Mapping Scientific Foraging

Keywords: science of science, network embedding, knowledge space, scientific foraging

Extended Abstract

Scientists are explorers of the vast expanse of knowledge space, driven by a multitude of factors that go beyond their intrinsic curiosity, including funding opportunities, third-party interests, and the popularity of certain topics or problems. This journey of scientific exploration is far from straightforward, and understanding how scientists navigate this space is crucial to unraveling the complexities of scientific progress. In this pursuit, the ability to conceptualize the ”location” of scientists in the knowledge space is essential [1].

Existing approaches to represent scientists' research interests, such as bag-of-topics models, fall short in capturing the true essence of scientific exploration. They fail to capture the relationships between topics, which are crucial in understanding how topics evolve over time. To address this limitation, we propose a geometric representation of scientists' research interests, which we consider as a model of the knowledge space.

Our approach begins with the construction of a citation network covering the whole science (see Fig. a). For that, we adopt the Microsoft Academic Graph [2] dataset, which includes citation data across all the scientific disciplines in addition to metadata and disambiguated authors. We obtain an embedding of papers (see Fig. b) using node2vec [3] with a modification to account for the directionality of citations. Then, we represent a scientist's research interest in year t as the centroid of the papers that the scientist cited during that year (see Figs. c-d). This approach captures the relationship between topics in the representation of a scientist's interests.

We demonstrate the effectiveness of our representation through visualization and evaluation of its potential in predicting the journals where papers are published and collaboration ties among authors. By using UMAP [4], we create an visualization that captures many global patterns of how the topics of science are organized. For instance, Physics is connected to Chemistry, which connects to Biological sciences, while Mathematics bridges Physics and Computer sciences. Similarly, Agricultural sciences bridge Biological and Earth & Ocean sciences, and Astronomy forms its own cluster, separated from Physics (see Fig. e).

To validate the embedding, we conducted a subject classification task where we predicted a paper's web of science subject category. More specifically, the prediction undertaken by calculating the most frequent categories in the k-nearest papers in the embedding. We measured the accuracy of the prediction using micro-F1 scores with 10-fold cross-validation. Results obtained from by using different embedding techniques are shown in Fig. f. We found that the node2vec performed best, followed by the SPECTER [5] embedding, with a large margin to the doc2vec prediction, which uses only the content in titles and abstracts.

To verify whether the location of scientists in the knowledge space is representative of their research interests, we conduct a validation test. We hypothesize that scientists who share similar research interests are more likely to collaborate with each other, and thus, they would be located closer in the knowledge space. We test this hypothesis by predicting the likelihood of the collaboration between two researchers based on their distance in the knowledge space. We calculate the likelihood of new collaboration for each pair based on their cosine distance in the knowledge space and evaluate the predictive power using AUC-ROC. We also use collaboration networks as a baseline and compute several similarity scores for predicting new collaborations.

The results are shown in Fig. g. For this test, the node2vec embedding method performs best, followed by DeepWalk [6] and SPECTER, outperforming all the network-based predictors by a large margin.

We also explore the associations between research trajectories with productivity and impact, demonstrating that our knowledge space can provide insights into the underlying mechanisms of scientific impact. In conclusion, our proposed knowledge space representation provides a more precise and detailed way to understand scientists' research interests and how they change over time. It captures the intricate relationships between topics and is a powerful tool for analyzing scientific exploration, allowing for better predictions of new collaborations and insights into the mechanisms of scientific impact. Our work contributes to the field of computational social science by providing a novel method to analyze the complex dynamics of scientific progress and the movement of scientists across the knowledge space.

References

[1] A. Zeng, Z. Shen, J. Zhou, Y. Fan, Z. Di, Y. Wang, H. E. Stanley, and S. Havlin, "Increasing trend of scientists to switch between topics," *Nature Communications*, vol. 10, p. 3439, July 2019.

[2] A. Sinha, Z. Shen, Y. Song, H. Ma, D. Eide, B.-j. P. Hsu, and K. Wang, "An overview of microsoft academic service (mas) and applications," in *Proceedings of the 24th international conference on world wide web*, pp. 243–246, ACM, 2015.

[3] A. Grover and J. Leskovec, "node2vec: Scalable feature learning for networks," in *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 855–864, 2016.

[4] L. McInnes, J. Healy, N. Saul, and L. Grossberger, "Umap: Uniform manifold approximation and projection," *The Journal of Open Source Software*, vol. 3, no. 29, p. 861, 2018.

[5] A. Cohan, S. Feldman, I. Beltagy, D. Downey, and D. S. Weld, "Specter: Document-level representation learning using citation-informed transformers," *arXiv preprint arXiv:2004.07180*, 2020.

[6] B. Perozzi, R. Al-Rfou, and S. Skiena, "Deepwalk: Online learning of social representations," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 701–710, 2014.

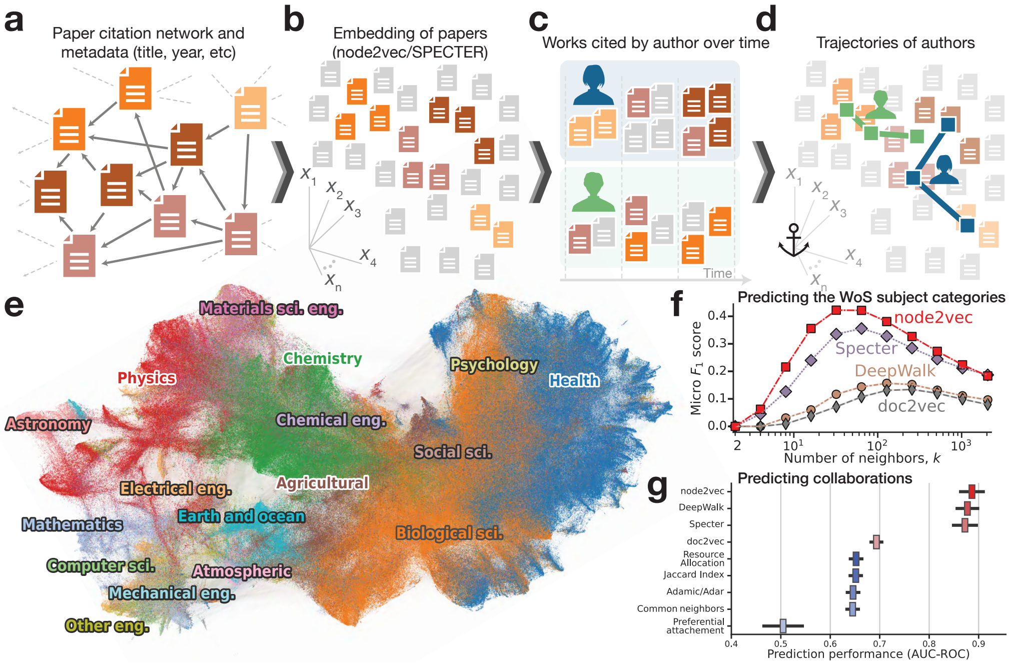

Figure 1: Schematic representation of the adopted methodology (a-d) and preliminary results (e-g). Starting from the citation network between papers and associated metadata (a), we construct an embedding of papers using node2vec or SPECTER (b). Authors are mapped according to their yearly centroids of cited works (c and d). The resulting embedding is visualized using UMAP (e), with colors indicating the NSF categories. We also present preliminary results on predicting the subject categories of papers (f) and future collaboration ties among authors (g).