Graph Data Science for Author Domain Classification and Link Prediction

Basil Ali Khan, Hamza Ansari, Hayyan Khan

Group 6
Graph Data Science, Spring 2025

I. Data and Preprocessing

Data source: [1]
The dataset comprises academic publications represented as a heterogeneous graph consisting of Author, Paper, and Citation entities. Preprocessing was conducted using R, where the raw data was cleaned, missing values were handled, and domain labels were assigned to authors. Subsequently, the data was imported into Neo4j and modeled using nodes: `Author`, `Paper`, and relationships: `WRITES`, `CITES`, and `CO_AUTHOR`.

A co-authorship graph was generated where authors sharing a paper are connected via `CO_AUTHOR` edges. Only authors with a known research domain (`domain_id`) were included in the node classification task. These authors were given an additional label `AuthorWithDomain`.

II. Methodology

A. Node Classification

A native projection of the co-authorship network was created for authors labeled `AuthorWithDomain`. Key features included:

- **Louvain Community ID:** Detected to capture cluster-level collaboration structure.
- **Graph Embedding:** FastRP embeddings generated for structural representation.
- **Domain ID:** Used as the target property for training.

A machine learning pipeline was created using the Neo4j Graph Data Science (GDS) library. A Random Forest classifier was trained with 5-fold cross-validation. Evaluation metrics included F1-Weighted, Accuracy, and Out-of-Bag (OOB) error.

B. Link Prediction

III. Results

A. Node Classification

B. Link Prediction

IV. Discussion

The classification model performed well in communities with distinct co-authorship patterns. However, misclassifications occurred where authors collaborated across multiple domains. Embeddings improved performance but were zero-vectors for disconnected nodes, requiring further filtering.

Link prediction success depended on network density and prior collaboration patterns. Future work may incorporate GNN-based predictors.

V. Future Work

- Use Graph Neural Networks for richer feature learning.
- Improve embedding quality by enriching node features.
- Incorporate topic modeling on paper abstracts.

VI. Conclusion

The project demonstrated the applicability of Graph Data Science methods for classifying authors by research domain and predicting future collaborations. While effective, several improvements can boost performance in complex, sparse graphs.

References

[1] L. Rothenberger, M. Q. Pasta, and D. Mayerhoffer, "Mapping and impact assessment of phenomenon-oriented research fields: The example of migration research," *Quantitative Science Studies*, vol. 2, no. 4, pp. 1466–1485, 12 2021. [Online]. Available: https://doi.org/10.1162/qss_a_00163