Author Name Disabiguation using Markov Chain Monte Carlo

Dhruv Singh Chandel
dhruv.singh.chandel@fit.fraunhofer.de
Fraunhofer FIT
Sankt Augustin, Germany

Nagaraj Bahubali Asundi
nagaraj.bahubali.asundi@fit.fraunhofer.de
Fraunhofer FIT
Sankt Augustin, Germany

Zeyd Boukhers
zeyd.boukhers@fit.fraunhofer.de
Fraunhofer FIT
Sankt Augustin, Germany

Sulayman K. Sowe
sowe@bdis.rwth-aachen.de
RWTH Aachen University
Aachen, Germany

Abstract

This paper presents a novel approach to the Incorrect Name Detection (IND) task as part of the KDD Cup 2024 Open Academic Graph Challenge (OAG-Challenge). We propose Author Name Disambiguation using Markov Chain Monte Carlo (AND-MCMC) algorithm to identify incorrectly assigned papers within author profiles in the WhoIsWho dataset [9]. Our method constructs graph structures or "graphlets" for each author and employs an iterative refinement process that prioritizes split actions over merge actions. The approach aims to effectively separate anomalous papers from those correctly attributed to the predominant author. Leveraging the dataset's structure that includes correctly and incorrectly assigned publications, the algorithm employed in this work processes one author's file at a time. We evaluate our method using a weighted Area Under the Receiver Operating Characteristic Curve (AUC) metric[2], which accounts for varying error distributions across authors. This work contributes to academic graph mining by addressing the challenges associated with detecting incorrect paper attributions in large-scale scholarly databases.

Keywords

Incorrect Name Detection using Author Name Disambiguation(IND-AND), Author Name Disambiguation, Graph Clustering, AND-MCMC Algorithm, Anomaly Detection, WhoIsWho Dataset, Weighted AUC, Graphlet Analysis, Scholarly Databases, Iterative Graph Refinement, OAG-Challenge, KDD Cup 2024

ACM Reference Format:

Dhruv Singh Chandel, Nagaraj Bahubali Asundi, Zeyd Boukhers, and Sulayman K. Sowe. 2024. Author Name Disabiguation using Markov Chain Monte Carlo. In *Proceedings of (KDD Cup 2024)*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

KDD Cup 2024, March 20th, 2024 - June 14th, 2024, Tsinghua University, China

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

The exponential growth of academic literature in recent years has led to an increasing need for accurate and efficient methods of organizing scholarly information[12]. One of the critical challenges in this domain is the correct attribution of academic papers to their respective authors, a task complicated by factors such as name ambiguity, inconsistent name spellings, and errors in database entries. The IND task, a crucial component of academic graph mining [3, 9], aims to address this challenge by identifying and rectifying instances where papers have been erroneously attributed to authors.

In the context of the OAG-Challenge at the KDD Cup 2024[1], this paper presents a novel approach to the IND task, leveraging the WhoIsWho dataset provided by AMiner.cn[2]. Our work builds upon the foundation laid by previous research in author name disambiguation and graph-based approaches to scholarly data analysis [11]. We propose the AND-MCMC algorithm, specifically tailored to detect incorrectly assigned papers within individual author profiles.

Our approach innovates by introducing an iterative refinement process that prioritizes the separation of anomalous papers from those correctly attributed to the predominant author. This is achieved through the construction and manipulation of graph structures, or "graphlets," for each author profile. ***By favoring split actions over merge actions in our algorithm***, we aim to effectively isolate incorrectly assigned papers, even in cases where they represent a small fraction of an author's overall publication record.

In the following sections, we provide a comprehensive overview of our methodology, including the data preprocessing steps, the AND-MCMC algorithm, and our evaluation framework. We then present and discuss our results, achieving a weighted AUC score of 0.499933, which indicates a moderate ability to distinguish between correct and incorrect paper assignments. Finally, we conclude with insights into the strengths and limitations of our approach, as well as potential avenues for future research in this critical area of academic graph mining.

This work contributes to the broader field of scholarly data analysis by addressing the specific challenge of incorrect paper attribution, a problem that has significant implications for the accuracy and reliability of academic databases and bibliometric studies. Our findings not only offer a novel solution to the Incorrect Name Detection

[1] https://www.biendata.xyz/kdd2024/
[2] https://www.aminer.cn/

(IND) task but also provide valuable insights into the complexities of author name disambiguation in large-scale academic datasets.

2 Background and Related Work

Author Name Disambiguation (AND) is a critical challenge in the field of academic graph mining, particularly as the volume of scholarly literature continues to grow exponentially[7] [1]. The task involves correctly attributing academic papers to their respective authors, a process complicated by factors such as name ambiguity, inconsistent name spellings, and errors in database entries. Accurate author name disambiguation is crucial for ensuring the reliability of bibliometric analyses, researcher evaluations, and literature discovery systems[10].

The AND task has been approached through various methodologies, broadly categorized into supervised, unsupervised, and hybrid approaches[4]. While supervised methods can achieve high accuracy, they often face challenges related to scalability and the need for substantial manually labeled data[6]. Unsupervised methods, on the other hand, offer better scalability but may sacrifice some accuracy compared to supervised approaches[8].

In recent years, graph-based approaches have gained significant traction in the field of AND. These methods leverage the complex relationships between authors, publications, and other metadata to improve disambiguation accuracy[1]. The use of co-authorship networks, in particular, has shown promise in distinguishing between authors with similar names but different collaboration patterns. The incorporation of advanced natural language processing techniques has also been explored in the context of AND. Word embeddings and topic modeling approaches have been employed to capture semantic similarities between paper titles and abstracts, providing valuable information for disambiguation purposes[5]. The use of these techniques allows for a more nuanced understanding of an author's research interests and how they evolve over time.

The concept of publication patterns over time has been identified as a valuable feature for disambiguation, especially in cases where authors have similar research interests but different career trajectories[6]. By modeling how an author's publications are distributed across topics and years, it becomes possible to distinguish between authors who may appear similar based on other metrics.

Despite these advancements, the AND problem remains challenging due to the dynamic nature of academic publishing, the increasing interdisciplinarity of research, and the global diversity of author names[8][1]. There is a continued need for methods that can effectively integrate multiple sources of information, handle large-scale datasets, and adapt to the evolving landscape of scholarly communication.

The AND-MCMC method proposed in this work aims to address these challenges by combining multiple features within a Markov Chain Monte Carlo framework. By leveraging domain similarity, co-authorship networks, publication patterns, and affiliation information, the method seeks to provide a comprehensive approach to author name disambiguation that can handle the complexities of modern academic databases.

3 Methodology

The data used for the IND task is collected from AMiner.cn and provided in the WhoIsWho dataset. The proposed approach transforms the dataset into a graph with nodes representing various features, and iteratively merges or splits these groups of nodes until clusters of homogeneous papers are generated. The entire approach used is shown in Figure 1.

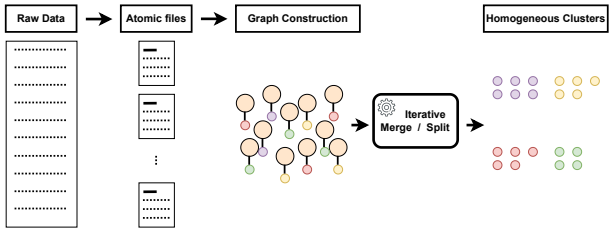

Figure 1: Architecture of the proposed approach.

For this task, we propose the AND-MCMC algorithm to detect the incorrectly assigned papers within each author's file. The key steps of the approach are as follows:

- **Data Loading and Transformation:** The process begins by loading the author's data from the provided dataset (e.g., `train_author.json`), which includes the author's ID, name, and paper IDs (both correctly and incorrectly assigned). This data is then transformed into an atomic name file format, creating a structured representation that facilitates subsequent graph construction and manipulation.
- **Graph Construction and Paper Sampling:** A graph structure, or "graphlet," is created with all paper nodes initially connected to a single atomic node representing the author. From this graphlet, a paper node is randomly selected for potential merge or split operations in the subsequent step.
- **Action Selection and Prioritization:** The algorithm selects an action to perform on the sampled paper node, either a merge or split operation. In the first iteration, the split action is favoured, as there are no other graphlets to merge with. Furthermore, throughout the iterative process, the split action is generally prioritised over the merge action to facilitate the separation of incorrectly assigned papers. an exaple of the split and merge actions can be seen in Figure 2.

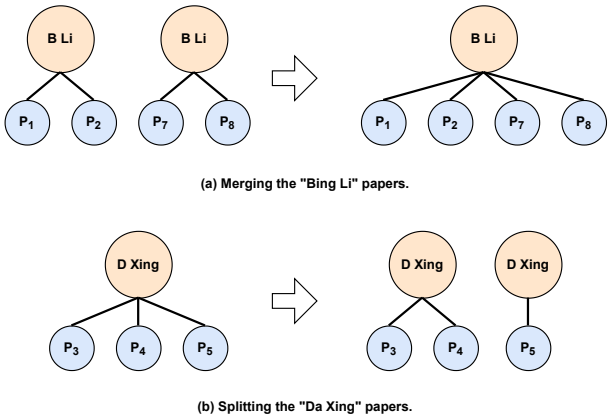

Figure 2: Illustrations of merge and split actions.

- **Action Execution and Iterative Refinement:** The selected merge or split action is applied to the sampled paper node, either combining it with an existing graphlet or creating a new one. This process is repeated iteratively until a stable graph topology is achieved, progressively refining the graphlets and separating incorrectly assigned papers from the predominant author's works.
- **Anomaly Identification:** After convergence, the graphlets with the highest number of paper nodes are considered to represent the correct author's papers. Conversely, the remaining graphlets with fewer paper nodes are flagged as anomalies or incorrectly assigned papers.

Acceptance Criterion

The algorithm evaluates proposed actions, such as merging or splitting, by first calculating an acceptance criterion. If this criterion exceeds a certain threshold, the proposed action is performed. The acceptance criterion is based on four factors.

(1) **Domain Similarity (α):** We assess the action by comparing the domain similarity of paper titles in the current graph state with the future state after the possible merge or split. For merging, a *target graphlet* is first sampled along with a *merging graphlet*. The decision to merge these graphlets is made based on the ratio of the similarity between the paper groups of the *target graphlet* and the *merging graphlet* to the similarity of the subgroups of papers within the *target graphlet*.
Let $\mathcal{P}(x)$ be a function that returns all papers in a graphlet x, $\texttt{Avg}(P)$ a function that returns the average of all embeddings of the paper set P, and $\texttt{Sim}(emb_1, emb_2)$ a function that returns the cosine similarity between the embedding vectors emb_1 and emb_2. Let g and h represent the target graphlet and the merging graphlet, respectively. Let g' and g'' be the interim splits of g such that $\mathcal{P}(g') \cup \mathcal{P}(g'') = \mathcal{P}(g)$. The ratio of domain similarity for the merge (α_{merge}) is given by:

$$\alpha_{merge} = \frac{\alpha^{(t+1)}}{\alpha^{(t)}} = \frac{\texttt{Sim}(\texttt{Avg}(\mathcal{P}(g)), \texttt{Avg}(\mathcal{P}(h)))}{\texttt{Sim}(\texttt{Avg}(\mathcal{P}(g')), \texttt{Avg}(\mathcal{P}(g'')))} \quad (1)$$

The interim splits g' and g'' are obtained by applying k-means clustering over the embeddings of all papers in g. The intuition behind computing α_{merge} is that we want to compare the similarity of the papers in graph state t with that in state $t+1$. Here, state t represents the status quo or the current state of the graph, and state $t+1$ represents the future state in which the target and merging graphlets are combined. Based on the ratio obtained, we decide whether it is worth merging the graphlets or not.
Unlike merging, we only need a target graphlet for splitting. Therefore, a target graphlet g is selected first. Then, a paper $\hat{p}$ is selected from all papers within g such that $\hat{p}$ is distant from the rest of the papers in terms of the domain. Now 3 interim splits g', g'', and g''' are obtained by applying k-means clustering to the papers of g such that $\mathcal{P}(g') \cup \mathcal{P}(g'') \cup \mathcal{P}(g''') = \mathcal{P}(g)$. Here g''' contains the paper $\hat{p}$. The ratio of domain similarity for the split (α_{split}) is given by:

$$\alpha_{split} = \frac{\alpha^{(t+1)}}{\alpha^{(t)}} = \frac{\texttt{Sim}(\texttt{Avg}(\mathcal{P}(g')), \texttt{Avg}(\mathcal{P}(g'')))}{\texttt{Sim}(\texttt{Avg}(\mathcal{P}(g') \cup \mathcal{P}(g'')), \texttt{Avg}(\mathcal{P}(g''')))} \quad (2)$$

(2) **Co-authorship Overlap (β):** Authors who have already published together are more likely to continue their collaboration and publish additional papers. Therefore, co-author networks are well-suited for author identification. In this work, the Jaccard index is used to measure the similarity between two graphlets.

(3) **Publication Pattern (γ):** The publication pattern models the domain of an author's papers over time. This is achieved by feeding the paper titles into a Gaussian Latent Dirichlet Allocation (LDA) model to extract latent topics. Once we have the topical distribution of papers, we determine the topical distribution over the duration of the author's publication period. For this, all the papers in a graphlet are collected, and the publication patterns are determined per topic. We model the distribution of the i^{th} topic over the years independently of the j^{th} topic, given there are n latent topics with $i \leq n$, $j \leq n$, and $i \neq j$. Determining the distribution per topic allows easier comparison of two sets of papers from independent graphlets.

(4) **Affiliation Overlap (κ):** While co-authorship overlap can help determine if both sets of co-authors belong to the same target author, there is still a possibility of ambiguity, as different authors may share the same co-author names. To address this, we also consider the affiliations of these co-authors. By checking for overlap in affiliations, we can better resolve co-author ambiguity.

The algorithm begins by sampling a target graphlet g. Then an appropriate action for g is selected. If the action is *merge*, another graphlet h is selected that could be a potential candidate to be merged with g. Also, two interim splits of g are obtained, namely g' and g''. The interim splits are just imaginary splits of g and have nothing to do with the *split* operation. The decision to merge g with h is based on the acceptance criterion A_{merge}, which is defined as follows:

$$A_{merge} = \alpha_{merge} \cdot \beta_{merge} \cdot \gamma_{merge} \cdot \kappa_{merge} \quad (3)$$

Once A_{merge} is calculated a value u is sampled from a uniform distribution $\mathcal{U}(0, 1)$. The proposal to merge g with h is accepted if $A_{merge} > u$. When the action is $split$, a paper $\hat{p}$ is selected within g that is farthest from the remaining papers with respect to the domain. Also, three interim splits of g are obtained, namely g', g'', and g'''. Here g''' contains the paper $\hat{p}$. The decision to separate g''' from g is based on the acceptance criterion A_{split} which is calculated similar to A_{merge}.

By favouring the split action and iteratively refining the graph structure, the proposed approach aims to separate the incorrectly assigned papers from the predominant author's papers within each file. The graphlets with the highest number of papers are considered to represent the correct author, while the remaining graphlets are flagged as anomalies or incorrect assignments.

This approach leverages the structure of the WhoIsWho dataset, where each author's file contains both correctly assigned papers (normal_data) and incorrectly assigned papers (outliers). By processing one author's file at a time and favouring the split action, the algorithm can effectively identify the anomalous papers that do not belong to the predominant author.

4 Evaluation Metrics

The performance of the proposed approach is evaluated using the Area Under the Receiver Operating Characteristic (ROC) Curve (AUC) metric

Here's a condensed, academic-style version of the information:

4.1 Author-wise AUC Calculation

The proposed approach is applied to each author's dataset, comprising both correctly and incorrectly assigned papers. Resultant graphlets are size-ranked, with the largest presumed to represent the author's legitimate works. Using ground truth labels, true positive and false positive rates are computed across various graphlet size thresholds. The ROC curve is then plotted, and the area under this curve (AUC) is calculated, yielding the author-specific AUC score.

4.2 Weighted AUC Calculation

To account for the varying importance of different authors in the dataset, a weighted AUC score is calculated as follows:

$$\text{Weight}_i = \frac{\text{\# Errors of the Author}_i}{\text{\# Total Errors}} \quad (4)$$

The weighted AUC is then calculated as the sum of the AUC scores for each author, multiplied by their respective weights.

$$\text{Weighted AUC} = \sum_{i=1}^{M} (\text{AUC}_i \times \text{weight}_i) \quad (5)$$

Where M is the total number of authors in the dataset.

The weighted AUC score accounts for each author's contribution to the dataset's total errors, emphasizing authors with higher incorrectly assigned paper counts. The proposed approach achieved a weighted AUC of 0.499933, indicating moderate success in distinguishing correct from incorrect assignments while highlighting potential for improvement. This metric offers a comprehensive evaluation across all authors, considering varied error distributions and appropriately weighting authors with higher error rates in the overall assessment.

5 Conclusion

This paper introduces the AND-MCMC algorithm, a novel approach to the Incorrect Name Detection (IND) task within the KDD Cup 2024 Open Academic Graph Challenge. Our method addresses the complexity of distinguishing correctly and incorrectly attributed papers in large-scale scholarly databases, achieving a weighted AUC score of 0.499933. While demonstrating potential, this result also indicates opportunities for further refinement. Our research contributes to creating reliable public benchmarks for academic graph mining, a critical gap in the field. The methodology serves as a foundation for future developments in author name disambiguation systems and showcases the potential of graph-based methods in scholarly data analysis. This work aims to stimulate further research in academic graph mining, addressing the limitation of suitable public benchmarks. While our approach shows promise, there remains significant scope for improvement. We anticipate this contribution will foster continued innovation in managing and analyzing scholarly data at scale, benefiting the broader KDD Cup 2024 audience and the academic community.

References

[1] Michele De Bonis, F. Falchi, and Paolo Manghi. 2023. Graph-based methods for Author Name Disambiguation: a survey. *PeerJ Computer Science* 9 (2023). https://api.semanticscholar.org/CorpusID:261793272

[2] Andrew P. Bradley. 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition* 30, 7 (1997), 1145–1159. https://doi.org/10.1016/S0031-3203(96)00142-2

[3] Başak Buluz and Burcu Yilmaz. 2017. Graph mining approach for modeling academic success. In *2017 25th Signal Processing and Communications Applications Conference (SIU)*. 1–4. https://doi.org/10.1109/SIU.2017.7960621

[4] Ijaz Hussain and Sohail Asghar. 2017. A survey of author name disambiguation techniques: 2010–2016. *The Knowledge Engineering Review* 32 (2017). https://api.semanticscholar.org/CorpusID:52175217

[5] Ayesha Manzoor, Sohail Asghar, and Tehmina Amjad. 2022. Toward a New Paradigm for Author Name Disambiguation. *IEEE Access* 10 (2022), 76055–76068. https://api.semanticscholar.org/CorpusID:250518707

[6] Andreas Rehs. 2021. A supervised machine learning approach to author disambiguation in the Web of Science. *J. Informetrics* 15 (2021), 101166. https://api.semanticscholar.org/CorpusID:236236516

[7] Neil R. Smalheiser and Vetle I. Torvik. 2009. Author name disambiguation. *Annu. Rev. Inf. Sci. Technol.* 43 (2009), 1–43. https://api.semanticscholar.org/CorpusID:205418775

[8] Alexander Tekles and Lutz Bornmann. 2019. Author name disambiguation of bibliometric data: A comparison of several unsupervised approaches1. *Quantitative Science Studies* 1 (2019), 1510–1528. https://api.semanticscholar.org/CorpusID:139102212

[9] Fanjin Zhang, Shijie Shi, Yifan Zhu, Bo Chen, Yukuo Cen, Jifan Yu, Yelin Chen, Lulu Wang, Qingfei Zhao, Yuqing Cheng, Tianyi Han, Yuwei An, Dan Zhang, Weng Lam Tam, Kun Cao, Yunhe Pang, Xinyu Guan, Huihui Yuan, Jian Song, Xiaoyan Li, Yuxiao Dong, and Jie Tang. 2024. OAG-Bench: A Human-Curated Benchmark for Academic Graph Mining. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

[10] Wenjing Zhang, Zhongmin Yan, and Yongqing Zheng. 2019. Author Name Disambiguation Using Graph Node Embedding Method. *2019 IEEE 23rd International Conference on Computer Supported Cooperative Work in Design (CSCWD)* (2019), 410–415. https://api.semanticscholar.org/CorpusID:199509475

[11] Yutao Zhang, Fanjin Zhang, Peiran Yao, and Jie Tang. 2018. Name Disambiguation in AMiner: Clustering, Maintenance, and Human in the Loop. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (2018). https://api.semanticscholar.org/CorpusID:207579405

[12] Ivan Zupic and Tomaž Čater. 2015. Bibliometric methods in management and organization. *Organizational research methods* 18, 3 (2015), 429–472.