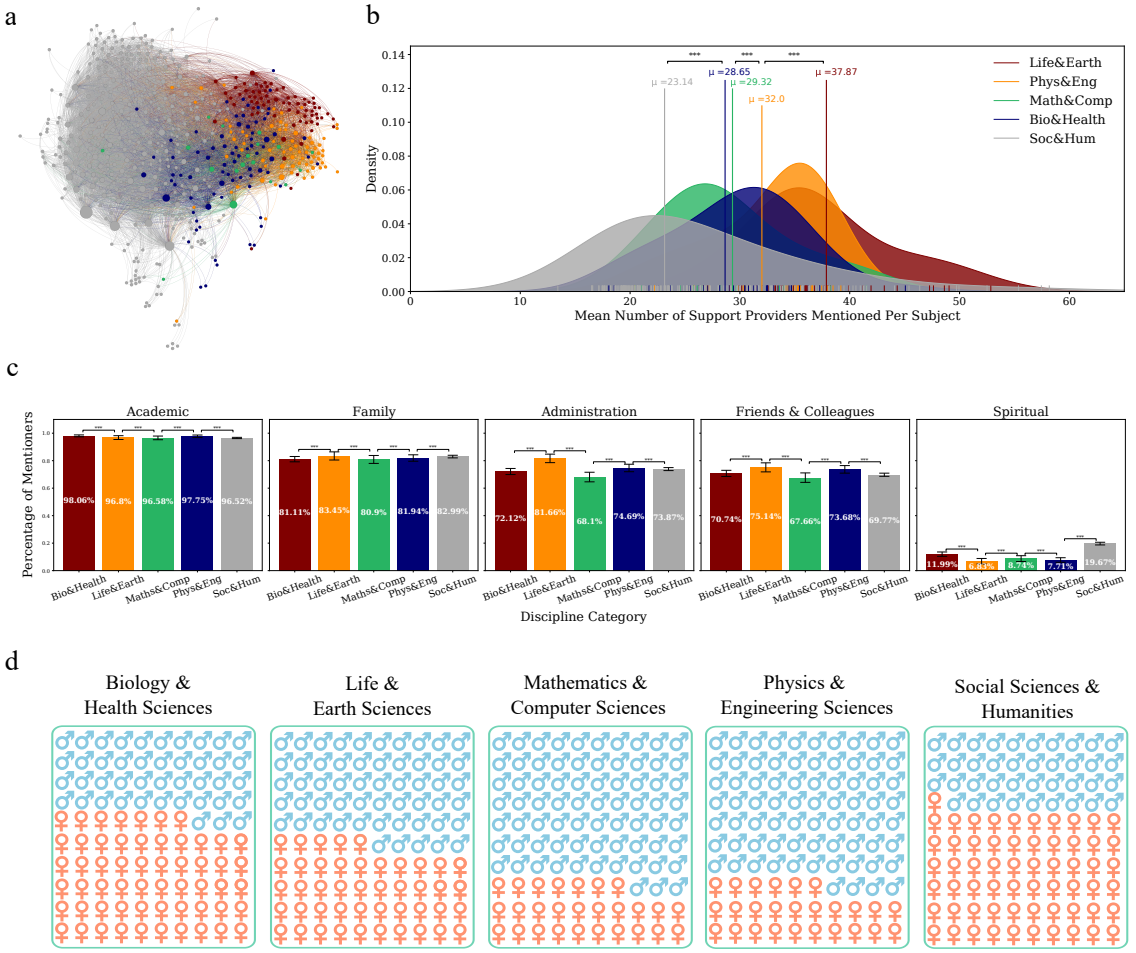

Figure 4.3 **Disciplinary differences.** Dissertation subjects are represented as nodes and the edges are formed by number of co-occurrences in the same document. This subject network reveals classification of field and the corresponding disciplines (a). Average number of support providers differ by disciplines ranging from individual to team sciences (b). Inclination to mention different support categories slightly diverge based on the discipline (c) and gender distributions observed to vary across disciplines (d). Individual groups were compared using the two-sided t-test. ***, $p \leq 0.001$; **, $p \leq 0.01$; *, $p \leq 0.05$

Sinatra & Barabási (2020); Way, Larremore & Clauset (2016)), we observed that female students are underrepresented in STEM fields. Female ratio is the lowest for Physics & Engineering Sciences with only 26%, which is followed by 27% in Mathematics & Computer Sciences. However, the majority of students in Social Sciences & Humanities are women, with a rate of 71%. These outcomes are also supported by previous work on intersectional inequalities, where it is demonstrated that there is homophily between identities and subject of research (Kozlowski, Larivière, Sugimoto & Monroe-White (2022)).

4.4 Social Determinants of the Academic Productivity

Research on academic performance and success focuses on metrics that are easy to quantify, accessible for research, and standardized across disciplines (Fortunato et al. (2018)). Efforts on quantifying academic performance at individual and group levels use productivity measures such as number of publications and impact indicators like citation counts (Sinatra, Wang, Deville, Song & Barabási (2016); Wu, Wang & Evans (2019)). Impact of academic mentorship and institutional quality for academic growth have been recently studied by using these bibliographic indicators (Ma, Mukherjee & Uzzi (2020); Sekara, Deville, Ahnert, Barabási, Sinatra & Lehmann (2018); Way et al. (2016); Way, Morgan, Larremore & Clauset (2019)).

We want to investigate academic productivity by utilizing the social aspect of doctoral studies. We investigated the publication records of students obtained from an online service, called Dimensions, for 2824 former doctoral students (Hook et al. (2018)). By conducting a regression analysis, we analyzed the role of academic support network by considering number of mentions and sentiment scores of acknowledgements to estimate publication count during the doctoral studies while controlling for gender variable. We employed an Inverse Gaussian regression model to estimate the parameters and their significance, since the target variable is the publication count and it approximately follows an Inverse Gaussian distribution (see SI Appendix, Section S5). To capture the productivity during doctoral studies, we used number of publications as a measure and considered the period of doctoral studies four years before graduation and four years after to account for the work in submission or in progress during the thesis defense. Results of the regression analysis are summarized in Table 4.1. By analyzing the regression coefficients and the significant variables, we assessed the factors influencing the academic productivity and the social determinants of the doctoral students performance.

We investigated the regression analysis to validate our earlier observations about the gender and disciplinary differences. It was shown in the literature on research outcomes that women have slightly less publication rates than men while the difference can be attributed to various systematic biases in academia (Fox & Faver (1985); Larivière, Ni, Gingras, Cronin & Sugimoto (2013); Lee & Bozeman (2005)). Especially for STEM fields, empirical data reveals considerable gender variations in number of citations, publication counts and the impact of their academic careers (Abramo, D'Angelo & Caprasecca (2009); Huang et al. (2020)). This phenomenon can be explained by several factors; it is possible to consider that women are un-

Table 4.1 **Regression results.** Inverse Gaussian model for explaining productivity by gender, discipline and textual features extracted from dissertation text.

	Publication Count		
	All Students	*Female Students*	*Male Students*
gender	-0.1839***		
	(0.038)		
life_earth	-0.1218*	-0.0943	-0.1544
	(0.058)	(0.086)	(0.079)
math_comp	-0.0032	0.0420	-0.0318
	(0.063)	(0.123)	(0.077)
phys_eng	0.2468***	0.1171	0.2623***
	(0.057)	(0.104)	(0.069)
soc_hum	-0.3968***	-0.3907***	-0.4404***
	(0.048)	(0.066)	(0.073)
family_sentiment	-0.0375	-0.0426	-0.0566
	(0.047)	(0.073)	(0.061)
academic_count	0.0044*	-0.0011	0.0071*
	(0.002)	(0.003)	(0.003)
family_count	0.0004	0.0044	-0.0027
	(0.004)	(0.006)	(0.005)
friends&colleagues_count	0.0037	0.0029	0.0073
	(0.003)	(0.005)	(0.005)
administration_count	0.0018	-0.0024	0.0048
	(0.002)	(0.003)	(0.003)
spiritual_count	0.0032	-0.0506	0.0083
	(0.210)	(0.040)	(0.025)
Constant	2.1818***	2.0914***	2.2291***
	(0.210)	(0.329)	(0.276)
Observations	2,824	1,099	1,725
AIC Score	16,206	5,831	10,383

Standard errors in parentheses $^{*}(p < 0.05)$; $^{**}(p < 0.01)$; $^{***}(p < 0.001)$

derrepresented in scientific cooperation and publishing and struggle with implicit biases since they are more likely to play a significant role in parenting (Kyvik & Teigen (1996)), obtain less institutional assistance and have more service duties (Duch, Zeng, Sales-Pardo, Radicchi, Otis, Woodruff & Nunes Amaral (2012)), or the systematic undervaluation of women's involvement and their invisibility in scientific research, known as the "Matilda Effect" (Rossiter (1993)). Consistent with the literature, our model have demonstrated that female productivity is lower than that of males when considering simply the number of publications (M = -0.1839, 95% CI [-0.258, -0.110]). These gender differences imply that studying the doctoral

process may help to better understand the above mentioned adverse conditions.

Another important aspect explaining the productivity is the academic discipline because publication counts vary from one field to another (Sabharwal (2013)), which is a key indicator of quality in higher education since the research performance has an influence on rewards, tenure, promotion decisions and staff recruitment (Blackburn & Lawrence (1995); Bland et al. (2006); Costas et al. (2010)). Therefore, it is essential to demonstrate and explain the alterations between scientific fields. When Biology & Health Sciences is taken as the reference group, our model indicates that while the Physics & Engineering students are associated with more publications (M = 0.2468, 95% CI [0.136, 0.358]), Life & Earth Sciences (M = -0.1218, 95% CI [-0.235, -0.009]) and Social Sciences & Humanities (M = -0.3968, 95% CI [-0.492, -0.302]) students are affiliated with less number of papers, as also suggested in the Becher's work on disciplinary differences (Becher (1994)). However, when we controlled for the gender variable, our findings showed that only being a Social Sciences & Humanities (M = -0.3907, 95% CI [-0.520, -0.262]) student is negatively correlated with publication counts for females. On the other hand, male Physics & Engineering (M = 0.2623, 95% CI [0.126, 0.398]) students are associated with more publications while Social Sciences & Humanities (M = -0.4404, 95% CI [-0.584, -0.297]) are associated with less. These results might indicate the under-representation of female doctoral students in Physics & Engineering fields.

Aside from the demographic aspects, our results demonstrated that the number of people from academic network mentioned in acknowledgements is associated with more publications (M = 0.0044, 95% CI [0.000, 0.009]). However, when controlled for genders, the regression analysis suggested that this statement holds only for male students (M = 0.0071, 95% CI [0.002, 0.013]).

Our models do not suggest a statistically significant relationship between the rest of the variables and the number of publications; however, they revealed the influence of gender and discipline on productivity.

Next, we normalized sentiment scores and publication counts between zero and one at the individual level by taking into account gender and discipline of a student. More clearly, we filtered out each gender-discipline pair from our data and normalized publication counts by min-max scaling. These values are then subtracted from the group mean to center around zero. Distributions of normalized sentiment and productivity scores are shown in Fig.4.4(a).

Empirical and visual evidence shows no sign of significant links between sentiment and productivity levels. Additionally, we compared the language characteristics

of extreme cases for both productivity and sentiment levels to help us understand the mindsets of people from upper and lower quantiles of the distributions. To achieve this, we inspect the word usage differences in two groups by using Jensen-Shannon (JS) divergence for words that are used more than 10 times in each group. These words are then represented as word-shift graphs as shown in Fig.4.4(b) for sentiment and Fig.4.4(c) for productivity (Gallagher, Frank, Mitchell, Schwartz, Reagan, Danforth & Dodds (2021)).

Sentiment scores depend on content and context of texts. Hence, there expected to be certain alterations between relatively more positive and negative acknowledgements. As seen on Fig.4.4(b), we observed that most contented 25% of PhD students emphasize gratitude by giving more space in their narratives to words such as *grateful*, *gratitude*, and *thankful*. These results also conformed to the past work which suggests that expressing gratitude helps to increase well-being (Emmonse & Mccullough (2003); Killen & Macaskill (2015)). Our results also demonstrated that both family and the academic environment are more frequently mentioned in the narrative of the most contented 25%. Fig.4.4(c) illustrates the JS divergences of words across most and least performing doctoral students. It is apparent that those who over-perform their counterparts emphasize more endeavour related concepts such as *productive*, *busy*, *internship*, and *article*.

4.5 Institutional Ranking and Student Performance

Since the most well-known university ranking organizations such as Quacquarelli Symonds (QS), Times Higher Education (THE), and Center for World University Rankings (CWUR) employ "number of research papers published" as a factor in their ranking, we assumed that productivity levels of doctoral students may have associations with the success of their institutions. We present analysis on CWUR ranking, since it provides a more granular and longer list, but our results are consistent for other ranking systems as well (see SI Appendix, Section S6).

We investigated the relationship between university rankings and productivity of graduate students. We found that university rankings have significant positive correlation with the number of publications (Fig.4.4(d)). Research environments in these institutes provide more opportunities to publish and introduce them a broader collaboration network as well, partially observed by total number of people mentioned

(Fig.4.4(e)). Number of people mentioned in dissertations have a higher correlation with institute ranking than the productivity levels, suggesting environment cultivate institutional success more than publications alone. However, there is no associations between sentiment of a doctoral student with respect to the ranking of their institutions (Fig.4.4(f)) meaning that the top-ranked institutes provide advantage in professional growth while well-being of the doctoral students mostly determined by their academic support networks.

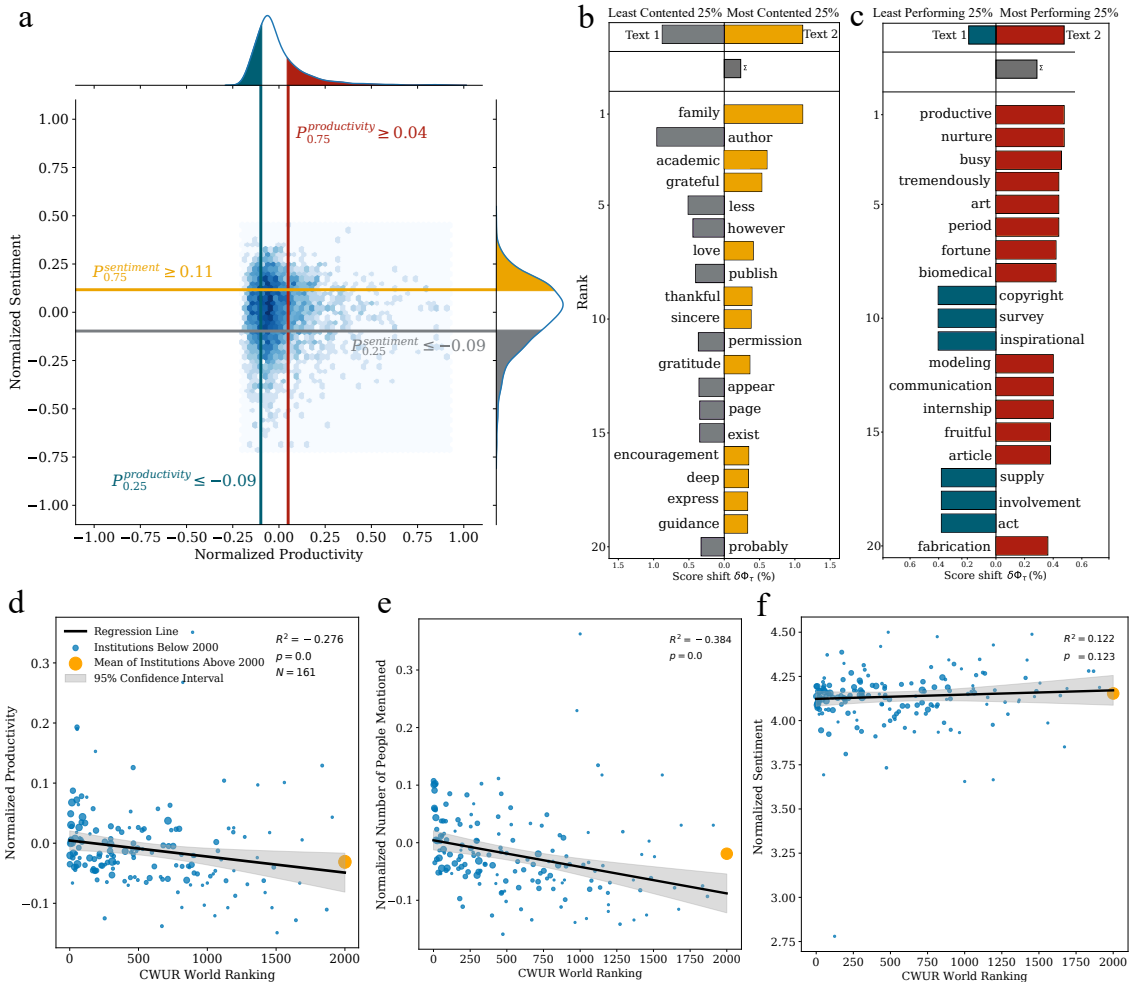

Figure 4.4 **Determinants of academic productivity and linguistic differences between extreme cases** Sentiment and productivity levels of students were normalized based on their disciplines and genders (a). Word usage differences are quantified by JS Divergence scores and compared students at the first and third quartiles based on normalized sentiment and productivity (b, c). Relationship between CWUR World rankings and normalized productivity levels (d). Relationship between CWUR World rankings and normalized number of mentions in acknowledgements (e). Relationship between CWUR World rankings and normalized sentiment scores in acknowledgements (f). R-squared values denote Spearman's rank correlation. Size of blue dots is proportional to the number of theses from these institutions.

5. CONCLUSION AND FUTURE WORK

Our research uncovered the network of support providers, assisting doctoral students in achieving their goals. We showed that there exist gender and disciplinary differences in acknowledging support providers and sentiment scores when mentioning different communities. Since acknowledgements often appear as the sole section in which students talk about their experience as doctoral candidates, it is noteworthy to observe that the link between productivity levels and their academic support networks can be revealed.

Our results showed that the number of publications among doctoral students varies by academic discipline, with students in social sciences and humanities publishing the least and students in physics and engineering publishing the most. Our data also suggested that productivity is positively correlated with the number of people mentioned from academic environment for male students when publication counts are normalized with regard to gender and discipline. We showed that female students are more likely to acknowledge each support provider group with a more positive sentiment. They did, however, mentioned fewer people from their workplace and published fewer academic publications, implying that women's professional support networks are limited, leading to lower productivity levels at the time they receive their doctoral degrees. Our results also demonstrated that schools with higher rankings provide PhD students with wider networks, in which academic environment is significantly bigger and productivity levels as a result are higher. In fact, it is shown in the literature that as the number of writers rises, so does the impact of the research (Larivière, Gingras, Sugimoto & Tsou (2015)), highlighting the importance of young scholars' academic networks.

Quantitative analysis of acknowledgement texts provided a deeper insight on social interactions and experiences of doctoral students as well. Our results suggested that the narrative of the most performing 25% is more centralized on endeavour-related content compared to the least performing 25%. Similarly, most contented 25% show more gratitude towards their family and academic environment compared to the least contented 25%.

It is crucial to note that while support providers from academic communities have a positive influence on productivity, overall well-being of a student require contributions from social interactions with family and friends and administrative support from their institute. Our results showed that higher university rankings or productivity levels do not lead to a higher sentiment reflected towards doctoral experience, but positive influence in their professional growth.

Analyzing thousands of acknowledgement sections, we created an alternative angle reflecting social aspects of the doctoral studies where friends, families, colleagues, and administrative staff have different roles to play ensuring utmost performance and well-being of the student. Therefore, instead of directly analyzing publication counts or number of citations to explain doctoral studies, it may be better to embrace a new approach where students' well-beings and academic support networks are also put forward. People compare themselves to those who are similar to them with regard to demographic and social proximity and how individuals evaluate their own subjective well-being and happiness depends on those of others (De la Garza, Mastrobuoni, Sannabe & Yamada (2012); Posel & Casale (2011)). It is also known that "success narratives" have an impact on the reader's judgements and decisions (Lifchits, Anderson, Goldstein, Hofman & Watts (2021)), which may imply that when doctoral students compare themselves with their counterparts, it would decrease their subjective well-being.

Future work may contribute to a profound understanding of how support networks influence productivity in late career stages and researchers' overall well-being by reaching out to people and possibly conducting a survey. It is important to collect theses published all around the world to improve the representativeness of the data and observe how cultural aspects influence the way of doctoral students acknowledge their support providers. It is also noteworthy to mention that our study leverages two online datasets, both of which fitting the description of "readymade" more than "custommade" as explained by Salganik (2019) because they do not consist of the answers directly for our questions but are rather repurposed to answer them. Therefore, it would be appropriate to use a "custommade" data by conducting a survey for PhD students to enrich the available data sources and measure the accuracy of our results.

BIBLIOGRAPHY

Abramo, G., D'Angelo, C. A., & Caprasecca, A. (2009). Gender differences in research productivity: A bibliometric analysis of the italian academic system. *Scientometrics*, *79*(3), 517–539.

Al-Ali, M. N. (2010). Generic patterns and socio-cultural resources in acknowledgements accompanying arabic ph. d. dissertations. *Pragmatics*, *20*(1), 1–26.

Alotaibi, H. S. (2018). Metadiscourse in dissertation acknowledgments: Exploration of gender differences in efl texts. *Educational Sciences: Theory and Practice*, *18*(4), 899–916.

Altbach, P. G. (2015). What counts for academic productivity in research universities? *International Higher Education*, (79), 6–7.

Bagga, A. & Baldwin, B. (1998). Algorithms for scoring coreference chains. In *The first international conference on language resources and evaluation workshop on linguistics coreference*, volume 1, (pp. 563–566). Citeseer.

Bazerman, C. et al. (1988). *Shaping written knowledge: The genre and activity of the experimental article in science*, volume 356. University of Wisconsin Press Madison.

Becher, T. (1994). The significance of disciplinary differences. *Studies in Higher Education*, *19*(2), 151–161.

Ben-Ari, E. (1987). On acknowledgements in ethnographies. *Journal of Anthropological Research*, *43*(1), 63–84.

Blackburn, R. T. & Lawrence, J. H. (1995). *Faculty at work: Motivation, expectation, satisfaction.* Johns Hopkins University Press.

Bland, C. J., Center, B. A., Finstad, D. A., Risbey, K. R., & Staples, J. (2006). The impact of appointment type on the productivity and commitment of full-time faculty in research and doctoral institutions. *The Journal of Higher Education*, *77*(1), 89–123.

Broström, A. (2019). Academic breeding grounds: Home department conditions and early career performance of academic researchers. *Research Policy*, *48*(7), 1647–1665.

Carpenter, C. R., Cone, D. C., & Sarli, C. C. (2014). Using publication metrics to highlight academic productivity and research impact. *Academic emergency medicine*, *21*(10), 1160–1172.

Casad, B., Franks, J., Garasky, C., Kittleman, M., Roesler, A., Hall, D., & Petzel, Z. (2020). Gender inequality in academia: Problems and solutions for women faculty in stem. *Journal of Neuroscience Research*, *99*.

Castelló, M., McAlpine, L., & Pyhältö, K. (2017). Spanish and uk post-phd researchers: writing perceptions, well-being and productivity. *Higher Education Research & Development*, *36*(6), 1108–1122.

Chan, T. H.-T. (2015). A corpus-based study of the expression of stance in dissertation acknowledgements. *Journal of English for Academic Purposes*, *20*, 176–191.

Cheng, S. W. et al. (2012). A contrastive study of master thesis acknowledgements by taiwanese and north american students. *Open Journal of Modern Linguistics*, *2*(01), 8.

Clark, K. & Manning, C. D. (2016a). Deep reinforcement learning for mention-ranking coreference models.

Clark, K. & Manning, C. D. (2016b). Improving coreference resolution by learning entity-level distributed representations. *arXiv preprint arXiv:1606.01323*.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46.

Collins, R. & Steffen-Fluhr, N. (2019). Hidden patterns: Using social network analysis to track career trajectories of women stem faculty. *Equality, Diversity and Inclusion: An International Journal*, *38*.

Costas, R., Van Leeuwen, T. N., & Bordons, M. (2010). A bibliometric classificatory approach for the study and assessment of research performance at the individual level: The effects of age on productivity and impact. *Journal of the American Society For Information Science and Technology*, *61*(8), 1564–1581.

Cronin, B., McKenzie, G., & Stiffler, M. (1992). Patterns of acknowledgement. *Journal of Documentation*.

De la Garza, A. G., Mastrobuoni, G., Sannabe, A., & Yamada, K. (2012). The relative utility hypothesis with and without self-reported reference wages.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding.

Duch, J., Zeng, X. H. T., Sales-Pardo, M., Radicchi, F., Otis, S., Woodruff, T. K., & Nunes Amaral, L. A. (2012). The possible role of resource requirements and academic career-choice risk on gender differences in publication rate and impact. *PloS one*, *7*(12), e51332.

Emmonse, R. & Mccullough, M. E. (2003). Counting blessings versus burdens: An experimental investigation of gratitude and subjective well-being in daily life. *Journal of Personality and Social Psychology*, *84*(2), 377–389.

Fortunato, S., Bergstrom, C. T., Börner, K., Evans, J. A., Helbing, D., Milojević, S., Petersen, A. M., Radicchi, F., Sinatra, R., Uzzi, B., et al. (2018). Science of science. *Science*, *359*(6379).

Fox, M. F. & Faver, C. A. (1985). Men, women, and publication productivity: Patterns among social work academics. *The Sociological Quarterly*, *26*(4), 537–549.

Gallagher, R. J., Frank, M. R., Mitchell, L., Schwartz, A. J., Reagan, A. J., Danforth, C. M., & Dodds, P. S. (2021). Generalized word shift graphs: a method for visualizing and explaining pairwise comparisons between texts. *EPJ Data Science*, *10*(1), 4.

García-Suaza, A., Otero, J., & Winkelmann, R. (2020). Predicting early career productivity of phd economists: Does advisor-match matter? *Scientometrics*, *122*(1), 429–449.

Girvan, M. & Newman, M. E. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, *99*(12), 7821–7826.

Harris, Z. S. (1954). Distributional structure. *Word*, *10*(2-3), 146–162.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National academy of Sciences*, *102*(46), 16569–16572.

Hook, D. W., Porter, S. J., & Herzog, C. (2018). Dimensions: Building context for search and evaluation. *Frontiers in Research Metrics and Analytics*, *3*, 23. https://www.frontiersin.org/articles/10.3389/frma.2018.00023/pdf.

Huang, J., Gates, A. J., Sinatra, R., & Barabási, A.-L. (2020). Historical comparison of gender inequality in scientific careers across countries and disciplines. *Proceedings of the National Academy of Sciences*, *117*(9), 4609–4616.

huggingface (2021). neuralcoref.

Hutto, C. & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, (pp. 216–225).

Hyland, K. (2003). Dissertation acknowledgements: The anatomy of a cinderella genre. *Written Communication*, *20*(3), 242–268.

Hyland, K. (2004). Graduates' gratitude: The generic structure of dissertation acknowledgements. *English for Specific purposes*, *23*(3), 303–324.

Hysenbegasi, A., Hass, S. L., & Rowland, C. R. (2005). The impact of depression on the academic productivity of university students. *Journal of mental health policy and economics*, *8*(3), 145.

Killen, A. & Macaskill, A. (2015). Using a gratitude intervention to enhance well-being in older adults. *Journal of Happiness Studies*, *16*(4), 947–964.

Kozlowski, D., Larivière, V., Sugimoto, C. R., & Monroe-White, T. (2022). Intersectional inequalities in science. *Proceedings of the National Academy of Sciences*, *119*(2).

Kyvik, S. & Teigen, M. (1996). Child care, research collaboration, and gender differences in scientific productivity. *Science, Technology, & Human Values*, *21*(1), 54–71.

Lamers, W. S., Boyack, K., Larivière, V., Sugimoto, C. R., van Eck, N. J., Waltman, L., & Murray, D. (2021). Measuring disagreement in science.

Larivière, V., Gingras, Y., Sugimoto, C. R., & Tsou, A. (2015). Team size matters: Collaboration and scientific impact since 1900. *Journal of the Association for Information Science and Technology*, *66*(7), 1323–1332.

Larivière, V., Ni, C., Gingras, Y., Cronin, B., & Sugimoto, C. R. (2013). Bibliometrics: Global gender disparities in science. *Nature News*, *504*(7479), 211.

Le, Q. V. & Mikolov, T. (2014). Distributed representations of sentences and documents.

Lee, S. & Bozeman, B. (2005). The impact of research collaboration on scientific productivity. *Social Studies of Science*, *35*(5), 673–702.

Levecque, K., Anseel, F., De Beuckelaer, A., Van der Heyden, J., & Gisle, L. (2017). Work organization and mental health problems in phd students. *Research Policy*, *46*(4), 868–879.

Lifchits, G., Anderson, A., Goldstein, D. G., Hofman, J. M., & Watts, D. J. (2021). Success stories cause false beliefs about success. *Judgment and Decision Making*, *16*(6), 1440.

Ma, Y., Mukherjee, S., & Uzzi, B. (2020). Mentorship and protégé success in stem fields. *Proceedings of the National Academy of Sciences*, *117*(25), 14077–14083.

Mantai, L. & Dowling, R. (2015). Supporting the phd journey: Insights from acknowledgements. *International Journal for Researcher Development*.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mu, W., Lim, K. H., Liu, J., Karunasekera, S., Falzon, L., & Harwood, A. (2022). A clustering-based topic model using word networks and word embeddings. *Journal of Big Data*, *9*(1), 1–38.

Nwala, A. C., Weigle, M. C., & Nelson, M. L. (2018). Bootstrapping web archive collections from social media. In *Proceedings of the 29th on Hypertext and Social Media* (pp. 64–72).

Pearson, M. & Brew, A. (2002). Research training and supervision development. *Studies in Higher education*, *27*(2), 135–150.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, (pp. 1532–1543).

Posel, D. R. & Casale, D. M. (2011). Relative standing and subjective well-being in south africa: The role of perceptions, expectations and income mobility. *Social Indicators Research*, *104*(2), 195–223.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*.

Rashed, A., Kutlu, M., Darwish, K., Elsayed, T., & Bayrak, C. (2020). Embeddings-based clustering for target specific stances: The case of a polarized turkey. *arXiv preprint arXiv:2005.09649*.

Rossiter, M. W. (1993). The matthew matilda effect in science. *Social studies of science*, *23*(2), 325–341.

Sabharwal, M. (2013). Comparing research productivity across disciplines and career stages. *Journal of Comparative Policy Analysis: Research and Practice*, *15*(2), 141–163.

Salganik, M. J. (2019). *Bit by bit: Social research in the digital age*. Princeton University Press.

Sarli, C. C. & Carpenter, C. R. (2014). Measuring academic productivity and changing definitions of scientific impact. *Missouri medicine*, *111*(5), 399.

Scrivener, L. (2009). An exploratory analysis of history students' dissertation acknowledgments. *The Journal of Academic Librarianship*, *35*(3), 241–251.

Sekara, V., Deville, P., Ahnert, S. E., Barabási, A.-L., Sinatra, R., & Lehmann, S. (2018). The chaperone effect in scientific publishing. *Proceedings of the National Academy of Sciences*, *115*(50), 12603–12607.

Serrano, M. Á., Boguná, M., & Vespignani, A. (2009). Extracting the multiscale backbone of complex weighted networks. *Proceedings of the National Academy of Sciences*, *106*(16), 6483–6488.

Sinatra, R., Wang, D., Deville, P., Song, C., & Barabási, A.-L. (2016). Quantifying the evolution of individual scientific impact. *Science*, *354*(6312).

Way, S. F., Larremore, D. B., & Clauset, A. (2016). Gender, productivity, and prestige in computer science faculty hiring networks. In *Proceedings of the 25th International Conference on World Wide Web*, (pp. 1169–1179).

Way, S. F., Morgan, A. C., Larremore, D. B., & Clauset, A. (2019). Productivity, prominence, and the effects of academic environment. *Proceedings of the National Academy of Sciences*, *116*(22), 10729–10733.

Weischedel, R., Hovy, E., Marcus, M., Palmer, M., Belvin, R., Pradhan, S., Ramshaw, L., & Xue, N. (2011). *OntoNotes: A Large Training Corpus for Enhanced Processing*.

Woolston, C. (2019). Phds: the tortuous truth. *Nature*, *575*(7782), 403–407.

Wu, L., Wang, D., & Evans, J. A. (2019). Large teams develop and small teams disrupt science and technology. *Nature*, *566*(7744), 378–382.

Yang, W. (2012a). Comparison of gratitude across context variations: A generic analysis of dissertation acknowledgements written by taiwanese authors in efl and esl contexts. *International journal of applied linguistics and English literature*, *1*(5), 130–146.

Yang, W. (2012b). A genre analysis of phd dissertation acknowledgements. *LSP Journal-Language for special purposes, professional communication, knowledge management and cognition*, *3*(2).

APPENDIX A

List of Support Providers

Table A.1 **List of support providing groups and words or phrases associated with these groups**

Support providers	**Keywords used for identification**
Academic	academic advisor, advisor, chair, chairman, chairperson, co-advisor, co-chair, committee, committee chair, committee member, dissertation chair, dissertation committee, doctoral committee, dr, guide, investigator, lab, mentor, prof, research group, researcher, role model, scientist, supervisor, thesis advisor, thesis committee
Family	aunt, baby, best friend, brother, child, cousin, dad, daddy, daughter, extended family, family, father, fiancé, girlfriend, grandchild, granddaughter, grandfather, grandma, grandmother, grandpa, grandparent, grandson, home, husband, in-law, kid, mama, mom, mommy, mother, nephew, niece, parent, partner, sibling, sister, son, spouse, uncle, wife, guardian
Administration	administrative staff, administrator, campus, college, council, dean, department, facility, faculty, faculty member, institute, institution, instructor, librarian, library, practitioner, secretary, staff, staff member, university, associate, king, educator, specialist, board, boss, coworker, employee, employer, government, hospital, leader, management, manager, officer, organization, president, worker, teacher, participant
Friends & Colleagues	buddy, cat, coach, dog, co-author, crew, group member, lab member, labmate, mate, officemate, team, teammate, technician, classmate, club, colleague, comrade, friend, peer, roommate, cheerleader, editor, motivator, sounding board
Spiritual	allah, almighty, christ, church, congregation, creator, god, jesus, lord, ministry, pastor, savior

Gender Inference & Gender Based Differences

List of Discipline Categories

Biology & Health Sciences: nursing (991), public health (670), molecular biology (583), neurosciences (517), mental health (448), cellular biology (388), genetics (370), biomedical engineering (298), microbiology (257), medicine (239), epidemiology (225), biology (210), health sciences (209), oncology (194), physiology (189), immunology (185), organic chemistry (158), pharmacology (152), nutrition (134),

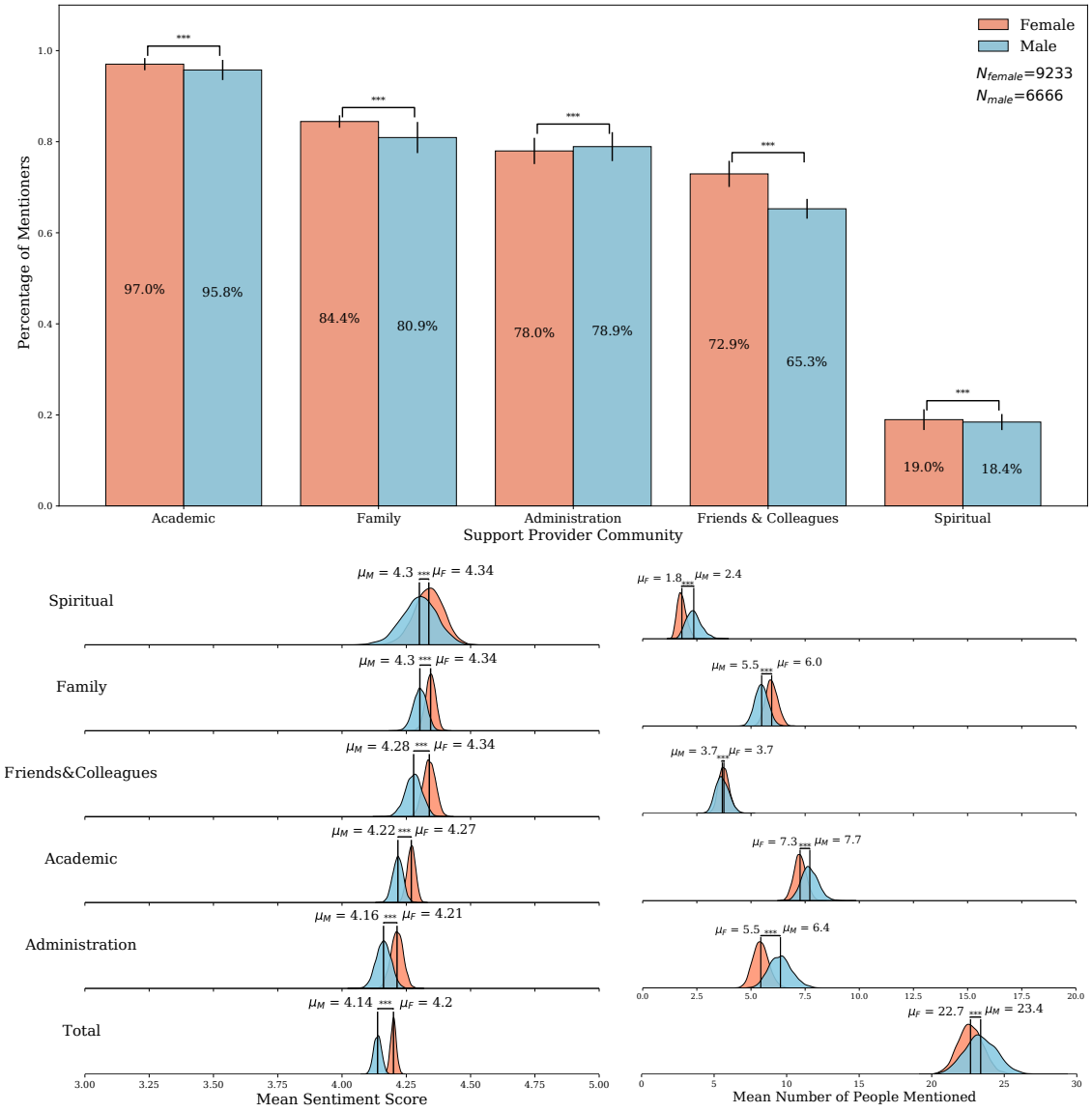

Figure A.1 **Gender Based Differences in Terms of Communities for Social Sciences and Humanities Students**

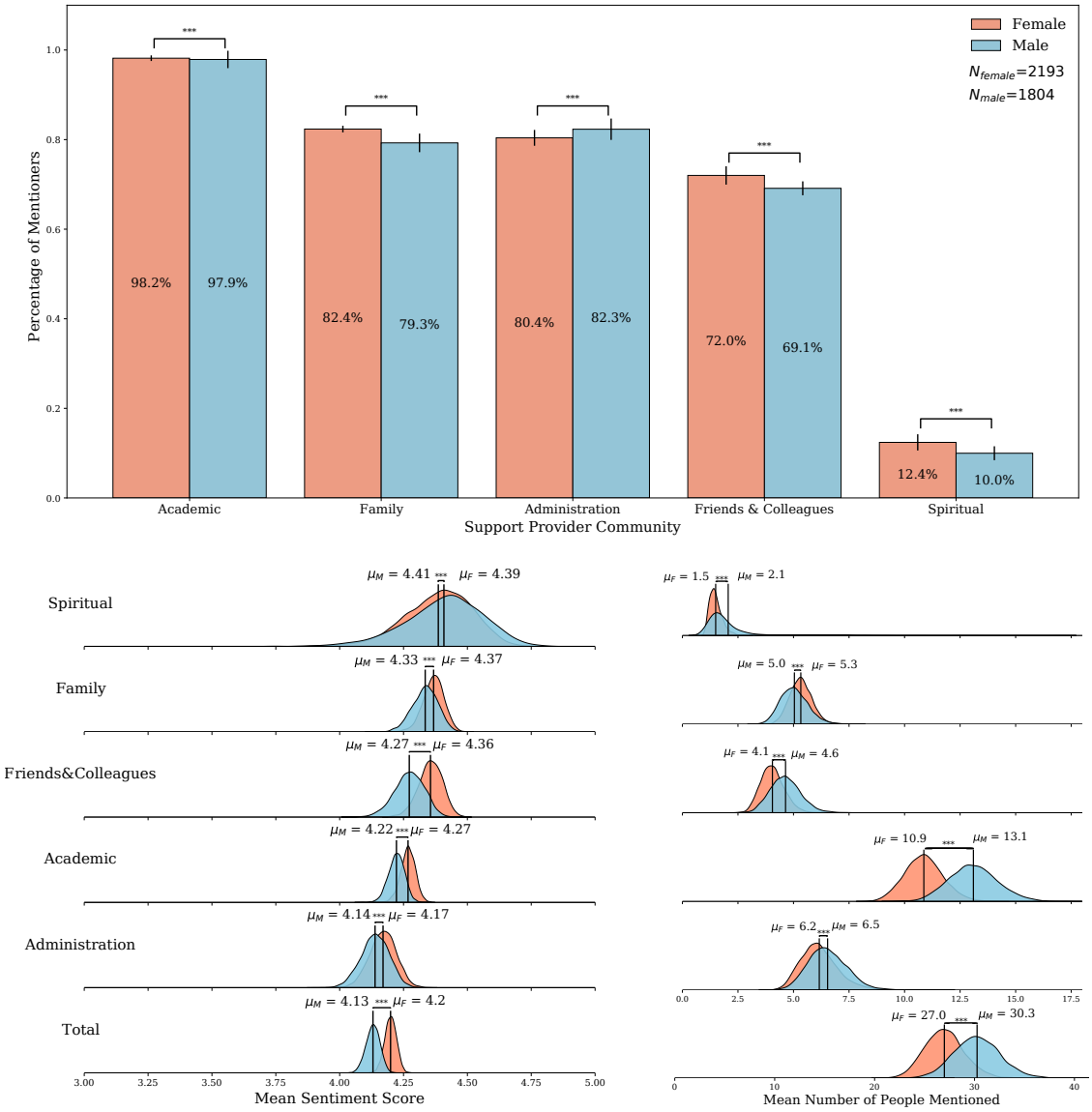

Figure A.2 **Gender Based Differences in Terms of Communities for Biology and Health Sciences Students**

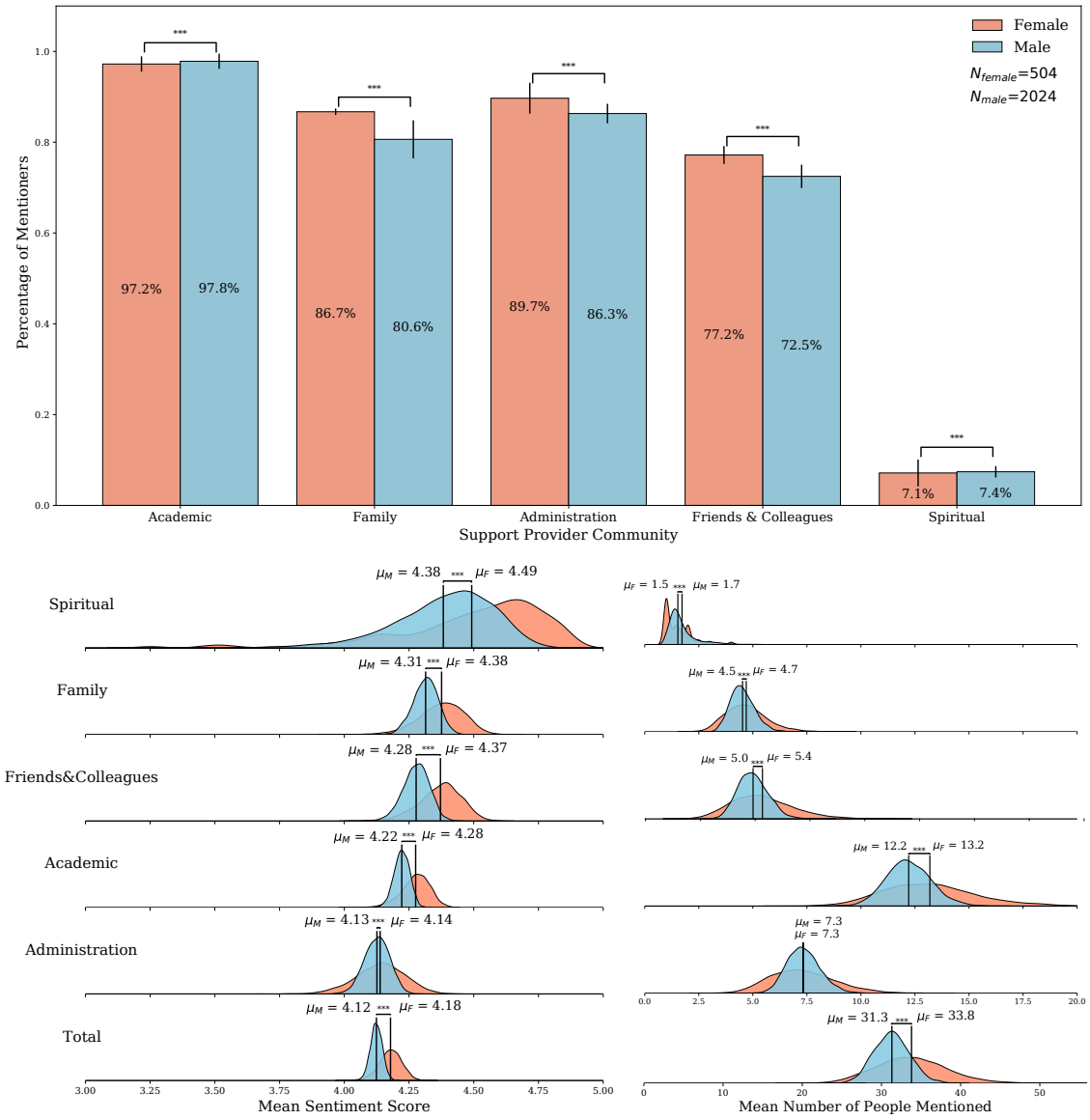

Figure A.3 **Gender Based Differences in Terms of Communities for Physics and Engineering Sciences Students**

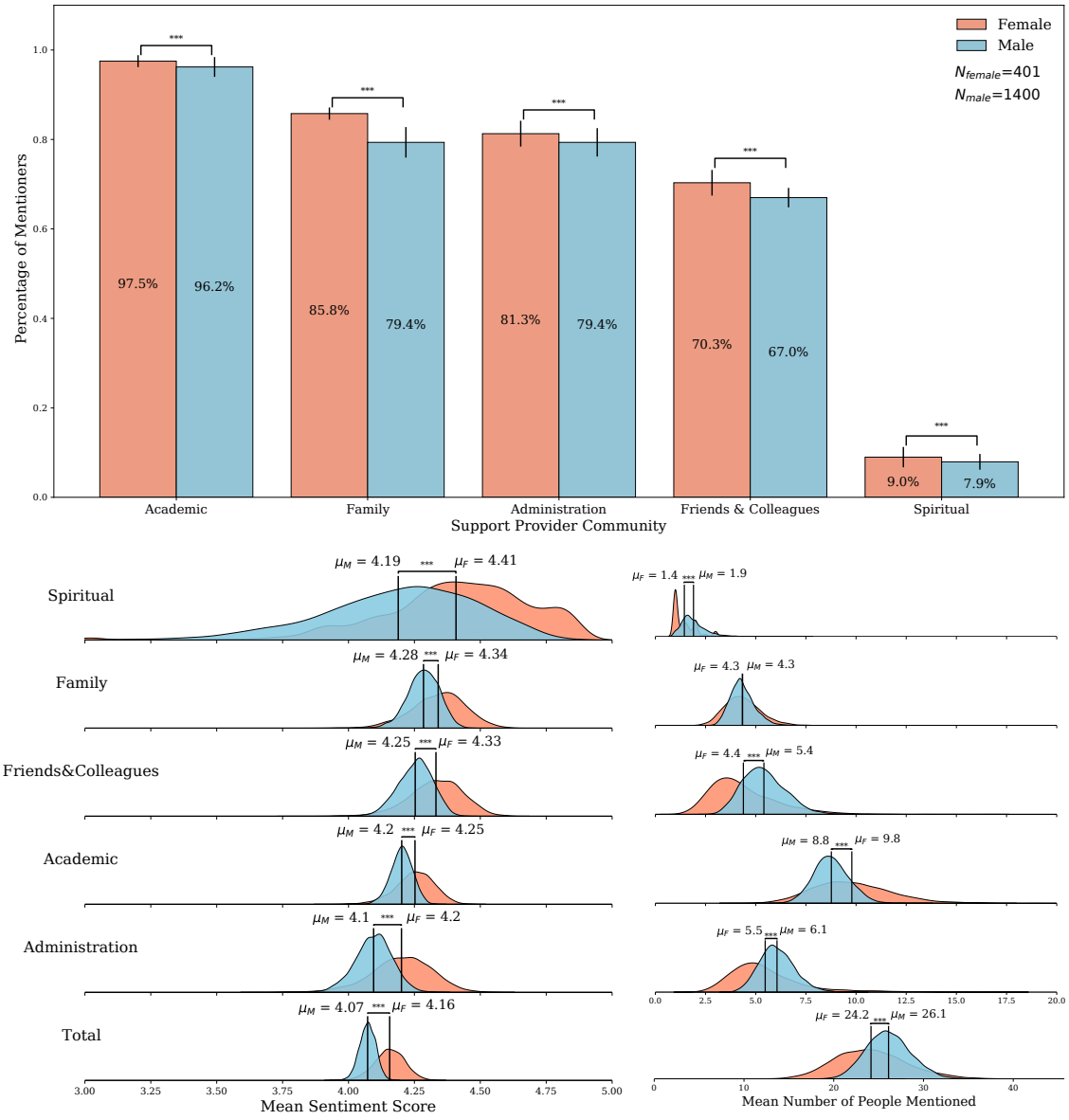

Figure A.4 **Gender Based Differences in Terms of Communities for Mathematics and Computer Science Students**

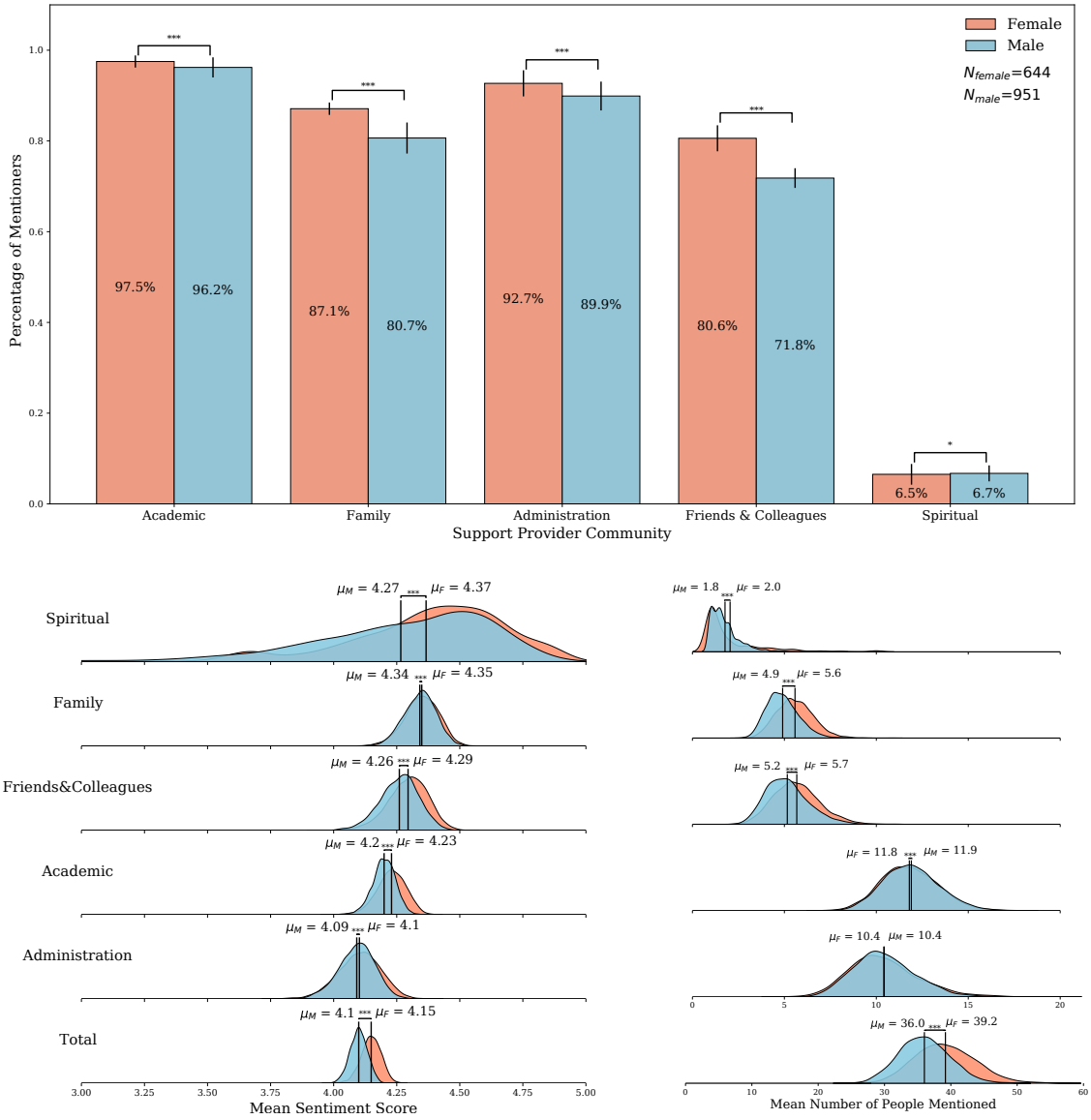

Figure A.5 **Gender Based Differences in Terms of Communities for Life and Earth Sciences Students**

evolution and development (125), virology (106), aging (102), developmental biology (82), surgery (82), kinesiology (76), pathology (74), physiological psychology (71), medical imaging (70), toxicology (63), biomechanics (62), pharmacy sciences (58), entomology (57), physical therapy (55), psychobiology (53), wildlife conservation (52), environmental health (49), animal sciences (48), alternative medicine (47), pharmaceutical sciences (45), parasitology (35), occupational health (34), food science (33), plant pathology (32), endocrinology (32), sexuality (31), audiology (30), systematic (26), veterinary services (26), obstetrics (26), ophthalmology (24), molecular chemistry (22), dentistry (22), medical personnel (22), molecular physics (21), organismal biology (20), systematic biology (16), neurology (15), histology (12), animal diseases (11), health care (10), biomedical research (10), pharmaceutical and medicine manufacturing (8), anatomy & physiology (5), rehabilitation (5), general medical and surgical hospitals (4), health (3), osteopathic medicine (2), radiology (2), optometry (2), pharmaceuticals industry (1), research and development in the physical (1), nursing care facilities (skilled nursing facilities) (1), plant propagation (1), sports medicine (1), home health care services (1), radiation (1), pharmaceuticals (1)

Life & Earth Sciences: ecology (429), biochemistry (428), environmental science (236), biophysics (228), chemistry (179), sustainability (165), geography (141), atmospheric sciences (120), geology (112), climate change (107), geophysics (96), hydrologic sciences (90), agriculture (87), plant sciences (84), environmental studies (81), geochemistry (71), environmental management (71), biological oceanography (67), natural resource management (66), zoology (66), water resource management (60), biogeochemistry (59), aquatic sciences (53), forestry (50), conservation (47), physical oceanography (45), plant biology (41), geographic information science (36), soil sciences (34), geomorphology (33), horticulture (30), meteorology (29), agronomy (29), atmospheric chemistry (27), physical geography (26), wildlife management (25), paleontology (25), limnology (23), land use planning (21), paleoclimate science (19), planetology (19), sedimentary geology (18), conservation biology (18), geotechnology (17), paleoecology (15), plate tectonics (14), chemical oceanography (14), botany (12), marine geology (12), water resources management (12), petroleum geology (10), mining engineering (10), geological (9), geophysical (8), wood sciences (8), macroecology (7), petrology (7), environmental geology (6), hydrology (6), aeronomy (6), mineralogy (5), atmosphere (5), petroleum production (4), geological engineering (4), urban forestry (3), geobiology (3), oceanography (3), geophysical engineering (2), and life sciences (1), polymers (1)

Mathematics & Computer Sciences: computer science (1048), information technology (438), mathematics (355), information science (274), computer engineer-

ing (235), bioinformatics (233), applied mathematics (227), statistics (226), artificial intelligence (202), library science (79), biostatistics (66), computational physics (28), theoretical mathematics (27), computational chemistry (19), logic (9), systems design (6), information systems (6), software & systems (1)

Physics & Engineering:electrical engineering (576), mechanical engineering (406), materials science (374), engineering (313), physics (278), civil engineering (254), chemical engineering (205), optics (196), aerospace engineering (174), condensed matter physics (169), systems science (169), physical chemistry (156), analytical chemistry (143), astrophysics (137), urban planning (133), operations research (128), energy (127), astronomy (125), environmental engineering (121), nanotechnology (104), robotics (100), industrial engineering (100), particle physics (92), nanoscience (90), remote sensing (87), theoretical physics (86), inorganic chemistry (82), plasma physics (79), quantum physics (78), nuclear physics (63), transportation planning (60), electromagnetics (59), alternative energy (55), architecture (53), polymer chemistry (52), acoustics (48), mechanics (46), nuclear engineering (32), transportation (30), atomic physics (28), bioengineering (28), fluid mechanics (26), petroleum engineering (23), ocean engineering (22), agricultural engineering (22), automotive engineering (17), low temperature physics (16), applied physics (16), textile research (14), statistical physics (14), thermodynamics (10), naval engineering (9), landscape architecture (9), nuclear chemistry (8), high temperature physics (6), plastics (6), architectural engineering (6), architectural (5), robots (5), mining (3), fluid dynamics (3), gases (3), condensation (3), hydraulic engineering (3), automobile and light duty motor vehicle manufacturing (2), aerospace materials (2), mining and oil and gas field machinery manufacturing (1), molecules (1), all other miscellaneous manufacturing (1)

Social Sciences & Humanities: educational leadership (2222), higher education (1560), management (1452), organizational behavior (989), clinical psychology (952), school administration (826), educational technology (806), education (757), teacher education (752), educational psychology (747), psychology (746), secondary education (730), womens studies (726), elementary education (600), health care management (585), education policy (571), educational evaluation (562), curriculum development (546), social psychology (538), special education (523), business administration (521), behavioral psychology (487), higher education administration (473), educational administration (471), african american studies (471), counseling psychology (451), public policy (449), adult education (449), music (415), community college education (398), health education (384), occupational psychology (374), individual & family studies (374), organization theory (374), gender studies (369), political science (357), cognitive psychology (355), educational tests & measurements (351), eco-

nomics (345), communication (341), mathematics education (341), developmental psychology (337), social research (332), public administration (321), reading instruction (321), cultural anthropology (320), educational sociology (314), middle school education (308), social work (303), linguistics (295), science education (291), religion (286), early childhood education (282), ethnic studies (279), black studies (276), philosophy (267), spirituality (260), criminology (259), multicultural education (256), pedagogy (247), english as a second language (247), hispanic american studies (233), military studies (232), sociology (225), american history (225), literacy (219), instructional design (207), lgbtq studies (199), music education (196), finance (191), school counseling (170), religious education (170), language arts (170), personality psychology (168), marketing (165), business education (165), theology (163), native american studies (163), behavioral sciences (151), asian studies (148), web studies (148), international relations (143), archaeology (141), mass communications (141), social studies education (140), gerontology (138), entrepreneurship (133), bilingual education (132), latin american studies (131), american studies (128), ethics (126), education philosophy (121), american literature (121), education finance (119), law (117), british and irish literature (116), social structure (115), african studies (115), rhetoric (111), art history (111), accounting (109), labor relations (108), vocational education (107), quantitative psychology (99), religious history (95), language (93), history (92), art education (91), disability studies (89), economic theory (87), south asian studies (87), middle eastern studies (86), european history (86), labor economics (83), education history (82), asian american studies (77), film studies (76), comparative literature (75), speech therapy (74), multimedia communications (72), recreation (69), continuing education (65), business and secretarial schools (65), agricultural economics (64), latin american history (64), design (63), experimental psychology (62), biblical studies (62), demography (57), theater (57), sociolinguistics (56), performing arts (56), public health education (56), caribbean studies (55), science history (55), sports management (54), modern literature (54), biographies (54), banking (53), foreign language education (53), international law (53), physical education (53), literature (52), sub saharan africa studies (51), gifted education (51), commerce-business (50), law enforcement (50), judaic studies (49), foreign language (48), therapy (48), physical anthropology (47), classical studies (45), peace studies (45), modern history (44), philosophy of science (43), environmental education (41), clerical studies (40), journalism (39), museum studies (39), folklore (39), environmental economics (39), cultural resources management (37), fine arts (37), epistemology (37), latin american literature (36), romance literature (36), military history (35), aesthetics (34), ancient history (34), black history (33), technical communication (33), medical ethics (33), south african studies (32), pastoral counseling (32), psychotherapy (32), clergy (31), asian literature (30), creative writing (29),

islamic studies (28), economic history (27), medieval literature (25), european studies (25), management consulting services (24), dance (24), middle eastern history (24), curricula (23), teaching (23), music history (23), occupational therapy (22), philosophy of religion (22), occupational safety (22), environmental justice (21), african history (21), musical composition (21), forensic anthropology (20), divinity (20), asian history (20), metaphysics (20), agricultural education (19), area planning and development (19), modern language (19), medieval history (18), music theory (18), comparative religion (17), middle eastern literature (17), cognitive therapy (17), near eastern studies (16), russian history (15), holocaust studies (15), theater history (15), morphology (15), range management (14), east european studies (14), alternative dispute resolution (14), regional studies (14), slavic literature (13), industrial arts education (13), art criticism (13), pacific rim studies (13), african literature (13), slavic studies (13), behaviorial sciences (13), performing arts education (12), home economics education (12), environmental philosophy (11), public finance activities (11), germanic literature (11), families & family life (11), personal relationships (11), academic guidance counseling (11), international affairs (10), world history (10), caribbean literature (10), music therapy (10), north african studies (9), ancient languages (9), canadian studies (9), german literature (8), administration of education programs (8), arts management (8), armed forces (8), experiments (8), mass media (8), composition (8), translation studies (8), southeast asian studies (8), fashion (7), area planning & development (7), experimental/theoretical (6), administration of general economic programs (6), community colleges (6), intellectual property (5), junior colleges (5), icelandic & scandinavian literature (5), engineering services (5), canadian history (5), restaurants and other eating places (4), managerial skills (4), canadian literature (4), organizational structure (4), home economics (4), monetary authorities-central bank (4), electronic shopping and mail-order houses (4), personality (4), french literature (4), musical performances (3), cinematography (3), united states (3), religious organizations (3), commercial banking (3), social trends & culture (3), capital & debt management (3), business associations (3), tax preparation (3), bookkeeping (3), and payroll services (3), scandinavian studies (3), native americans (3), native studies (3), minority & ethnic groups (3), canon law (2), social policy (2), schools and educational services (2), classical literature (2), document preparation services (2), credit bureaus (2), environmental consulting services (2), small business (2), administration of urban planning and community and rural development (2), regulation (2), licensing (2), and inspection of miscellaneous commercial sectors (2), business to business electronic markets (2), direct life (2), and medical insurance carriers (2), colleges (2), universities (2), and professional schools (2), supermarkets and other grocery (except convenience) stores (2), interior design (2), middle ages (2), motion pictures (2), company specific (1), fine arts schools (1),

patent law (1), acquisitions & mergers (1), service industries not elsewhere classified (1), telephone call centers (1), public relations (1), transportation equipment industry (1), french canadian literature (1), french canadian culture (1), national security (1), history of oceania (1), africa (1), multinational corporations (1), media buying agencies (1), counseling education (1), multilingual education (1), direct mail advertising (1), other direct selling establishments (1), psychological tests (1), demographics (1), asia & the pacific (1), boards of directors (1), hispanic americans (1), hotels (except casino hotels) and motels (1), all other schools and instruction (1), museums (1), civic and social organizations (1), educational support services (1), industrial design services (1), labor unions and similar labor organizations (1), theater companies and dinner theaters (1), residential mental health and substance abuse facilities (1), Italian literature (1), environmental law (1)

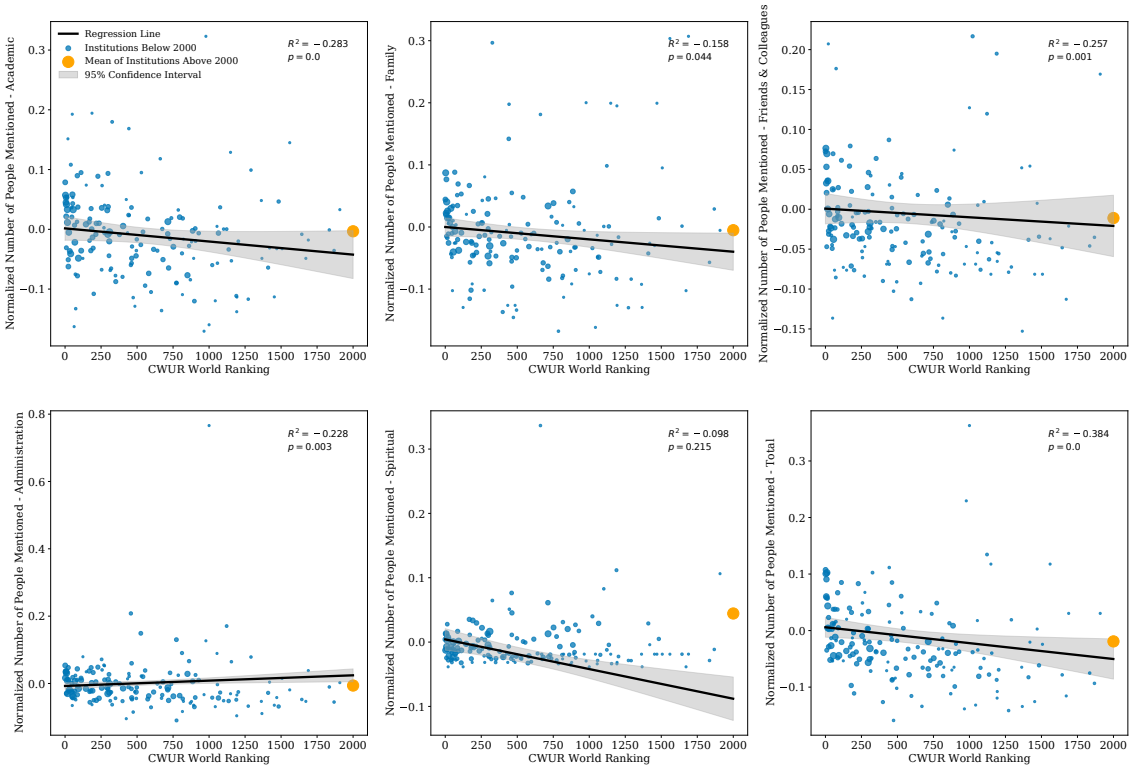

Figure A.6 **Total Number of Mentions and CWUR World Rankings**

Detailed Illustrations of University Rankings

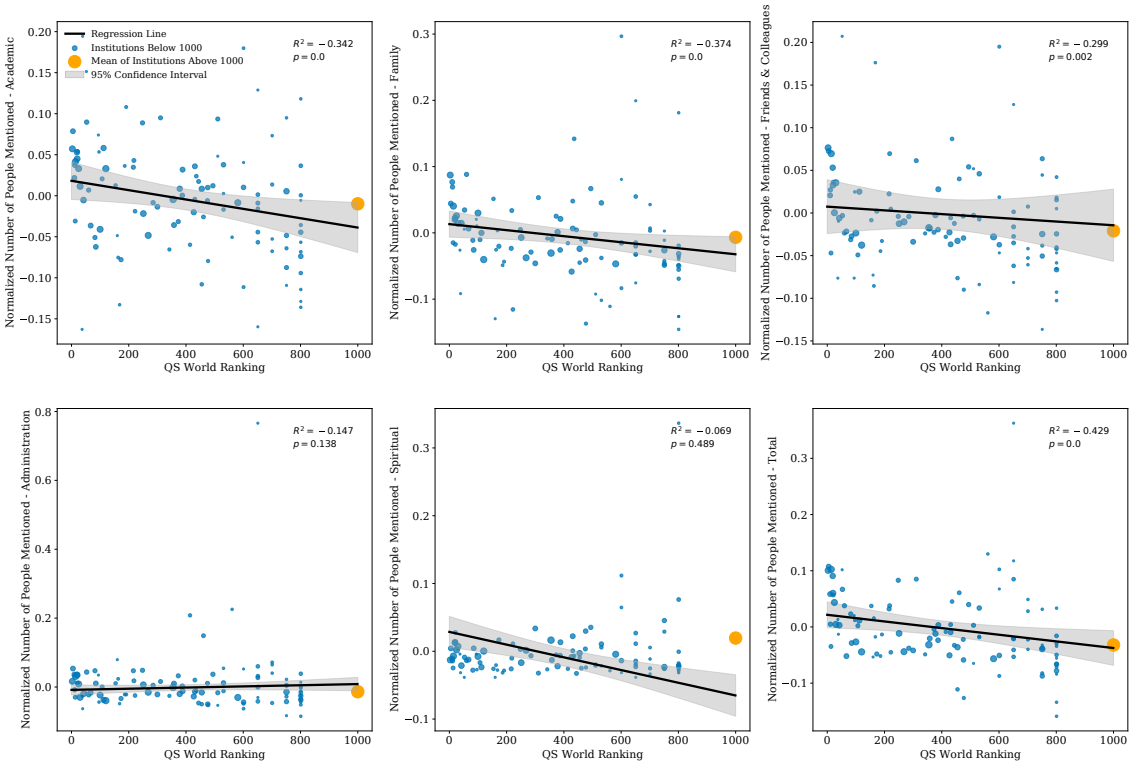

Figure A.7 **Total Number of Mentions and QS World Rankings**

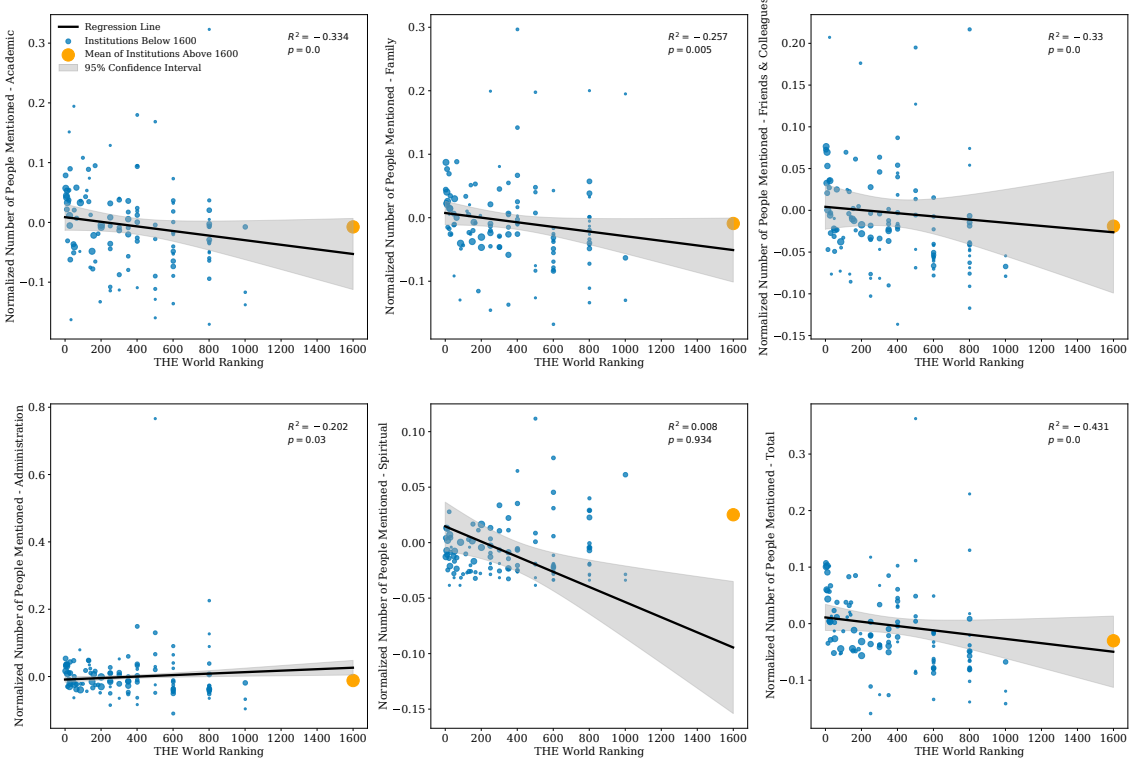

Figure A.8 **Total Number of Mentions and THE World Rankings**

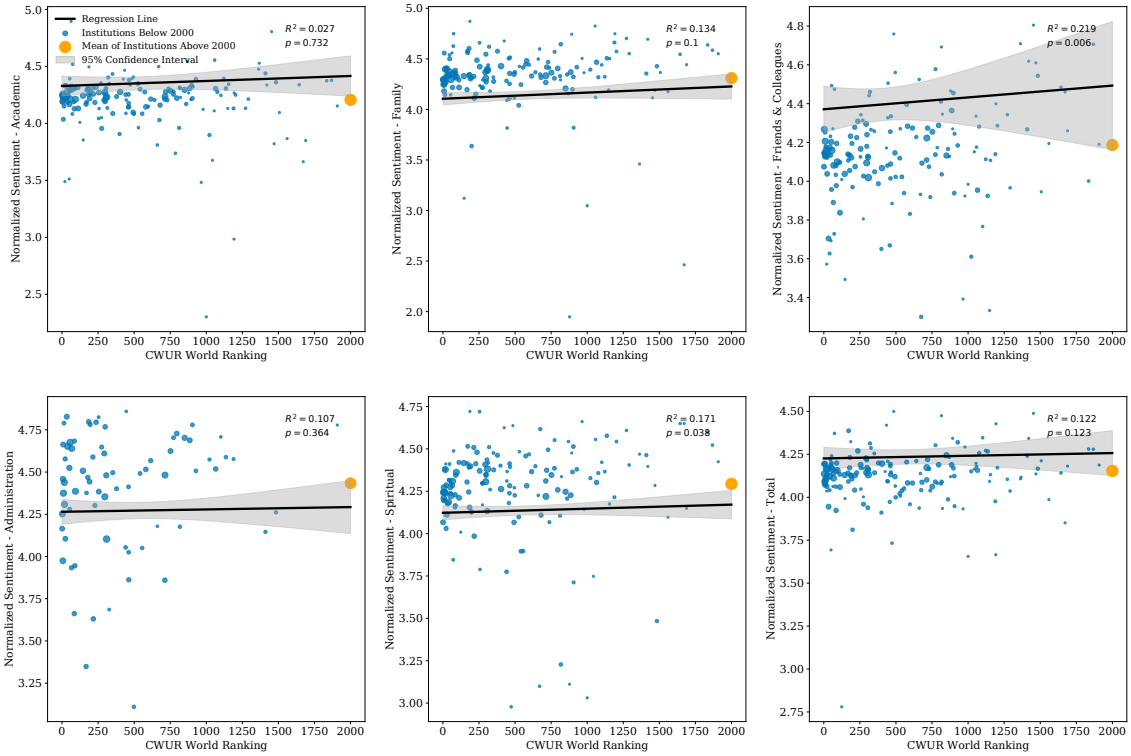

Figure A.9 **Sentiment Scores for Each Support Provider Category and CWUR World Rankings**

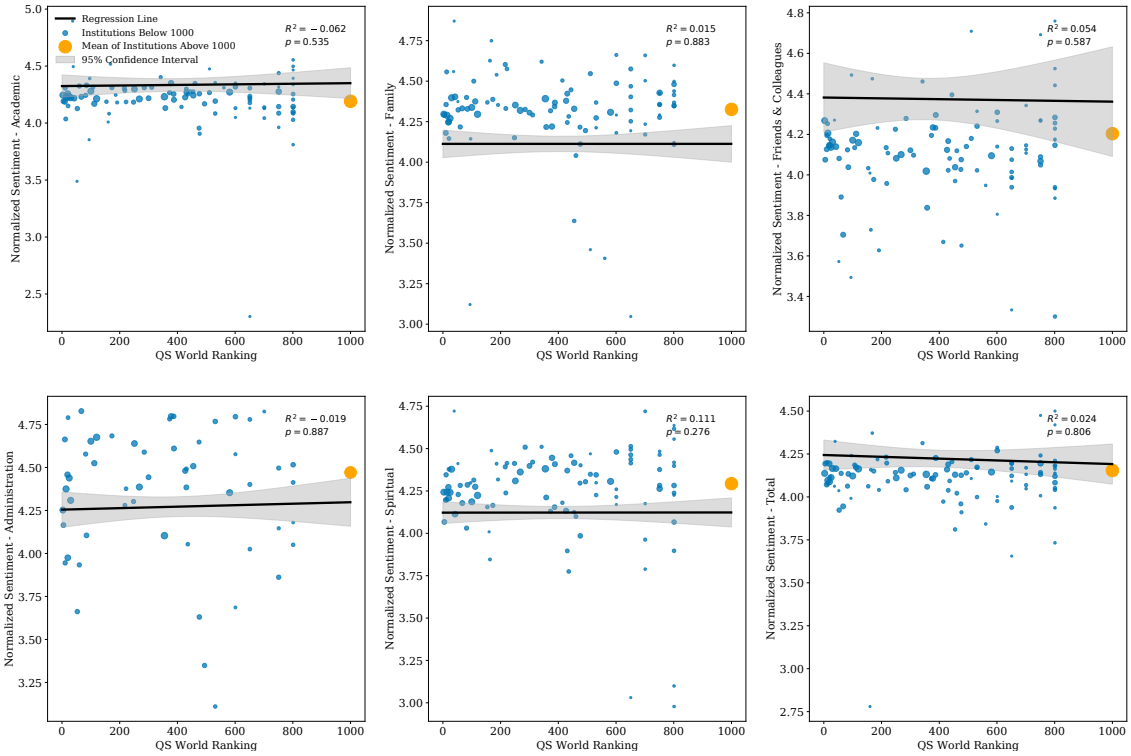

Figure A.10 **Sentiment Scores for Each Support Provider Category and QS World Rankings**

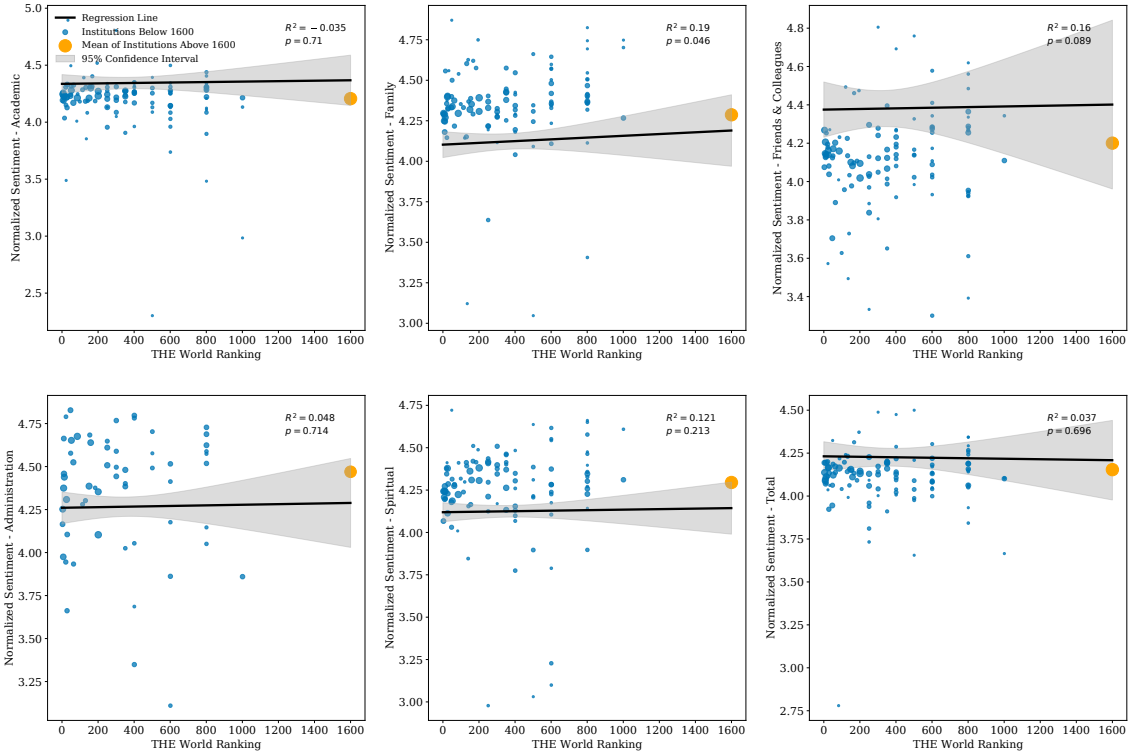

Figure A.11 **Sentiment Scores for Each Support Provider Category and THE World Rankings**