Contrasting Global and Refined Views for Author Name Disambiguation in Heterogeneous Networks

Victor Gonçalves Lima[1], **Reinaldo Silva Fortes**[1], **Anderson A. Ferreira**[1]

[1]Postgraduate Program in Computer Science, Federal University of Ouro Preto.
victor.gl@aluno.ufop.edu.br
{reifortes,anderson.ferreira}@ufop.edu.br

Abstract. *Author Name Disambiguation (AND) as a clustering problem aims to correctly group publications sharing ambiguous author names into real author clusters. This work approaches the AND task using Heterogeneous Information Networks (HINs) to capture complex relationships between publications. Our method, AND-GloRe, employs a Graph Neural Network encoder trained via unsupervised contrastive learning to generate dual-view representations: a global view capturing overall relationships among publications, and a refined view capturing meta-path specific relationships. The effectiveness of our method is evaluated on the WhoIsWho datasets, achieving a pairwise F1-Score of 92.28%.*

1. Introduction

The exponential growth of academic publications in large bibliographic systems, such as Google Scholar and OpenAlex, has exacerbated the author name ambiguity problem. Multiple individuals often share similar or identical names, leading to the erroneous aggregation of their works and making it difficult to retrieve accurate author profiles. This ambiguity undermines the reliability of citation analysis, hinders the construction of scholarly networks, and impacts research evaluation and historical documentation. In this context, Author Name Disambiguation (AND) is a fundamental task in Text Mining and Information Retrieval, aiming to correctly group publications by their true authors. Effective AND is essential for improving search precision, enabling robust scientometric indicators, and supporting the integrity of academic knowledge graphs. The complexity of the task arises from diverse naming conventions, incomplete metadata, and the dynamic nature of scholarly collaboration. As a result, AND remains a critical area of research, with ongoing efforts to develop more accurate and scalable solutions.

An academic publication is generally associated with one or more authors, and each author may have an ambiguous name that refers to multiple people. To correctly group publications by author, it is usually necessary to capture relationships such as co-authorship, publication venue, and organizational affiliation. Furthermore, publications possess important semantic attributes related to the research area, such as the work title, abstract, and keywords, which can assist in disambiguation. The AND task can be formulated in different ways, depending on the available resources and the specific context [Ferreira et al. 2012, Amado Olivo et al. 2025]. Recently, AND has been widely treated as a clustering problem, where the objective is to group publications into clusters corresponding to the works of the same real author [Rodrigues et al. 2024]. A formal definition of the AND problem as a clustering problem is presented in Section 3.1.

In recent years, AND research has focused on approaches that build publication-based networks using overlapping relational attributes [Huang et al. 2024, Liu et al. 2024,

Zheng et al. 2021, Ye et al. 2025]. The network constructed for an ambiguous name set can be either homogeneous, in which all nodes represent publications and edges represent overlap in common attributes, or heterogeneous, in which nodes represent different entity types, such as publications, authors, venues, and organizations, and edges represent different types of relationships between these entities. In both cases, learned representations of publications derived from the networks are clustered to achieve disambiguation.

To break down the complexity of the AND task, it is common to construct an ambiguous name graph in which publications are associated with a specific ambiguous name. The construction of these graphs is often susceptible to noisy links, which can severely degrade model performance. Specifically, ambiguous features such as common co-author names can create spurious connections between otherwise unrelated publications, leading to incorrect disambiguation. In addition, ambiguity levels can vary significantly across ambiguous names, with some real authors having many publications and others having only a few, or even a single, publication. Also, publications in closely related fields can share similar features, making it difficult to distinguish between authors working on the same topic. These challenges highlight the need for more sophisticated modeling approaches to effectively capture the complex relationships between publications.

Recent advances in self-supervised learning and contrastive learning have shown great potential for improving the performance of graph learning models, including node clustering [Chen et al. 2020, Wang et al. 2019, Yu et al. 2024]. The objective is to discriminate between positive nodes within the same class and negative nodes from different classes using a learned representation. The effectiveness of these methods relies heavily on the quality of the positive and negative samples used during training. Recent techniques have focused on graph augmentation and multi-view learning, aiming to compare a node's role across different graph views to learn more robust representations.

In AND, approaches leveraging graph contrastive learning have shown promising results [Liu et al. 2024, Ye et al. 2025]. However, the effectiveness of those methods is often hindered by oversimplifying the relationships between publications. Those works use an initial heuristic step to construct a publication-based network into a homogeneous graph, where a defined set of features (e.g., co-authors, co-venues, co-organizations) is used to create edges between publication nodes. Although such a method captures the global relationships between publications by treating all connections equally, it can fail to capture the nuanced patterns of connectivity in real information networks.

Based on the discussed challenges, this work aims to propose an unsupervised contrastive learning solution based on Heterogeneous Information Networks (HINs), by constructing ambiguous heterogeneous graphs with different types of nodes specifically designed for the AND task: publications, authors, venues and organizations. The proposed method enables us to capture complex relationships between two publications via meta-paths. Unlike previous works that use heterogeneous networks for AND [Pooja et al. 2022, Zheng et al. 2021], we propose a contrastive learning framework that leverages a global view and a refined view based on relationship types to improve the quality of the learned representations for disambiguation. In summary, the main contributions of this work are: (1) present a HIN representation for the AND task that can capture the complex relationships between publications inside an ambiguous name graph; (2) propose AND-GloRe, a GNN-based unsupervised contrastive learning framework for

AND that leverages a global view and a refined view to learn effective representations for disambiguation; and (3) evaluate the proposed solution on commonly used AND datasets and compare its performance against other available baselines.

The remainder of this paper is organized as follows: Section 2 presents a review of related work in the field of AND. Section 3 describes our proposed method, including the problem formulation, data preprocessing, network construction, and the GNN-based model. Section 4 details the experimental evaluation including a discussion of the results obtained from the experiments and a comparison with baseline methods. Finally, Section 5 concludes the paper and suggests directions for future work.

2. Related Work

The development of the Author Name Disambiguation (AND) methods has evolved significantly over the past decades, as documented in recent surveys [Ferreira et al. 2020, Cappelli et al. 2025]. To better consolidate the AND research, several works have proposed ways to categorize the different approaches based on the type of evidence used and the methodological approach. [Ferreira et al. 2012] introduced a taxonomy for categorizing AND methods by evidence type (e.g., web information, citation relationships) and methodological approach (author grouping vs. author assignment). [Hussain and Asghar 2017] grouped AND techniques into supervised, unsupervised, semi-supervised, graph-based, and heuristic-based methods, with graph-based approaches leveraging co-authorship networks and similarity measures to enhance disambiguation.

Early solutions for AND addressed foundational challenges, including homonymy, name changes, spelling variations, incomplete metadata, and the impracticality of manual annotation at scale. The first automatic methods were based on clustering techniques and supervised learning models, such as Naive Bayes and Support Vector Machines (SVMs), which used publication attributes, such as co-authorship, title, and venue, to group or distinguish authors [Han et al. 2004, Huang et al. 2006, Cota et al. 2007, Elliott 2010, Rehs 2021]. Following, the AND research evolved, exploring a wider range of features and employing techniques such as deep learning and graph-based models [Levin and Heuser 2009].

The features and resources available for disambiguation can vary across datasets and domains, changing how the problem is approached. For instance, in some cases, we may have access to rich metadata, such as authors' affiliations with specific organizations and document abstracts, while in others, we may only have access to publication bibliographic information. The choice of features can impact the performance of AND models, and it is crucial to consider the specific characteristics of the dataset when designing and evaluating disambiguation models [Amado Olivo et al. 2025]. Recent advancements in machine learning, as in graph-based models, have shown promise in addressing the complexities of AND [De Bonis et al. 2023]. Techniques such as Graph Random Walks [Wang et al. 2020, Zhou et al. 2024b] and Graph Neural Networks (GNNs) [Zheng et al. 2021, Zhou et al. 2024a] have been employed to capture the intricate relationships between publications.

[Ma et al. 2019] proposes a meta-path channel-based heterogeneous network representation learning method to capture the complex semantics underlying academic graphs. Their approach defines neighboring and structural similarities to jointly learn

and update the embeddings of various node types directly from the entire heterogeneous network. Building upon meta-path concepts, [Xie et al. 2022] addresses the challenge of redundant feature selection by synthesizing both structural and semantic representations. They extract semantic features and introduce a meta-path-level self-attention mechanism, along with skip-gram embeddings, to automatically learn the weights of structural features. To address large-scale publication challenges, [Zhou et al. 2024a] adopts an ensemble framework. Their approach utilizes an extended MiniLMfor semantic feature extraction and applies an unsupervised feature fusion algorithm that synthesizes multi-order connectivity information via lightweight convolutions on the paper relational network, effectively modeling the discriminability differences across features.

Recently, contrastive learning has emerged as a promising approach to improving the performance of AND methods. [Liu et al. 2024] proposed a contrastive learning framework that adopts an iterative graph structure refinement process to continuously reduce the noise connections in the graph based on cosine similarity between learned representations. By employing supervised contrastive learning, the model improves the quality of the final publication representations, which are then clustered with a hierarchical algorithm to achieve disambiguation. However, similarity-based refinement may introduce false-positive links, degrading the method's performance. [Ye et al. 2025] also proposed a contrastive learning framework for AND, but with a focus on a self-supervised approach. The framework consists of multi-view contrastive learning by dynamically augmenting the graph structure. The model also incorporates a cluster-aware contrastive loss using the pseudo-labels generated by a density-based clustering algorithm . The final pseudo-label representations are then used to group the publications into groups corresponding to real authors.

Despite the advancements in contrastive learning for AND, current state-of-the-art frameworks [Liu et al. 2024, Ye et al. 2025] primarily operate on homogeneous networks derived from heuristic feature overlaps. Such simplification collapses the multi-modal nature of bibliographic data into a single edge type, potentially discarding discriminative structural noise. To mitigate this, our approach explicitly adopts a Heterogeneous Information Network (HIN) architecture that leverages meta-paths to preserve the semantic integrity of distinct relational types.

3. The AND-GloRe Method

We propose an Author Name Disambiguation method via a Global and Refined contrastive learning (AND-GloRe). Our method leverages the strengths of both global and refined views of the heterogeneous network to learn effective representations for disambiguation. While the global view captures the overall structure and relationships among publications in a simplified form, the refined view focuses on specific types of relationships among publications using selected meta-paths from the HIN representation.

Next, we formulate the AND problem as a clustering task and describe the three stages of AND-GloRe. First, AND-GloRe preprocesses the data to extract relevant features and construct the HIN representation for each ambiguous name. Next, we define a GNN-based model that includes a GAT encoder and a contrastive process. Finally, AND-GloRe applies a hierarchical agglomerative clustering algorithm on the learned representations to group publications into clusters corresponding to the

real authors. The source code for the proposed method is made publicly available at https://github.com/Victorgonl/AND-GloRe.

3.1. Problem Formulation

The AND task can be formulated as an open-set clustering problem. Let $P_a = \{p_1^a, p_2^a, \ldots, p_n^a\}$ be the set of publications associated with an ambiguous name a. Each publication $p_i^a \in P_a$ is formally represented by a feature tuple $\mathbf{x}_i = \langle S_i, R_i \rangle$, where S_i denotes the semantic attributes (e.g., work title, abstract, keywords) and R_i denotes the relational features (e.g., co-authors, co-venues, co-organizations). The objective is to partition the set P_a into a set of clusters $C = \{C_1, C_2, \ldots, C_{\hat{k}}\}$, such that ideally each cluster corresponds to the works of a single distinct real author sharing the ambiguous name a. Formally, we aim to find a clustering function $f : P_a \rightarrow \{1, 2, \ldots, \hat{k}\}$ that maps each publication p_i^a to a specific cluster based on its features $\mathbf{x}_i$. Because this is an open-set clustering problem, the number of clusters $\hat{k}$ is not a fixed prior, rather, $\hat{k}$ is an unknown output variable that must be jointly inferred with the cluster assignments in order to maximize intra-cluster similarity and minimize inter-cluster similarity.

3.2. Preprocessing Step

To construct the HIN representation for the AND task, we first preprocess each publication in P_a to extract relevant relational features. This involves extracting co-authors' names, their respective organizations, and publication venues. After, we standardize the authors' names by creating a unique identifier for different variations of the same name (e.g., converting "Barry Smith" to "barry_smith"). The venue and organization names are cleaned and normalized by converting to lowercase, removing punctuation, and removing stop words. After preprocessing, we have a set of author names, venue titles, and organization names associated with the publications in P_a.

As semantic attributes, we use the work title, abstract, and keywords from each publication to create a corpus for a Word2Vec model [Mikolov et al. 2013]. The resulting word embeddings are averaged to create a single semantic feature vector for each publication, which is then used as node features in the HIN representation. Notably, only the title and keywords are used when generating the embeddings. The choice to use Word2Vec is motivated by its ability to capture local contextual information in text, which can be beneficial for distinguishing between publications with similar topics [Ye et al. 2025]. [Cheng et al. 2024] demonstrates the model's effectiveness at generating representations specifically for AND, compared with pre-trained Transformer-based language models.

3.3. Heterogeneous Information Networks Construction

After the preprocessing step, for each ambiguous name a, we construct an HIN. The nodes of the HIN can be of four types: (1) **Publication (P)**; (2) **Author (A)**; (3) **Venue (V)**; and (4) **Organization (O)**. The ambiguous name a is not included as an author node, only as a co-author. The publication nodes are related to their respective relational attributes. Because co-authors' names can also be ambiguous, the authors' nodes are related to organizations, improving the capture of affiliation information. The four types of relationships between these nodes are defined as follows: (1) **Publication-Author (P-A)**: An edge exists between a publication node and an author node if the author is listed as one of the co-authors of the publication. The ambiguous name a is not included;

(2) **Publication-Organization (P-O)**: An edge exists between a publication node and an organization node if the ambiguous author name a is affiliated with the organization at the time of the publication; (3) **Publication-Venue (P-V)**: An edge exists between a publication node and a venue node if the publication was published in that venue; (4) **Author-Organization (A-O)**: An edge exists between a co-author node and an organization node if the author is affiliated with the organization at the time of the publication.

Based on the relationships among these nodes, we can indirectly capture relationships among publications via meta-paths. A meta-path is a sequence of node types and edge types that defines a path between two target nodes in a heterogeneous graph. In our case, we want to capture relationships between publication nodes, so we use four types of meta-paths based on these relationships. These meta-paths are defined as follows: (1) **P-A-P**: Connects two publication nodes through a common co-author node. This meta-path captures the co-authorship relationships between publications, since publications written with the same collaborators are more likely to belong to the same real-world author; (2) **P-V-P**: Connects two publication nodes through a common venue node. This meta-path was selected to capture topical consistency and academic habits. Researchers tend to specialize in a specific subfield and routinely publish their work in a concentrated, recurring set of journals or conferences. (3) **P-O-P**: Connects two publication nodes through a common organization node. This meta-path captures the relationships between publications that are affiliated with the same organization, since authors tend to maintain institutional affiliations over multiple publications; (4) **P-A-O-A-P**: Connects two publication nodes through authors who share a common organization node. This meta-path captures relationships among publications written by authors affiliated with the same institution or laboratory, allowing the model to exploit organizational proximity and indirect collaboration patterns even when direct co-authorship is absent.

3.4. Global and Refined Method

Our proposed method consists of four main components: (1) hub-targeted graph augmentation; (2) graph network encoder; (3) global and refined representation learning; and (4) adaptively weighted contrastive loss. Described individually in the next subsections.

3.4.1. Hub-Targeted Graph Augmentation

Contrastive learning highly depends on the quality of the positive and negative samples used during training. To enhance the robustness of the learned representations and prevent the model from over-relying on dominant nodes, we generate graph augmentations using hub-targeted strategies. By analyzing the degree distributions of nodes in the constructed HINs, we identify hub nodes with a disproportionately high number of connections, which can skew the learning process. By calculating the degree centrality of nodes in the constructed HINs, we identify hub entities, including highly ambiguous co-author names, popular publication venues, and large organizations. To mitigate the disproportionate influence of these hubs, we apply a targeted edge dropout strategy that selectively removes connections to these nodes with an adjustable probability proportional to their structural centrality. We utilize a parallel random masking strategy to drop features within publication nodes. The parameters for these augmentations must be carefully tuned to balance preserving essential structural information with sufficient variability to enhance the model's generalization.

3.4.2. Graph Network Encoder

For the graph encoder, we use a Graph Attention Network (GAT) variant proposed by [Wang et al. 2019] that is specifically designed for graph clustering tasks. This GAT incorporates higher-order structural information by extending the standard attention mechanism beyond immediate neighbors. Specifically, instead of restricting message passing to one-hop neighborhoods as in conventional GAT, the model aggregates information from multi-hop neighbors through a topology-aware proximity matrix. This enables the encoder to better capture global graph structure and long-range dependencies among nodes.

3.4.3. Global and Refined Representation Learning

To effectively capture the complex structural dependencies within HINs, we propose an AND method driven by a contrastive learning objective that integrates both a global view and a refined view. This dual-view approach was proposed by [Yu et al. 2024] for heterogeneous graph contrastive learning on HINs. Specifically, the framework constructs a coarse view to determine which objects are connected by meta-paths, alongside a fine-grained view that leverages meta-path contexts to characterize the specific details of their connections.

In the global view, the model constructs a comprehensive topological representation by averaging the multiple meta-path specific adjacency matrices into a single, unified adjacency structure. The raw target node features, representing the papers, are first projected into a hidden-dimensional space using a linear transformation, then passed through an ELU activation function. We process this combined global graph structure and the projected target features by the GAT encoder to learn generalized, high-level structural embeddings. Then we pass the resulting representations through a non-linear projection head that uses a $\tanh$ function and L2-Normalization to produce the final global representations for the contrastive objective.

The refined meta-path view captures fine-grained semantic nuances by distinctly aggregating neighborhood features for each specific relation type. The model employs sparse matrix multiplication to aggregate transformed neighbor features, such as connected authors, venues, and organizations, and integrates them with the target publication features via a dedicated agglomerative neural layer, one distinct for each meta-path. To construct robust contrastive pairs, the framework applies the hub-targeted masking strategy. Both the original and these augmented meta-path graphs are encoded using the shared GAT module. Finally, a semantic-level attention mechanism dynamically computes weights for each meta-path and combines them into a single fine-grained representation, which is then projected and normalized.

3.4.4. Adaptively Weighted Contrastive Loss

To align both perspectives, the global and the refined representations are projected into a common latent space. This is achieved via a linear projection head followed by a $\tanh$ activation and L2-Normalization. A weighted *InfoNCE* contrastive objective is then employed to maximize the agreement between these two views.

Unlike standard contrastive loss, this formulation incorporates a learnable weight-

ing function to dynamically adjust the contribution of negative samples, enabling more informative and discriminative supervision. Specifically, a Multi-Layer Perceptron (MLP) with a hidden layer using $\tanh$ activation and an output layer using Sigmoid activation computes soft-valued weights for every pairwise node combination. The input to this MLP is the element-wise addition of the projected node representations, yielding a weight matrix $W \in \mathbb{R}^{N \times N}$ bounded between 0 and 1.

The adaptively weighted contrastive loss for the i-th node is defined as:

$$\mathcal{L}_i = -\log \frac{\exp(z_{refined,i} \cdot z_{global,i}/t)}{\sum_{j=1}^{N} W_{i,j} \exp(z_{refined,i} \cdot z_{global,j}/t)} \tag{1}$$

where t is a temperature hyperparameter, $z_{refined}$ and z_{global} are the projected representations from the refined and global views, respectively, and $W_{i,j}$ represents the adaptive weight for the negative sample pair (i, j) assigned by the MLP. The final objective is calculated as the average loss across all N nodes in the batch.

3.5. Hierarchical Agglomerative Clustering

After the learning process, we use the final refined representation $z_{refined}$ to cluster publications into groups corresponding to the real authors' works. We use a bottom-up Hierarchical Agglomerative Clustering (HAC) algorithm, along with the cosine similarity distance metric and average linkage criteria. Due to the unknown number of clusters and the different publication graphs that can have different optimal representations, we use a dynamic stopping criterion based on the silhouette score on top of the hierarchical clustering dendrogram.

4. Experimental Evaluation

We conduct experiments on commonly used AND datasets described in Section 4.1 to validate our proposed method. The experimental setup is detailed in Section 4.2. Section 4.3 describes the evaluation metrics for AND. Section 4.4 describes available baselines for comparison with our proposed method. Finally, Section 4.5 presents and discusses the experimental results and compares them with the baseline methods.

4.1. Datasets

Two widely used datasets were selected for the evaluation of the proposed WhoIsWho-v1 and WhoIsWho-v2 datasets [Chen et al. 2023]. Both are part of the WhoIsWho-AND data collection[1]. The datasets contain a large number of publications from different research areas, with varying degrees of ambiguity across the ambiguous names. The WhoIsWho-v1 dataset contains 321 ambiguous names, of which 50 are used for testing. The WhoIsWho-v2 dataset contains 165 ambiguous names, of which 34 are used for testing. Table 1 summarizes the main statistics of these datasets.

4.2. Experimental Setup

We conduct the experiments using the generated HIN representations for each ambiguous name. The input semantic features, averaged from the Word2Vec model, have a dimensionality of 100. The initial feature transformations map to a 128-dimensional hidden

[1]https://www.aminer.cn/open/article?id=5de9efd2530c707ed8b87d99

Table 1. Author Name Disambiguation Datasets. Paper per name and author per name columns refer to min/max/average, respectively. The total number of unique papers documents can be less than the sum shown in the table because a paper can be associated with multiple ambiguous authors.

	WhoIsWho-v1	WhoIsWho-v2
Papers	292,448	173,548
Names	319	165
Real Authors	26,093	10,479
Authors per Name	2/588/81.80	3/483/63.51
Papers per Name	33/5,682/916.76	99/6,142/1,051.81

space, which serves as the input to the GAT encoder. The GAT encoder employs a hidden dimension of 256. The projection head for both the global and refined views consists of a linear layer that maps the 64-dimensional embeddings output into a 64-dimensional projection space. The MLP for the adaptive weighting function has a hidden layer of 16 units.

For the training loop, we use AdamW to optimize the model parameters. The hyper-parameters were selected per dataset based on a grid search over the validation set, with the following ranges: (1) number of train epochs from 30 to 100 with step 10, (2) learning rate from 5e-05 to 5e-04 with step 5e-05, (3) temperature t from 0.1 to 1.0 with step 0.1, (4) hub-targeted edge dropout probability and (5) feature dropout probability from 0.1 to 0.5 with step 0.1. The selected hyperparameters for WhoIsWho-v1 were 50 epochs, a learning rate of 1e-04, a temperature t of 0.1, an edge dropout probability of 0.2, and a feature dropout probability of 0.1. For the WhoIsWho-v2, there were 70 epochs, a learning rate of 8e-05, a temperature t of 0.1, an edge dropout probability of 0.1, and a feature dropout probability of 0.2. The experiments were conducted using CUDA-enable acceleration on a 16GB GPU. All experiments were repeated 5 times, and the reported results correspond to the mean $\pm$ 95% confidence interval across runs.

4.3. Evaluation Metrics

The most widely adopted evaluation metrics for AND are pairwise Precision, Recall, and F1-Score. To provide a more robust assessment of clustering performance and account for varying cluster sizes, complementary metrics such as B-Cubed [Bagga and Baldwin 1998] and the Adjusted Rand Index (ARI) [Hubert and Arabie 1985] are also frequently reported.

Pairwise metrics evaluate clustering quality by examining the co-assignment of publication pairs, where Precision (pP) measures cluster purity and Recall (pR) measures completeness. In contrast, B-Cubed metrics evaluate performance at the individual publication level to prevent large clusters from disproportionately skewing the results, it calculate Precision (bP) and Recall (bR) for each publication and average them across the data. For both approaches, the F1-Score ($pF1$, $bF1$) provides the harmonic mean, offering a balanced measure of Precision and Recall.

The Adjusted Rand Index (ARI) measures the agreement between two data clusters, mathematically adjusted to account for chance groupings. It evaluates overall structural similarity by examining all pairs of publications, rewarding both true positives (pairs correctly clustered together) and true negatives (pairs correctly separated). The ARI score ranges from -1 to 1 where a score of 1 denotes perfect agreement, 0 indicates agree-

Table 2. AND-GloRe results. Average metrics across all ambiguous test names.

	WhoIsWho-v1	WhoIsWho-v2
pP	87.41 ± 1.87	91.64 ± 1.14
pR	91.21 ± 1.59	94.03 ± 1.76
pF1	88.10 ± 1.64	92.28 ± 1.31
bP	88.48 ± 1.39	91.45 ± 0.96
bR	91.59 ± 1.18	93.73 ± 1.30
bF1	89.47 ± 1.14	92.29 ± 0.95
ARI	84.83 ± 2.08	88.46 ± 1.83

ment expected by chance, and negative values imply agreements worse than by chance.

4.4. Baseline Methods

To evaluate the performance of our proposed method, we compare it against baseline models available in the literature in the Author Name Disambiguation domain, including: (1) **AGAND** [Hussain and Asghar 2017], (2) **AMiner** [Zhang et al. 2018], (3) **SDN-all** [Chen et al. 2023], (4) **EAND** [Zhou et al. 2024b], (5) **BOND** [Cheng et al. 2024] and (6) **MCCG** [Ye et al. 2025]. We specifically focus our comparative analysis on MCCG, which is the best performing baseline in terms of pairwise Precision, Recall, and F1-Score on both datasets when compared to the other baselines. Also, this model employs an unsupervised contrastive learning framework with multi-view learning, which is the most similar approach to our proposed method.

4.5. Results and Discussion

Table 2 shows the performance of AND-GloRe on the datasets based on the evaluation metrics described in Section 4.3. The method achieves pairwise F1-Scores of 88.10% and 92.28% on the WhoIsWho-v1 and WhoIsWho-v2 datasets, respectively, alongside B-Cubed F1-Scores of 89.47% and 92.29%. These pairwise metrics highlight a strong balance between cluster purity (high Precision) and the comprehensive grouping of relevant publications (high Recall). Furthermore, the B-Cubed results validate the model's robustness at the individual publication level, demonstrating its ability to maintain clustering quality across varying cluster sizes. Finally, ARI scores of 84.83% and 88.46% emphasize the structural alignment between our predictions and the ground truth, confirming AND-GloRe's capacity to accurately capture the underlying real author clusters.

We compare the performance of our method against the baselines available in the literature in terms of pairwise Precision, Recall and F1-Score. The results are summarized

Table 3. AND-GloRe results comparison with other baselines.

	WhoIsWho-v1			WhoIsWho-v2		
Model	pP	pR	pF1	pP	pR	pF1
AGAND	67.97	52.89	57.64	76.12	52.99	61.08
AMiner	81.01	60.78	69.45	84.33	61.97	71.44
SDN-all	76.69	81.96	76.57	84.90	87.38	85.25
EAND	78.12	77.98	78.05	79.98	85.61	82.70
BOND	83.75	89.67	84.77	90.22	93.19	90.89
MCCG	87.02	89.96	87.46	**92.60**	93.45	**92.57**
AND-GloRe	**87.41**	**91.21**	**88.10**	91.64	**94.03**	92.28

Table 4. AND-GloRe results comparison with a baseline for each ambiguous name in WhoIsWho-v1. Best results are highlighted in bold.

			AND-GloRe			MCCG		
Name	Papers	Authors	pP	pR	pF1	pP	pR	pF1
biao_huang	784	14	83.81	**99.78**	**91.10**	**83.82**	99.56	91.02
bin_zhou	1529	95	**80.14**	92.54	**85.85**	67.61	**97.85**	79.97
bing_xu	1236	68	**76.04**	92.58	83.39	74.21	**96.31**	**83.83**
fei_chen	1412	79	**92.74**	96.26	**94.45**	90.06	**97.28**	93.53
feng_wang	741	19	79.65	82.70	80.38	**90.40**	**88.92**	**89.65**
gang_li	3228	157	**74.07**	84.92	**79.09**	63.19	**85.63**	72.72
hongjie_zhang	626	12	99.39	**67.37**	**80.31**	**99.46**	64.94	78.58
jian_yang	2863	120	**81.89**	87.16	**84.40**	68.92	**92.71**	79.06
jie_gao	1254	102	**70.46**	90.80	**79.33**	67.28	**95.07**	78.80
jie_tang	1198	43	92.48	**92.85**	**92.62**	**92.93**	90.87	91.89
jie_wu	2590	84	**96.85**	**95.08**	**95.94**	88.23	90.11	89.16
rui_wang	2643	143	**76.46**	78.04	**77.22**	60.15	**81.51**	69.22
tao_li	2161	138	**52.86**	79.54	63.30	52.82	**84.43**	**64.98**
weimin_liu	1434	30	97.33	**99.38**	98.34	**99.33**	97.54	**98.43**
yao_wang	1406	77	**80.24**	60.63	69.04	67.21	**72.83**	**69.91**

in Table 3. AND-GloRe outperforms all models in terms of pairwise Precision, Recall and F1-Score for the WhoIsWho-v1 dataset, surpassing the best baseline by 0.64% in pairwise F1-Score. For the WhoIsWho-v2 dataset, our method achieves a competitive pairwise F1-Score of 92.28%, which is slightly lower than the best baseline, MCCG, by 0.29%. In Author Name Disambiguation, achieving high Precision is crucial to prevent publications from different authors from being incorrectly grouped together. In this regard, we can observe that both MCCG and AND-GloRe improve Precision while maintaining a competitive Recall.

Looking at the performance on individual ambiguous names, we may have insights into how the model performs across different levels of ambiguity. Table 4 shows the results for fifteen highly ambiguous names in the WhoIsWho-v1 dataset. Also, we compare side-by-side the performance of our method with the MCCG model, which is the best performing baseline in terms of pairwise F1-Score. As observed, our method consistently outperforms MCCG on most ambiguous names in terms of pairwise Precision and F1-Score. As mentioned before, achieving a high Precision is crucial in AND when maintaining a relatively high Recall. This is because a model with high Precision ensures that publications from different authors are not incorrectly grouped together, which, in real-world scenarios, is more desirable than a model that clusters most publications together while producing many false positives.

We can also observe how our model outperforms the MCCG for large networks. This is the case for the ambiguous names "Gang Li" (3,228 publications and 157 real authors) and "Jie Wu" (2,590 publications and 84 real authors). For both those names, our method achieves the highest pairwise F1-Score, with a significant improvement of 6.37% and 6.78%, respectively. This suggests that our method can capture complex relationships by leveraging both global and refined views of an HIN representation. In contrast, the MCCG model, which also uses a contrastive learning framework with multi-view learning, does not consider the specific types of relationships in the name network when creating a homogeneous graph for contrastive learning.

However, there are cases where our method performs worse on the disambiguation

task. For instance, for the ambiguous name "Feng Wang", our method decreases pairwise F1-Score by around 9.27% compared with MCCG. In the WhoIsWho-v2, our method struggles with the ambiguous name "Weihua Wang", achieving a pairwise F1-Score of 84.43%, which is dramatically lower than the 93.43% achieved by MCCG. Analyzing ambiguous name networks and publication feature spaces in this and similar cases reveals an architectural vulnerability in AND-GloRe. Because the model relies heavily on raw node features to initialize both global and refined embeddings, initial noise propagates directly through the target-neighbor aggregation step. Inheriting this baseline ambiguity before any contrastive optimization occurs, the GAT encoder computes flawed structural coefficients, and the semantic attention module struggles to effectively weigh meta-path views, ultimately trapping the contrastive loss in a suboptimal local minimum.

5. Conclusion

We presented AND-GloRe, an unsupervised Author Name Disambiguation method that leverages Heterogeneous Information Networks and a dual-view contrastive learning strategy to learn robust publication representations for disambiguation. By combining a global coarse view with a refined meta-path-aware view, encoding both with a shared GAT encoder, and aligning them with an adaptively weighted contrastive objective, the method effectively mitigates noisy links and captures fine-grained relational semantics in ambiguous name graphs. Experiments on standard AND datasets show that AND-GloRe produces strong disambiguation results, achieving a pairwise F1 of 88.10% and 92.28% on the WhoIsWho-v1 and WhoIsWho-v2 datasets, achieving competitive performance when compared with the baselines.

Future work will focus on mitigating the impact of noisy features by addressing the model initialization vulnerability. We plan to experiment with our model in additional datasets to evaluate its performance on different domains. Furthermore, we intend to conduct comprehensive ablation studies to thoroughly evaluate the impact of varying the different components of our proposed method. Finally, we plan to explore the integration of additional information into the HIN representation, such as temporal information by inserting the publication year as a weighted feature, authors' positions in the author list of collaborations, as well as external information, including the geographical location of affiliations, serving as additional components of the HIN representation.

Acknowledgments

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001, the Universidade Federal de Ouro Preto (UFOP), CNPQ, FAPEMIG, and INCT-TILDIAR (grant # 408490/2024-1).

Statement on the use of Artificial Intelligence

The authors declare the use of Artificial Intelligence (AI) tools in the preparation of this work. Specifically, Gemini and Grammarly were used to assist with spelling correction, text condensation, and the improvement of textual consistency.

References

Amado Olivo, V., Kerzendorf, W., Lu, B., Shields, J. V., Flörs, A., and Chen, N. (2025). Practical author name disambiguation under metadata constraints: A contrastive learning approach for astronomy literature. *PASP*, 137(12):124503.

Bagga, A. and Baldwin, B. (1998). Entity-based cross-document coreferencing using the vector space model. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 79–85. Association for Computational Linguistics.

Cappelli, F., Colavizza, G., and Peroni, S. (2025). Deep learning approaches to author name disambiguation: A survey. *IJDL*, 26(4):21.

Chen, B., Zhang, J., Zhang, F., Han, T., Cheng, Y., Li, X., Dong, Y., and Tang, J. (2023). Web-scale academic name disambiguation: The whoiswho benchmark, leaderboard, and toolkit. In *KDD '23*, pages 3817–3828. ACM.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In *37th International Conference on Machine Learning*, volume 119, pages 1597–1607. PMLR.

Cheng, Y., Chen, B., Zhang, F., and Tang, J. (2024). BOND: Bootstrapping from-scratch name disambiguation with multi-task promoting. In *ACM Web Conference 2024*, pages 4216–4226. ACM.

Cota, R. G., Gonçalves, M. A., and Laender, A. H. F. (2007). A heuristic-based hierarchical clustering method for author name disambiguation in digital libraries. In *Anais do XXII SBBD*, pages 20–34. SBC.

De Bonis, M., Falchi, F., and Manghi, P. (2023). Graph-based methods for author name disambiguation: a survey. *PeerJ Computer Science*, 9:e1536.

Elliott, S. (2010). Survey of author name disambiguation: 2004 to 2010. *LPP*, 443.

Ferreira, A. A., Gonçalves, M. A., and Laender, A. H. (2012). A brief survey of automatic methods for author name disambiguation. *SIGMOD Rec.*, 41(2):15–26.

Ferreira, A. A., Gonçalves, M. A., and Laender, A. H. F. (2020). *Automatic Disambiguation of Author Names in Bibliographic Repositories*. Springer International Publishing.

Han, H., Giles, L., Zha, H., Li, C., and Tsioutsiouliklis, K. (2004). Two supervised learning approaches for name disambiguation in author citations. In *JCDL '04*, page 296–305. ACM.

Huang, J., Ertekin, S., and Giles, C. L. (2006). Efficient name disambiguation for large-scale databases. In *Knowledge Discovery in Databases: PKDD 2006*, pages 536–544. Springer Berlin Heidelberg.

Huang, Z., Zhang, H., Hao, C., Yang, H., and Wu, H. (2024). A cross-domain transfer learning model for author name disambiguation on heterogeneous graph with pre-trained language model. *Knowledge-Based Systems*, 305:112624.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of Classification*, 2(1):193–218.

Hussain, I. and Asghar, S. (2017). A survey of author name disambiguation techniques: 2010-2016. *The Knowledge Engineering Review*, 32:e22.

Levin, F. H. and Heuser, C. A. (2009). Evaluating the use of social networks in author name disambiguation in digital libraries. In *Anais do XXIV SBBD*, pages 46–60. SBC.

Liu, D., Zhang, R., Chen, J., and Chen, X. (2024). Author name disambiguation via paper association refinement and compositional contrastive embedding. In *WWW '24*, page 2193–2203. ACM.

Ma, X., Wang, R., and Zhang, Y. (2019). Author name disambiguation in heterogeneous academic networks. In *WISA*, pages 126–137. Springer International Publishing.

Mikolov, T., Chen, K., Corrado, G. S., and Dean, J. (2013). Efficient estimation of word representations in vector space. In *ICLR (Workshop Poster) 2013*.

Pooja, K., Mondal, S., and Chandra, J. (2022). Exploiting higher order multi-dimensional relationships with self-attention for author name disambiguation. *TKDD*, 16(5).

Rehs, A. (2021). A supervised machine learning approach to author disambiguation in the web of science. *Journal of Informetrics*, 15(3):101166.

Rodrigues, N. S., Mariano, A. M., and Ralha, C. G. (2024). Author name disambiguation literature review with consolidated meta-analytic approach. *IJDL*, 25(4):765–785.

Wang, C., Pan, S., Hu, R., Long, G., Jiang, J., and Zhang, C. (2019). Attributed graph clustering: a deep attentional embedding approach. In *IJCAI-19*, page 3670–3676.

Wang, H., Wan, R., Wen, C., Li, S., Jia, Y., Zhang, W., and Wang, X. (2020). Author name disambiguation on heterogeneous information network with adversarial representation learning. In *AAAI-20*, volume 34, pages 238–245.

Xie, W., Liu, S., Wang, X., and Jia, T. (2022). Author name disambiguation via heterogeneous network embedding from structural and semantic perspectives. In *ICTAI 2022*, pages 245–250.

Ye, F., Xia, Z., Ling, Z., and Wu, L. (2025). Multi-view contrastive and cluster-guided learning for author name disambiguation. *Expert Systems with Applications*, 289:128324.

Yu, J., Ge, Q., Li, X., and Zhou, A. (2024). Heterogeneous graph contrastive learning with meta-path contexts and adaptively weighted negative samples. *IEEE Transactions on Knowledge and Data Engineering*, 36(10):5181–5193.

Zhang, Y., Zhang, F., Yao, P., and Tang, J. (2018). Name disambiguation in AMiner: Clustering, maintenance, and human in the loop. In *KDD '18*, pages 1002–1011. ACM.

Zheng, X., Zhang, P., Cui, Y., Du, R., and Zhang, Y. (2021). Dual-channel heterogeneous graph network for author name disambiguation. *Information*, 12(383).

Zhou, A., Shi, M., and Yuan, R. (2024a). An effective author name disambiguation framework for large-scale publications. *IEEE Access*, 12:182086–182100.

Zhou, Q., Chen, W., Zhao, P.-P., Liu, A., Xu, J.-J., Qu, J.-F., and Zhao, L. (2024b). Towards effective author name disambiguation by hybrid attention. *JCST*, 39(4):929–950.