Investigating Academic Graph-Based Factors behind Funding Success in National Institutes of Health

Lin, Tianqianjin Zhejiang University, China | lintqj@zju.edu.cn
Wang, Qian Worcester Polytechnic Institute, USA | qwang18@wpi.edu
Jiang, Zhuoren Zhejiang University, China | jiangzhuoren@zju.edu.cn
Yuan, Weikang Zhejiang University, China | yuanwk@zju.edu.cn
Huang, Cui Zhejiang University, China | huangcui@zju.edu.cn
Mabry, Patricia HealthPartners Institute, USA | Patricia.L.Mabry@healthpartners.com
Liu, Xiaozhong Worcester Polytechnic Institute, USA | xliu14@wpi.edu

ABSTRACT

While major funding agencies are striving for diversity and fairness, the mechanisms behind funding success have yet to be fully elucidated. Existing studies reveal valuable evidences about the effect of the applicant's individual attributes, e.g., gender and age, on the funding success. However, the relationship between funding success and academic activities, e.g., collaborator's characteristics, remains underexplored. This work collects massive scholarly data from open academic graphs and public data about National Institutes of Health awards to investigate the effect of various academic graph-based factors on the "K to R" success. Leveraging a heterogeneous graph model for predicting the "K to R" success, we regard the gain in the model performance of a factor as a proxy variable for the magnitude of its effect on the "K to R" success. Our preliminary results suggest that interest by peers in the applicant's research and the timing of the interest are strongly correlated with the outcome. Meanwhile, the applicant's social connections, e.g., their collaborators, can also contribute to the outcome.

KEYWORDS

Research Funding; National Institutes of Health; Academic Graph-based Factor

INTRODUCTION

Research funding has been viewed as a key determinant of scientific activity (Rusu et al., 2022), and can have a long-term and vital impact on the trajectory of researchers' careers. Funding affords researchers essential resources and opportunities to establish themselves in their fields and pursue their research goals (Jacob & Lefgren, 2011). Moreover, the research funding success has been shown to be consistent with the Matthew effect and the Halo effect (Liao, 2021). For example, in The Netherlands, early funding success can introduce a growing rift with narrow-win applicants accumulating more than twice much research funding during the following eight years as near-miss applicants (Bol et al., 2018; Liao, 2021).

To ensure the career development of the researchers and protect the entire research system from inefficiency caused by inappropriate funding allocation (Kulczycki et al., 2017; Sandström & Van den Besselaar, 2018), it's imperative to understand the factors that affect the funding decisions in order to ensure that the fairness and impartiality are not compromised. With the increasing number of publicly available funding data, it's possible to conduct in-depth and quantitative analyses of the current mechanisms underlying the funding success, thus informing the current funding management, enhancing the credibility of scientific funds, and fostering a healthy academic atmosphere.

Prior studies have investigated the factors that potentially contribute to the researchers' funding success. While the quality of the research proposal and the relevance of the proposal to the funding's priorities are reasonable considerations (Ayoubi et al., 2019; Boyack et al., 2018), factors such as the researchers' gender or affiliation, which ought to have no bearing on funding decisions, may nonetheless be incorrectly taken into count and impose substantial impact on the application results (Viner et al., 2004; Witteman et al., 2019). For example, Van der Lee and Ellemers (2015) found that female researchers have lower success rates when applying for research funding compared to their male counterparts. It is important to recognize such potential biases in the funding application process and ensure equal opportunities for all researchers to advance their careers.

Although these efforts have yielded valuable insights and raised attention among researchers and funding managers regarding the issues, they have two major limitations. **First**, their focus is limited to the attributes of the researchers themselves, neglecting the potential impact of other factors inherent in the academic graph, e.g., social capitals conveyed by collaborators. **Second**, their empirical results are usually based on linear regression/correlation analysis, overlooking possible interactions between factors and factors with nonlinear relationships to the funding success (Armstrong, 2019).

To address the limitations of previous studies, in this work, we conduct a systematic investigation on the effect of a variety of academic graph-based attributes in the context of biomedical research funding. First, we collected data on

86th Annual Meeting of the Association for Information Science & Technology | Oct. 27 – 31, 2023 | London, United Kingdom.

11,358 Principal Investigators (PIs) who had won the Career Development Awards ("K") from the National Institutes of Health (NIH) and whether they subsequently became PI on an R01-equivalent awards ("R"), which is an explicit expectation of K funding. The PIs were further mapped to entities of authors in two open academic graphs (OAGs), PubMed Central database (PMC) and AMiner database. Second, leveraging emerging deep learning techniques, we developed a heterogeneous graph model and utilize it to predict whether the PIs ever received the R project based on the OAGs. To assess the effect of various graph semantics and their associated attributes on funding success, we evaluated the change in model performance when incorporating them into the prediction model.

Our contributions are threefold: **(a)** We introduce the academic graph-based factors into research on funding success, utilizing open academic graphs. **(b)** Leveraging a graph neural network, we design a simple yet effective approach to estimate the association between predictive factors and funding success, without any assumption on the form of the association. **(c)** We comprehensively evaluated the potential effects of the academic graph-based factors on the "K to R" success in NIH, which can provide empirical evidences to support effective funding management.

DATA AND RESEARCH DESIGN

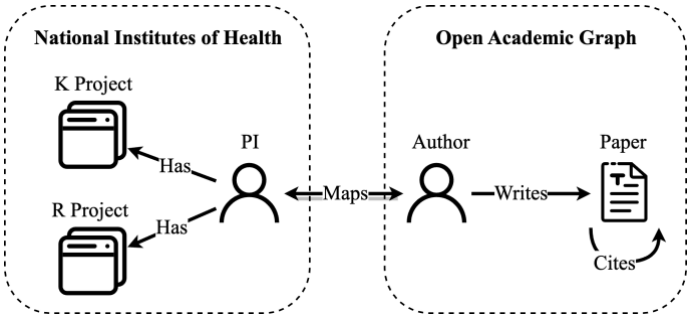

Figure 1. Schema of the Composed Academic Graph

Data Collection and Preprocessing

As shown in Figure 1, our academic graph integrated two resources. On one hand, we manually collected information on 11,358 K projects from NIH and their associated PIs from the NIH RePORTER database. Among these PIs, 5,382 individuals obtained subsequent R projects. On the other hand, we use open academic graphs (OAGs), i.e., PMC and AMiner, to obtain the academic activities of these PIs, e.g., their publications and collaborations. To map the PIs to author entities in the OAGs, we identified publications of the PIs in the OAGs by searching NIH project numbers, and further aligned the PIs with the author entities based on the same-named authors within the identified publications. As a result, a composed academic graph was obtained and the statistics of the number of nodes and edges of this graph is shown in Table 1.

Node		Edge	
Type	**Count**	**Type**	**Count**
PI	11,358	PI –(IsMappedTo)→ Author	16,663
Author	301,218	Author –(Writes)→ Publications	717,999
Publication	91,837	Publication –(Cites)→ Publication	16,899
K Project	11,358	PI –(Has)→ K Project	11,358
R Project	5,382	PI –(Has)→ R Project	5,382

Table 1. Data Statistics of the Composed Academic Graph

Additionally, we engineered features for each node in the composed academic graph. For PIs and authors, we inferred their gender and race based on their names and calculated their number of publications and the citation count of their publications. For K projects and publications, in addition to their original metadata such as application (or publication) year, we further assigned a topic distribution vector to them via the LDA algorithm (Hoffman et al., 2010). Moreover, we labeled the ranking of the PI's agency organizations as an attribute of the K project and labeled the impact factor of journals as an attribute of the publication according to the Scimago ranking database. In this work, whether a PI obtained an R project is considered as the dependent variable of interest.

Academic Graph-based Factors

To construct factors associated with the academic activities, we first predefine a set of intelligible semantics derived from the relations/meta-paths (Dong et al., 2017; Wang et al., 2019) within the composed academic graph. For example, the relation "PI –(IsMappedTo)→ Author –(Writes)→ Publication" represents the PIs' publications and

the meta-path "PI –(IsMappedTo)→ Author –(Writes)→ Publication ←(Writes)– Author" represents the collaborators of the PIs. Based on these semantics, the semantic-conditioned attributes, e.g, publication's topic or collaborator's gender, are defined as academic graph-based factors which we examine in this work.

Effect Magnitude Evaluation Approach

Motivated by Zhang et al. (2023), we adopt a prediction-based approach to measure the effect of the factors on funding success. Assuming we have a baseline model that only utilizes the applicant's individual attributes to predict the funding success, we can evaluate the effect of a factor (or a group of factors) on the application result by incorporating it (or them) into the baseline model and regard the change in model performance as a proxy variable for the magnitude of its effect. The greater the improvement in performance, the greater the effect of this factor. For example, in Figure 2, we can estimate the effect of PI's citations by calculating the incremental improvement in model performance when citation semantics are included in the input, versus performance of the base model. Similarly, if we need to measure the effect of a specific academic graph-based factor separately, we can solely integrate the corresponding semantic with only the single attribute into the base model.

To handle such (heterogeneous) graph data, technically, we first jointly utilize three pooling-based aggregators (Hamilton et al., 2017; Xu et al., 2019), which are max-pooling, mean-pooling and sum-pooling, to encode the neighbor sets of various semantics. Then we leverage a Transformer-based merger (Vaswani et al., 2017; Yang et al., 2023) to fuse the encoded neighbor sets and the PI's attributes into one single fixed-size representation. Finally, we predict the application result based on the representation by a multi-layer perceptron (Hornik et al., 1989). These modules can thoroughly capture the features of the semantic-based neighbor sets, facilitate sufficient interactions between the PI's attributes and the academic graph-based factors, and automatically fit their potential relationship with the application results. This approach avoids erroneous judgments caused by the preconceived assumptions about the relationship between the academic graph-based factor and outcome, e.g., linear assumption.

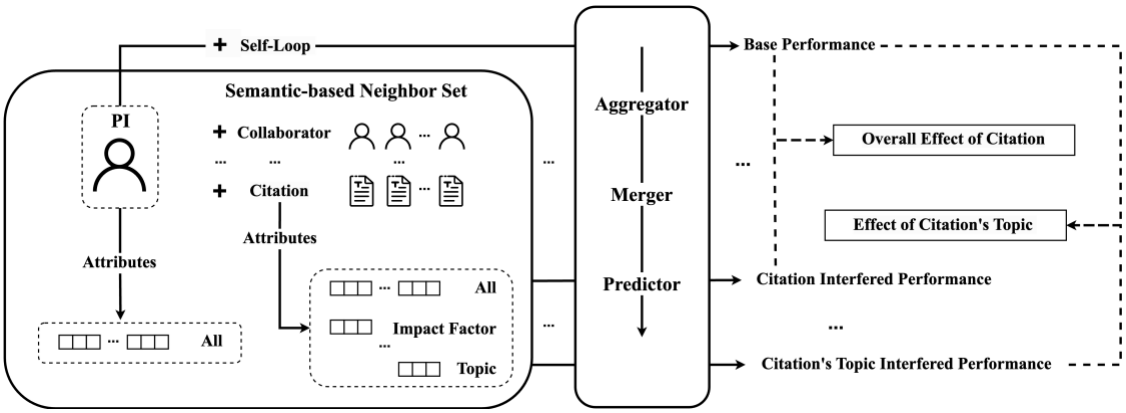

Figure 2. Framework for Effect Evaluation

PRELIMINARY RESULTS

In this work, our method is capable of handling all available semantics and associated factors. For the preliminarily analysis, we focused on only two semantics derived from the composed academic graph, which are "PI –(IsMappedTo)→ Author –(Writes)→ Publication ←(Writes)– Author" (denoted as collaborators) and "PI –(IsMappedTo)→ Author –(Writes)→ Publication ←(Cites)– Publication" (denoted as citations). We identified 389,161 collaborators and 85,257 citations of the PIs. To ensure the robust prediction performance in each setting, we conducted 10-fold cross-validation and reported the mean and standard deviation of F1 score, precision, and recall for successful applications.

As shown in Table 2, we first estimated the overall improvement in model performance over the base model by integrating the two semantics separately into the base model. It can be observed that either of the two semantics can enhance the model performance, and the integration of citation can even improve the performance exceed 10%.

At a fine-grained level, we further evaluated the performance gain associated with each specific factor, and the results are shown in Table 3 (where "#" means "the number of"). Specifically, among the collaborator's attributes, the number of citations and publications are the most influential factors. Additionally, the gender and race of collaborators, especially race, can increase the recall for successful applications while only slightly decreasing the precision. As for the attributes of citation, the number of citations is the strongest predictor, which can significantly improve all three metrics (9.45%, 6.00%, and 12.61% for F1, precision, and recall, respectively). The second most influential factor is the publishing year, which significantly improved F1, precision, and recall by 8.68%, 4.34%, and 12.90% respectively. It is worth mentioning that the influence of these two factors can exceed the impact factor of

the journal, suggesting that the interest by peers and timing of the interest may be more important than quality. However, the topic of the citations seems to have little effect and it only marginally improve the precision by 2.43%.

Input	F1	Precision	Recall
Only PI (Base)	59.88±2.57	64.32±1.75	56.14±4.02
with Collaborator	61.53±1.81 (+2.76%)	65.58±2.40 (+1.96%)	58.10±3.28 (+3.49%)
with Citation	67.81±1.20 (+13.24%)	71.08±1.17 (+10.51%)	64.88±2.22 (+15.57%)

Table 2. Semantic Interfered Performance and Their Percentage of Change.

FINDINGS AND DISCUSSION

By aligning semantic features based on the OAGs with social concepts or theories, we can draw further insights from our experimental results regarding the underlying mechanisms behind NIH funding success of the R project.

Our findings suggest that the characteristics of an applicant's personal publications are the most important factor. Firstly, experimental results show that publishing year of the applicant's citations strongly impacts the model's performance. Since the time when the citation occurred can reflect the time period when the applicant's work has attracted attention in the field, we can speculate that the recent value and attention given to the applicant's work may be an important factor in funding success. Secondly, the impact factors and the number of citations of the applicant's citations have a significant positive effect on model's predictive performance. As the impact factors and the number of citations of the applicant's citations not only reflect the quality of the citations but also imply that the quality of the applicant's publications, we can infer that high-quality work will increase the probability of success.

Input		F1	Precision	Recall
Only PI (Base)		59.88±2.57	64.32±1.75	56.14±4.02
with Collaborator	#Publication	61.48±1.87 (+2.67%)	63.29±2.07 (-1.60%)	59.93±3.55 (+6.75%)
	#Citation	62.07±2.71 (+3.66%)	64.87±2.08 (+0.86%)	59.90±5.57 (+6.70%)
	Gender	60.11±2.13 (+0.38%)	63.97±2.19 (-0.54%)	56.83±3.45 (+1.23%)
	Race	60.66±3.99 (+1.30%)	63.17±1.97 (-1.79%)	58.77±6.92 (+4.68%)
with Citation	Year	65.08±1.87 (+8.68%)	67.11±2.42 (+4.34%)	63.38±3.73 (+12.90%)
	#Citation	65.54±2.05 (+9.45%)	68.18±1.00 (+6.00%)	63.22±3.90 (+12.61%)
	Impact Factor	62.70±2.07 (+4.71%)	64.88±2.53 (+0.87%)	60.82±3.47 (+8.34%)
	Topic	60.24±2.54 (+0.60%)	65.88±2.54 (+2.43%)	55.79±4.73 (-0.62%)

Table 3. Academic Graph-based Factor Interfered Performance and Their Percentage of Change.

Our results demonstrate that integrating collaborators' information can significantly increase the recall of successful applications at the cost of sacrificing precision. This suggests the possibility of a phenomenon where funding applications initially deemed unqualified have succeeded, and this outcome can be attributed to the consideration of characteristics about the applicants' collaborators. Specifically, it's observed that the number of publications and the number of citations of the collaborators, as well as their race, are the main influential factors. The former can indicate the applicant's social capital in academia, while the latter may reflect the academic circle of the applicants. Consequently, we can infer that applicants' social capital in academia and the academic circle they belong, may contribute to the outcome.

CONCLUSION AND FUTURE WORK

This work introduces academic graph-based factors into the research of funding mechanisms by collecting data from OAGs. We design a prediction-based effect magnitude estimation scheme based on emerging techniques of graph neural networks. We further empirically analyze the effects of two factors: collaborators and citations. While this work provides valuable supplementary information for funding research, it has several limitations. For example, it currently cannot reveal the specific relationship law between a factor and the funding success nor can it tell whether the factor is a causal factor. These issues will be addressed in our future work.

ACKNOWLEDGMENTS

We gratefully acknowledge support from NSF Award (# 2122232-SCISIPBIO) and appreciate all publications support and the anonymous reviewers' helpful and insightful comments.

REFERENCES

Armstrong, R. A. (2019). Should pearson's correlation coefficient be avoided? Ophthalmic & physiological optics: the journal of the British College of Ophthalmic Opticians (Optometrists), 39(5), 316–327.

Ayoubi, C., Pezzoni, M., & Visentin, F. (2019). The important thing is not to win, it is to take part: What if scientists benefit from participating in research grant competitions? Research Policy, 48(1), 84–97.

Bol, T., Vaan, M., & Rijt, A. (2018). The matthew effect in science funding. Proceedings of the National Academy of Sciences, 115(19), 201719557.

Boyack, K. W., Smith, C., & Klavans, R. (2018). Toward predicting research proposal success. Scientometrics, 114, 449–461.

Dong, Y., Chawla, N. V., & Swami, A. (2017). Metapath2vec: Scalable representation learning for heterogeneous networks. Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, August 13 - 17, 2017, 135–144. https://doi.org/10.1145/3097983.3098036

Hamilton, W. L., Ying, Z., & Leskovec, J. (2017). Inductive representation learning on large graphs. In I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, & R. Garnett (Eds.), Advances in neural information processing systems 30: Annual conference on neural information processing systems 2017, december 4-9, 2017, long beach, ca, USA (pp. 1024–1034). https:// proceedings. neurips. cc/ paper/ 2017/hash/5dd9db5e033da9c6fb5ba83c7a7ebea9-Abstract.html

Hoffman, M., Bach, F., & Blei, D. (2010). Online learning for latent dirichlet allocation. advances in neural information processing systems, 23.

Hornik, K., Stinchcombe, M. B., & White, H. (1989). Multilayer feedforward networks are universal approximators. Neural Networks, 2(5), 359–366. https://doi.org/10.1016/0893-6080(89)90020-8

Jacob, B. A., & Lefgren, L. (2011). The impact of research grant funding on scientific productivity. Journal of public economics, 95(9-10), 1168–1177.

Kulczycki, E., Korzeń, M., & Korytkowski, P. (2017). Toward an excellence-based research funding system: Evidence from poland. Journal of Informetrics, 11(1), 282–298. https://doi.org/https://doi.org/10.1016/j.joi. 2017.01.001

Liao, C. H. (2021). The matthew effect and the halo effect in research funding. Journal of Informetrics, 15(1), 101108. https://doi.org/https://doi.org/10.1016/j.joi.2020.101108

Rusu, V. D., Mocanu, M., & Bibiri, A.-D. (2022). Determining factors of participation and success rates in research funding competitions: Case study. Plos one, 17(7), e0272292.

Sandström, U., & Van den Besselaar, P. (2018). Funding, evaluation, and the performance of national research systems. Journal of Informetrics, 12(1), 365–384. https://doi.org/https://doi.org/10.1016/j.joi.2018.01.007

Van der Lee, R., & Ellemers, N. (2015). Gender contributes to personal research funding success in the netherlands. Proceedings of the National Academy of Sciences, 112(40), 12349–12353.

Viner, N., Powell, P., & Green, R. (2004). Institutionalized biases in the award of research grants: A preliminary analysis revisiting the principle of accumulative advantage. Research Policy, 33(3), 443–454.

Wang, X., Ji, H., Shi, C., Wang, B., Ye, Y., Cui, P., & Yu, P. S. (2019). Heterogeneous graph attention network. In

L. Liu, R. W. White, A. Mantrach, F. Silvestri, J. J. McAuley, R. Baeza-Yates, & L. Zia (Eds.), The world wide web conference, WWW 2019, san francisco, ca, usa, may 13-17, 2019 (pp. 2022–2032). ACM. https://doi.org/10.1145/3308558.3313562

Witteman, H. O., Hendricks, M., Straus, S., & Tannenbaum, C. (2019). Are gender gaps due to evaluations of the applicant or the science? a natural experiment at a national funding agency. The Lancet, 393(10171), 531– 540.

Xu, K., Hu, W., Leskovec, J., & Jegelka, S. (2019). How powerful are graph neural networks? 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019. https://openreview.net/forum?id=ryGs6iA5Km

Yang, X., Yan, M., Pan, S., Ye, X., & Fan, D. (2023). Simple and efficient heterogeneous graph neural network.

Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023.

Zhang, Y., Jin, B., Zhu, Q., Meng, Y., & Han, J. (2023). The effect of metadata on scientific literature tagging: A cross-field cross-model study. WWW'23.