Predicting future collaborations in a scientific community using graph neural networks

Nachyn Dorzhu[1], Tatiana Sukhomlinova[1], Lijing Luo[2] and Sergey Kovalchuk[1]

[1] ITMO University, Saint Petersburg, Russia
[2] Independent researcher, Netherlands
kovalchuk@itmo.ru

Abstract. Graph-based machine learning models have gained significant attention in predicting the emergence of new relationships in evolving networks. In this work, we present a study on forecasting scientific collaborations using a Graph Attention Network (GAT) with L2 regularization and dropout. We construct yearly co-authorship graphs based on historical publication data and analyze the evolution of these graphs over time with the International Conference on Computational Science (ICCS) as an example of a living scientific community. Our approach involves training on past yearly graphs to predict the formation of new edges in future graphs. We assess the model's performance by varying the prediction window and evaluating results using link prediction metrics. The proposed method demonstrates the feasibility of utilizing deep learning techniques for predicting future collaborations based on past scientific interactions.

Keywords: graph neural networks, graph attention network, link prediction, complex networks, coauthors graphs, temporal graph.

1 Introduction

Scientific collaboration drives innovation, and co-authorship networks reveal how ideas spread [1,2]. Predicting future researcher collaborations, such as grants and lab space, is key for optimizing resource allocation. Traditional link-prediction methods (e.g., common neighbors, Jaccard similarity) use static features and ignore changes in research interests and trends [3]. Most graph neural network methods also fail to capture temporal dynamics, which is essential for predicting new collaborations [4]. Graph Attention Networks (GATs) overcome this by dynamically weighing evolving neighbor influences [5]. In this study, we build temporal co-authorship graphs from the International Conference on Computational Science (ICCS)[1] publications since 2001 and use GAT-based link prediction with varying forecast windows. Our results show how temporal attention mechanisms enhance prediction accuracy and provide actionable insights for strategic academic planning.

[1] https://www.iccs-meeting.org/

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97635-3_45

2 Modelling collaboration in scientific events

We model collaboration as a link prediction task in a co-authorship network. While focused on ICCS, the approach generalizes to other scientific domains. Here we consider the following definition of emergent scientific collaboration. Two researchers are labeled as target emergent collaboration pair appeared after an event in year Y if they: (1) don't have any publications in common prior to this event in any source (no co-authorship recorded during the years before the event ($y \leq Y$)); (2) come to the event presenting own research (publication in the event proceeding); (3) publish together in any source after the year (co-authorship recorded in $y > Y$).

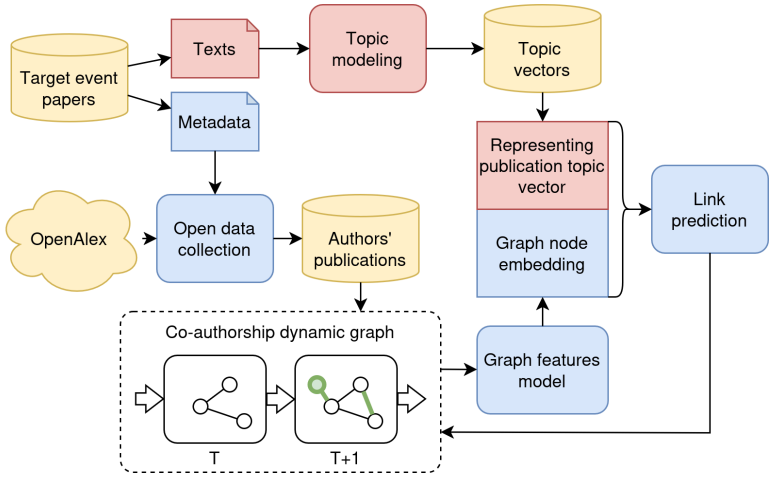

Fig. 1. General pipeline of the collaboration prediction process: (1) data acquisition from ICCS + OpenAlex, (2) yearly co-authorship graph construction, (3) emergent collaboration labeling, (4) GAT-based link prediction, and (5) evaluation.

The definition used in this study may be biased, as it does not account for other forms of collaboration, such as meetings or situations where researchers plan collaborations without publishing together (e.g., sharing affiliations). Additionally, presenting at an event under a single author's name may also obscure collaborations. However, the initial experiment assumes the primary collaboration trends are captured by this definition.

Our approach uses two main data sources (see Fig. 1). First, we utilize the ICCS publication corpus, which is structured and vectorized through topic modeling and manuscript vectorization based on topic relevance [7]. Second, we supplement this with metadata from OpenAlex[2], which provides structured information about authors and their works. We also extend topic analysis using OpenAlex topics. By using these data sources, we reconstruct a yearly co-authorship graph from ICCS history, forming a dynamic graph for link prediction. Both data sources provide features for the

[2] https://openalex.org/

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97635-3_45

GNN-based model. Detailed implementation steps are provided in the following subsections.

The construction of co-authorship graphs is a crucial step in modeling scientific collaborations. This process involves extracting data from publication metadata and representing it as a graph, where nodes are authors and edges represent co-authored publications. Authorship data is extracted from the dataset, and co-authorship graphs are constructed yearly using NetworkX,[3] where nodes are authors and edges are shared publications.

The G_{all} graph represents a network of scientific collaborations across all years, where each node corresponds to an author, and edges between nodes denote co-authored publications. Multiple edges between the same nodes reflect repeated collaborations over different years. This graph serves as the foundation for tasks such as link prediction, temporal analysis, and network evolution modeling.

To analyze the evolution of scientific collaborations, yearly co-authorship graphs are constructed, allowing us to study how co-authorship patterns change over time and serving as the basis for link prediction experiments. The starting point for creating temporal graphs is the full co-authorship graph Gall, which includes all collaborations from all years. For each year, from 2023 to 2001, we construct cumulative graphs by excluding all publications after year Y. Each graph G_Y (for Y from 2001 to 2023) represents the cumulative collaboration network at the end of year Y. These graphs allow us to analyze the growth, stability, and evolution of the co-authorship network, including the emergence of new research clusters and future collaboration predictions based on past trends. These temporal graphs serve as the foundation for evaluating link prediction models, assessing how well historical data can forecast new scientific partnerships.

To train a machine learning model for link prediction, we need to create a dataset of positive and negative examples. Positive examples are new edges that appear in the co-authorship network from year Y to $Y+1$. For each yearly graph G_Y, we compare it with G_{Y+1} to identify new edges. Each tuple (author1, author2, year) in positive examples represents a new collaboration.

Negative examples represent potential collaborations that didn't happen but could have. We sample pairs of authors who haven't collaborated yet but are in close network proximity. The dataset is balanced to ensure an equal number of positive and negative examples. These datasets will train and evaluate the graph-based neural network model (GAT) for predicting future collaborations.

The GAT model architecture captures structural and relational patterns within the co-authorship network. The input graph $G = (V, E)$ consists of nodes (authors) and edges (co-authorship relationships). The model uses an attention mechanism to assign varying importance to neighboring nodes, enhancing the representation of author interactions.

For link prediction, node embeddings are processed through multiple GAT layers to capture high-order dependencies. The final step involves combining node embeddings for a given author pair (e.g., through concatenation, element-wise multiplication, or absolute difference), followed by a fully connected layer with a sigmoid activation to generate a probability score for a future collaboration.

[3] https://networkx.org/

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97635-3_45

The training uses a binary cross-entropy loss function, comparing predicted probabilities with actual labels (positive for new collaborations, negative for non-existent links). Dropout and L2 regularization are applied to prevent overfitting and improve generalization. Dropout is applied to both the attention coefficients and the hidden node representations, ensuring the model does not overly depend on specific nodes or features. Through experimentation, we found that a dropout rate of 0.3 and a weight decay of 5e-4 effectively balance overfitting and model capacity, contributing to stable and robust training.

The attention mechanism in GAT enables the model to assign different importance scores to neighboring nodes during message passing. In co-authorship networks, recurring co-authors with shared topics offer a more predictive signal. By learning attention weights, the model amplifies informative connections and suppresses noisy ones, improving link prediction performance.

We use binary cross-entropy loss and evaluate performance with ROC-AUC and F1-score. Fig. 2 shows the dynamics of the selected metrics during the model training. All three curves in Fig. 2 show rapid improvement during the first 20 epochs, after which the loss stabilizes and both ROC-AUC and F1-score plateau. This indicates that the model converges early and maintains stable generalization throughout training, with no evidence of overfitting. The model achieves a stable ROC-AUC of ~0.86 and F1-score of ~0.78, indicating a reliable balance between precision and recall. These results suggest that the GAT effectively captures both thematic similarity and structural proximity in co-authorship graphs. The training loss steadily converges without overfitting, validating the regularization strategy. By leveraging these loss and evaluation metrics, we ensure that the model learns meaningful representations and generalizes well to unseen author pairs, effectively predicting future scientific collaborations.

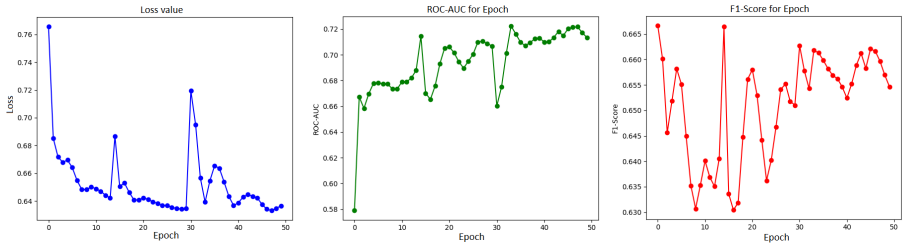

Fig. 2. Loss value (left), ROC-AUC (center), and F1 score (right) during model training.

Abstracts from ICCS proceedings are processed by cleaning, converting to TF-IDF vectors, and reducing them to 128-dimensional semantic embeddings using SVD. Simultaneously, topic identifiers for each paper are retrieved from OpenAlex, and a frequency count of each topic across an author's publications is compiled. The low-dimensional abstract embedding is then combined with this topic-frequency profile to form a rich feature vector for each author.

Co-authorships are represented as a weighted graph, where authors are nodes, and edges reflect the frequency and recency of collaborations. To predict new collaborations, a two-layer Graph Attention Network (GAT) is employed. In the first layer, each author's feature vector is projected into an intermediate representation and

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97635-3_45

combined with those of neighboring authors through parallel attention mechanisms. This enables the model to focus on relevant collaborators by assigning higher weights to similar neighbors. After applying a nonlinear activation function and dropout for regularization, the enriched information flows into a second attention layer, generating a final 32-dimensional embedding for each author.

For link prediction, the embeddings of two authors are fed into a small multilayer perceptron, which outputs the likelihood of future collaboration. The model is trained on known emerging collaborations versus random author pairs, optimized using binary cross-entropy loss. This end-to-end pipeline—combining TF-IDF reduction, topic counting, graph attention, and a lightweight classifier—provides a clear framework for understanding how shared thematic interests drive scientific collaboration.

3 Prediction of collaboration in computational science community

We retrieved metadata for all ICCS proceedings via OpenAlex and collected 16178 authors and 321848 papers. We identified 3623 "emerging" collaborations by finding author pairs who first met at ICCS (i.e., had no prior joint publications) and later co-authored any paper. The co-authorship network contains a core of 1,687 authors (65.6%) and multiple smaller clusters. The top 30 contributors show distinct collaboration patterns over time. Fig. 3 visualizes this by plotting, for each year, the count of new co-author links each author brought into the network, with larger bubbles indicating more partnerships. Most core contributors join ICCS, spark fresh collaborations, and then sustain a steady rate with gradual tapering in new ties, while a few exhibit intermittent yet significant bursts of activity. Beyond this core lies a multitude of smaller components – 186 isolated author pairs and 109 clusters of 3-16 authors – representing collaborations that emerged largely independently of the main ICCS hub.

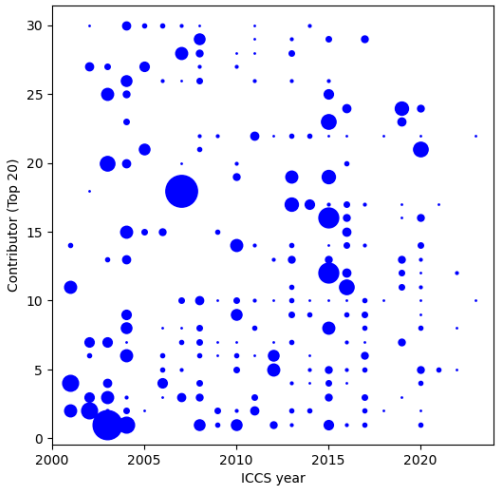

Fig. 3. Top contributors to the collaborating process.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97635-3_45

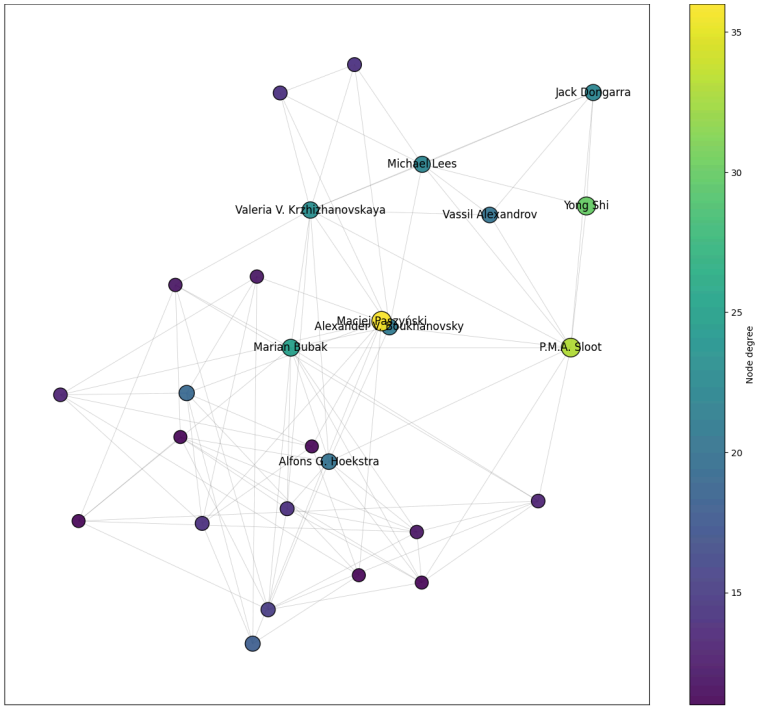

Fig. 4. Co-authorship graph: key authors and significant connections.

Topic similarity correlates with collaboration frequency: dense clusters form around Machine Learning and Data Science, while niche topics form smaller, isolated components.

A key aspect of our study is the effect of varying the time window parameter on predicting future scientific collaborations. The time window defines how much historical data is used when constructing the training graph. We experiment with different values of N, the number of years of past data included, before predicting the next year's collaborations.

By varying N, we can assess how historical data impacts prediction accuracy, such as whether a longer collaboration history improves performance or adds noise. This analysis also explores whether recent collaborations are more predictive than older ones, highlighting temporal dynamics in co-authorship patterns. The optimal time window strikes a balance between using enough historical data and avoiding outdated collaborations (see Fig. 5). A shorter N may capture recent trends more sharply, while a longer N captures long-term relationships at the cost of outdated interactions. We compare the performance of multiple training graphs using different values of N.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97635-3_45

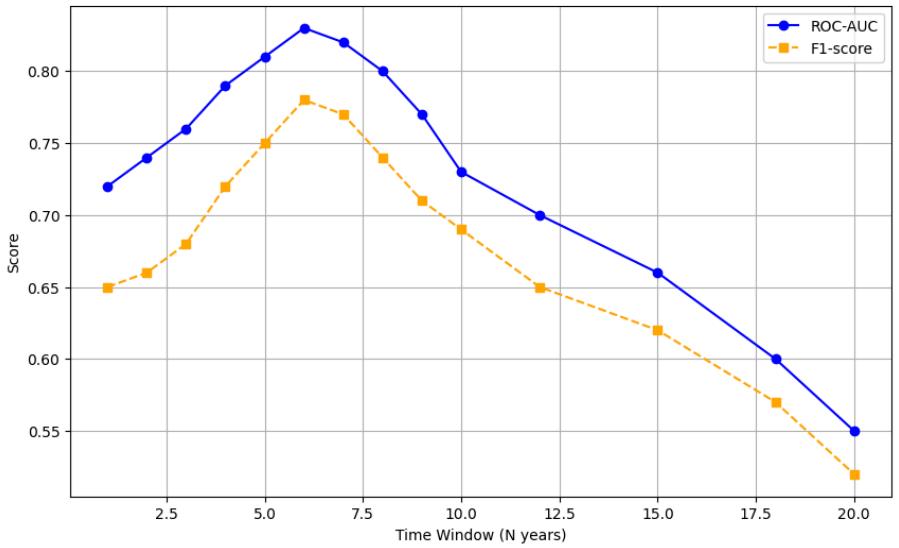

Fig. 5. Model predictive performance depending on the time window.

Evaluating the predictive performance of a model is crucial for ensuring its reliability in forecasting future collaborations. In this study, we assess the GAT-based model by comparing predicted links with actual collaborations that emerge in subsequent years. A robust evaluation framework is used to validate the model's ability to identify meaningful connections and distinguish between likely and unlikely co-authorships.

To measure performance, multiple key evaluation metrics are employed. By considering domain-specific variations, the analysis offers insights into the model's strengths and areas for improvement, guiding refinements in model architecture, feature selection, and data preprocessing to enhance predictive accuracy across scientific networks.

The choice of the N-year time window is vital for predicting future collaborations. We conducted experiments with varying N values, from short-term (1–3 years) to long-term (over 10 years) data, and trained models on these windows. Based on AUC results, we identified 5–7 years as optimal for balancing recency and historical depth.

The results show distinct trends depending on the time window. Performance drops for very long histories (N>10), even with controlled regularization. This decline is primarily due to noise from outdated collaborations, not overfitting, as no instability or divergent training loss was observed.

Using a short-term history (1–3 years) captures recent trends but overlooks long-term relationships, yielding higher precision but lower recall. A medium-term history (4–7 years) strikes a balance, incorporating both recent and long-term patterns, and provides the best performance across evaluation metrics. A long-term history (>10 years) includes outdated collaborations, reducing precision as older ties become less relevant.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97635-3_45

Fig. 5 illustrates the trade-off: short windows focus predictions narrowly, while excessive history dilutes the signal with irrelevant past interactions. An optimal time window of 5-7 years offers the best balance between recency and depth. These findings suggest that recent collaborations play a dominant role in predicting future co-authorship, while very old collaborations contribute less to accuracy.

4 Conclusions

Compared to classic heuristics, the attention mechanism in GAT more effectively captures higher-order patterns in scientific collaboration. Key limitations of the current approach include incomplete publication records and the reliance on a fixed historical window. Our experiments show that a 5-7-year time window provides the best overall performance; however, the optimal span is likely to vary across different fields. Further enhancements could consider integrating author-level features (e.g., citation metrics, transformer-based abstract embeddings), employing adaptive time-window selection, and expanding the dataset by incorporating external sources such as Google Scholar, ORCID, ArXiv, or citation networks. These improvements would help reduce data sparsity and enhance coverage, leading to better prediction accuracy and generalization.

Acknowledgments. The research was supported by the Russian Science Foundation, agreement No. 24-11-00272, https://rscf.ru/project/24-11-00272/.

References

1. Luo, L., Bochenina, K., Abuhay, T.M., Dorzhu, N., Kampis, G., Kovalchuk, S., Krzhizhanovskaya, V., Paszynski, M., Mulatier, C.D., Dongarra, J., Sloot, P.: Evolution of Computational Science Community: The Dynamics of Topics Analysis and Authors Collaboration in 24 Years of Iccs and Jocs Publications, https://www.ssrn.com/abstract=5060728, (2024). https://doi.org/10.2139/ssrn.5060728.
2. Fortunato, S., Bergstrom, C.T., Börner, K., Evans, J.A., Helbing, D., Milojević, S., Petersen, A.M., Radicchi, F., Sinatra, R., Uzzi, B., Vespignani, A., Waltman, L., Wang, D., Barabási, A.-L.: Science of science. Science. 359, eaao0185 (2018). https://doi.org/10.1126/science.aao0185.
3. Arrar, D., Kamel, N., Lakhfif, A.: A comprehensive survey of link prediction methods. J. Supercomput. 80, 3902–3942 (2024). https://doi.org/10.1007/s11227-023-05591-8.
4. Zare, G., Jafari Navimipour, N., Hosseinzadeh, M., Sahafi, A.: Network link prediction via deep learning method: A comparative analysis with traditional methods. Eng. Sci. Technol. Int. J. 56, 101782 (2024). https://doi.org/10.1016/j.jestch.2024.101782.
5. Gu, W., Gao, F., Lou, X., Zhang, J.: Link Prediction via Graph Attention Network, https://arxiv.org/abs/1910.04807, (2019). https://doi.org/10.48550/ARXIV.1910.04807.
6. Soundarajan, S., Hopcroft, J.: Using community information to improve the precision of link prediction methods. In: Proceedings of the 21st International Conference on World Wide Web. pp. 607–608. ACM, Lyon France (2012). https://doi.org/10.1145/2187980.2188150.
7. Luo, L., Kovalchuk, S., Krzhizhanovskaya, V., Paszynski, M., Mulatier, C. de, Dongarra, J., Sloot, P.M.A.: Trends in Computational Science: Natural Language Processing and Network Analysis of 23 Years of ICCS Publications. Lect. Notes Comput. Sci. 14833, 19–33 (2024). https://doi.org/10.1007/978-3-031-63751-3_2.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97635-3_45