Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

NETS

Social Network Analysis Lab Report

SNA Lab: Coauthorship Networks and Author Success

Ilayda Söz Yilmaz

Julius-Maximilians-Universität Würzburg
Center for Artificial Intelligence and Data Science
Lehrstuhl für Informatik - Machine Learning for Complex Networks

Abstract

Academic success is shaped by various factors, including collaboration networks and social influence among researchers. Inspired by Sarigöl et al. [7], this study explores how an author's position in the coauthorship network relates to their likelihood of publishing in top-tier venues. Using data from DBLP [3], this project constructs a time-resolved network and applies machine learning (ML) models to analyze the relationship between centrality measures and success. The results highlight the strengths and limitations of network-based success prediction. They suggests that while collaboration can offer meaningful insights, academic success depends on a wider range of factors.

1 Introduction

The academic publishing world is fiercely competitive, and identifying the factors that influence a researcher's success is a key challenge. Previous research, conducted by Sarigöl et al. [7], has examined the role of coauthorship networks using citation counts as a metric for success. Their findings revealed that network characteristics, such as an author's position within the coauthorship network, and centrality measures can be powerful indicators of future achievement.

In this project, I analyze the coauthorship network of researchers using data from the DBLP [3] dataset. The goal is to investigate whether publishing in top-tier venues affects an author's future success. My approach is inspired by the work [7], which studies the evolution of academic careers using citation counts. However, since the DBLP [3] dataset does not include citation information, I could not replicate their exact methodology. Instead, I devised an alternative approach: defining a surrogate success label based on whether a paper appears in a predefined set of top-tier venues. This allows to study author success without relying on citation data. To achieve this, I constructed a time-resolved coauthorship network, computed centrality measures for authors, and trained machine learning models to predict an author's likelihood of publishing in a top-tier venue.

The main objectives of this project are:

- Construct a time-evolving coauthorship network from DBLP [3] data and analyze its structural properties.
- Extract key network-based features (centrality measures) for authors and aggregate them at the paper level.
- Train Random Forest [5] and Multi-Layer Perceptron [4] to assess whether these network features can predict an author's success in publishing in top-tier venues.
- Evaluate these models and discuss potential limitations and improvements.

While this approach is different from [7]'s citation-based method, it still allows me to explore the role of network position in academic success.

2 Dataset and Preprocessing

DBLP [3] dataset contains metadata about academic publications, it provides a rich source of coauthorship information. However, as the dataset does not contain citation counts, I had to derive an alternative way to measure success.

2.1 Extracting Data

The original DBLP [3] XML dataset contains information about millions of publications, but for this project, I filtered and extracted only the relevant records. The key filtering steps were:

- Parsing publications from 2019 to 2025 to analyze up-to-date data. This results in 2,873,652 publications.
- Keeping only entries that contain authors, a publication year, and a venue.
- Removing duplicate records based on paper title and year.
- Standardizing venue names (converting them to uppercase) to avoid mismatches.

2.2 Defining the Success Label

Since citation counts were unavailable, I introduced a binary success label for each paper:

- Papers published in top-tier venues (ICLR, ICML, NeurIPS, CVPR, SIGCOMM, IEEE Access) were assigned success = 1.
- All other papers were labeled success = 0.

This follows the assumption that publishing in prestigious venues correlates with an author's impact in their field. Although this approach does not capture all aspects of academic success, it provides a meaningful approximation.

2.3 Creating a Sampled Dataset

Due to computational limitations, especially during constructing networks and computing centralities for millions of data, I worked with subsets of the 2019-2025 DBLP [3] dataset. I created smaller samples (100K, 200K, 500K, and 1M papers) to experiment with model performance at different scales.

3 Building the Coauthorship Network

After the dataset was preprocessed, the next step was to build the coauthorship network. The coauthorship network was modeled as an undirected graph:

- Nodes represent authors.
- Edges exist between coauthors who have at least one paper together.
- Edge weights represent the number of coauthored papers.

To study how coauthorship patterns evolved, I used a sliding time window approach like in [7]. I created 2-year networks (e.g., 2019-2020, 2020-2021, etc.).

Also, since academic networks can be large and sparse, I removed self-loops (where an author is linked to themselves), filtered out isolated nodes and extracted the largest connected component to ensure meaningful analysis.

4 Computing Network Features

To quantify an author's influence in the network, I computed the following centrality measures. These are chosen based on [7]'s findings, which identified as the most meaningful indicators. Eigenvector Centrality was excluded, since it has the smallest influence according to their findings.

- **Betweenness Centrality:** Measures how often an author appears on the shortest paths between other authors. Higher values suggest that an author plays a bridging role.
- **K-Core Number:** Represents how embedded an author is in densely connected subgroups.

These metrics were computed using Networkit [1], which provides efficient algorithms for large-scale graphs.

Since predictions were made at the paper level, I aggregated author-level centrality values. Each paper was assigned the maximum betweenness and k-core value among its authors. This assumes that the most central author in a paper plays a key role in its success. After feature extraction, each paper was represented as Max Betweenness, Max K-Core and Success Label (0 or 1).

5 Predictive Modeling

With the processed dataset and extracted network features, I trained machine learning models to predict whether a paper would be published in a top-tier venue. Given the class imbalance in the dataset (with far fewer successful papers), I experimented two different techniques to improve prediction performance: Random Forest [5] classifier and Multi-Layer Perceptrons (MLP) [4]. More detailed results will be provided in the section 6.

5.1 Random Forest Classifier

The first model was a Random Forest [5] classifier, chosen for its robustness and ability to handle small datasets. Also, it was used in the original paper [7] as well.

Results show high precision for non-successful papers (above 75%). But, low recall for successful papers due to class imbalance.

5.2 Multi-Layer Perceptron (MLP)

To explore deep learning, I implemented an MLP [4] classifier with Feature Scaling (StandardScaler [6]) and Class Balancing (SMOTE oversampling [2]).

Results show slightly improved recall for successful papers but still struggled for minority class due to limited features.

6 Results

The performance of the predictive models was evaluated using multiple dataset sizes to understand how features correlate with an author's success in publishing in top-tier venues.

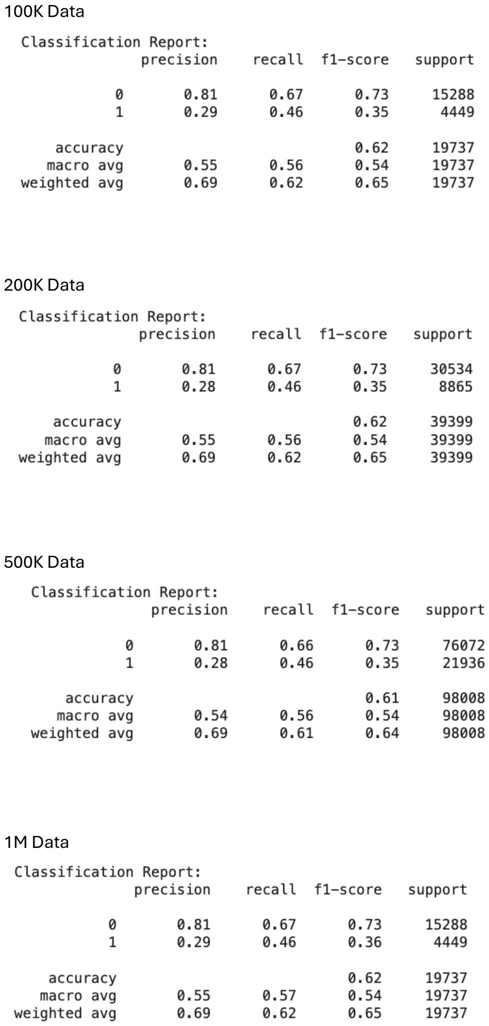
Both Random Forest [5] and MLP [4] classifiers trained on 100K, 200K, 500K, and 1M subsets of the DBLP [3] dataset. The goal was to observe whether increasing data size improves prediction accuracy and whether network-based features remain informative as more data is added.

6.1 Random Forest Performance

The Random Forest [5] classifier performed well in distinguishing between successful and non-successful authors but suffered from class imbalance (far fewer successful authors).

- Precision for the majority class (non-successful papers) remained high (above 75%), while precision for successful papers remained significantly lower.

- As dataset size increased from 100K → 1M, there was only a minor improvement in recall for successful authors. This shows that the features alone might not be enough.



100K Data

```
Classification Report:
              precision    recall  f1-score   support

           0       0.81      0.67      0.73     15288
           1       0.29      0.46      0.35      4449

    accuracy                           0.62     19737
   macro avg       0.55      0.56      0.54     19737
weighted avg       0.69      0.62      0.65     19737
```

200K Data

```
Classification Report:
              precision    recall  f1-score   support

           0       0.81      0.67      0.73     30534
           1       0.28      0.46      0.35      8865

    accuracy                           0.62     39399
   macro avg       0.55      0.56      0.54     39399
weighted avg       0.69      0.62      0.65     39399
```

500K Data

```
Classification Report:
              precision    recall  f1-score   support

           0       0.81      0.66      0.73     76072
           1       0.28      0.46      0.35     21936

    accuracy                           0.61     98008
   macro avg       0.54      0.56      0.54     98008
weighted avg       0.69      0.61      0.64     98008
```

1M Data

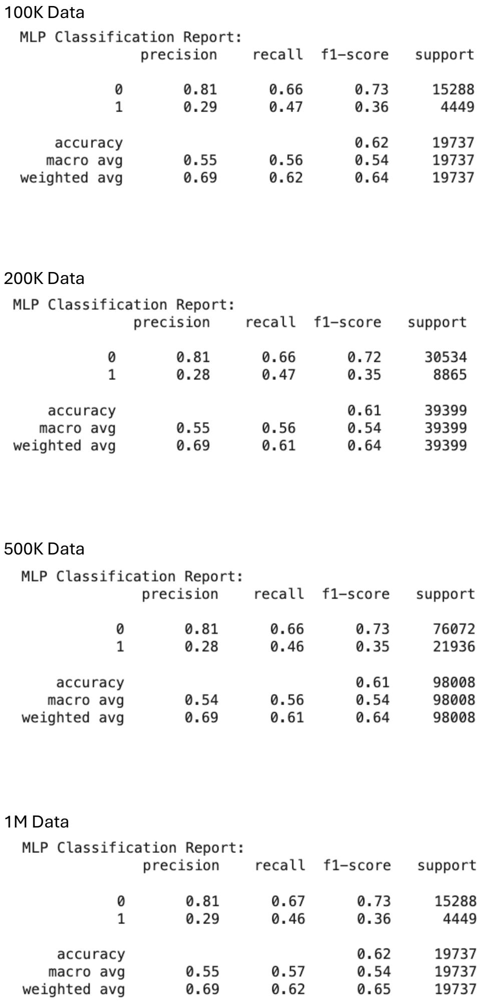
```
Classification Report:
              precision    recall  f1-score   support

           0       0.81      0.67      0.73     15288
           1       0.29      0.46      0.36      4449

    accuracy                           0.62     19737
   macro avg       0.55      0.57      0.54     19737
weighted avg       0.69      0.62      0.65     19737
```

Figure 1: Classification report for Random Forest model for different dataset sizes.

6.2 MLP Performance

The MLP [4] classifier was tested with StandardScaler [6] for feature normalization and SMOTE [2] for oversampling the minority class to handle class imbalance.

- Compared to Random Forest [5], the MLP [4] slightly improved recall for successful authors but still struggled to significantly improve F1-score for the minority class.
- As dataset size increased, overfitting was less of a concern, but the model still struggled with the feature limitations.



100K Data

```
MLP Classification Report:
              precision    recall  f1-score   support

           0       0.81      0.66      0.73     15288
           1       0.29      0.47      0.36      4449

    accuracy                           0.62     19737
   macro avg       0.55      0.56      0.54     19737
weighted avg       0.69      0.62      0.64     19737
```

200K Data

```
MLP Classification Report:
              precision    recall  f1-score   support

           0       0.81      0.66      0.72     30534
           1       0.28      0.47      0.35      8865

    accuracy                           0.61     39399
   macro avg       0.55      0.56      0.54     39399
weighted avg       0.69      0.61      0.64     39399
```

500K Data

```
MLP Classification Report:
              precision    recall  f1-score   support

           0       0.81      0.66      0.73     76072
           1       0.28      0.46      0.35     21936

    accuracy                           0.61     98008
   macro avg       0.54      0.56      0.54     98008
weighted avg       0.69      0.61      0.64     98008
```

1M Data

```
MLP Classification Report:
              precision    recall  f1-score   support

           0       0.81      0.67      0.73     15288
           1       0.29      0.46      0.36      4449

    accuracy                           0.62     19737
   macro avg       0.55      0.57      0.54     19737
weighted avg       0.69      0.62      0.65     19737
```

Figure 2: Classification report for MLP model across different dataset sizes.

6.3 Observations on Model Effectiveness

The classification reports from both models suggest the following:

Class Imbalance Issue:

- Both models struggled to predict success accurately due to the severe class imbalance in the dataset.
- SMOTE [2] improved recall slightly in MLP [4], but it did not significantly improve the overall predictive accuracy.
- One possible reason for this imbalance is the predefinition of ”top-tier” venues, which is inherently **subjective**. For example, redefining what qualifies as ”top-tier” could provide a more balanced dataset, potentially leading to better predictive performance.

Feature Limitations:

- The betweenness centrality and k-core number were useful features, but they alone did not fully capture the complexity of an author’s success.
- More advanced graph-based embeddings or collaboration history patterns could potentially improve the prediction.

Scaling Effects:

- Increasing the dataset size had a minimal impact on classification performance. Even with 1M data, recall for successful papers remained low, suggesting the need for additional features.

Potential Improvements:

- Adding publication count trends per author or venue reputation scores could provide a richer feature set.
- Using Graph Neural Networks (GNNs) or Node Embeddings could better capture coauthorship dynamics.

7 Conclusion

This study was inspired by [7]'s work on predicting academic success using **coauthorship networks**. Their approach relied on **citation counts** to define success, but since DBLP [3] does not provide citation information, I had to adapt the methodology. Instead, I assigned a **surrogate success label** based on whether a paper appeared in a **predefined set of top-tier venues**. This alternative definition allowed me to investigate author influence and career trajectories using network-based features.

The results indicate that:

- **Network-based features provide meaningful predictive signals:** Both betweenness centrality and k-core number helped distinguish successful from non-successful authors.
- **Models performed reasonably well despite class imbalance:** The MLP [4] model with SMOTE [2] oversampling improved recall for successful authors, while the Random Forest [5] classifier maintained robust precision for non-successful papers.
- **Venue selection influences results:** Top-tier venues were predefined and subjective, which may have influenced classification performance. Adjusting the venues or using a more dynamic ranking system could improve balance.

7.1 Future Work

While the models performed reasonably well within the scope of this study, several directions can be explored:

1. **Refining the success definition:** Instead of a fixed set of top-tier venues, future work could incorporate venue rankings or acceptance rates.
2. **Integrating additional features:** Factors such as publication trends or venue impact scores might improve predictions.
3. **Exploring graph embedding techniques:** Methods like Node2Vec could capture richer network dynamics beyond classical centrality measures.

This study demonstrates that coauthorship networks contain valuable predictive information about academic success. While challenges remain, the findings suggest that network-based features can serve as a strong foundation for future research in academic impact prediction.

References

[1] Eugenio Angriman, Alexander van der Grinten, Michael Hamann, Henning Meyerhenke, and Manuel Penschuck. Algorithms for large-scale network analysis and the networkit toolkit. In *Algorithms for Big Data*, pages 3–20. Springer Nature Switzerland, 2023. doi: 10.48550/arXiv.2209.13355. URL `https://arxiv.org/abs/2209.13355`.

[2] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[3] dblp Team. dblp computer science bibliography – Monthly Snapshot XML Release. URL `https://doi.org/10.4230/dblp.xml`.

[4] Simon Haykin. *Neural networks: a comprehensive foundation.* Prentice Hall PTR, 1994.

[5] Tin Kam Ho. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pages 278–282. IEEE, 1995.

[6] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

[7] Emre Sarigöl, René Pfitzner, Ingo Scholtes, Antonios Garas, and Frank Schweitzer. Predicting scientific success based on coauthorship networks. *EPJ Data Sci.*, 3(1): 9, 2014. doi: 10.1140/epjds/s13688-014-0009-x. URL `https://doi.org/10.1140/epjds/s13688-014-0009-x`.