Analyzing Multi-Principal Investigators Collaboration Networks in NIH Grants

Cuiran Shi, Shuying Han, Shreya Kusumanchi, Mia Zhou, Didong Li

Department of Biostatistics, University of North Carolina, Chapel Hill, NC, USA.

Contributing authors: cuiran@unc.edu;

Abstract

The abstract serves both as a general introduction to the topic and as a brief, non-technical summary of the main results and their implications. Authors are advised to check the author instructions for the journal they are submitting to for word limits and if structural elements like subheadings, citations, or equations are permitted.

1 Introduction

This project aims to explore multi-principal investigator (PI) collaboration networks of 30,127 NIH R01-equivalent grants over the past 23 years. It offers a valuable resource for researchers, policy analysts, and funding agencies by enabling rapid and reproducible retrieval of meaningful public health research areas.

1.1 NIH R01-equivalent Multi-PI Grants

National Institues of Health (NIH) grants are a primary source of funding for biomedical and health-related research in the United States, supporting innovative studies that advance scientific knowledge and improve public health. Over the past two decades, the size of NIH grants has expanded significantly, reflecting trends toward greater collaboration and the inclusion of more public health domains. Understanding these trends in NIH grant funding is crucial, as it can guide researchers in statistics and biostatistics to focus on health research areas with promising funding opportunities.

Among NIH grants, R01-equivalent grants are highly competitive and prestigious which provide substantial resources for independent investigator-led projects. They are used to support a discrete, specified, circumscribed research project, and generally awarded for 3-5 years. R01-equivalent grants are defined as activity codes DP1, DP2, DP5, R01, R37, R56, RF1, RL1, U01 and R35, either funded by NIAMS, or another NIH Institute, Office, or Center (ICO).

Multi-Principal investogator (multi-PI) grants, a variant of the R01 mechanism, are particularly worth studying because they foster interdisciplinary collaboration by allowing multiple leading researchers to jointly oversee a project. By allowing two or more leading researchers to jointly direct a project, multi-PI grants integrate diverse expertise, enhancing innovation and addressing complex research questions that may be beyond the scope of a single investigator. Unlike traditional single-PI grants, multi-PI grants span different displines or methodologies. For instance, one PI may contribute to computational biology while another specializes in translational medicine, enabling a more robust and multifaceted research approach. This collaborative model not only strengthens the scientific rigor of proposals but also facilitates resource-sharing, risk distribution, and broader impact. Examining multi-PI

R01-equivalent grants offers valuable insights into how collaborative funding structures can optimize research outcomes, making them a compelling subject for further investigation within NIH funding dynamics.

1.2 Current Research

NIH offers information on awarded grants via the RePORTER website (http://projectreporter.nih.gov/). For each award, the NIH RePORTER provides detailed information including PI names, organizations, project title, project abstract, NIH spending categorization, keywords, etc. However, NIH RePORTER searches are not optimal for various information needs and analyses,where those keywords and categorization tags are only intended to meet specific NIH reporting requirements, rather than to comprehensively characterize the entire NIH research portfolio (Talley et al., 2011).

To facilitate navigation and discovery of NIH-funded research, Talley et al. (2011) developed a database using unsupervised machine learning techniques, such as topic modeling and graph-based clustering, to categorize and visualize NIH grants by their titles and abstracts. This method provided a more contextually relevant framework for analyzing NIH-funded research, helping users discern funding patterns and emerging trends. Nevertheless, the reliance of their clustering methods on textual data from grant titles and abstracts may not adequately reflect the scope and accuracy of biomedical research, as their algorithms might not be ideally suited for biomedical text data.

To tackle the challenges of capturing rare biomedical word distributions, Zhang et al. (2023) proposes a novel approach to analyzing NIH grant texts using a time-aware neural topic model named Turtling. This model utilizes pretrained biomedical word embeddings to enhance topic extraction and employs a probabilistic time-series model to track topic evolution smoothly over time, complemented by multitask losses to

enrich topic quality and facilitate funding institute prediction. However, the reliance on text data alone may not capture the full scope of the research funded, and the current algorithms might not fully accommodate the nuances of biomedical terminology and institute-specific information. This could potentially limit the model?s ability to generalize across different types of biomedical text data.

Apart from topic modeling on NIH grants text data, there is abundant previous research on co-authorship networks which is crucial to analyzing the co-PI collaboration networks in NIH grants. Ji, Jin, Ke, and Li (2022) constructed co-citation and co-authorship networks from a comprehensive dataset spanning over 41 years from statistical journals. They applied unsupervised machine learning techniques to investigate research interests and community structures within the statistics field. Their findings indicate a significant evolution in research interests, with areas such as biostatistics and high-dimensional data analysis becoming more prominent. However, the authors acknowledged potential biases in their dataset and the challenges associated with interpreting large-scale networks.

2 Methods

In this project, we aim to: 1) detect clusters in co-PI collaboration networks; 2) interpret characteristics of these clusters; 3) determine representative PIs for each cluster; and 4) explore promising health research areas and PI collaboration patterns over time. We applied the Leiden algorithm (Traag, Waltman, & Van Eck, 2019) for community detection in co-PI collaboration networks and utilized the dynamic BERTopic model (Grootendorst, 2022) combined with BioSentVec (Chen, Peng, & Lu, 2019) biomedical sentence embeddings. Our methods offer several advantages, including the ability to generate higher quality communities, fast convergence, suitability for biomedical topic modeling, and incorporation of temporal dynamics in topic modeling.

2.1 Description of Dataset

Our analysis focuses on NIH R01-equivalent grants with multiple PIs awarded between 2006 and 2023, comprising 30,133 PIs and 148,360 grants. Notably, the multi-PI grant mechanism was introduced in 2006, so our dataset covers its entire history. We restrict our study to the largest connected component of the multi-PI collaboration network, which includes 13,873 PIs and 30,127 grants. The grant-level data was obtained from NIH RePORTER and includes attributes such as PI names, organizations, locations, project titles, abstracts, NIH spending categories, keywords, and fiscal years. For PI-level data, we scraped PubMed based on each investigator?s most recent affiliation, collecting information on departments, organizations, and locations.

2.2 Community Dection on PI Collaboration Networks

To identify communities with similar collaboration patterns among PIs, we constructed a collaboration network using the largest connected component of the multi-PI grant dataset. In this network, nodes represent PIs, and edges indicate co-participation in the same grant. We represented the network as an unweighted adjacency matrix and applied the Leiden algorithm for community detection. The Leiden algorithm, introduced by Traag, Waltman, and Van Eck (2019) as an improvement over the Louvain method, optimizes modularity while ensuring well-connected communities. The algorithm iteratively refines partitions through three phases?local node movement, partition refinement, and network aggregation?until modularity convergence is achieved. Following community detection, we identified representative PIs for each cluster based on normalized node degrees. Additionally, we performed summary statistical analyses on community-level attributes, including NIH spending categories, departmental affiliations, organizations, and geographic locations.

2.3 Topic Modeling on Grant Information

To better characterize the detected communities, we included edge information from the collaboration network. We incorporated textual information from grant titles and abstracts using topic modeling. Specifically, we employed the BERTopic model (Grootendorst, 2022), which combines contextual word embeddings with class-based TF-IDF (c-TF-IDF) to generate interpretable topics while preserving semantically meaningful terms in topic representations. For word embeddings, we utilized BioSentVec (Chen et al., 2019), a 200-dimensional embedding model pre-trained on biomedical corpora with a controlled biomedical vocabulary. This choice ensures optimal semantic representation of NIH grant texts compared to generic embedding approaches. Following topic modeling, we quantified the distribution of topics across communities to facilitate comparative interpretation of their research themes.

2.4 Dynamic Analysis

We are interested in seeing the time trend of communities and topics over the past two decades. We stratified the dataset into 3 time intervals: 2006-2015, 2016-2020, and 2020-2023, and conducted community dection for the three stratified datasets. We then analyzed the evolution communities in the dynamic PI collaboration network. We also incorporated dynamic BERTopic model to analyze the evolution of topics over time. BERTopic allows for dynamic topic modeling by calculating the topic representation at each timestep without the need to run the entire model several times.

Acknowledgement

Statements and Declarations

Conflict of interest The authors have declared no conflict of interest.

References

Chen, Q., Peng, Y., Lu, Z. (2019). Biosentvec: creating sentence embeddings for biomedical texts. *2019 ieee international conference on healthcare informatics (ichi)* (pp. 1–5).

Grootendorst, M. (2022). Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*, ,

Ji, P., Jin, J., Ke, Z.T., Li, W. (2022). Co-citation and co-authorship networks of statisticians. *Journal of Business & Economic Statistics*, *40*(2), 469–485,

Talley, E.M., Newman, D., Mimno, D., Herr, B.W., Wallach, H.M., Burns, G.A., ... McCallum, A. (2011). Database of nih grants using machine-learned categories and graphical clustering. *Nature Methods*, *8*(6), 443–444,

Traag, V.A., Waltman, L., Van Eck, N.J. (2019). From louvain to leiden: guaranteeing well-connected communities. *Scientific reports*, *9*(1), 1–12,

Zhang, R., Duan, Z., Lee, C., Riffle, D., Min, M.R., Zhang, J. (2023). Turtling: a time-aware neural topic model on nih grant data. *Bioinformatics Advances*, *3*(1), vbad096,