MAVAC: Mapping and Visualization of Academic Collaborations

Maya Zeng

California Polytechnic State University, San Luis Obispo, CA 93401, USA
maya_zeng@elcamino.edu

Abstract

To analyze equity amongst academic researchers, we construct and analyze visualizations to better understand collaboration networks, represented by collaborative experiences such as publications and geospatial trends. Data gathered for this project includes processes such as web scraping to create an accurate construction of the network, and technologies such as SQL and NetworkX are used to view the visualizations and geospatial structure of these networks.

1. Introduction

We aim to discover key insights into academic equity, for example, analyzing the evolution overtime of a female researcher's average degree, by building an academic collaboration network. The network consists of nodes representing researchers and edges representing the collaboration between authors of a publication. As a starting point, the Scopus database is used to aggregate publication information as the database provides one of the largest abstract and citation database of peer-reviewed literature [1]. The database is then be applied for the visualization of our network using technologies such as Google Maps API and Python.

We then further refine the construction of our network by web scraping university websites for their faculty, more specifically in the departments of electrical engineering, computer science, math, psychology, statistics, chemistry, economics, agricultural science, and food science and nutrition, reason being Elsevier's gender report highlighted these majors. As of now, the universities of our interest are schools within the California State University and University of California network. Overall, we aim to find distinct patterns relating to gender collaboration in research with our end goal to ultimately aid and promote successful collaborations for all.

2. Related Work

An overview of collaborative network graphs is described by Wood in the realms of computer science and electrical engineering departments from the University of California system [2], and a collaborative network within Cal Poly SLO's colleges from Migler [3]. Further research has been made by both Migler and Wood with analyzing collaborations between Cal Poly SLO and the University of California system [4].

To expand upon their research and further explore average trends within female researcher's degrees, our research entails the creation of a collaborative network that enmeshes faculty from both the University of California system and the California State University system. Additional departments in consideration in this exploration include mathematics, psychology, statistics, chemistry, economics, agricultural science, and food sciences. Additional universities considered in this phase of research include all the schools under the California State University system.

3. Refining the Database

As of now, the network heavily relies on the Scopus database to retrieve and aggregate information of co-authors through their provided API. From there, a database was created with the following tables:

Author	AuthorField	CalibratedProb	Collaboration	Department	Ethnicity
Field	Gender	NewSBAuthors	Organization	PairwiseOrgDistances	Publication

To create an accurate construction of the network, we would use technologies such as the Python Beautiful Soup package to visit university websites and scrap departments for their faculty. From there, we would parse pertinent information such as the first and last name of the author and their emails to a csv file to upload in our database.

Given the fact that universities have different standards of HTML layouts and user interface, we had to adjust our web scraping scripts in accordance with what is provided on each website's faculty directory. With that, web scraping pertinent information requires a human element to check the HTML code provided and apply the necessary algorithms needed to extract information.

Other factors were considered if we were to construct a network with the following data, such cases include the fact that the professor has either retired or passed and their emails are no longer active. Not only that, but we also had to considered if it was pertinent to parse the researcher's middle name as it can provide more accurate results.

During this time, we also realized that the Scopus database does not reflect the departments or professors accurately as some professors are missing or there were multiple entries for the same professor in the database, and some departments could represent two fields for example, department of mathematics and statistics. Discussion has been held to generate a 9-digit ID for new professors compared to the 10-digit ID provided by the Scopus database; however, further discussion is needed on the matter.

Naming conventions for the universities names are a concern as well, thus the database also needs a human element to make inferences and to find data created for us to visualize. For example, in the database provided, schools that are part of the California State University system would be denoted as "California State University, Northridge";

however, in some cases, they would be denoted as “California State University Channel Islands” or “San Jose State University”. Essentially, the naming of schools is varied.

We will continue to attend and attempt the development of the database as it is integral to providing an accurate representation of researchers and their collaborations.

4. Conclusions and Future Work

To construct an accurate representation of the network, our research entails web scrapping professors from different universities that are related to our research interests. Ideally, we would have scrapped all faculty data and have constructed a database that reflects the departments accurately provided in the university’s website, standardized wording schemas, and resolved other foundational issues that arise. From there, we can make proper adjustments and additions that can give us accurate visualizations and degrees between academic researchers.

5. Acknowledgements

The author would like to express their sincerest gratitude to Theresa Migler, Zoë Wood, and the Computing Research Association for giving them the chance to partake in this project. I would also like to thank the following student researchers who took the time to mentor me along this process: Brandon Tat, Logan McNichols, and Leticia Siqueira.

References

[1] Scopus. https://service.elsevier.com/app/answers/detail/a_id/15534/supporthub/scopus/#tips, accessed: 2021-08-03

[2] Carroll, C., Garg, N., Migler, T., Walker, B., Wood, Z.: Mapping and visualization of publication networks of public university faculty in computer science and electrical engineering. CATA (2), 1–12 (2020)

[3] McNichols, L., Medina-Kim, G., Nguyen, V. L., Rapp, C., & Migler, T. (2019). Gender’s Influence on Academic Collaboration in a University-Wide Network. Studies in Computational Intelligence, 94–104. doi:10.1007/978-3-030-36683-4_8

[4] McNichols L. , Pineda S., Sauerborn E., Tat B., Yoo K., Lehr J., Wood W., Migler T. (2021) MAVAC: Mapping and Visualization of Academic Collaborations with a Focus on Diversity. In: Teixeira A.S., Pacheco D., Oliveira M., Barbosa H., Gonçalves B., Menezes R. (eds) Complex Networks XII. CompleNet-Live 2021. Springer Proceedings in Complexity. Springer, Cham. https://doi.org/10.1007/978-3-030-81854-8_8