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 Preprocessing and Graph Modelling

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 and missing values were handled.

The graph model was designed to represent the relationships between authors, papers, topics, journals, and publishers in the academic domain. The data was modeled as a heterogeneous graph in Neo4j, with the following node types and relationships:

A. Nodes

- **Author**: Represents researchers or authors. Key properties include:
  - `id`: Unique identifier for the author.
  - `name`: Name of the author.
  - `total_citations`: Total number of citations received by the author's papers.
  - `primary_topic`: The most researched topic by the author.
  - `last_collab_year`: The most recent year of collaboration with other authors.
- **Paper**: Represents academic papers. Key properties include:
  - `id`: Unique identifier for the paper.
  - `title`: Title of the paper.
  - `year`: Year of publication.
  - `citationCount`: Number of citations received by the paper.
- **Topic**: Represents research topics. Key properties include:
  - `id`: Unique identifier for the topic.
  - `name`: Name of the topic.
- **Journal**: Represents academic journals. Key properties include:
  - `name`: Name of the journal.
- **Publisher**: Represents publishers of journals. Key properties include:
  - `name`: Name of the publisher.

B. Relationships

- **WROTE**: Connects an `Author` to a `Paper` they authored.
- **CO_AUTHORED**: Connects two `Author` nodes who collaborated on the same paper. Includes:
  - `collaboration_year`: The year of collaboration.
- **HAS_TOPIC**: Connects a `Paper` to a `Topic` it addresses.
- **RESEARCHES**: Connects an `Author` to a `Topic` they have researched. Includes:
  - `paper_count`: Number of papers the author has written on the topic.
- **SHARED_TOPIC**: Connects two `Author` nodes who have researched the same topic. Includes:
  - `shared_topics`: Number of shared topics between the authors.
- **SHARED_JOURNAL**: Connects two `Author` nodes who have published in the same journal. Includes:
  - `shared_journals`: Number of shared journals between the authors.
- **PUBLISHED_IN**: Connects a `Paper` to the `Journal` it was published in.

- **PUBLISHED_BY**: Connects a `Journal` to its `Publisher`.
- **CITES**: Connects one `Paper` to another paper it cites.

C. Feature Engineering

To enhance the graph model, additional features were computed:

- **Total Citations**: The total number of citations received by an author's papers.
- **Primary Topic**: The topic most frequently researched by an author.
- **Last Collaboration Year**: The most recent year an author collaborated with others.

D. Neo4j Script For Graph Creation

The Neo4j script to model the graph can be found in the github repository in the $graph\_model\_creation\_script.txt$ file

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