KENNESAW STATE UNIVERSITY
COLLEGE OF COMPUTING AND SOFTWARE ENGINEERING
School of Data Science and Analytics

Charting the Course: A Data-Driven Approach to Ph.D. Program Fit

Nathalie Jones & Alexis Lowery

REFERENCES

INTRODUCTION

Previous research and conference experience influences students' motivations for pursuing a Ph.D. (Guerin, C. et al., 2014), but successful completion of a Ph.D. relies on perception of support from advisors, faculty, and other students (Litalien, D., & Guay, F., 2015). Conversely, social isolation can motivate Ph.D. candidates to drop out of the program (Young, S. N., et al., 2019), suggesting that prospective Ph.D. candidates' success depends on their fit in an institution's social atmosphere. Not every Ph.D. hopeful lives near an institution that offers a program within their interests, and travel to assess atmosphere can be costly. Without firsthand experience, Ph.D. applicants must somehow answer the question:

"Which school, program, and atmosphere fit me best?"

This project presents a personalized method to analyze fit between PhD applicants and Ph.D. programs. To analyze programs, faculty, and aspects of the application process, this example includes a set of inclusion criteria personalized to the author's research interests.

METHODS

CollegeScorecard.
The CollegeScorecard dataset includes data on 6,543 post-secondary institutions with 233,979 different fields of study at various academic levels. This analysis uses the following variables:

INSTNM. The institution's name
INSTURL. The URL to the institution's webpage
CITY. The city in which an institution is located
STABBR. The state in which an institution is located
CREDLEV. The credential level of the program
CIPDESC. The field of study the program offers

Inclusion Criteria.
Institutions included in the final dataset must:

- Be accredited & offer at least one graduate degree (11,411 programs from 1,828 institutions)
- Offer a doctoral degree program in statistics, data science, computer science, math, or neuroscience (8,012 programs from 1,219 institutions)
- Foster interdisciplinary research opportunities & offer courses in graph theory, probability, machine learning, or neuroscience (220 programs from 60 institutions)

Web Scraping Program & Faculty Data.
The data included faculty publication data scraped from Google Scholar and institutional webpages and directories using the following methods:

- *Beautifulsoup.* A Python library that calls/extracts webpage HTML. Used to collect program and application information (e.g., degree outlines, statement prompts, faculty member information, application deadlines) using official institution URLs from the CollegeScorecard. This project used this library to scrape program webpages and institutional directories to collect potential faculty members of interest
- *SerpAPI.* A Python API service that automatically scrapes Google webpages (e.g., Google Scholar). Used to collect publication information (i.e., title, abstract, year, co-author information)
- The final dataset included 31,109 publications by 383 faculty members from 8 schools

RESULTS

Figure 2: Frequency of Application Fees by Cost

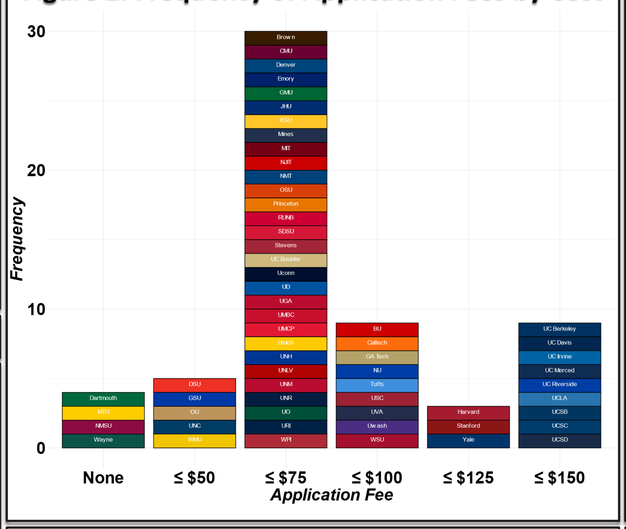


Figure 3: Fee Waiver Availability & GRE Requirements

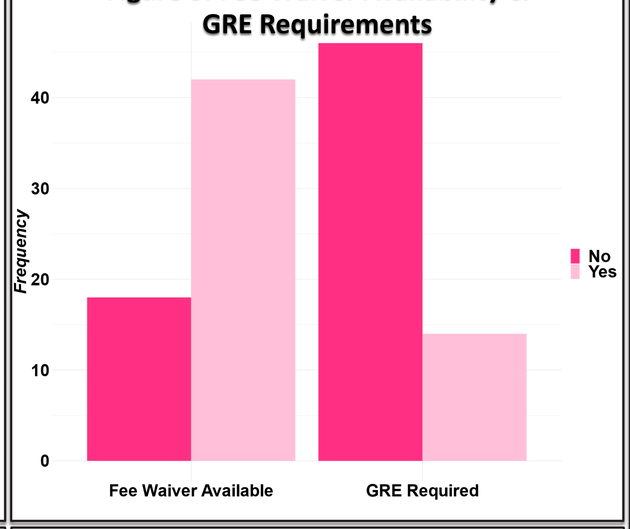


Figure 4: Median Monthly Housing Payments

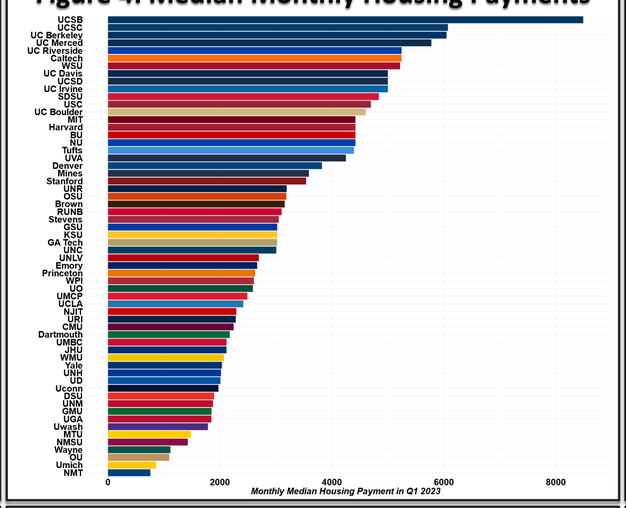


Figure 5: Number of Graduate Fields of Interest

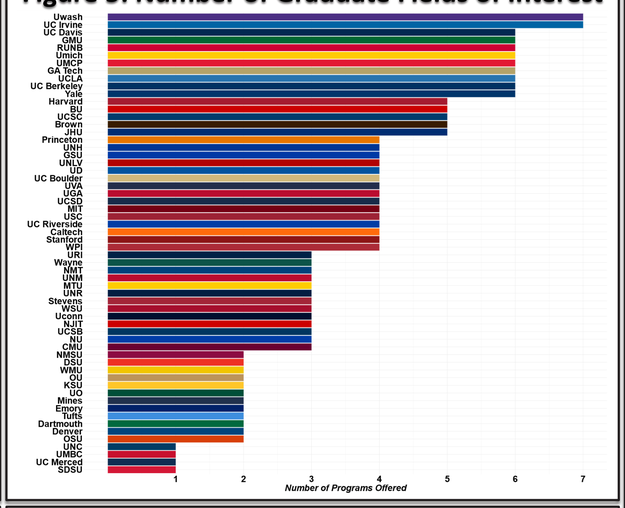


Figure 1: Concentration of Application Deadlines

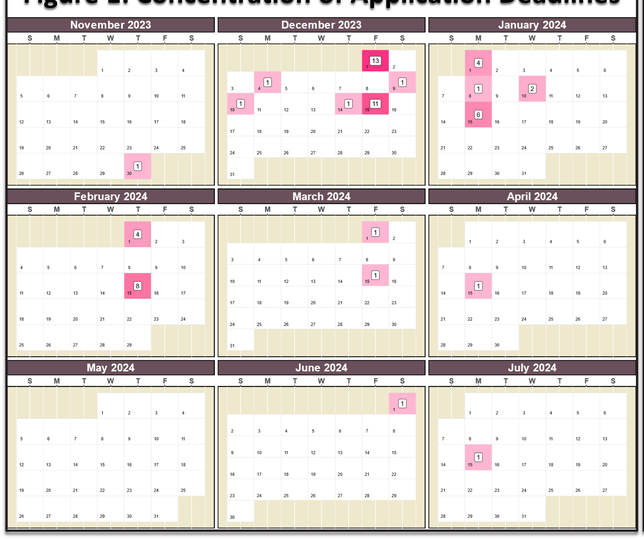


Figure 6: Number of Faculty Members of Interest

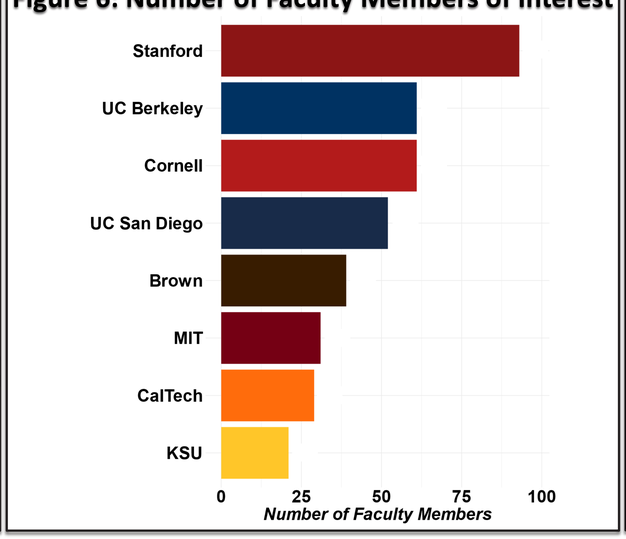
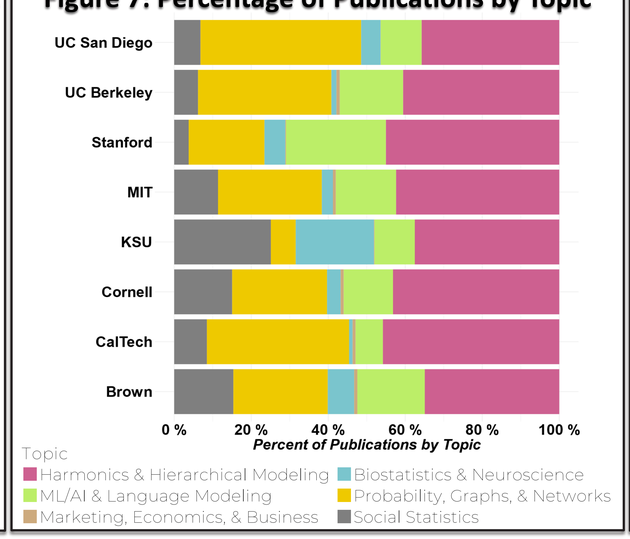


Figure 7: Percentage of Publications by Topic



CONCLUSIONS

Concentration of Application Deadlines (Figure 1).
Most institutions schedule their graduate application deadlines in the winter and early spring, with two schools extending their deadlines into the summer. For 2023, December hosts the greatest number of application deadlines at 28 in total, with 13 of those occurring on 1 Dec and 11 occurring on 15 Dec. January has the next greatest number of deadlines, with a total of 13, and February follows closely behind with 12. This scheduling pattern aligns with the fact that most schools only admit graduates in the fall semester.

Frequency of Application Fees by Cost (Figure 2).
Pursuing a Ph.D. can become costly even before admission into a program. This includes application and exam fees, which add up quickly as the number of programs of interest increases. Most of the 60 schools included required an application fee between $50 and $75 and an ambitious applicant can apply to all 60 for $4,700.

Application fees between $100 and $125 appeared the least frequently, with Harvard, Stanford, and Yale as the only institutions in this category. Dartmouth, MTU, NVSU, and WSU all required no application fee at all. Notably, the UC institutions all exhibited the highest application fees at greater than $125 per application.

Fee Waiver Availability & GRE Requirements (Figure 3).
To help applicants save on cost, some schools may choose to offer either need-based or merit-based application fee waivers, in addition to dropping requirements for traditional standardized testing like the GRE. For some students, cost-prohibition can determine the programs they apply to or choose in the end.

Of the 60 schools included in the analysis, 42 of them offered waivers for their application fees, potentially saving those who apply to all 60 institutions $3,505. The remaining 18 institutions did not offer a fee waiver, leaving those applicants responsible for the remaining $1,195.

Additionally, 14 of the schools required GRE scores, while 46 did not. Though programs do not often require GRE scores as part of the application process, those that do add another layer of planning and expense to applicants' experiences.

Median Monthly Housing Payments (Figure 4).
UC Santa Barbara resides in the most expensive location of all schools analyzed, with median monthly housing payments at nearly $9,000 per month. The institution in the second most expensive area, UC Santa Cruz, exhibits median monthly housing payments of about $6,000. Similar median monthly housing payments occur around UC Berkeley and UC Merced, with prices beginning to drop for UC Riverside, with a median monthly housing payment closer to $5,500.

On the less expensive end of the data, New Mexico Institute of Mining and Technology, followed by the University of Michigan, make up the only two institutions in locations with median monthly housing payments under $1,000. Institutions in the next least expensive areas include other New Mexico and Michigan schools such as Oakland University, Wayne State University, and New Mexico University.

Number of Graduate Fields of Interests (Figure 5).
Schools offering a higher number of programs can potentially offer students the ability to personalize their degree to fit their career goal. Some schools, such as the University of Washington, offer multiple programs per field of study (i.e., Pure Math vs Applied Math). These programs vary by coursework and teaching requirements

Number of Faculty Members of Interest (Figure 6).
An institution's faculty of interest include those faculty members whose research falls within one of the areas listed in the inclusion criteria. Figure 6 presents the number of these faculty at each institution. The data included information on 387 faculty members of interest across the 60 institutions. Stanford hosts the greatest number of faculty of interest, at 52 individuals within four possible degree programs.

Percentage of Publications by Topic (Figure 7).
Figure 7 presents the proportion of publications across six areas of interest as determined by a topic model for the 8 institutions with the earliest application deadlines. Caltech has the greatest proportion of publications related to Harmonics & Hierarchies, which appears most frequently across institutions' publications. Caltech's proportion of publications related to Probability, Graphs, & Networks ranks second only to UC San Diego. For Neuroscience & Biostatistics, as well as Social Statistics, KSU exhibits the highest proportion of related publications. Finally, Stanford ranks first for the proportion of publications related to Machine Learning, Language Modeling, & AI. The proportion of publications related to Marketing, Economics, & Business remained miniscule across all 8 schools.