EXPLORING GENDER DISPARITIES IN COLLABORATION NETWORKS: AN ANALYSIS OF H-INDICES AND COLLABORATOR PROXIMITY

A Thesis

presented to

the Faculty of California Polytechnic State University,

San Luis Obispo

In Partial Fulfillment

of the Requirements for the Degree

Master of Science in Computer Science

by

Andrew Estrada

June 2024
COMMITTEE MEMBERSHIP

TITLE: Exploring Gender Disparities in Collaboration Networks: An Analysis of h-Indices and Collaborator Proximity

AUTHOR: Andrew Estrada

DATE SUBMITTED: June 2024

COMMITTEE CHAIR: Theresa Migler, Ph.D.
Assistant Professor of Computer Science

COMMITTEE MEMBER: Zoë J. Wood, Ph.D.
Professor of Computer Science and
Associate Dean for Diversity and Student Success,
College of Engineering

COMMITTEE MEMBER: Mugizi Rwebangira, Ph.D.
Assistant Professor of Computer Science
ABSTRACT

Exploring Gender Disparities in Collaboration Networks: An Analysis of h-Indices and Collaborator Proximity

Andrew Estrada

Research is crucial for expanding the boundaries of what is known, driving innovation, and solving problems faced by communities. It is carried out across all sectors of society by all manner of institutions. Academic research is one such sector that contributes to a plethora of disciplines. Research often compels collaboration among researchers, and as with any team, the dynamics and outcomes are affected by the individuals who contribute to the research. For instance, researchers can be from different institutions; therefore research teams can differ in collaboration distance — the geographic distance between the organizations of authors. In a similar vein, facets of individual researchers may impact their collaboration patterns. Two such known sources of difference are gender and an author’s measured impact.

This thesis investigates differences in geographic collaboration distance and correlations between impact and network metrics based on the inferred gender of authors from the California Public University system. In particular, this thesis uses publication data primarily from the area of computing with contributions from authors of California Polytechnic State University, San Luis Obispo and University of California schools. From this data, two collaboration networks are constructed with one used to calculate two measurements of collaboration distance for each author — distance of individual collaborations and reach of collaborations over time — and the other to calculate network metrics by author impact.

This thesis provides evidence suggesting a differences in collaboration distance over time and network metrics of inferred female and male authors. These differences tend to favor male authors.
ACKNOWLEDGMENTS

Thank you to:

• My advisor, Dr. Theresa Migler, and her family. Dr. Migler’s support and overwhelming confidence in me and my abilities has been incredibly meaningful throughout my time at Cal Poly. I am so glad I was able to work on my thesis with her; her excitement for the work we were doing was contagious and invigorating. Both she and Prof. Andrew Migler are clearly so passionate and they both helped equip me with the skills needed to contribute to this project. I must also thank Gabriel and Felicity for their warm smiles and outstanding baked treats. I wish Dr. Migler all the best in Spain.

• My family for their love and support — my parents and my girlfriend for always believing in me along this journey, and my sister for being my reason to be a good role model.

• The BEACoN Program for its funding and support throughout this project.

• All previous contributors to this project: Theresa Migler, Mugizi Rwebangira, Zoë Wood, Colin Chun, Mitashi Parikh, Logan McNichols, Steven Pineda, Emma Sauerborn, Brandon Tat, Kevin Yoo, Lauren Nakamichi, Conor Carroll, Nupur Garg, Gabriel Medina-Kim, Viet Lien Nguyen, Christian Rapp, Leticia Siqueira, Polina Volnuhina, Meghan Tran, Riley Mae Badnин, Justin Brunings, Bill Chan, Alex Liang, Eeshan Mishra, Drew Soderquist, and Alex Johnson.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>LIST OF TABLES</th>
<th>viii</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIST OF FIGURES</td>
<td>ix</td>
</tr>
</tbody>
</table>

## CHAPTER

1. **Introduction**  
2. **Background**  
   2.1 Cal Poly Research  
   2.2 Collaboration Networks  
   2.3 Impact and h-Index  
3. **Related Work**  
   3.1 California Public Universities  
   3.2 Gender  
   3.3 Distance  
4. **Methods**  
   4.1 Data Collection  
   4.2 Network Building and Analysis  
     4.2.1 California Public University Network  
     4.2.2 Cal Poly Network  
   4.3 Distance Calculation  
5. **Findings**  
   5.1 Distance  
     5.1.1 Average Distance of Collaborations  
     5.1.2 Average Distance to Collaborators  
     5.1.3 Movement of Authors  
   5.2 Impact  

vi
5.2.1 h-Index ................................................. 27
5.2.2 Degree ................................................. 27
5.2.3 Clustering Coefficient ............................... 29
5.2.4 Betweenness Centrality ............................. 29
5.2.5 Average Distance to Collaborators ................ 31
  5.2.5.1 Cal Poly ........................................ 32
  5.2.5.2 UC Davis ........................................ 33
6  Discussion .................................................. 37
7  Conclusions and Future Work ............................ 39
BIBLIOGRAPHY ............................................... 42
APPENDICES
  A  Histograms .............................................. 46
  B  SQL Database .......................................... 47
<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.1</td>
<td>Statistics for Average Distance of Collaborations</td>
<td>23</td>
</tr>
<tr>
<td>5.2</td>
<td>Statistics for Average Distance to Collaborators</td>
<td>24</td>
</tr>
<tr>
<td>5.3</td>
<td>Statistics for Average Distance to Collaborators Across Different Time Periods</td>
<td>25</td>
</tr>
<tr>
<td>5.4</td>
<td>Average Number of Affiliated Publication-Specific Organizations by Gender for Base Authors</td>
<td>26</td>
</tr>
<tr>
<td>5.5</td>
<td>h-Index Across All Cal Poly Authors and Among the Top 10%</td>
<td>27</td>
</tr>
<tr>
<td>5.6</td>
<td>Degree Across All Cal Poly Authors and Among the Top 10% By h-Index</td>
<td>27</td>
</tr>
<tr>
<td>5.7</td>
<td>Clustering Coefficient Across All Cal Poly Authors and Among the Top 10% By h-Index</td>
<td>29</td>
</tr>
<tr>
<td>5.8</td>
<td>Betweenness Centrality Across All Cal Poly Authors and Among the Top 10% By h-Index</td>
<td>30</td>
</tr>
<tr>
<td>5.9</td>
<td>Average Distance to Collaborations (km) Across All Cal Poly Authors and Among the Top 10% By h-Index</td>
<td>32</td>
</tr>
<tr>
<td>5.10</td>
<td>h-index Across All UC Davis Authors and Among the Top 10%</td>
<td>34</td>
</tr>
<tr>
<td>5.11</td>
<td>Average Distance to Collaborations (km) Across All UC Davis Authors and Among the Top 10% By h-index</td>
<td>34</td>
</tr>
</tbody>
</table>
## LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1 Distribution of All Publication Years</td>
<td>15</td>
</tr>
<tr>
<td>4.2 Authors By Gender</td>
<td>16</td>
</tr>
<tr>
<td>4.3 Collaborations on a Geographic Map for an Example Author A</td>
<td>19</td>
</tr>
<tr>
<td>5.1 Average Distance of Collaborations Across All Authors</td>
<td>24</td>
</tr>
<tr>
<td>5.2 Average Distance to Collaborators Across All Authors</td>
<td>25</td>
</tr>
<tr>
<td>5.3 Average Distance to Collaborators Across Different Time Periods</td>
<td>25</td>
</tr>
<tr>
<td>5.4 Average Degree for All Cal Poly Authors by h-Index</td>
<td>28</td>
</tr>
<tr>
<td>5.5 Average Degree for Female Cal Poly Authors by h-Index</td>
<td>28</td>
</tr>
<tr>
<td>5.6 Average Degree for Male Cal Poly Authors by h-Index</td>
<td>28</td>
</tr>
<tr>
<td>5.7 Average Clustering Coefficient for All Cal Poly Authors by h-Index</td>
<td>29</td>
</tr>
<tr>
<td>5.8 Average Clustering Coefficient for Female Cal Poly Authors by h-Index</td>
<td>30</td>
</tr>
<tr>
<td>5.9 Average Clustering Coefficient for Male Cal Poly Authors by h-Index</td>
<td>30</td>
</tr>
<tr>
<td>5.10 Average Betweenness Centrality for All Cal Poly Authors by h-Index</td>
<td>31</td>
</tr>
<tr>
<td>5.11 Average Betweenness Centrality for Female Cal Poly Authors by h-Index</td>
<td>31</td>
</tr>
<tr>
<td>5.12 Average Betweenness Centrality for Male Cal Poly Authors by h-Index</td>
<td>31</td>
</tr>
<tr>
<td>5.13 Average 'Avg. Distance to Collaborators' (km) for All Cal Poly Authors by h-Index</td>
<td>33</td>
</tr>
<tr>
<td>5.14 Average 'Avg. Distance to Collaborators' (km) for Female Cal Poly Authors by h-Index</td>
<td>33</td>
</tr>
<tr>
<td>5.15 Average 'Avg. Distance to Collaborators' (km) for Male Cal Poly Authors by h-Index</td>
<td>34</td>
</tr>
<tr>
<td>5.16 Average 'Avg. Distance to Collaborators' (km) for All UC Davis Authors by h-index</td>
<td>35</td>
</tr>
<tr>
<td>5.17 Average 'Avg. Distance to Collaborators' (km) for Female UC Davis Authors by h-index</td>
<td>35</td>
</tr>
</tbody>
</table>
Chapter 1

INTRODUCTION

Academic research is a crucial driver of change and progress. Research usually requires more than a team of one and collaborations can take many forms, from two people who share an office to tens or hundreds of people who are spread across the globe. The backgrounds and interactions of researchers profoundly influence the trajectory and results of investigation; the notoriety of authors influences how audiences respond to their research. Subsequently, understanding the interactions of research can provide meaningful insight into how such influential knowledge is created and received by others as well as suggest areas for improvement.

It has been shown that demographics of individuals can have varying impacts. For instance, researchers of underrepresented race and gender are more likely to produce novel or innovative connections between ideas, yet their work is less likely to be impactful and they are less likely to sustain careers in research at rates of 5% to 25% [14]. Even though it has been shown that diversity of perspectives [26] and, by extension, attributes such as gender [29], is beneficial to problem-solving and research, research and academia in the U.S. remains predominantly male [32].

Taking a wider perspective however, trends can take shape over the relationships between people, trends that with some analysis can yield important findings. Bearman and Moody looked at friendship data for over 13,000 middle and high school students, and found that the density of friendship networks (how likely an individual’s friends are also friends) had an effect on suicide ideation for females and on the likelihood of suicide attempt for males [4].

One method of analyzing relationships is the use of social network graphs. This type of analysis has been used to identify trends in collaboration — a specific type of relationship. It
was used to show that among Italian researchers, female researchers have a greater tendency to collaborate than males except at the international level by Abramo et al. [2] as well as prove that, within the field of life science, the tendency to collaborate with individuals of the same gender, known as gender homophily, still exists within academia by Holman and Morandin [15].

In 2020, Carroll et al. investigated the geography of research collaboration from computing-adjacent publication data collected for authors from University of California (UC) schools. They created a web-based visualization tool which utilized the latitude and longitude coordinates of organizations, allowing them to examine organization collaborations on a world map. They found a high tendency for the UC schools to work on publications with European organizations. This work, while insightful, only looked at organizations as a whole; it didn’t concern the demographics of individual researchers.

With the goal of further understanding collaboration patterns, this thesis intends to investigate what trends exist within research collaboration with regard to gender. In particular, what trends exist in terms of collaboration distance and impact with regard to gender? This is the driving question behind the investigation into the California Polytechnic State University (Cal Poly) and UC research collaboration network with an eye toward computing.

- **Collaboration Distance** - For this thesis, ‘collaboration distance’ is defined as a measure of the geographic distance between an author and everyone they have published with. International collaboration, from a city-to-city perspective, has increased over the last three decades along with the average distance to the strongest collaboration partners [9]. The geographic distance of research collaborations has mostly been investigated on a larger scale, either by city [9] or by university [22]. This thesis aims to investigate whether differences in collaboration distance at the scale of individual authors exist with regard to gender.
• Impact - The impact of authors has long been studied and with good reason – authors deserve recognition for their contributions and audiences benefit from knowing the notoriety behind the work they read. One measure of impact, using bibliometric data, is the h-index. This means that an author’s impact is measured in relation to the frequency at which their work has cited by other work. This thesis aims to identify trends in an author’s position in a collaboration network, both mathematically (network metrics) and geographically (collaboration distance), with respect to gender and impact as measured by h-index.

Computing as a topic was chosen as the area of interest because it defies departmental ties (ie. bioinformatics, mathematical algorithms, etc. fall within the area of computing) allowing for a broader base of authors and publications from a greater number of fields. This thesis uses preexisting data that was gathered from Scopus, an extensive database of authors and publications from Elsevier. Scopus was used to retrieve author data and publication data from publications surrounding the topic of computing with contributions from California public university authors primarily from 1970 to 2021.

From this data, two social networks for researchers were created and analyzed with regard to collaboration distance and impact. The creation of the networks follow a few key definitions:

- **Seed Schools:** Public California University
- **Vertices:** Seed authors and all authors who have collaborated with a seed author
- **Edges:** Connect authors who have contributed to the same publication
  - This ignores the publications with only one contributor.

This thesis contributes findings that

1. Within the realm of computing, reflect a difference in collaboration distance between female and male authors. Since no previous work investigates the same issue, the findings were rationalized using previous related findings.
2. Suggest a difference in network characteristics between male and female researchers and that these characteristics correlate with h-index.

The rest of the thesis proceeds as follows. Chapters 2 and 3 provide important preliminary knowledge of the topics, tools, and prior work. Chapter 4 details the processes followed to obtain findings. Chapter 5 reports what was found from the data. Lastly, Chapters 6 and 7 recount what this could mean and reiterates its importance.
Chapter 2

BACKGROUND

Section 2.1 details Cal Poly efforts in relation to research; Section 2.2 provides background knowledge on the topic of collaboration networks; Section 2.3 explains h-index.

2.1 Cal Poly Research

Cal Poly San Luis Obispo is a well respected university known for its “Learn by Doing” approach to education. Cal Poly is ranked as the best public master’s-level university in the West for the 31st time in a row and ranked the best overall master’s-level university in the West for the first time in the U.S. News & World Report’s annual Best Colleges guidebook for 2024 among 120 universities in the region. The university has also ranked number one in the West for Most Innovative Schools since 2018 [31]. These rankings, among numerous others, are a testament to the university’s commitment to excellence and continuous improvement.

In 2011, Cal Poly adopted a framework for the teacher-scholar model in AS-725-11 Resolution on Defining and Adopting the Teacher-Scholar Model with the goal of creating a more engaging academic environment by further applying the university’s flagship “Learn by Doing” approach to faculty [33][5]. The teacher-scholar model has since been incorporated into retention, promotion and tenure standards for instructional faculty university wide, solidifying a new emphasis [5].

In 2019-2020, Cal Poly launched two research-focused programs, the Strategic Research Initiatives (SRI) program and the BEACoN Research Scholars program [6][7]. The SRI program is partnership between Academic Affairs, University Development, and the Division of Research whose research goal is to beneficially influence economic and social issues of the
Central Coast and Beyond [6]. The program has funded 11 projects across the following five broader themes:

1. Central Coast Place-Based Research
2. Community Health
3. Data Science & Analytics
4. Environment of California and Beyond
5. Technology Workforce

BEACoN, which stands for Believe, Educate & Empower, Advocate, Collaborate, Nurture, is a program run by Cal Poly’s Office of University Diversity and Inclusion whose goal is to help provide underrepresented students with research experience by connecting them to faculty mentors across all colleges. BEACoN scholars work with their mentors for 2 quarters and receive a total stipend of $4,000. The BEACoN program also has a travel fund to encourage current and previous BEACoN scholars to take part in academic conferences related to their work [7].

2.2 Collaboration Networks

A graph is composed a set of nodes (a.k.a. vertices) and a set of edges, in which each edge is a connection between two nodes. A collaboration network is a type of graph in which nodes represent actors (authors, researchers, inventors, etc.) and edges connect actors who have worked together in some capacity. Collaboration networks act as representation of how actors within a certain population interact and can reveal a large variety of trends.

A few data points of interest that are observable within a network for identifying trends include:

1. **Degree** – the number of edges connected to a node; reflects the number of people a certain actor has worked with.
2. **Clustering Coefficient** – a measure of how likely a node’s neighbors are also connected; reflects the extent to which an actor’s collaborators create a tightly knit community.

3. **Betweenness Centrality** – a measure of how frequently a node lies in the shortest path between other pairs of nodes; reflects the degree to which an actor acts as a mediate between distinct communities.

Using these metrics and more, K. Whittington studied the differences in the network positioning of female and male inventors within the realm of life science. The study found that the largest difference was that women tend to have more dense sub-networks (higher clustering coefficient), whereas men were more likely to connect otherwise unconnected inventors (higher betweenness centrality) [34]. The study also found that men were more likely to be rewarded for their connections.

### 2.3 Impact and h-Index

Measuring the impact of scholarly work is important because while very few researchers are awarded the distinction of a Nobel Prize, there are many researchers whose work is valuable and influential. Measuring impact provides a metric for assessing the quality and significance of a researcher’s work and therefore gives rise to the ability to recognize influential/notable authors in different fields.

The impact of authors, publications, journals, etc. is most commonly measured using bibliometric data since it is based on objective, calculable values. One naive measure is simply taking citation counts for face value, such as an author’s total citation count. A more nuanced and commonly accepted measure is an author’s **h-index** which was introduced in 2005 by Jorge E. Hirsch [13]. The h-index is included as a statistic for authors in many citation databases like Scopus, Web of Science, and Google Scholar.

The h-index was proposed by J. E. Hirsch in 2005 as a way to measure the impact of a physicist with the idea that it could be used across other disciplines [13].
The h-index is calculated using the citation counts of each of an author’s papers and can be defined as the largest value \( h \) such that \( h \) of an author’s papers have \( h \) or more citations each. For example:

- An author who has published 10 papers which each have 7 citations, has an h-index of 7.
- An author who has published 7 papers which each have 10 citations, also has an h-index of 7.

The h-index avoids many of the pitfalls of other metrics such as over weighing a small number of highly-cited papers, penalizing high productivity, and requiring the use of an arbitrary boundary [13], however, it faces criticism and debate surrounding its validity [3] and bias [16, 10].

As with many metrics, the h-index has progressed to use as a heuristic for decision making. The h-index influences hiring, promotion, and funding decisions [1, 21, 12]. It is not a big jump to then say that the h-index sways the landscape of research and its evolution.

Unfortunately, impact measured by bibliometric data has been found to be biased. For instance, the share of all citations held by the highest 1\% of authors increased dramatically from 14\% to 21\% from 2000 to 2015 [24]. Moreover, within academic medicine, it has been shown that female authors have significantly lower h-indexes than male authors across most fields and ranks [10].
3.1 California Public Universities

In 2020, Nakamichi et al. compiled collaboration data from the Microsoft Academic Knowledge API, MathSciNet, IEEE Xplore API and publication lists provided by researchers to analyze collaboration trends in 4 departments of Cal Poly – Biology, Computer Science, Electrical Engineering and Math [23]. They investigated claims about gender differences in research patterns from previous findings for each department, using Gender API to infer binary gender of the authors.

The following claims were investigated: Men Tend to Have More Collaborators, Women Tend to Repeatedly Collaborate with the Same Collaborators, Researchers Tend to Collaborate with Authors of the Same Gender, and Women Tend to Collaborate More Intramurally.

Nakamichi et al. found evidence agreeing with the previous findings for certain departments, but no claim had all departments in agreement. In particular, the Math Department was often an outlier. Interestingly, however, all claims held true for the Computer Science department [23].

In 2019, McNichols et al. investigated 3 nested scopes of the Cal Poly collaboration network – the Computer Science Department, the College of Engineering, and university-wide – drawing information from Google Scholar, Microsoft Academic, arxiv, Grants applications from the university’s grants office and, in the case of faculty members of the Computer Science Department, personal web-pages and curricula vitae [19]. Only collaborative experiences with 7 or fewer collaborators were included. McNichols et al. used Gender API to infer gender of authors an investigate claims of gender differences similar to those of
Nakimichi et al. Notably, the paper echos that men have more collaborators at all scopes, but opposes Nakimichi et al.’s findings that women have a higher tendency to collaborate with the same collaborators. The authors found evidence supporting the claim that a “higher fraction of women’s co-authors are co-authors with each other” university-wide and within the College of Engineering. They also found that gender homophily across all three nested networks was greater in 2019 than it was in 2009 [19].

In 2020, Carroll et al. created a web-based tool for visualizing university-to-university collaborations and used it for geospatial analysis using publication data rooted in faculty members in University of California computer science and electrical engineering departments. Publication data from 1967 to 2017 was aggregated from the Scopus database [8]. Utilizing their tool they were able to examine two-step neighborhoods about institutions, random walks, and rank universities based off willingness to collaborate. They found that high collaborator count publications were the cause of the rankings heavy skew towards Europe and invariability of the rankings despite changes in time period thresholds. By limiting analysis to publications from teams up to 200 people, the rankings shifted to favor California schools as would be expected since the network was built from UC school collaborations. Irregardless, the paper showed evidence for a strong tendency of UC schools to collaborate with European institutions [8].

In 2021, McNichols et al. expanded on the visualization work of Carroll et al. with an upgraded UI and additional analyzation features [20]. The paper used publication data surrounding Cal Poly authors and University of California “faculty with a research interest in Computing” from the Scopus database. The authors chose to exclude publications containing more than 10 authors from the analysis. Latitude and longitude coordinates were obtained from data fields available in Scopus using the open source library GeoPy, and the NamSor API was used for gender and ethnicity inference. With this compiled data, the authors were then able to compare a networks:

- Cal Poly vs. UC Computing: Cal Poly authors had fewer collaborators on average compared to UC authors.
• Cal Poly Computing vs. UC Santa Barbara Computing: Co-authors of Cal Poly Computing authors are co-authors of each other at a substantially higher rate than those of UCSB Computing authors, meaning the Cal Poly network is denser.

• Cal Poly vs. Cal Poly Computing: The Cal Poly Computing subnetwork had a higher average collaborator count than the full Cal Poly network.

• Cal Poly Female Subnetwork vs. Cal Poly Male Subnetwork: The male network had a larger span of collaborators while the female network had stronger collaboration ties. Interestingly, within Cal Poly Computing, the inferred female subnetwork had a higher average collaborator count than the inferred male subnetwork.

UC Santa Barbara was singled out because it is of comparable size to Cal Poly and both are public, but it is PhD-level research institution [20].

3.2 Gender

Spoon et al., looking at employment census data and survey responses, found that not only do women leave academia at higher rates than men, they do so (or are considering doing so) for gendered reasons across the entire U.S. university system [28]. The findings provide more evidence that despite the fact that the rate at which women receive doctoral degrees is increasing, there remains a gender disparity in academia because the retention rate of female faculty is lower.

With more focus on collaboration, Whittington used global life-science patent information to analyze the collaboration network of 216,000 inventors. The analysis showed that women are less likely to be in positions of strategic advantage within the network – they are less likely to have “brokerage” ties where they connect two otherwise unconnected inventors [34].

In 2016, Charisse Madlock-Brown and David Eichmann constructed the collaboration network of 17 medical research institutions that are Clinical and Translational Science Awardees [17].
They were then able to investigate, for each institution, differences in clustering coefficients and betweenness centrality with regard to gender, productivity (number of papers), and h-index. When investigating clustering coefficients, the study found evidence that women tend to have higher clustering coefficients and that clustering coefficient is negatively correlated with both productivity and h-index. When investigating betweenness centrality, they found:

- The average betweenness centrality was greater for men across all institutions,
- Betweenness centrality was highly correlated with productivity for both men and women, and
- There was a much stronger correlation between h-index and betweenness centrality for women (almost twice that of men).

The paper suggests that men and women differ in collaboration network characteristics with betweenness centrality being strongly tied to impact and professional advancement, especially for women [17].

### 3.3 Distance

Not much research exists comparing geographic distance along gender although, Abramo et al., showed that among Italian researchers, female researchers have a lower tendency to collaborate internationally than males, despite overall having a greater propensity to collaborate [2].

More generally in regards to collaboration distance, the 2020 study by Csomos et al. looked at Web of Science (WoS) data for three 2-year spans from 1994-2016 to compile a city-to-city collaboration network. They found that international collaboration has increased along with the average distance to the strongest collaboration partners. While it is the case across all three decades that the farther the distance, the weaker the collaboration ties,
high-impact collaborations tend to span large distances, suggesting impact can be gained from a larger geographic reach [9].
Chapter 4

METHODS

Section 4.1 details how data was obtained; Section 4.2 details what restrictions were enforced when using the data to analyze the collaboration data; and Section 4.3 details how distance was calculated.

4.1 Data Collection

This thesis uses data that was previously gathered from Scopus [20], an extensive database of authors and publications from Elsevier. Scopus was used to retrieve author and publication data from publications with contributions from California public university authors primarily from 1970 to 2021. See Figure 4.1 for the distribution of publications by year. For Cal Poly, all authors were included, whereas for the UC schools, authors were only included if they had a subject of ‘computer science’ [20].

From Scopus, the name, organization, document count, citation count, and field of each author was collected. For publications, the title, year, venue, language, cited by count, DOI, and contributors list was collected. The contributor list consists of the authors and the organizations they were affiliated with at the time of publication.

There were two separate instances in which data was pulled from Scopus. Firstly, most of the data was pulled during the Covid-19 pandemic when Scopus was free and data could be retrieved directly. There was then a desire for more information after the pandemic, however at this point a license was needed and the request had to be sent through an intermediary.

Latitude and longitude coordinates were obtained from data fields available in Scopus using the open source library GeoPy [20]. NamSor was used to infer both the binary gender of
The resulting database has collaboration information on 555,827 publications spanning 851,077 authors. The portion of the database surrounding Cal Poly has collaboration information on 13,686 publications spanning 35,953 individual authors, 3,752 of whom are from Cal Poly.

4.2 Network Building and Analysis

Two networks are used for investigation: one created with the authors of multiple California public universities, and one with just Cal Poly authors. The schools under analysis in each network are referred to as base schools and the authors affiliated with them as base authors. Organizations with misspellings and extensions of schools, such as specific departments, are included as base schools. The larger network is for investigating distance while the Cal Poly network is for investigating impact. In both networks, vertices represent authors and two
authors are connected with an edge if they have collaborated on a publication. This ignores publications with only one contributor. Each author is either a base author or someone who has directly collaborated with a base author.

Publications with more than 20 authors were ignored when creating the network (and consequently when conducting analysis) to avoid skewing of the data. The rationale behind a threshold of 20 is that it is unlikely that an author would be able to meaningfully engage with each collaborator in a larger group and that is what the findings of this thesis are meant to capture. The restriction of at most 20 authors brings the number of analyzed publications to 535,178 spanning 619,804 authors.

From this data, the collaboration networks were created and validated to appropriately reflect the true networks by crosschecking with publicly available data regarding the gender distributions of research within California public universities [25][30].
4.2.1 California Public University Network

The base schools for this network are the public California universities for which there are more than 3,000 authors in the database – Cal Poly, UC Berkeley, UC Davis, UC Irvine, UC Los Angeles, UC Riverside, UC San Diego, UC San Francisco, and UC Santa Barbara. This is in an attempt to ensure only universities for which data was specifically retrieved are analyzed, as opposed to schools whose authors only exist in the database because they collaborated on a publication.

The network was constructed implicitly through database queries connecting base authors to their collaborators. The queries were also used to analyze collaboration distance as described in Section 4.3.

4.2.2 Cal Poly Network

The single base school for this network is Cal Poly, and the Python package, NetworkX [11], was used for graph creation and analysis. For each publication with a Cal Poly affiliated contributor, a clique was created connecting all pairs of contributing authors.

For instance, suppose Cal Poly affiliated author A worked on a publication with authors B, C, and D. The edge set added to the network from a publication is all pairwise combinations of authors, in this case:

\{(A, B), (A, C), (A, D), (B, C), (B, D), (C, D)\} \rightarrow

Further suppose author A is a collaborator of another publication with authors B and E. The new edges added to the graph would be:
Lastly, suppose authors C and E worked together on another publication where none of the contributing authors are affiliated with Cal Poly. The edge \{(C, E)\} is not included in the network since the publication as a whole has no relation to Cal Poly.

Cal Poly was chosen for the impact network because it has the lowest percentage of authors missing an h-index in the database (just 0.45%). These authors are excluded from the network which brings the number of analyzed publications to 12,636 spanning 19,369 authors including 3,670 from Cal Poly. Cal Poly authors who only worked independently (do not have any collaborators) are also excluded from the graph. The resulting collaboration network has 19,157 vertices (3,456 of which represent base authors) and 75,857 unique edges.

The use of h-index in this analysis is not in support of the metric. It is used because it is:

1. A commonly available metric in citation databases.
2. Used in important decisions regarding hiring, promotion, and funding and is therefore impactful in the academic community.
3. Already present in the database for a portion of authors from Scopus from the time of data retrieval. This means that it is more representative of the gathered collaboration data than if we were to get another metric for the authors now.

In the gender analysis of the network, only base authors with a gender inference probability \( \geq 75\% \) from NamSor are included to more accurately compare female and male authors. This restriction mainly excludes authors whose first name in Scopus is just an initial (ie. “F. Lastname”).
When obtaining statistics regarding the Cal Poly authors with the highest impact measured by h-index, we chose authors with h-index \( \geq 7 \) since it was approximately the top 10% of the population and no secondary metric would be needed to separate authors with the same h-index. 11.44% of all base authors, 10.37% of female base authors, and 11.95% of male base authors have a h-index \( \geq 7 \).

### 4.3 Distance Calculation

Distance calculation for an individual author is a measure of the geographic distance between their affiliated intuition and the affiliated institutions of their collaborators. There are two pieces of location data associated with authors within the database – the affiliated organization at the time data was retrieved (which will be referred to as *current*), and the organization they were affiliated with at the time of each publication (which will be referred to as *publication-specific*). With this information, there were two ways in which distance was calculated:

![Collaborations on a Geographic Map for an Example Author A](image)

**Figure 4.3: Collaborations on a Geographic Map for an Example Author A**

1. **Average Distance of Collaborations** - A collaboration is two authors working together for a specific publication. The distance of a collaboration is the distance between the authors’ publication-specific organizations. This calculation looks specifically at publications for which an author was at a base school.
For example, let us suppose there is an author $A$ who worked on three publications $P_1$, $P_2$, and $P_3$. Further suppose that for $P_1$ author $A$ was at base school $S_1$ working with author $B$; for $P_2$ author $A$ was at institution $I$ working with author $B$; and for $P_3$ author $A$ was at base school $S_2$ working with authors $B$ and $C$. Meanwhile, authors $B$ and $C$ remained at organizations $O_1$ and $O_2$, respectively, for all publications. Refer to Figure 4.3 for the corresponding visual – note that this does not represent the network construction but instead the location and connection of authors for the publications. Author $A$’s average distance of collaborations, assuming no others exist, is

$$\frac{(\text{distance from } S_1 \text{ to } O_1) + (\text{distance from } S_2 \text{ to } O_1) + (\text{distance from } S_2 \text{ to } O_2)}{3}$$

(4.1)

Notice that $P_2$ is ignored since author $A$ was not affiliated with a base school when contributing.

Collaboration data from 317,655 publications, all of which have at least one author who was affiliated with a base school at the time of publication was analyzed in this manner.

2. **Average Distance to Collaborators** - This is a bit more straightforward using the current organization to calculate distance. For each author whose current organization is a base school, the average the distance to unique collaborators is calculated.

Using the same example above (Figure 4.3), let’s suppose that author $A$’s current organization is base school $S_2$; their average distance to collaborators is:

$$\frac{(\text{distance from } S_2 \text{ to } O_1) + (\text{distance from } S_2 \text{ to } O_2)}{2}$$

(4.2)

Collaboration data from 438,576 publications, all of which have at least one author who is currently affiliated with a base school, was analyzed in this manner.
The Average Distance of Collaborations gives information to the effect of “at any given
time, how far are an author’s collaborators?”, while the Average Distance to Collaborators
is more a measure of an author’s reach over time.

Currently, distances to organizations for which there is not already longitude and latitude
data (NULL values within the database) are ignored. Of all organizations in the database
6.03% did not have location data. In both versions of the distance calculations, there exist
authors who have no distance data, either because they did not collaborate with anyone or
because they only collaborated with authors from organizations which there is no location
data for.

The Haversine formula was chosen to calculate distance between organizations of authors
[18]. While the use of an existing distance calculation API was considered, there were
paywalls, usage limitations, and most gave information irrelevant to the investigation. For
example, many gave distance and time computed for commuting between two locations.
This is way beyond the level of required detail since no assumptions were made about
means of collaboration or travel. A simple “as the crow flies” distance is appropriate.
The Haversine formula calculates distance between two latitude and longitude coordinates
converted to radians \((\phi_1, \lambda_1)\) and \((\phi_2, \lambda_2)\) as follows:

\[
\text{dist} = 2r \arcsin \sqrt{\sin^2 \left(\frac{\phi_2 - \phi_1}{2}\right) + \cos \phi_1 \cos \phi_2 \sin^2 \left(\frac{\lambda_2 - \lambda_1}{2}\right)}
\]  

(4.3)

As a simple example, the distance calculated by the Haversine formula between Cal Poly
and UC Berkeley using coordinates in the database is 317.97 km. The travel distance
according to Google is approximately 369 km. This is expected as there is no way to travel
in a straight line between the two locations.

Haversine and Vincenty’s are the 2 formulas for calculating great-circle distance (point A
straight to point B distance). The Haversine formula models the Earth as a sphere, whereas
Vincenty’s formula models the Earth as an ellipsoid. While both formulas are approxima-
tions, Vincenty’s formula uses a more accurate model of Earth’s irregular ellipsoid shape.
Vincenty’s formula has a worst case percent error of just 0.1113%, while the Haversine formula has a worst case percent error of 0.334%. The accuracy of Vincenty’s formula, however, comes with a higher calculation time cost that can be as much as double that of Haversine’s [18].

Haversine was chosen over the more accurate Vincenty’s formula since both levels of accuracy are more than sufficient for the analysis purposes of this thesis and the larger calculation overhead would be noticeable with the number of distances being calculated [18].
Chapter 5

FINDINGS

Section 5.1 presents findings related to collaboration distance from California Public University network; Section 5.2 presents findings related to impact from the Cal Poly network.

5.1 Distance

Section 5.1.1 details the results of analyzing individual collaborations – average distance of collaborations – and Section 5.1.2 details the results of analyzing reach – average distance to collaborators. Section 5.1.3 presents data regarding the number of affiliated organizations for female and male researchers.

5.1.1 Average Distance of Collaborations

From analyzing the collaboration data of all publications which have at least one author who was affiliated with a base school at the time of publication, there was little difference between genders. See Figure 5.1 for the distributions and Table 5.1 for aggregate data. This suggests that for any given collaboration, the distance will be the same regardless of the author’s gender.

<table>
<thead>
<tr>
<th>Data Point</th>
<th>Female Authors</th>
<th>Male Authors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average (km)</td>
<td>987.0</td>
<td>996.8</td>
</tr>
<tr>
<td>Max (km)</td>
<td>15051.8</td>
<td>15567.0</td>
</tr>
<tr>
<td>Median (km)</td>
<td>271.7</td>
<td>220.9</td>
</tr>
<tr>
<td>Mode</td>
<td>0 km (31.0%)</td>
<td>0 km (35.1%)</td>
</tr>
<tr>
<td>Total Author Count</td>
<td>36,313</td>
<td>81,819</td>
</tr>
<tr>
<td>Authors Without Collaborators</td>
<td>199</td>
<td>533</td>
</tr>
</tbody>
</table>
5.1.2 Average Distance to Collaborators

From analyzing the collaboration data of all publications which have at least one author who is currently affiliated with a base school, male authors tend to have a slightly further reach by approximately 144 km further than female authors. Note this shows female author’s reach being 92% that of male authors. See Figure 5.2 for the distributions and Table 5.2 for aggregate data.

<table>
<thead>
<tr>
<th>Data Point</th>
<th>Female Authors</th>
<th>Male Authors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average</td>
<td>1556.7 km</td>
<td>1700.7 km</td>
</tr>
<tr>
<td>Max</td>
<td>14979.6 km</td>
<td>15569.3 km</td>
</tr>
<tr>
<td>Median</td>
<td>965.5 km</td>
<td>1098.8 km</td>
</tr>
<tr>
<td>Mode</td>
<td>0 km (14.3%)</td>
<td>0 km (15.7%)</td>
</tr>
<tr>
<td>Total Author Count</td>
<td>43,715</td>
<td>87,527</td>
</tr>
<tr>
<td>Authors Without Collaborators</td>
<td>198</td>
<td>500</td>
</tr>
</tbody>
</table>

Three separate decades of data (1970–1980, 1990–2000, and 2010–2020) was subsequently analyzed in the same fashion. This trend (i.e. male authors having a slightly further reach than female authors) was consistent over the decades even though the average distance (reach) increased for both genders over time. See Figure 5.3 and Table 5.3.
Figure 5.2: Average Distance to Collaborators Across All Authors

Table 5.3: Statistics for Average Distance to Collaborators Across Different Time Periods

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Average</td>
<td>Female</td>
<td>Male</td>
<td>Female</td>
</tr>
<tr>
<td></td>
<td>1042.1 km</td>
<td>1302.1 km</td>
<td>1508.7 km</td>
</tr>
<tr>
<td>Max</td>
<td>12741.3 km</td>
<td>15569.3 km</td>
<td>14404.0 km</td>
</tr>
<tr>
<td>Median</td>
<td>174.0 km</td>
<td>358.6 km</td>
<td>844.2 km</td>
</tr>
<tr>
<td>Mode</td>
<td>0 km (30.8%)</td>
<td>0 km (27.3%)</td>
<td>0 km (18.4%)</td>
</tr>
<tr>
<td>Total Author Count</td>
<td>909</td>
<td>4,994</td>
<td>6,436</td>
</tr>
<tr>
<td>Authors Without Collaborators</td>
<td>44</td>
<td>211</td>
<td>83</td>
</tr>
</tbody>
</table>

Figure 5.3: Average Distance to Collaborators Across Different Time Periods
5.1.3 Movement of Authors

Since the relocation of authors (the difference in where they were for a publication vs. where they are now) is perhaps the biggest contributor to differences in the two distance metrics above, we investigated information to this effect. We calculated the average number of distinct organizations authors currently at base schools were affiliated with when contributing to publication — refer to Table 5.4.

Table 5.4: Average Number of Affiliated Publication-Specific Organizations by Gender for Base Authors

<table>
<thead>
<tr>
<th>Gender</th>
<th>Average Number of Affiliated Organizations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female Authors</td>
<td>1.406</td>
</tr>
<tr>
<td>Male Authors</td>
<td>1.688</td>
</tr>
</tbody>
</table>

The population of base authors and publications included in this calculation slightly differ from those in the average distance to collaborators metric because it includes publications with a solo contributor and includes publications in which a base author solely collaborated with authors from organizations with NULL location data in the database. While this calculation counts different affiliations from related organizations (due to reasons such as funding or department ties) it also counts misspellings as distinct and may therefore double count some ties.

5.2 Impact

Section 5.2.1 describes statistics of h-index across the population of Cal Poly authors. Sections 5.2.2 through 5.2.5 detail the relation between network characteristics and impact as measured by h-index. Plots were chosen to include aggregate data (averages) rather than all raw data for cleanliness and clarity of trends. They include a logarithmic color scale to indicate the number of authors included in the average for each h-index value.
5.2.1 h-Index

Over the whole population of Cal Poly authors, female authors have a lower average h-index; only \( \sim 87\% \) that of male authors with a difference of 0.367. Similarly, among the top \( \sim 10\% \) of authors (those with h-index \( \geq 7 \)), the average h-index of female authors is \( \sim 94\% \) that of male authors with a difference of 0.817. See Table 5.5 for statistics regarding h-index.

<table>
<thead>
<tr>
<th>Data Point</th>
<th>Whole Population</th>
<th>Top ( \sim 10% )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average</td>
<td>2.790</td>
<td>13.599</td>
</tr>
<tr>
<td>Female Avg.</td>
<td>2.541</td>
<td>13.021</td>
</tr>
<tr>
<td>Male Avg.</td>
<td>2.908</td>
<td>13.838</td>
</tr>
</tbody>
</table>

5.2.2 Degree

Looking at the distribution of degree – the number of unique collaborators – across the network, there is a consistent difference between men and women overall (\(-4.0\% \) for women), and among the most impactful authors (\(-4.9\% \) for women). Notably the average degree among the top authors is much higher than the overall average, meaning that top authors have a higher number of collaborators. Figures 5.4, 5.5, and 5.6 echo this trend as they all show a positive correlation between h-index and degree. See Table 5.6 for statistics regarding degree.

<table>
<thead>
<tr>
<th>Data Point</th>
<th>Whole Population</th>
<th>Top ( \sim 10% )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average</td>
<td>10.176</td>
<td>40.469</td>
</tr>
<tr>
<td>Female Avg.</td>
<td>9.895</td>
<td>39.042</td>
</tr>
<tr>
<td>Male Avg.</td>
<td>10.310</td>
<td>41.061</td>
</tr>
<tr>
<td>Median</td>
<td>5</td>
<td>27.5</td>
</tr>
<tr>
<td>Female Max.</td>
<td>222</td>
<td>222</td>
</tr>
<tr>
<td>Male Max.</td>
<td>345</td>
<td>345</td>
</tr>
</tbody>
</table>
Figure 5.4: Average Degree for All Cal Poly Authors by h-Index

Figure 5.5: Average Degree for Female Cal Poly Authors by h-Index

Figure 5.6: Average Degree for Male Cal Poly Authors by h-Index
5.2.3 Clustering Coefficient

Looking at Table 5.7 for clustering coefficient data in the network, women have a higher average clustering coefficient overall (+6.8%) and a lower average clustering coefficient within the top 10% of impactful authors (-5.6%). Unlike degree however, clustering coefficient seems to be negatively correlated with h-index as shown in Figures 5.7, 5.8, and 5.9.

Table 5.7: Clustering Coefficient Across All Cal Poly Authors and Among the Top 10% By h-Index

<table>
<thead>
<tr>
<th>Data Point</th>
<th>Whole Population</th>
<th>Top ~10%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average</td>
<td>0.645</td>
<td>0.265</td>
</tr>
<tr>
<td>Female Avg.</td>
<td>0.674</td>
<td>0.255</td>
</tr>
<tr>
<td>Male Avg.</td>
<td>0.631</td>
<td>0.270</td>
</tr>
<tr>
<td>Median</td>
<td>1</td>
<td>0.146</td>
</tr>
</tbody>
</table>

Figure 5.7: Average Clustering Coefficient for All Cal Poly Authors by h-Index

5.2.4 Betweenness Centrality

There is a large disparity in between centrality between female and male authors. Women have an average betweenness centrality only 66.6% that of men across the whole network (a difference of 0.000301) and only 61.1% that of men among the most impactful authors (a difference of 0.001966). See Table 5.8 for statistics regarding betweenness centrality. Figures 5.10, 5.11, and 5.12 suggest a positive correlation between h-index and betweenness centrality.
Figure 5.8: Average Clustering Coefficient for Female Cal Poly Authors by h-Index

Figure 5.9: Average Clustering Coefficient for Male Cal Poly Authors by h-Index

Table 5.8: Betweenness Centrality Across All Cal Poly Authors and Among the Top 10% By h-Index

<table>
<thead>
<tr>
<th>Data Point</th>
<th>Whole Population</th>
<th>Top ~10%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average</td>
<td>0.000806</td>
<td>0.004485</td>
</tr>
<tr>
<td>Female Avg.</td>
<td>0.000602</td>
<td>0.003095</td>
</tr>
<tr>
<td>Male Avg.</td>
<td>0.000904</td>
<td>0.005061</td>
</tr>
<tr>
<td>Median</td>
<td>0</td>
<td>0.001</td>
</tr>
<tr>
<td>Female Max.</td>
<td>0.031</td>
<td>0.031</td>
</tr>
<tr>
<td>Male Max.</td>
<td>0.055</td>
<td>0.055</td>
</tr>
</tbody>
</table>
5.2.5 Average Distance to Collaborators

We hypothesized that the measure of Average Distance to Collaborators may be different between Cal Poly and PhD-level R1 institutions due to factors such as funding. The same measure for UC Davis, a public California R1 institution, was then calculated for comparison in Section 5.2.5.2. UC Davis was chosen because among R1 institutions in the database, it has the lowest percentage of authors with NULL h-indexes (only 3.17% of the 22420 UC Davis authors). This was chosen over trying to web-scrape missing h-indexes since the
h-indexes would then be from different points of time, would be from different sources, and would be less reflective of the existing database data.

5.2.5.1 Cal Poly

In the Cal Poly network, women have a higher average distance than men overall (+2.6%) and more significantly among the top ∼10% authors (+11.0%). See Table 5.9 for statistics related to Cal Poly author reach. Figures 5.13, 5.14, and 5.15 all show a positive correlation between h-index and average distance to collaborators.

Table 5.9: Average Distance to Collaborations (km) Across All Cal Poly Authors and Among the Top 10% By h-Index

<table>
<thead>
<tr>
<th>Data Point</th>
<th>Whole Population</th>
<th>Top ∼10%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average</td>
<td>1,303.2 km</td>
<td>2,626.2 km</td>
</tr>
<tr>
<td>Female Avg.</td>
<td>1,326.3 km</td>
<td>2,824.7 km</td>
</tr>
<tr>
<td>Male Avg.</td>
<td>1,292.2 km</td>
<td>2,543.8 km</td>
</tr>
<tr>
<td>Median</td>
<td>369.7 km</td>
<td>2,268.1 km</td>
</tr>
<tr>
<td>Female Max.</td>
<td>12,548.5 km</td>
<td>8,411.8 km</td>
</tr>
<tr>
<td>Male Max.</td>
<td>12,985.2 km</td>
<td>10,897.8 km</td>
</tr>
</tbody>
</table>

Figure 5.12: Average Betweenness Centrality for Male Cal Poly Authors by h-Index
5.2.5.2 UC Davis

The final UC Davis collaboration network has 21,271 UC Davis authors. UC Davis differs slightly from Cal Poly in its demographic makeup, with roughly 40% women and 60% men. When looking at the make up of the top 10% of UC Davis authors by h-index, however, representation is very different. For Cal Poly, the top 11% represents 10% of women and 12% of men. For UC Davis, however, the top 11% (authors with h-index $\geq 23$) represents only 6% of women and 13% of men. See Table 5.10 for h-index statistics and Table 5.11 for distance statistics for UC Davis. Women have an average distance 92% that of men
Figure 5.15: Average 'Avg. Distance to Collaborators' (km) for Male Cal Poly Authors by h-Index

overall and 94% that of men in the top 10%. Figures 5.16, 5.17, and 5.18 suggest a positive correlation between h-index and average distance to collaborators.

Table 5.10: h-index Across All UC Davis Authors and Among the Top 10%

<table>
<thead>
<tr>
<th>Data Point</th>
<th>Whole Population</th>
<th>Top ~10%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average</td>
<td>8.694</td>
<td>37.231</td>
</tr>
<tr>
<td>Female Avg.</td>
<td>6.567</td>
<td>34.124</td>
</tr>
<tr>
<td>Male Avg.</td>
<td>10.027</td>
<td>38.140</td>
</tr>
<tr>
<td>Median</td>
<td>4</td>
<td>32</td>
</tr>
<tr>
<td>Female Max.</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Male Max.</td>
<td>164</td>
<td>164</td>
</tr>
</tbody>
</table>

Table 5.11: Average Distance to Collaborations (km) Across All UC Davis Authors and Among the Top 10% By h-index

<table>
<thead>
<tr>
<th>Data Point</th>
<th>Whole Population</th>
<th>Top ~10%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average</td>
<td>1,565.896</td>
<td>2,244.906</td>
</tr>
<tr>
<td>Female Avg.</td>
<td>1,488.940</td>
<td>2,141.885</td>
</tr>
<tr>
<td>Male Avg.</td>
<td>1,614.084</td>
<td>2,275.055</td>
</tr>
<tr>
<td>Median</td>
<td>928.483</td>
<td>1,900.919</td>
</tr>
<tr>
<td>Female Max.</td>
<td>14,979.646</td>
<td>14,824.026</td>
</tr>
<tr>
<td>Male Max.</td>
<td>14,959.888</td>
<td>14,782.219</td>
</tr>
</tbody>
</table>
Figure 5.16: Average 'Avg. Distance to Collaborators' (km) for All UC Davis Authors by h-index

Figure 5.17: Average 'Avg. Distance to Collaborators' (km) for Female UC Davis Authors by h-index
Figure 5.18: Average 'Avg. Distance to Collaborators' (km) for Male UC Davis Authors by h-index
Chapter 6

DISCUSSION

The lack of difference seen between female and male researchers in the distance of individual collaborations (Section 5.1.1) is encouraging. It suggests that female authors are afforded the same opportunities as male authors when it comes to collaboration outside their affiliated university.

On the other hand, we hypothesize that the disparity in geographic reach over time observed in Section 5.1.2 may be a reflection of previous findings that may compound. Since female researchers are less likely to be rewarded for their position in a collaboration network [34], they are more likely to leave academia, less likely to be pulled towards better opportunities, and less likely to relocate [28]. Section 5.1.3 also suggests that female researchers are less likely to relocate as they have a lower average number of affiliated organizations. This might help explain why the disparity exists in reach over time and not in the distance of individual collaborations. Because male authors move, or their male collaborators move — thanks to gender homophily [15] — their reach over time can be greater than the distance when the collaboration actually occurred. The tendency for women to collaborate less internationally during their academic careers [2], may intensify the difference.

Many other factors may be involved as well. For instance it may be the case that women are less likely to take on remote opportunities/collaborations which is measured in reach as well.

The greater tendency of women to leave academia may also contribute to the disparities observed in Section 5.2. They are less likely to grow a large collaboration base (smaller degree and higher clustering coefficient) and are subsequently less likely to connect authors who haven’t already collaborated (lower betweenness centrality), all because they are in academia for shorter time spans.
The significantly larger difference between female and male authors when analyzing betweenness centrality in the Cal Poly network is a bit puzzling. While it is possible that because the network is so large and the numbers are so small that the percent difference isn’t that meaningful, these researchers are all drawn from the same institution, are held to the same expectations, and most likely following similar processes of collaboration. We would therefore expect a small disparity in any given metric.

This large difference in betweenness centrality is not a isolated occurrence. K. Whittington in 2018 looked at patent collaboration information within the life sciences and found a difference in average betweenness centrality of 45% (favoring men) across the network even though there was little difference in the position of authors in the network (center/periphery) [34]. This differs from our findings, however, since the lowest disparity, just 19%, was found in university collaboration. The disparities found as a part of this thesis as well as in Whittington’s work might indicate a significant and fundamental pattern in collaboration networks. It may be the case that women and men are rewarded differently (men for breadth of connections and women for tightness of communities) and therefore each tend to follow those patterns. It may also be the case that women are somehow systematically being excluded from holding “central” connections. It may be something else entirely, but there seems to be a pattern in regards to betweenness centrality.

Overall, these findings speak to the gravity behind making the academic space more accepting of non-male researchers. This may include changing current measures of success within the field, such as abandoning h-index.
Chapter 7

CONCLUSIONS AND FUTURE WORK

Academic research has long been an important source of contribution to a wide variety of fields; it has impacted technology, industry, policy, and so much more. Impact, as measured by h-index has been shown to have influence over decisions made in the pivotal realm of academic research. It is interesting, particularly when geared towards such an emerging field as computing, to analyze patterns in the work being done. Using collaboration data from public California universities, we analyze, at the scale of individual authors, measures of collaboration distance and network metrics and their relation to h-index.

In terms of collaboration distance, we find evidence that in regard to gender, a difference lies in the geographic reach of an author over all their publications. On the side of network metrics, we find evidence suggesting network differences between female and male authors that favor men. These findings are relevant to discussions surrounding university-level research objectives with the aim of a more equitable academic environment.

Women on average have a lower average degree and betweenness centrality which are both correlated with higher h-indexes. Clustering coefficient is negatively correlated with h-index; men have a lower average clustering coefficient overall, while among the most impactful authors, women have a lower average clustering coefficient. This is in support of the work of both and C. Madlock-Brown and D. Eichmann [17] and K. Whittington [34] which suggest the tendency for women to work with small groups of people is penalized in collaboration networks. The greatest disparity between male and female authors was in betweenness centrality which was shown to be highly correlated with h-index for women [17]. Surprisingly, there seems to be little difference in average distance to collaborators at UC Davis and female authors even out perform male authors at Cal Poly.
We acknowledge that there are limitations to the results. One limitation of this thesis is that it relies on preexisting data built over time from other students and researchers, meaning there must be some level of blind trust that the data is sufficiently accurate. Many “sanity checks” were conducted at many stages about many facets of the data, but fully ensuring accuracy would require starting from scratch.

The collaboration data available is not fully representative of every author’s work which can skew the data. For instance, there are cases in which there only exists one publication in the database affiliated with an author that has an h-index of 35 which is not indicative of all the work they have produced. Therefore their true position and metrics in the network are different in this thesis’s representation. For the distance metric specifically, the data may be skewed by the lack of location data for some institutions.

Refinement of this work, as well as future work, can be done in two lanes:

For the two measures of distance, refinements may include analyzing the two distance types for the same author population (currently the populations overlap but are not the same) as well as excluding graduate students who are likely to have a collaboration distance of 0 which skews the average. This can be seen in the histograms of Figure A.1 and A.2 which both show a large number of authors with a collaboration distance of 0 or close to 0. Further work could explore distance patterns of international collaboration, distance patterns within specific subject areas, and distance patterns between schools vs. industry.

For the network metrics, refinements may include conducting error analysis on the data to account for fewer authors with high h-indexes as well as calculating the strength of correlations. Further work could explore combining the collaboration networks of more institutions to see the relations of metrics beyond the scale of a single institution. Trends with respect to ethnicity and non-binary gender can also be explored. Currently, in many of the plots there is an obvious trend among the lower h-indexes that begins to scatter more and more with higher h-indexes with less authors. It would be interesting to see if the number of h-indexes showing strong trends increases with the inclusion of more authors.
Lastly, the difference in betweenness centrality between genders is something worth further investigation. It may beneficial to look at it from a rewards standpoint — are male researchers more likely to be promoted, be awarded grants or distinctions, and be involved in impactful research if they have a high betweenness centrality? And does the same hold true for women to the same extent? It would be interesting to see how the metrics for an author change over time if we have a more dynamic social network that takes when authors collaborate into account — can we see betweenness centrality increase as an author’s notability increases?


Appendix A

HISTOGRAMS

Figure A.1: Distribution of Average Distance to Collaborators for Female Authors Currently at Base Schools

Figure A.2: Distribution of Average Distance to Collaborators for Male Authors Currently at Base Schools
Appendix B

SQL DATABASE

Figure B.1: SQL Database Schema Including the Tables Used

Below is an example SQL query used to obtain the Average Distance to Collaborators for female base authors.

\[
\text{AVG DISTANCE TO COLLABORATORS}
\]

With BaseUniv as ( 

\[
\text{Universities with > 3,000 authors}
\]
Select o.org_id
from Organization o
inner join OrgMappings om on o.mapping = om.mapping_id
where (om.school = "Cal Poly San Luis Obispo"
    or om.school = "UC Berkeley" or om.school = "UC Davis"
    or om.school = "UC Irvine" or om.school = "UCLA"
    or om.school = "UC Riverside" or om.school = "UC San Diego"
    or om.school = "UC San Francisco"
    or om.school = "UC Santa Barbara"
    or om.school = "UC Santa Cruz")
    and o.lat IS NOT NULL and o.lng IS NOT NULL
),
Pubs as (  
    -- Publications with at most 20 contributors.
    select pub_id, count(distinct auth_id) as NumCollaborators
    from Collaboration c
    group by pub_id
    having NumCollaborators <= 20
    ),
TempCollabs as (  
    -- All pairs of authors who have worked together on publications
    -- with at most 20 contributors.
    select distinct c1.auth_id as auth1, c2.auth_id as auth2
    from Collaboration c1
    inner join Collaboration c2 on c1.pub_id = c2.pub_id
    and c1.auth_id != c2.auth_id
    where c1.pub_id in (Select pub_id from Pubs)
    ),
UniqueCollabs as (
— *The distance between authors who have worked together where author 1 is a female base author.*

```sql
SELECT auth1, o1.org_name AS org1, auth2, o2.org_name AS org2,
       haversine(o1.lat, o1.lng, o2.lat, o2.lng) AS dist
FROM TempCollabs c
INNER JOIN Author a1 ON c.auth1 = a1.auth_id
  AND a1.gender = "female"
  AND a1.current_org IN (SELECT * FROM BaseUniv)
  AND a1.gender_prob > = .75
INNER JOIN Organization o1 ON a1.current_org = o1.org_id
INNER JOIN Author a2 ON c.auth2 = a2.auth_id
INNER JOIN Organization o2 ON a2.current_org = o2.org_id
WHERE o1.lat IS NOT NULL AND o1.lng IS NOT NULL
  AND o2.lat IS NOT NULL AND o2.lng IS NOT NULL
)

— *Average the distances for each author*

SELECT auth1, COALESCE(SUM(dist),0) AS total_dist,
       COUNT(*) AS collab_count,
       COALESCE(SUM(dist),0)/COUNT(*) AS avg_dist
FROM UniqueCollabs
GROUP BY auth1
```

49