Received: 14 September 2021 | Revised: 28 June 2022 | Accepted: 3 October 2022

DOI: 10.1002/asi.24720

RESEARCH ARTICLE

JASIST WILEY

LAGOS-AND: A large gold standard dataset for scholarly author name disambiguation

Li Zhang | Wei Lu | Jinqing Yang

School of Information Management, Wuhan University, Wuhan, China

Correspondence
Wei Lu, School of Information Management, Wuhan University, 299, Bayi Street, Wuchang District, Wuhan, China.
Email: weilu@whu.edu.cn

Funding information
National Key Research and Development Program of China, Grant/Award Number: 2019YFB1404702

Abstract

In this article, we present a method to automatically build large labeled datasets for the author ambiguity problem in the academic world by leveraging the authoritative academic resources, ORCID and DOI. Using the method, we built LAGOS-AND, two **la**rge, **go**ld-**s**tandard sub-datasets for author name disambiguation (**AND**), of which LAGOS-AND-BLOCK is created for clustering-based AND research and LAGOS-AND-PAIRWISE is created for classification-based AND research. Our LAGOS-AND datasets are substantially different from the existing ones. The initial versions of the datasets (v1.0, released in February 2021) include 7.5 M citations authored by 798 K unique authors (LAGOS-AND-BLOCK) and close to 1 M instances (LAGOS-AND-PAIRWISE). And both datasets show close similarities to the whole Microsoft Academic Graph (MAG) across validations of six facets. In building the datasets, we reveal the variation degrees of last names in three literature databases, PubMed, MAG, and Semantic Scholar, by comparing author names hosted to the authors' official last names shown on the ORCID pages. Furthermore, we evaluate several baseline disambiguation methods as well as the MAG's author IDs system on our datasets, and the evaluation helps identify several interesting findings. We hope the datasets and findings will bring new insights for future studies. The code and datasets are publicly available.

1 | INTRODUCTION

Author name ambiguity is a well-known issue in academic literature databases/digital libraries. The name ambiguities in the real world reflect authors represented by name variants (synonyms), and some authors share the same name (homonyms) (Aman, 2018; Kim & Kim, 2020; Shoaib et al., 2020). This problem is very challenging in literature databases because, for example, there were about 40 K citations[1] authored by "Wei Wang" in Microsoft Academic Graph (MAG) as of March 2019, and the name ambiguity problem is even more pronounced for the abbreviated names. Although some databases such as MAG and AMiner[2] have provided disambiguated author identifiers (IDs), the performance of the created author ID systems based on author name disambiguation (AND) approaches for million-scale databases is far from satisfactory (Zhang et al., 2020).

Identifying author uniqueness is crucial for many studies and applications. For instance, in the field of bibliometric research, a recent high-impact study used disambiguated author IDs to meet a larger goal of examining gender inequality in scientific careers (Huang et al., 2020). In digital library management research, Zhang et al. (2018) claimed that AND is a core component of AMiner, which is a free online service for academics.

To address the name ambiguity problem, the research community has developed many labeled AND datasets in recent years (see Supplemental material A for the list of the AND datasets) to help develop supervised or semi-

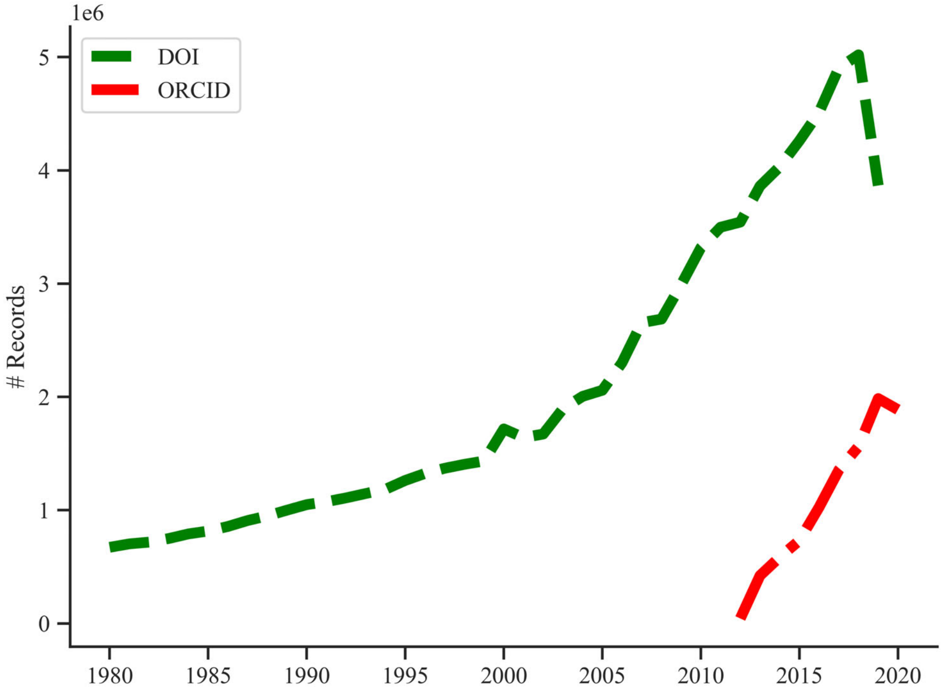

FIGURE 1 The number of ORCID and DOI records each year, obtained from database dumps of ORCID and OpenAIRE's DOIBoost.

supervised disambiguation methods (Louppe et al., 2016; Mihaljevic & Santamaría, 2021), as well as test the performance of various disambiguation methods (Kim & Kim, 2020; Tekles & Bornmann, 2020). However, we find that existing datasets suffer from several issues or limitations. To be specific, the issues or limitations are as follows. (1) Unclear dataset creation process. Most datasets such as GS-PubMed (Vishnyakova et al., 2019) and SCAD-zbMATH (Müller et al., 2017) are created manually, meaning that many annotators are involved and a great deal of effort has to be invested. In addition, for some datasets such as Han-DBLP (Han, Xu, et al., 2005), the details of creation are not thoroughly described, for example, the quality assurance measures, which may raise concerns about the quality of the datasets. (2) Limited scale. Existing datasets are mostly limited in size (see the Supplemental material A); however, in literature databases, the name ambiguity problem is generally more complex than that in small datasets (Xiao et al., 2020). This problem may make the data-driven disambiguation methods perform poorly in real literature databases. (3) Limited number of unbiased datasets. Existing datasets are unable to reach the level of the gold standards. One example that indicates the biases is that most datasets are designed to address the name homonym problem (Sanyal et al., 2021); however, the name ambiguities include synonyms as well as homonyms.

All of these issues and limitations not only bias the performance of disambiguated author ID systems but, more importantly, hinder the development of effective AND methods. Motivated by the credibility and increasing popularity of the two academic resources Open Researcher Contributor Identification[3] (ORCID) and Digital Object Identifier[4] (DOI), we herein propose a method to automatically build improved datasets for AND. Our proposed method and the created datasets can address the above issues or limitations appropriately. Specifically, (1) Because ORCID iDs and DOIs are able to identify authors and scientific papers unambiguously, the publication history of an author (query DOIs by ORCID iD) and the authorship of a paper (query ORCID iDs by DOI) can be accurately identified. Thus, relying on this information it is feasible to develop an automatic method that can effortlessly build AND datasets. In addition, as the rationale of the developed method is simple and clear, it is also feasible to regenerate the datasets and create a new version of the datasets. (2) The two academic resources have gained increasing popularity in recent years. As shown in Figure 1, the two resources have been growing at a high and constant speed recently, and the number of ORCID records and DOI records in 2018 was 1,561,789 and 5,020,071, respectively. Therefore, such a large amount of labeled data made it possible to build a large AND dataset. (3) The proposed method is manageable and controllable due to its simplicity, meaning that, by adjusting the created datasets and comparing them with a real literature database, we can improve the quality of the datasets in several aspects, such as the covered ambiguity patterns.

In summary, our contributions are as follows:

- We develop a method that can automatically build large labeled datasets for the author name disambiguation research. The method is clearly presented and can be reused to generate new versions of the datasets.

- We have built LAGOS-AND based on the proposed method, which contains two **la**rge, **go**ld-**s**tandard sub-datasets for **AND**. To the best of our knowledge, our datasets are the world's largest AND datasets. The technical validation demonstrates that the two datasets show close similarities to the whole MAG across validations of six facets. Our datasets are available at https://zenodo.org/record/7313380.
- We calculate the degree of last name variation in building the datasets. Evaluation results for three large literature databases show that the degrees range from 5.80% to 6.34%, and the variation degrees are even higher if (1) a popular name-parsing tool is used to extract the last names from full names for name comparison or (2) the accented alphabets are not transliterated to the standard characters (e.g., "á" → "a").
- We evaluate several baseline disambiguation methods and the author ID system of MAG on our datasets. The experimental results indicate that MAG's author IDs show poor performance on the two gold standards and that incorporating a semantic relatedness feature of citations boosts the performance of disambiguation. The code is available at https://github.com/carmanzhang/LAGOS-AND.

2 | RELATED WORKS

In this section, we review the most important datasets created for AND research.[5] According to our survey, there are at least 12 datasets available so far. These datasets have been widely adopted to develop disambiguation methods in various scenarios or for different objectives. However, we have identified a number of unresolved and even undiscovered issues with the datasets.

First, most of the existing datasets were created manually (Han, Zha, & Giles, 2005; Vishnyakova et al., 2019; Xiao et al., 2020), which means that a great deal of effort needed to be invested in the data creation process. For example, a large group consisting of 22 annotators was involved in creating a dataset for AMiner (Wang et al., 2011), and each publication was annotated by at least three annotators to ensure a high annotation accuracy. Some datasets were created in a crowdsourcing fashion on platforms such as Amazon Mechanical Turk (MTurk). The method of building labeled datasets is usually considered effective (Zhang et al., 2016); however, the method appeared to be ineffective when it was used to create AND datasets, as a recent study (Vishnyakova et al., 2019) found that it was hard to control the data quality because the annotation tasks were distributed to many untrusted annotators who may try to guess the class labels rather than find the ground truths.

Second, current datasets are either limited in size or limited in scope. As shown in Supplemental material A, most datasets contain fewer than 10,000 citations. However, small datasets such as Han-DBLP (Han, Xu, et al., 2005; Han, Zha, & Giles, 2005), Qian-DBLP (Qian et al., 2015), Kim-DBLP (Kim, 2018), Tang-AMiner (Tang et al., 2012; Wang et al., 2011), Culotta-REXA (Culotta et al., 2007), Cota-BDBComp (Cota et al., 2010), Song-PubMed (Song et al., 2015), and GS-PubMed (Vishnyakova et al., 2019) may be unable to adequately reflect the real complexity of name ambiguities, as a recent study (Xiao et al., 2020) pointed out that the patterns of name ambiguities in large literature databases exceed those represented in a small dataset. Due to the limited name patterns, a small dataset will restrict the exploration of some data-driven techniques. Note that, although some datasets such as GESIS-DBLP,[6] SCAD-zbMATH (Müller et al., 2017), and Kim-PubMed (Kim & Owen-Smith, 2021) have decent numbers of instances, they are limited in scopes (covered domains). For example, SCAD-zbMATH is designed specifically for a mathematical domain database, zbMATH.[7] Such domain-specified datasets pose a frequently encountered problem in machine learning (ML): a model trained for a domain may be unsuitable for application to another domain because different domains cover different scopes of knowledge.

Among all the datasets, WhoisWho (Xiao et al., 2020) is the one that not only has a decent size but also covers a wide domain (large scope). However, a prominent drawback with the dataset is that it shows clear biases with respect to author ethnicity and name variation, which we refer to as the third limitation. Specifically, in WhoisWho, most last names (53 out of 65) are Chinese last names, which is inconsistent with the fact that the authors are from all over the world. Note that this issue is nontrivial because many studies have confirmed that different ethnicities have different levels of ambiguities (Louppe et al., 2016), and, based on this idea, some ethnicity-based disambiguation methods have been successfully developed (Kim et al., 2021; Louppe et al., 2016; Subramanian et al., 2021). Name variation is another frequently ignored aspect in building AND datasets. Author names presented in literature databases may differ from their actual names for many reasons (Gomide et al., 2017). This issue is known to AND researchers, but remains unsolved because addressing this issue is still very challenging; some studies have pointed out this issue in their research limitations (Zhang et al., 2021) or mentioned it in relation to future works (Sanyal et al., 2021). Unfortunately, existing datasets, including

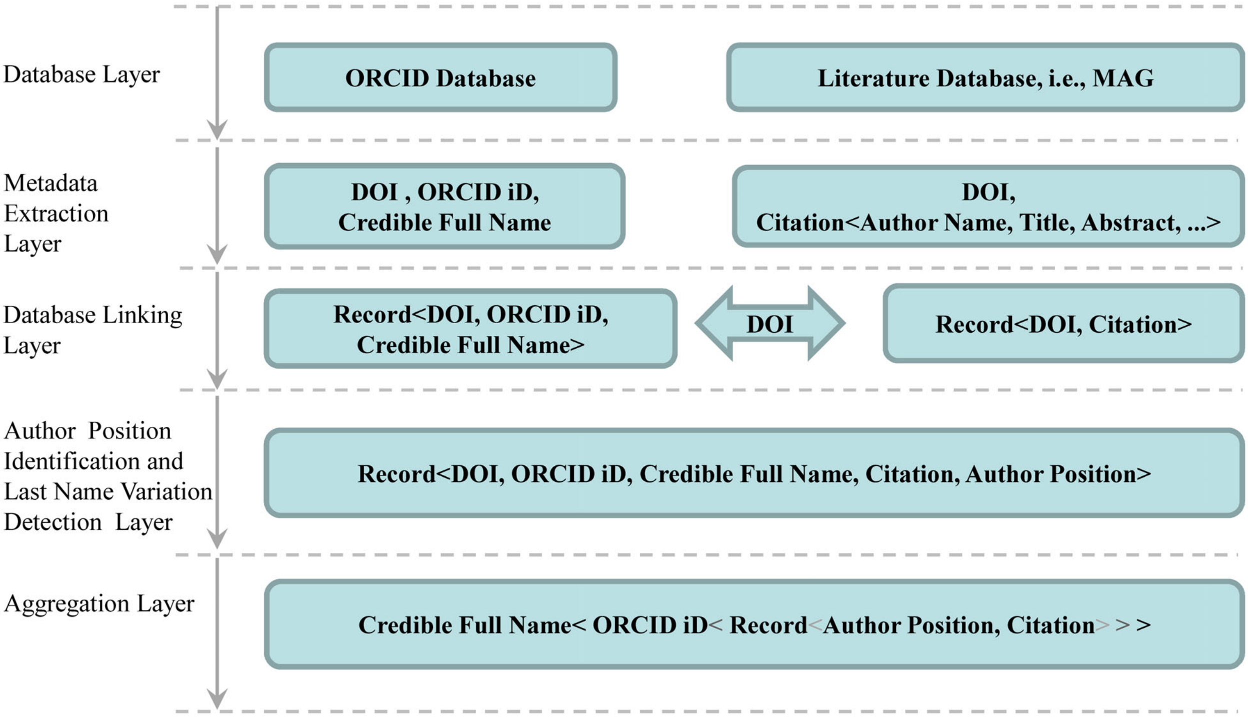
Database Layer
ORCID Database
Literature Database, i.e., MAG
Metadata Extraction Layer
DOI , ORCID iD, Credible Full Name
DOI, Citation<Author Name, Title, Abstract, ...>
Database Linking Layer
Record<DOI, ORCID iD, Credible Full Name>
DOI
Record<DOI, Citation>
Author Position Identification and Last Name Variation Detection Layer
Record<DOI, ORCID iD, Credible Full Name, Citation, Author Position>
Aggregation Layer
Credible Full Name< ORCID iD< Record<Author Position, Citation> > >

FIGURE 2 Dataset building pipeline

WhoisWho, can scarcely touch this issue or represent this aspect adequately. For WhoisWho, the variation degree of the last names is 0.37%,[8] which we believe is lower than that of real literature databases (see the RESULTS section for a formal investigation of the problem).

In response to these limitations, we propose a practical method to automatically build labeled datasets for author name disambiguation by leveraging the two large authoritative resources, ORCID and DOI. ORCID is dedicated to reducing the risk of errors in professional-related resources by providing a persistent identifier (ORCID iD) that authors can control and manage.[9] DOI is a persistent interoperable identifier for digital objects and is developed for use on digital networks.[10] The two academic resources have hosted a considerable number of records, which provide valuable labeled information as well as sufficient research materials to construct a large labeled AND dataset that integrates more complicated name patterns. Such a high-quality dataset has the potential to meet the requirement of building more effective disambiguation methods.

3 | DATASET BUILDING APPROACH

3.1 | Method overview

We developed an automatic method to build our AND datasets based on the ORCID database and a literature database, MAG.[11] The reasons for choosing MAG are as follows. First, MAG has a high demand for disambiguated authors because MAG has been used in many applications. Some well-known examples are Bing, Cortana, Microsoft Word, and Microsoft Academic.[12] Second, MAG is a heterogeneous graph containing a variety of publication-relevant metadata such as citation networks, and institutions, which may be useful for the development of disambiguation methods.

The main steps of our method are shown in Figure 2. In the database layer, we retrieved the ORCID data (baseline version of October 2020) and the MAG data (baseline version of March 2019) from the respective repositories.[13,14] Then, in the second layer, we extracted those metadata that are only related to the final AND datasets from the two databases in order to reduce the storage and computational overhead. For the literature database, we extracted DOI and other citation-related metadata required by disambiguation approaches such as article title and venue. For the ORCID data, we extracted author-related metadata such as ORCID iD and the authors' credible full names (CFNs) shown on the ORCID pages, as well as the DOIs of citations claimed by the authors. In the database linking layer, we employed the DOIs to connect the two databases. As DOI is an ambiguity-free indicator for digital objects, it is, therefore, safe to connect the databases. However, the ORCID system does not specify the positions of a user (author) in the claimed citations, meaning that the author-level metadata such as affiliation that are frequently used in prior AND studies cannot be obtained from the linked MAG citations if the positions are not identified. To address this issue, we designed an algorithm to identify the author positions, which corresponds to the fourth layer. In addition, in this layer, we also investigated the name variation problem. In the last layer, we conducted several aggregation operations on the ORCID-MAG linked data to build our AND datasets. The following subsections elaborate on the key steps of the method.

3.2 | Author position identification

The heuristic algorithm for identifying author positions is shown in Algorithm 1. For the two inputs, a CFN and the n author names $[FN_1, FN_i, ..., FN_n]$ of a citation, the algorithm firstly maps the CFN and a FN_i to the character-level 2-grams features $CFN^{(f)}$ and $FN_i^{(f)}$, respectively. For example, "John Smith" is represented by the 2-grams list of [*jo*, *oh*, *hn*, ..., *th*]. The reason for using the 2-grams measure is that it not only considers the order of characters but also is insensitive to name variants (e.g., reversed author names). Thus, the measure can be used to detect name variants. Then, the algorithm calculates the similarity S_i of the CFN and FN_i by measuring the number of intersections between $CFN^{(f)}$ and $FN_i^{(f)}$ to the length of the concatenated 2-grams lists. Afterward, for the citation with n authors, the algorithm sorts the original author positions $I = [1, ..., n]$ by the corresponding similarity scores $S = [S_1, ..., S_n]$ in descending order, and the first element $P1 = I_1^{(r)}$ of the ranked author positions $I^{(r)}$ is likely to be the correct position. Note that, in some citations where more than one FN appears to be similar to a CFN, the algorithm may tend to incorrectly identify the author positions. For example, the two similar names "M.C. Ciornei" and "F.C. Ciornei" in the MAG citation (article ID: 2742497971)[15] have the same similarity score of 0.44 as compared to the CFN "Florina Carmen Ciornei." To handle this issue, the algorithm tries to exclude such citations by applying the rule: $P1 = I_1^{(r)}$ can be considered as the final author position only if S_{p1} is higher than the second-best similarity score S_{p2} ($P2 = I_2^{(r)}$, if available) by a threshold. In this study, the threshold is empirically determined to be 0.2.

Algorithm 1 **A heuristic algorithm for author position identification**

```
Input: ORCID credible full name, CFN
            Author names list of a MAG citation, $FNs = (FN_1, ..., FN_n)$
    Output: Identified author position, P
    $CFN^{(f)} = 2\text{-grams}(CFN)$
    $I = ()$
    $S = ()$
    For each $FN_i \in FNs$ Do
            $FN_i^{(f)} = 2\text{-grams}(FN_i)$
            $s_i = 2 * intersection(CFN^{(f)}, FN_i^{(f)}) / (|CFN^{(f)}| + |FN_i^{(f)}|)$
            $I \leftarrow i$
            $S \leftarrow s_i$
    End
    // sorts the original author positions I by the similarity scores S in descending order,
    // and return the ranked author positions $I^{(r)}$.
    $I^{(r)} = sort\_index(I, S)$
    $P1 = I_1^{(r)}$
    $P2 = I_2^{(r)}$
    If $S_{P1} - S_{P2} > 0.2$ Then
            $P = P1$
    Else Then
            // 0 means that the author's position could not be identified.
            $P = 0$
    End
    return P
```

3.3 | Last name variation detection

Through author position identification, we obtained a large number of CFN-FN matches, which were used to detect name variants. Among all kinds of name variations, the last name variation is the most influential one because it is used to create the last name (LN)-based or last name and first initial (LNFI)-based blocks, which is a widely used disambiguation framework in AND studies (Levin et al., 2012; Louppe et al., 2016; Schulz, 2016) and even in production environments (Kim et al., 2016; Torvik & Smalheiser, 2009). By assuming that author names are consistent in all the authored publications, the ambiguous authors are grouped into a particular block, and they are only compared within the block. Therefore, the framework can reduce the computational complexity. However, this assumption is idealized because there are many kinds of name discrepancies resulting from various reasons (see Supplemental material B). Based on the analysis, the name variation problem will eventually result in a performance reduction of AND methods as the citations of the same author may be divided into different blocks.

In view of this, we herein show how to detect the last name variants and measure the degree of the variation. Specifically, we compared the last names recorded in literature databases to the authors' official last names shown on the ORCID pages. It should be noted that many literature databases such as MAG do not provide the last name field,[16] making the name comparison unfeasible. We developed three measures to address this issue. The first one is "Endwith," representing whether an FN *ends with* the credible last name. The second measure extracts the last names from FNs using Joshfraser,[17] which is a popular name parser working with complex, language-independent names. The criterion for determining a name variant is whether the extracted last name is identical to the ORCID last name. The third measure follows the same criterion but adopts another name parser, Derek73.[18] This tool has attracted many developers to continuously improve for it over 10 years and had been used by over 700 applications as of May 2022.

3.4 | Block-based AND dataset building

We built our block-based AND dataset (LAGOS-AND-BLOCK) with Algorithm 2. Based on the connected database DB_{ocbib} between the ORCID database DB_{oc} and the literature database DB_{bib} (i.e., MAG in this study), we aggregated the connected citations at the author level and then at the block level to build the dataset.

At the author level, we aggregated those citations belonging to the same author into a citation group (CG) by ORCID iD. This exercise aims to restore the publication history of authors unambiguously.[19,20] At the block level, we further aggregated CGs into blocks by CFNs so that a specific block could contain multiple CGs. It is important to note that, instead of the commonly used LNFIs or FNs, we used CFNs to group the CGs because the method of building the block-based dataset has the following advantages. First, CFN is more authoritative in terms of representing blocks than LNFI or FN as the CFNs are maintained by the authors and are displayed directly on the ORCID pages without changes. In contrast, the author names presented in literature databases are error-prone. Second, it is more meaningful to disambiguate on a full-name-based dataset. In LN-based or LNFI-based datasets, authors who are apparently different persons may exist in a block. For example, the two different authors with the different names "Richard Freyman" and "Robin Freyman" can exist in the LNFI block "Freyman_R." In comparison, the blocks of our dataset are represented by full names, meaning that all ambiguous authors included in a particular block have the same name. This design makes our dataset more meaningful to disambiguate. Third, as shown in Figure 3, the created dataset considers synonymous names and homonymous names simultaneously. On the one hand, the author names shown in a CG may be different from a CFN, and therefore they constitute the synonymous names. On the other hand, a block usually consists of multiple CGs;

Algorithm 2 **An algorithm for automatically building LAGOS-AND-BLOCK**

```
Input: ORCID database, $DB_{OC}$
          A literature database, $DB_{bib}$
    Output: LAGOS-AND-BLOCK
    // extract required metadata from databases, "rec" stands for a record in a database.
    $DBMD_{oc} = \{<rec.DOI, rec.ORCID, rec.CFN> \mid rec \in DB_{oc}\}$
    $DBMD_{bib} = \{<rec.DOI, rec.Citation> \mid rec \in DB_{bib}\}$
    // linking databases
    $DB_{ocbib} = \{<rec_{oc}, rec_{bib}> \mid rec_{oc}^{DOI} \equiv rec_{bib}^{DOI}, rec_{oc} \in DBMD_{oc}, rec_{bib} \in DBMD_{bib}\}$
    $LAGOS\text{-}AND\text{-}BLOCK = \emptyset$
          For each <DOI, ORCID, CFN, Citation> $\in DB_{ocbib}$ Do
          // identify the author position using Algorithm 1.
          $P = AuthorPositionIdentification(CFN, Citation.AuthorNameList)$
          $Item = (DOI, ORCID, CFN, Citation, P)$
          // find the block where this Item should be merged into
          $BLK = \{blk \mid blk.CFN \equiv CFN, blk \in LAGOS\text{-}AND\text{-}BLOCK\}$
          // find the citation group where this Item should be merged into
          $CG = \{cg \mid cg.ORCID \equiv ORCID, cg \in BLK\}$
          // update this citation group
          $CG = CG \cup \{Item\}$
          // update this block
          $BLK = BLK \cup \{CG\}$
          // update the dataset
          $LAGOS\text{-}AND\text{-}BLOCK = LAGOS\text{-}AND\text{-}BLOCK \cup \{BLK\}$
    End
    return LAGOS-AND-BLOCK
```

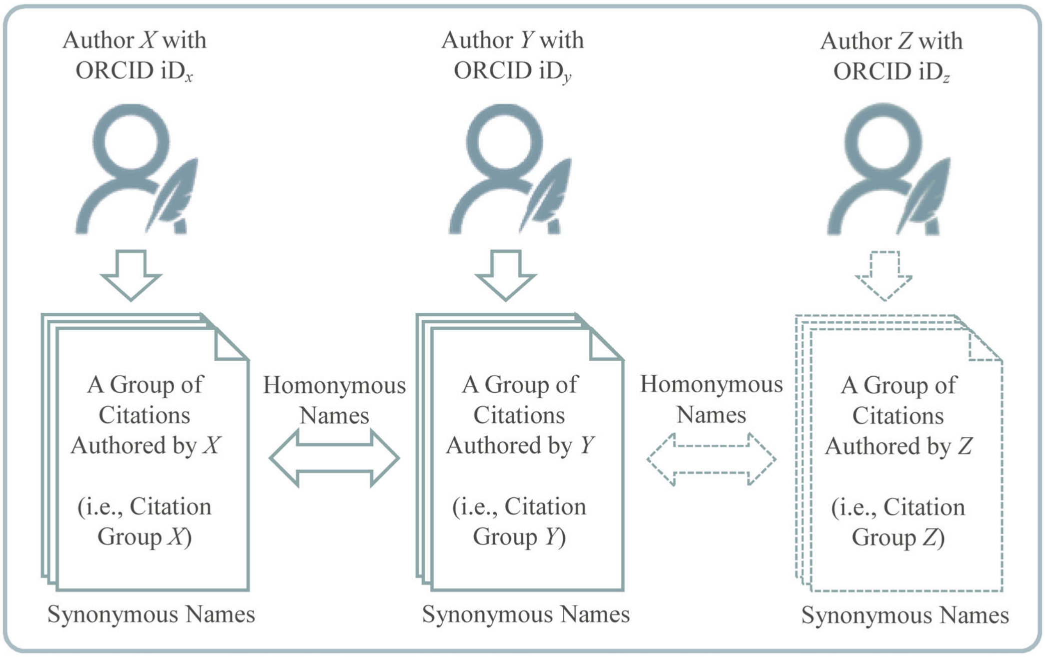
Author X with ORCID iD$_x$
Author Y with ORCID iD$_y$
Author Z with ORCID iD$_z$
A Group of Citations Authored by X (i.e., Citation Group X)
Homonymous Names
A Group of Citations Authored by Y (i.e., Citation Group Y)
Homonymous Names
A Group of Citations Authored by Z (i.e., Citation Group Z)
Synonymous Names
Synonymous Names
Synonymous Names

FIGURE 3 The structure of the block-based dataset

thus, the names across different CGs but within the same block constitute the homonymous names.

3.5 | Pairwise-based AND dataset building

Existing AND datasets are either block-based or pairwise-based, both of which are important because they play different roles in AND research: the block-based datasets were created for clustering-based disambiguation approaches, and the pairwise-based datasets were created for classification-based disambiguation approaches. In this study, we also considered the classification-based evaluation scenario, we created a paired-citation-based AND dataset (LAGOS-AND-PAIRWISE) based on our LAGOS-AND-BLOCK dataset following the idea of randomly sampling paired citations over blocks (Song et al., 2015; Zhang et al., 2021). Similar to LAGOS-AND-BLOCK, LAGOS-AND-PAIRWISE is also a large gold-standard dataset, which is demonstrated in the Result section.

4 | RESULTS

4.1 | Author position identification

We test the performance of author position identification on the three databases: PubMed, MAG, and Semantic Scholar (S2).[21] For each database, we randomly selected 2,000 matched-name instances (FNs-CFNs) and manually examined the identified author positions. As shown in Supplemental material C, the accuracies on PubMed, MAG, and S2 are 100%, 99.95%, and 99.80%, respectively, demonstrating the reliability of the method.

4.2 | Last name variation

Based on a considerable number of FN-CFN matches, we calculated the degree of variation in last names with the three mentioned measures. It should be pointed out that name variation is more pronounced for languages using some non-Western characters (e.g., "á") (Müller et al., 2017). To reflect this problem comprehensively, we reported both the character-sensitive variation degree (CSVD) and the character-insensitive variation degree (CIVD), the latter was achieved by transliterating the special characters into the standard characters (e.g., "á" → "a").

The results are presented in Table 1. We observed that the results yielded by Joshfraser and Derek73 are very similar and both are higher than the degree achieved by Endwith on MAG and S2. Although the tools may introduce parsing errors, they are indeed useful for developing AND methods because most methods rely on an explicit first name or last name field for feature computation (Wu et al., 2017) and name instance blocking (Kim, 2018; Kim et al., 2016). However, many databases such as MAG do not provide such fields. In addition, Endwith yields a CSVD of ~9% and a CIVD of ~6% on the three databases, such high

TABLE 1 The variation degrees of last names in three large literature databases

Database	# Citations (C) # Authors (A) # Linked authors (LA)	Measure	# Variants	CSVD (%)	CIVD (%)
PubMed	C: 30,128,785	–	489,147	8.04	5.80
	A: 121,251,488				
	LA: 6,082,042				
MAG	C: 213,972,535	Derek73	1,469,738	11.65	9.05
	A: 561,517,211	Joshfraser	1,499,691	11.91	9.33
	LA: 12,613,771	Endwith	1,170,892	9.28	6.34
S2	C: 179,590,271	Derek73	1,691,295	12.35	9.36
	A: 476,379,238	Joshfraser	1,718,657	12.55	9.59
	LA: 13,697,566	Endwith	1,302,345	9.51	6.13

degrees demonstrate that the last name variation problem is nontrivial in literature databases. To facilitate a better understanding, we have manually examined name variants in an attempt to summarize the typical types of variation and identify the possible reasons. From Supplemental material B, we found that there can be many reasons for the name discrepancies. The provided reasons explain why last name variation is prevalent in literature databases.

4.3 | Multi-faceted evaluation of LAGOS-AND

We present evidence to demonstrate that the two LAGOS-AND datasets can be regarded as standard resources for author name disambiguation. For this purpose, we performed an evaluation to present the closeness between LAGOS-AND and the whole MAG in multiple facets. Note that the evaluation was performed after pruning those citations that are over-presented on a certain facet from our datasets to approximate the real distribution of MAG in that facet.

Before conducting the multi-faceted evaluation, we show the accuracy of authorship in the generated LAGOS-AND datasets. To do this, we randomly selected 1,000 instances (paired citations) from LAGOS-AND-PAIRWISE and determined the authorship of the paired citations manually. The results show that the accuracy is 98.2% (99.7% for the v2.0 dataset), demonstrating that our datasets are very accurate in terms of labeled authorship.

4.4 | Last name variation

We used the "Endwith" measure to calculate the variation degree of the last name for our datasets. The CSVD and CIVD for LAGOS-AND-BLOCK were identified at 9.63% and 6.46%, respectively; and for LAGOS-AND-PAIRWISE, the CSVD and CIVD were identified at 9.72% and 6.55%, respectively. Because the variation degrees are very close to the degrees of MAG (9.28% and 6.34% shown in Table 1), the two LAGOS-AND datasets are able to represent MAG in terms of this aspect.

4.4.1 | Publication date distribution

Figure 4a shows the number of publications each year. In comparison to MAG, LAGOS-AND reflects the tendency of the number of publications before 2010; however, the two curves of LAGOS-AND grow faster than MAG after 2010.[22] We attribute this to the creation timeframe of the ORCID system. ORCID launched its registry service in 2012[23] (see Figure 1). As a result, the papers published earlier than this timeframe may be underrepresented in ORCID. Additionally, the increasing popularity of ORCID iDs also exacerbates the under-representation problem of "older papers." A simple measure to handle this would be pruning those citations published after 2010. However, according to our experiments, such a measure would not only significantly reduce the size of our datasets but would also deprive our datasets of many valuable name ambiguity patterns, which is detrimental to the development of effective AND methods. Due to this, we decided not to make adjustments for those citations published after 2010.

4.4.2 | Author position distribution

Some datasets focus only on a particular author position, for example, Song-PubMed (Song et al., 2015) was created to disambiguate the first author. Note that over-focusing on a position will bias the datasets because the

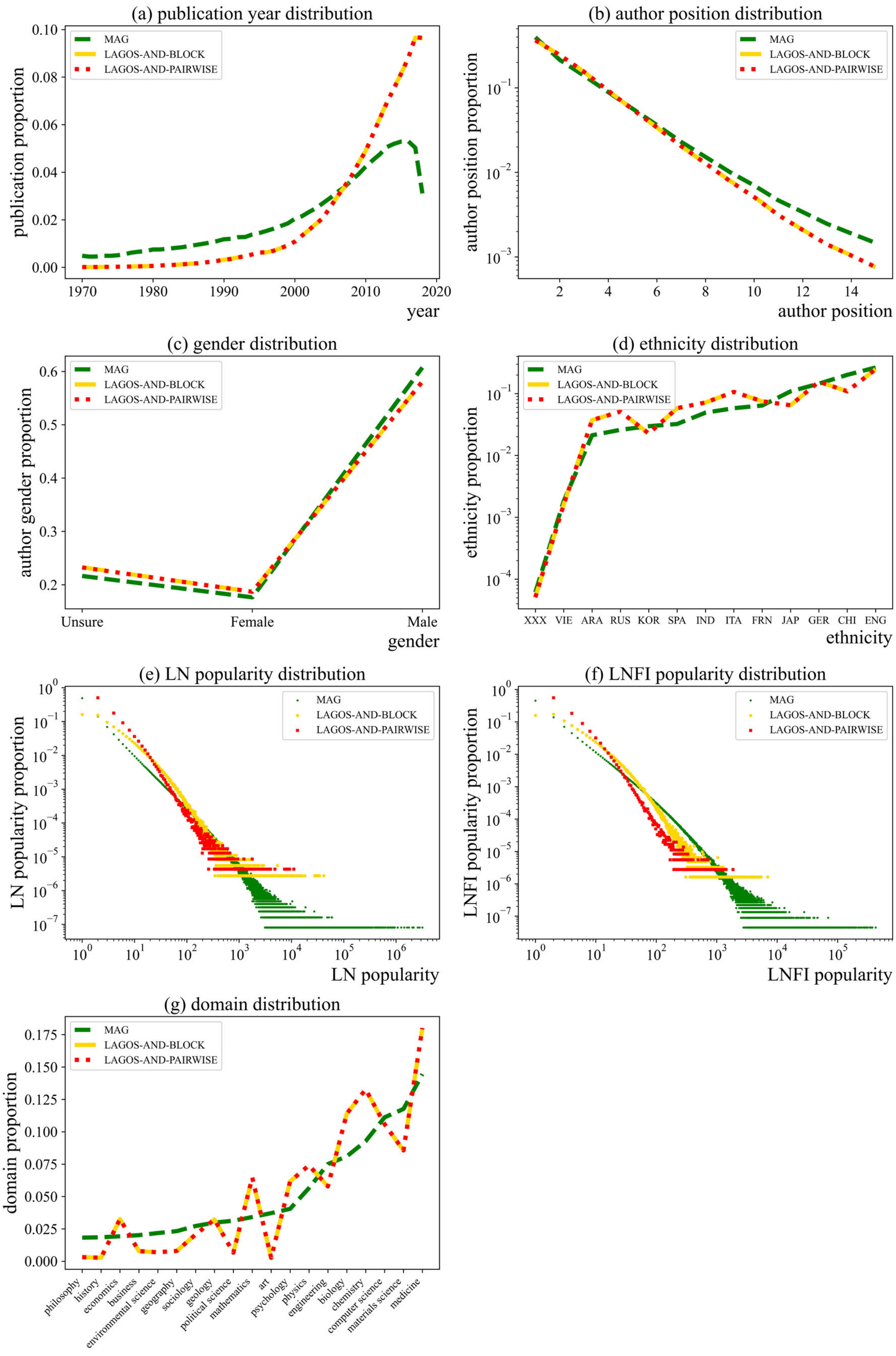

FIGURE 4 Multi-faceted evaluation of the LAGOS-AND datasets

underlying information of the first author such as author affiliation may be richer than that of other positions in some databases (Song et al., 2015). In this study, we considered all author positions equally. As shown in Figure 4b, the closeness of the three curves demonstrates that our datasets can represent the whole MAG in this aspect.

4.4.3 | Gender distribution

Gender distribution is an important facet to examine the quality of AND datasets because there is a correlation between name patterns and genders (Jia & Zhao, 2019; To et al., 2020; Wais, 2016). However, this facet has not received enough attention in previous datasets. To examine the gender distribution, we used Genni+Ethnea (Smith et al., 2013; Torvik & Agarwal, 2016), a widely used gender dataset containing 4,934,974 distinct names collected from PubMed (Kim & Owen-Smith, 2021; Subramanian et al., 2021). We queried genders from Genni+Ethnea by author names to obtain the gender predictions of LAGOS-AND and MAG. As shown in Figure 4c, the closeness of the curves suggests that LAGOS-AND can represent MAG in terms of gender distribution.

4.4.4 | Ethnicity distribution

Person names of different ethnicities usually have different levels of ambiguities. For instance, Chinese authors are more difficult to disambiguate than other ethnicities (Gomide et al., 2017; Kim & Diesner, 2016; Kim, Kim, & Owen-Smith, 2019). Here, we present an ethnicity distribution to demonstrate that there is no significant bias in LAGOS-AND. Similar to gender detection, detecting ethnicity from names also has a high confidence level because person names are highly culturally related, and many ethnicities have their own naming conventions (Treeratpituk & Giles, 2012). Specifically, we used an ethnicity prediction dataset EthinicSeer, a part of Genni +Ethnea, to associate the ethnicity predictions to the name instances of LAGOS-AND and MAG.[24] Because the three curves shown in Figure 4d are very close, our LAGOS-AND datasets can therefore represent the whole MAG in this aspect.

4.4.5 | Name popularity distribution

Another way to examine how our dataset represents MAG is to compare the name popularity, defined as the frequencies of a name in a database. We considered two kinds of name popularity, LN popularity and LNFI popularity, the proportions of which are illustrated in Figure 4e and 4f, respectively. From the results, we found that the minimum percentages of the two LAGOS-AND curves are different from that of MAG. The reason is that MAG covers a wider range of name popularity (LN: [1–3,193,636], LNFI: [1–424,493]) in comparison with the name popularity of LAGOS-AND-BLOCK (LN: [1–41,764], LNFI: [1–6,945]) and the name popularity of LAGOS-AND-PAIRWISE (LN: [2–11,096], LNFI: [2–1,864]). Therefore, the minimum percentages of LN and LNFI popularity in the three databases (MAG LN: 8.03 e-8, LAGOS-AND-BLOCK LN: 2.75 e-6, LAGOS-AND-PAIRWISE LN: 4.28 e-6; MAG LNFI: 4.45 e-8, LAGOS-AND-BLOCK LNFI: 1.62 e-6, LAGOS-AND-PAIRWISE LNFI: 2.74 e-6) are different. Despite the dissimilarities, the tendency of the three curves is similar, which demonstrates that our datasets can overall represent the whole MAG in this facet.

4.4.6 | Domain distribution

Another prominent difference between our dataset and others is domain coverage. To reflect this, we adopted the level-0 field of study (FoS) of MAG, a set of 19 concepts for classifying the full disciplines of science (Shen et al., 2018), as a proxy to describe the domain distribution. From the close similarities of the curves in Figure 4g, we made the inference that LAGOS-AND not only covers a variety of domains but is also representative of the whole MAG in this facet.

5 | DISAMBIGUATION METHODS EVALUATION ON LAGOS-AND

In this section, we provide an evaluation of several disambiguation methods and the MAG's author ID system on the two LAGOS-AND sub-datasets, which will serve as baselines for future AND studies interested in LAGOS-AND.

5.1 | Evaluation datasets

We used the LAGOS-AND-BLOCK dataset to evaluate clustering-based AND methods and used the LAGOS-AND-PAIRWISE dataset to evaluate classification-based AND methods. To facilitate the development of disambiguation methods, we split the two datasets into training, validation, and test folds following the same ratios of 50:25:25 (for LAGOS-AND-BLOCK, instance = block; for

LAGOS-AND-PAIRWISE, instance = paired-citations), and aligned the two datasets in each data fold to ensure instances with the same CFN belonging only to a fixed fold. For example, the instance (MAG paper IDs:2,093,242,633 and 2,464,285,545) with the CFN "Paulo Silva" in the test set of LAGOS-AND-PAIRWISE can only be found in the test set of LAGOS-AND-BLOCK.

5.2 | Baseline methods

Depending on the disambiguation scenarios, the disambiguation methods have different implementations. In the classification-based scenario, the disambiguation methods will try to predict the authorship of paired citations, Here, we followed the commonly used supervised learning ideology with handcrafted features (Song et al., 2015; Zhang et al., 2021) to develop our methods on LAGOS-AND-PAIRWISE. In contrast, in the clustering-based scenario, the disambiguation methods will try to attribute citations to the right authors. Therefore, we followed the semi-supervised ideology to develop our methods on LAGOS-AND-BLOCK. The semi-supervised methods disambiguate authors by applying the supervised AND methods developed in the classification-based scenario to derive the author similarities and then using a clustering algorithm (Cen et al., 2013; Cota et al., 2010; Ferreira et al., 2014; Louppe et al., 2016; Qian et al., 2015; Smith et al., 2013; Torvik & Smalheiser, 2009) such as Hierarchical Agglomerative Clustering (HAC) to group citations into disjoint clusters based on the calculated author similarities.

The baseline methods developed in this study are simple as well as generic because the underlying metadata are available in most literature databases (see the Supplemental material A), and the logic of feature extractions is straightforward. Table 2 itemizes all the features used in the baseline methods. We divide the features into two groups: the base feature group *BF* and the content feature group *CF* according to the metadata being used. For two particular citations, the *BF* feature group contains four basic features: the similarity of author names, the gap years between the two publication dates, the similarity of publication venues, and the similarity of author affiliations, which are calculated by the measures shown in Table 2. The *CF* feature group contains four kinds of content-based features extracted by different measures on the same article content (i.e., title and abstract). The reasons why we paid special attention to the content similarity are that content similarity is helpful for improving disambiguation (Kim, Rohatgi, & Giles, 2019) and, more importantly, authors can be intuitively disambiguated by judging the closeness of the research topic of two citations when other metadata such as affiliation are missing or the author names provide little discriminative information (e.g., the ambiguous authors share the same full name in our datasets).

As shown in Table 2, the similarity measures for the CF group are Jaccard index, TFIDF, Doc2vec (Le & Mikolov, 2014), and a simple neural network that can capture the content similarity at the semantic level (see Supplemental material D for the diagram and the parameter settings of the network).[25] Based on all the *BF* features and a *CF* feature, we developed several baseline methods, which can be found in Tables 3 and 4. We evaluated these methods on both LAGOS-AND-BLOCK and LAGOS-AND-PAIRWISE datasets. Additionally, we evaluated the MAG's author ID system (denoted by MAG-Author-ID), which was created by the Microsoft Academic team for the over 560 million authorship in MAG. We evaluated MAG-Author-ID because the ID system has been widely used for many downstream tasks (Färber, 2019; Huang et al., 2020). However, it is unclear whether the actual performance of the ID system represents the uniqueness of the authors.

5.3 | Metrics and parameter settings

We report precision (P), recall (R), F1, and Macro-F1 metrics for the classification-based AND approaches. The reason for using the additional metric Macro-F1 is that the LAGOS-AND-PAIRWISE is naturally skewed (95.56% of instances are positives), and Macro-F1 is a more suitable metric for evaluation on such an imbalanced dataset. It is not surprising that most instances in LAGOS-AND-PAIRWISE are positives because most blocks of LAGOS-AND-BLOCK contain citations belonging to a single author (see the Supplemental material E), therefore, sampling over LAGOS-AND-BLOCK likely yields positive samples. To measure the performance of clustering-based AND approaches, we use B-cubed (B3) precision (B3-P), B-cubed recall (B3-R), and B-cubed F1 (B3-F1) as the metrics have been widely used in prior clustering-based AND studies (Han et al., 2017; Qian et al., 2015).

In terms of model settings, we used the Random Forest (RF) algorithm to predict the similarity of paired authors and used the HAC algorithm to cluster the ambiguous authors, because RF has achieved robust performance in prior studies (Sanyal et al., 2021), and HAC is also a commonly adopted clustering algorithm (Han et al., 2015; Wu & Ding, 2013). Here, the number of tree components of the RF classifiers is set to 100. Note that, in contrast to some supervised algorithms such as RF showing a robust performance, the performance of

TABLE 2 Features list and feature extraction measures

Feature group	Feature name	Metadata	Measure
Base feature (BF) group	Name similarity	Full author name	Char-level Jaccard index (2-grams)
	Publication year gap	Publication year	Absolute difference
	Venue similarity	Venue	Word-level Jaccard index
	Affiliation similarity	Affiliation	Word-level Jaccard index
Content feature (CF) group	$CF_{jaccard}$	Title and abstract	Word-level Jaccard index
	CF_{tfidf}	Title and abstract	TFIDF
	$CF_{doc2vec}$	Title and abstract	Doc2vec
	CF_{nn}	Title and abstract	Neural network

TABLE 3 Evaluation results on LAGOS-AND-PAIRWISE

Method	P	R	F1	Macro-F1
Random	95.46	50.01	65.64	36.9
MAG-Author-ID	**98.82**	70.16	82.06	51.12
Name Similarity	95.8	87.57	91.5	50.08
BF	95.55	**99.56**	97.51	50.16
BF + $CF_{jaccard}$	95.62	99.31	97.43	51.53
BF + CF_{tfidf}	95.67	98.53	97.08	52.35
BF + $CF_{doc2vec}$	95.67	98.65	97.14	52.46
BF + CF_{nn}	96.57	98.57	**97.56**	**65.21**

Note: Bolded value indicates the best performance of each metric.

clustering algorithms is often largely affected by the built-in parameters. Thus, a tuning process should be conducted to identify the optimal clustering parameter instead of using an empirical value. For HAC, the distance threshold is the only parameter that needs to be tuned. We tuned the parameter for all the baseline methods requiring clustering on the validation set of LAGOS-AND-BLOCK by searching the parameter in the range of [0, 1] with the incremental step being set to 0.05, and we determined the optimal parameter when the B3-F1 metric reached the maximal. Finally, the optimal distance thresholds of all the semi-supervised baselines were surprisingly identified at the same value 1.0, and the B3 metrics achieved by these baseline methods are the same (see Table 4). This finding implies that the B3-F1 metrics are only maximized when all the citations are merged into the same cluster. A deeper analysis suggests that these are normal behaviors, because most LAGOS-AND-BLOCK blocks consist of citations belonging to a single author (see the Supplemental material E), and therefore simply merging them into one cluster would yield the best performance.

The investigation also identifies a limitation of the LAGOS-AND-BLOCK dataset, namely, it is not suitable to focus the dataset on developing clustering-based AND methods because the overwhelming “single author blocks” do not support parameter tuning for clustering algorithms. However, this does not mean that LAGOS-AND-BLOCK is completely useless. We will discuss this in the Discussion section.

To obtain meaningful clustering results, we trimmed those blocks containing only one author from LAGOS-AND-BLOCK, leaving all the blocks containing at least two real-life authors. After this step, the trimmed LAGOS-AND-BLOCK dataset (denoted by LAGOS-AND-BLOCK-TRIMMED) contains 39,528 blocks (9,950 test blocks) and 758,584 citations. Although the step significantly reduced the size of LAGOS-AND-BLOCK, LAGOS-AND-BLOCK-TRIMMED is still a very large dataset: it outperforms 11 out of 12 datasets in terms of dataset size, as shown in the Supplemental material A. With the trimmed dataset, we conducted the tuning process and developed our clustering-based methods on it. Finally, the optimal parameters of the baseline methods shown in Table 3 are identified at 0.45, 0.25, 0.2, 0.2, 0.25, and 0.2, respectively.

6 | RESULTS

Tables 3 and 4 show the respective evaluation results on our datasets. Our observations are as follows.

First, we found that F1 and Macro-F1 of MAG-Author-ID are 82.06% and 51.12% on LAGOS-AND-PAIRWISE, and similarly, the achieved B3-F1 score on LAGOS-AND-BLOCK is 70.59%. It is surprising to see that the performance of the disambiguated ID system is much lower than expected, given that it has been widely used by many studies (Huang et al., 2020). In addition, we found that MAG-Author-ID is high in precision but low in recall. This can be explained by the method of building the author ID system (Wang et al., 2020). The Microsoft Academic engine harvests scientific articles

	Method	B3-P	B3-R	B3-F1
LAGOS-AND-BLOCK	MAG-Author-ID	99.88	64.73	70.59
	All learnable baselines	97.79	100	98.52
LAGOS-AND-BLOCK-TRIMMED	MAG-Author-ID	**97.68**	71.11	77
	Name Similarity	70.37	87.63	74.78
	BF	75.85	86.62	77.4
	BF + $\text{CF}_{jaccard}$	77.6	89.07	79.61
	BF + CF_{tfidf}	77.27	90.09	79.93
	BF + $\text{CF}_{doc2vec}$	74.14	**91.62**	78.69
	BF + CF_{nn}	79.68	89.59	**81.16**

TABLE 4 Evaluation results on LAGOS-AND-BLOCK and LAGOS-AND-BLOCK-TRIMMED[a]

Note: Bolded value indicates the best performance of each metric.

[a]Note that the "Random" baseline is not applicable to the clustering-based evaluation due to the method cannot assign ambiguous authors to a specific cluster before knowing the number of the unique authors (clusters) in a block.

online, thus, it can find many personal websites and public curricula vitae containing the author's publication list. Since the author–article relationship in the publication list is very accurate, the author ID system created by the method achieved a high level of precision. However, a critical issue with the method is that the crowdsourced publication list of authors is often not complete. To deal with this issue, the research team of MAG developed a machine learning approach to merge other possible articles to the authors when the predictions by the approach exceeded a 97% confidence threshold. The method indeed improved the incompleteness, however, this conservative method inevitably split the articles belonging to the same author into multiple clusters because a high confidence score needed to be met. This approach makes MAG-Author-ID achieve a low call.

Second, we found that combining a content-based feature with all *BF* features significantly improved the disambiguation performance and that different content features have different contributions. The method BF + CF_{nn} was proven to be the best performer, which improved BF by a wide margin. This evidence suggests that the content information is very helpful for disambiguation methods in our datasets.

7 | DISCUSSION

7.1 | Insights into LAGOS-AND

We performed a feature analysis to help understand the characteristics of LAGOS-AND, as feature contributions can reflect the importance of the metadata in our dataset. The feature contributions of the best-performing method BF + CF_{nn} are shown in Figure 5, where the vertical lines and dots denote the standard deviations and means of the feature contributions across all RF ensemble trees. We found that the neural network-based content similarity has the highest contribution, which demonstrates that article content can be effective for disambiguation. Moreover, an interesting finding is that, in contrast to other AND datasets, the name metadata in our dataset is less discriminative. This can be understood by the fact that, in a given block, the ambiguous authors have the same full name, that is, CFNs, and therefore very limited discriminative information can be obtained from names.

7.2 | Drawback of LAGOS-AND

The parameter tuning process has demonstrated that the LAGOS-AND-BLOCK dataset is not suitable for developing AND methods requiring clustering. However, this does not mean that the dataset is completely useless. At least, LAGOS-AND-BLOCK provides a platform for the evaluation of disambiguated author IDs in a gold standard manner as disambiguated author IDs do not require clustering and parameter tuning. This point is important for two reasons. First, *evaluating disambiguated ID systems is as important as evaluating disambiguation methods*. Many existing methods have achieved a high-performance score of more than 90% on test datasets (Vishnyakova et al., 2016; Zeng & Acuna, 2020). However, we argue that these methods may encounter performance reduction if they are applied to real literature databases because the blocks of whole databases are usually larger than those of the evaluation datasets. For example, the largest block, "David Smith," in our dataset contains 1,067 citations while the corresponding block in MAG contains 4,033 citations, suggesting that name disambiguation on literature databases such as MAG is much more difficult than disambiguation on the evaluation datasets. Thus, we can infer that the performance achieved by a disambiguation method is likely higher

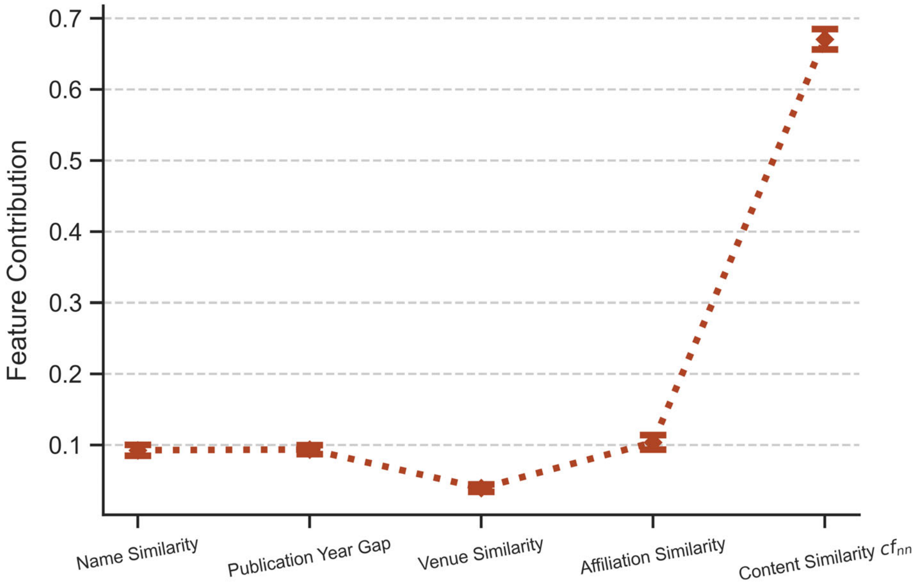

FIGURE 5 Feature contribution analysis for $BF+CF_{nn}$; scores are voted by the ensemble trees of the Random Forest classifier

than the performance of the disambiguated author IDs created by the disambiguation method on the same test set. In this sense, an independent evaluation of disambiguated ID systems is greatly important for AND research. Second, *it is also important to use a gold standard dataset to evaluate disambiguated author IDs*. Given that existing datasets are more or less biased, the gold standard LAGOS-AND-BLOCK dataset can reflect the performance of the disambiguated author IDs in a realistic scenario.

Despite the above, we have shown that the drawback is relatively easy to overcome. By simply removing the blocks containing a single author from LAGOS-AND-BLOCK, we created another block-based dataset, LAGOS-AND-BLOCK-TRIMMED, which can be used to develop clustering-based methods. This implies that the two block-based datasets play different roles in AND studies. Specifically, we suggest that future studies that are interested in LAGOS-AND use LAGOS-AND-BLOCK to *test* the disambiguated author IDs and use LAGOS-AND-BLOCK-TRIMMED to *develop and evaluate* the disambiguation methods requiring clustering.

7.3 | Error analysis for LAGOS-AND

Although we followed rigorous procedures to build the LAGOS-AND datasets, we realize that our datasets are not error-free. Here, we summarize several reasons that could result in the errors according to our intensive observations. (1) Reversed Names. In the name management interface of the ORCID system, the input boxes of the first name and the last name are explicitly distinguished to ensure that authors (users) will enter the right name components into the boxes. Though this kind of error is extremely rare, we still observed outliers. For example, when writing this article, we found that the author named "Ruixue, Sun" with ORCID iD "0000-0003-2495-0433" has reversed her/his name to "Sun, Ruixue" on the author's ORCID page. (2) Author with Multiple ORCID iDs. As the ORCID team claims, there might be some authors who have created multiple ORCID iDs.[26] Fortunately, the ORCID team has developed several measures to prevent such errors from occurring or to eliminate them if they do occur. For instance, when a new registration is received, the ORCID system will attempt to block the registration by searching the registration database for a matching existing account/accounts. If possible accounts are found, the system will return the alternatives to the author for selection. The system also allows authors to manage the already created duplicates in case they have been unintentionally created, that is, marking one iD as the primary and deprecating others.[27] These measures are indeed helpful for eliminating these kinds of errors, however, we suspect that there are still undetected duplicates inside the ORCID data. (3) Incorrect Author Position Identification. The author position identification algorithm does not necessarily guarantee perfect performance. As shown in Supplemental material C, the algorithm fails for 0.05% of MAG author names, and thus this step will introduce errors into our datasets.

7.4 | Implications of the performance of MAG author IDs

Our evaluation shows that the MAG's author ID system only achieved a 70.59% B3-F1 score, an 82.06% F1 score, and a 51.12% Macro-F1 score on our gold standard datasets. Such low performance may lead to distorted results for those studies drawn on this basis. In view of this, we suggest that future studies or applications should be more careful in using MAG author IDs. Additionally, we found that there is a significant gap between the performance of MAG author IDs and many disambiguation methods. For example, many methods have achieved a performance of more than 90% (Song et al., 2015; Vishnyakova et al., 2019). The discrepancies in performance highlight an important research question about the practicality of AND methods: many studies approaching AND used fancy techniques such as heterogeneous graphs (King et al., 2014) and adversarial learning (Peng et al., 2019), however, most of them are limited in terms of being used on large literature databases such as MAG and OpenAlex[28] due to the high computational complexity (Xiao et al., 2020). The lessons learned from the significant gaps will help better understand the name ambiguity problem and the performance of disambiguation methods in real-world large-scale literature databases.

7.5 | Implications of last name variation

By connecting the ORCID data to three large literature databases, the variation degrees in last names were identified at 8.04%–12.55% (CSVD) and 5.80%–9.59% (CIVD). Notably, the problem is nontrivial because it plays an important role in the widely accepted block-based disambiguation framework, in which the author's last name is assumed to be consistent across all the author's publications, and the last name (or the last name and first initial) is used to group name instances into blocks so that disambiguation for large literature databases will be more computationally efficient. However, the high variation degrees suggest that an author's publications may be divided into multiple blocks and thus assigned directly to different authors. Based on the analysis, this finding is important in revealing the limitation of the classic block-based disambiguation framework, as well as helping future studies develop a better disambiguation framework.

7.6 | Research limitations

The first limitation is the potential errors of the ORCID names (CFNs). Although the names are maintained by the authors themselves, we indeed find reversed names (very rare). However, it is difficult to detect the reversed names because determining whether author names are reversed is often confusing without strong background knowledge about the naming conventions of different groups of people. The second limitation is that we have not fully considered the ORCID iD duplicates. Although the measures provided by the ORCID team are effective to eliminate the duplicates, we failed to find a way to identify and remove all the possible duplicates. Third, we calculated the last name variation degree for three literature databases by comparing the author name instances inside these databases to CFNs. It should be pointed out that the degrees can be influenced by many factors, for example, whether authors have uploaded all their publications to the ORCID system. Unfortunately, we are unable to examine the impact of these factors because the underlying information is not available.

8 | CONCLUSIONS

In this article, we described a method that can automatically build large labeled datasets for the author name disambiguation research. Based on the method and the academic resources ORCID and DOI, we built two AND datasets: LAGOS-AND-BLOCK and LAGOS-AND-PAIRWISE, which not only have a large size but also show close similarities to the whole Microsoft Academic Graph across validations of six facets. In building the dataset, we investigated the last name variation problem and revealed the variation degrees in three considerable literature databases. Furthermore, we evaluated the MAG's author ID system and several baseline methods on the created datasets; the analyses for the datasets and the experimental results are also presented in the article.

AUTHOR CONTRIBUTIONS

Li Zhang: Conceptualization, Investigation, Methodology, Software, Formal analysis, Resources, Data curation, Writing - Original Draft, Review & Editing. **Wei Lu**: Supervision, Project administration, Funding acquisition, Methodology, Review & Editing. **Jinqing Yang**: Language Editing.

ACKNOWLEDGMENTS

This work was supported by the National Key Research and Development Program of China (No. 2019YFB1404702). The authors would like to express their sincere gratitude to the editors and the anonymous reviewers for their valuable comments and suggestions.

CONFLICT OF INTEREST

None declared.

DATA AVAILABILITY STATEMENT

The initial versions of our datasets are available at https://zenodo.org/record/7313380, and the code for the baseline methods as well as the analysis presented in this article is available at https://github.com/carmanzhang/LAGOS-AND

ORCID

Li Zhang https://orcid.org/0000-0003-2104-0194

ENDNOTES

[1] In this study, we use *citation* rather than *paper* or *article* to represent the published papers, as most literature databases only contain article metadata.

[2] https://www.aminer.cn/

[3] https://orcid.org

[4] https://www.doi.org

[5] Note that we eliminated the KISTI-AD-E-01 dataset created by (Kang et al., 2011) from this review because it is not retrievable according to the given link.

[6] https://doi.org/10.7802/1234

[7] https://www.zbmath.org

[8] Note that WhoisWho contains two name writing styles because a large number of citations are Chinese citations in which Chinese authors prefer to write their last name first. To accurately calculate the degree of variation for WhoisWho, we identified the Chinese papers and converted the writing style of the Chinese names to the standard Western name style.

[9] https://info.orcid.org/researchers/

[10] https://www.doi.org/

[11] Note that all experimental results reported in this paper are based on the version v1.0 of the LAGOS-AND datasets. By rerunning our dataset creation pipeline on the OpenAlex database, we have created the second version of the datasets, available at https://zenodo.org/record/7313353. We built them on OpenAlex instead of MAG because MAG was discontinued on December 31, 2021, and OpenAlex not only positions itself as a drop-in replacement for MAG but also keeps evolving by aggregating academic resources from other repositories. We plan to release the third versions in 2023–2024.

[12] https://www.microsoft.com/en-us/research/project/microsoft-academic-graph

[13] https://orcid.figshare.com/articles/dataset/ORCID_Public_Data_File_2020/13066970/1

[14] https://zenodo.org/record/2628216#.YBI2KtUzaUk

[15] Only first name initials are available in this citation.

[16] There is no explicit field for last name in MAG, only full names are available.

[17] https://github.com/joshfraser/PHP-Name-Parser

[18] https://github.com/derek73/python-nameparser

[19] https://members.orcid.org/api/tutorial/reading-xml

[20] https://support.orcid.org/hc/en-us/articles/360006973853

[21] https://www.semanticscholar.org/

[22] Note that the MAG curve declines after 2018, which probably caused by the incomplete indexing of citations published after 2018.

[23] https://info.orcid.org/orcid-launches-registry

[24] The ethnicity predictions included in EthinicSeer are Vietnamese (VIE), Arabian (ARA), Russian (RUS), Korean (KOR), Columbian-Spanish-Venezuelan (SPA), Indian (IND), Italian (ITA), French (FRN), Japanese (JAP), German (GER), Chinese (CHI), British (ENG), and others (XXX).

[25] It should be noted that we also tried to incorporate the handcrafted features into the neural network. However, we did not obtain a better result.

[26] https://info.orcid.org/managing-duplicate-orcid-ids/

[27] https://support.orcid.org/hc/en-us/articles/360006971593-Do-you-have-more-than-one-account-

[28] https://openalex.org/

REFERENCES

Aman, V. (2018). A new bibliometric approach to measure knowledge transfer of internationally mobile scientists. *Scientometrics*, *117*(1), 227–247. https://doi.org/10.1007/s11192-018-2864-x

Cen, L., Dragut, E. C., Si, L., & Ouzzani, M. (2013). Author disambiguation by hierarchical agglomerative clustering with adaptive stopping criterion. In G. J. F. Jones, P. Sheridan, D. Kelly, M. de Rijke, & T. Sakai (Eds.), *The 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'13, Dublin, Ireland—July 28–August 01, 2013* (pp. 741–744). ACM. https://doi.org/10.1145/2484028.2484157

Cota, R. G., Ferreira, A. A., Nascimento, C., Gonçalves, M. A., & Laender, A. H. F. (2010). An unsupervised heuristic-based hierarchical method for name disambiguation in bibliographic citations. *Journal of the Association for Information Science and Technology*, *61*(9), 1853–1870. https://doi.org/10.1002/asi.21363

Culotta, A., Kanani, P., Wick, M., & Mccallum, A. (2007). *Author disambiguation using error-driven machine learning with a ranking loss function*. AAAI Workshop—Technical Report.

Färber, M. (2019). The microsoft academic knowledge graph: A linked data source with 8 billion triples of scholarly data. In C. Ghidini, O. Hartig, M. Maleshkova, V. Svátek, I. F. Cruz, A. Hogan, J. Song, M. Lefrançois, & F. Gandon (Eds.), *The Semantic Web—ISWC 2019—18th International Semantic Web Conference, Auckland, New Zealand, October 26–30, 2019, Proceedings, Part II* (Vol. 11779, pp. 113–129). Springer. https://doi.org/10.1007/978-3-030-30796-7_8

Ferreira, A., Gonçalves, M., & Laender, A. (2014). Disambiguating author names using minimum bibliographic information. *World Digital Libraries—An International Journal*, *7*(1), 71–84. https://doi.org/10.3233/WDL-120115

Gomide, J., Kling, H., & Figueiredo, D. (2017). Name usage pattern in the synonym ambiguity problem in bibliographic data. *Scientometrics*, *112*(2), 747–766.

Han, D., Liu, S., Hu, Y., Wang, B., & Sun, Y. (2015). ELM-based name disambiguation in bibliography. *World Wide Web*, *18*(2), 253–263.

Han, H., Xu, W., Zha, H., & Giles, C. L. (2005). A hierarchical naive bayes mixture model for name disambiguation in author citations. In H. Haddad, L. M. Liebrock, A. Omicini, & R. L. Wainwright (Eds.), *Proceedings of the 2005 ACM Symposium on Applied*

Computing (SAC), Santa Fe, NM, March 13–17, 2005 (pp. 1065–1069). ACM. https://doi.org/10.1145/1066677.1066920

Han, H., Yao, C., Fu, Y., Yu, Y., Zhang, Y., & Xu, S. (2017). Semantic fingerprints-based author name disambiguation in Chinese documents. *Scientometrics*, *111*(3), 1879–1896. https://doi.org/10.1007/s11192-017-2338-6

Han, H., Zha, H., & Giles, C. L. (2005). Name disambiguation in author citations using a k-way spectral clustering method. In M. Marlino, T. Sumner, & F. M. S. III (Eds.), *ACM/IEEE Joint Conference on Digital Libraries, JCDL 2005, Denver, CO, June 7–11, 2005, Proceedings* (pp. 334–343). ACM. https://doi.org/10.1145/1065385.1065462

Huang, J., Gates, A. J., Sinatra, R., & Barabási, A.-L. (2020). Historical comparison of gender inequality in scientific careers across countries and disciplines. *Proceedings of the National Academy of Sciences*, *117*(9), 4609–4616.

Jia, J., & Zhao, Q. (2019). Gender prediction based on Chinese name. In J. Tang, M.-Y. Kan, D. Zhao, S. Li, & H. Zan (Eds.), *Natural Language Processing and Chinese Computing—8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9–14, 2019, Proceedings, Part II* (Vol. 11839, pp. 676–683). Springer. https://doi.org/10.1007/978-3-030-32236-6_62

Kang, I.-S., Kim, P., Lee, S., Jung, H., & You, B.-J. (2011). Construction of a large-scale test set for author disambiguation. *Information Processing & Management*, *47*(3), 452–465. https://doi.org/10.1016/j.ipm.2010.10.001

Kim, J. (2018). Evaluating author name disambiguation for digital libraries: A case of DBLP. *Scientometrics*, *116*(3), 1867–1886. https://doi.org/10.1007/s11192-018-2824-5

Kim, J., & Diesner, J. (2016). Distortive effects of initial-based name disambiguation on measurements of large-scale coauthorship networks. *Journal of the Association for Information Science and Technology*, *67*(6), 1446–1461. https://doi.org/10.1002/asi.23489

Kim, J., & Kim, J. (2020). Effect of forename string on author name disambiguation. *Journal of the Association for Information Science and Technology*, *71*(7), 839–855. https://doi.org/10.1002/asi.24298

Kim, J., Kim, J., & Owen-Smith, J. (2019). Generating automatically labeled data for author name disambiguation: An iterative clustering method. *Scientometrics*, *118*(1), 253–280. https://doi.org/10.1007/s11192-018-2968-3

Kim, J., Kim, J., & Owen-Smith, J. (2021). Ethnicity-based name partitioning for author name disambiguation using supervised machine learning. *Journal of the Association for Information Science and Technology*, *72*(8), 979–994.

Kim, J., & Owen-Smith, J. (2021). ORCID-linked labeled data for evaluating author name disambiguation at scale. *Scientometrics*, *126*(3), 2057–2083.

Kim, K., Khabsa, M., & Giles, C. L. (2016). *Random forest DBSCAN for USPTO inventor name disambiguation*. CoRR. http://arxiv.org/abs/1602.01792

Kim, K., Rohatgi, S., & Giles, C. L. (2019). Hybrid deep pairwise classification for author name disambiguation. In W. Zhu, D. Tao, X. Cheng, P. Cui, E. A. Rundensteiner, D. Carmel, Q. He, & J. X. Yu (Eds.), *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019, Beijing, China, November 3–7, 2019* (pp. 2369–2372). ACM. https://doi.org/10.1145/3357384.3358153

King, B., Jha, R., & Radev, D. R. (2014). Heterogeneous networks and their applications: Scientometrics, name disambiguation, and topic modeling. *Transactions of the Association for Computational Linguistics*, *2*, 1–14. https://doi.org/10.1162/tacl_a_00161

Le, Q. V., & Mikolov, T. (2014). Distributed representations of sentences and documents. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21–26 June 2014* (pp. 1188–1196). JMIR.org. http://proceedings.mlr.press/v32/le14.html

Levin, M., Krawczyk, S., Bethard, S., & Jurafsky, D. (2012). Citation-based bootstrapping for large-scale author disambiguation. *Journal of the Association for Information Science and Technology*, *63*(5), 1030–1047. https://doi.org/10.1002/asi.22621

Louppe, G., Al-Natsheh, H. T., Susik, M., & Maguire, E. J. (2016). Ethnicity sensitive author disambiguation using semi-supervised learning. In A.-C. Ngonga Ngomo & P. Křemen (Eds.), *Knowledge engineering and semantic web* (pp. 272–287). Springer International Publishing.

Mihaljevic, H., & Santamaría, L. (2021). Disambiguation of author entities in ADS using supervised learning and graph theory methods. *Scientometrics*, *126*(5), 3893–3917. https://doi.org/10.1007/s11192-021-03951-w

Müller, M. C., Reitz, F., & Roy, N. (2017). Data sets for author name disambiguation: An empirical analysis and a new resource. *Scientometrics*, *111*(3), 1467–1500. https://doi.org/10.1007/s11192-017-2363-5

Peng, L., Shen, S., Li, D., Xu, J., Fu, Y., & Su, H. (2019). Author disambiguation through adversarial network representation learning. *International Joint Conference on Neural Networks, IJCNN 2019 Budapest, Hungary, July 14–19, 2019*, 1–8. https://doi.org/10.1109/IJCNN.2019.8852233

Qian, Y., Zheng, Q., Sakai, T., Ye, J., & Liu, J. (2015). Dynamic author name disambiguation for growing digital libraries. *Information Retrieval*, *18*(5), 379–412. https://doi.org/10.1007/s10791-015-9261-3

Sanyal, D. K., Bhowmick, P. K., & Das, P. P. (2021). A review of author name disambiguation techniques for the PubMed bibliographic database. *Journal of Information Science*, *47*(2), 227–254. https://doi.org/10.1177/0165551519888605

Schulz, J. (2016). Using Monte Carlo simulations to assess the impact of author name disambiguation quality on different bibliometric analyses. *Scientometrics*, *107*(3), 1283–1298. https://doi.org/10.1007/s11192-016-1892-7

Shen, Z., Ma, H., & Wang, K. (2018). A web-scale system for scientific knowledge exploration. In F. Liu & T. Solorio (Eds.), *Proceedings of ACL 2018, Melbourne, Australia, July 15–20, 2018, System Demonstrations* (pp. 87–92). Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-4015

Shoaib, M., Daud, A., & Amjad, T. (2020). *Author name disambiguation in bibliographic databases: A survey*. CoRR. https://arxiv.org/abs/2004.06391

Smith, B. N., Singh, M., & Torvik, V. I. (2013). A search engine approach to estimating temporal changes in gender orientation of first names. In J. S. Downie, R. H. McDonald, T. W. Cole, R. Sanderson, & F. Shipman (Eds.), *13th ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL '13, Indianapolis, IN, July 22–26, 2013* (pp. 199–208). ACM. https://doi.org/10.1145/2467696.2467720

Song, M., Kim, E. H. J., & Kim, H. J. (2015). Exploring author name disambiguation on PubMed-scale. *Journal of Informetrics*, *9*(4), 924–941. https://doi.org/10.1016/j.joi.2015.08.004

Subramanian, S., King, D., Downey, D., & Feldman, S. (2021). S2AND: A benchmark AND evaluation system for author name disambiguation. In *2021 ACM/IEEE Joint Conference on Digital Libraries (JCDL)* (pp. 170–179). IEEE.

Tang, J., Fong, A. C. M., Wang, B., & Zhang, J. (2012). A unified probabilistic framework for name disambiguation in digital library. *IEEE Transactions on Knowledge and Data Engineering*, *24*(6), 975–987. https://doi.org/10.1109/TKDE.2011.13

Tekles, A., & Bornmann, L. (2020). Author name disambiguation of bibliometric data: A comparison of several unsupervised approaches. *Quantitative Science Studies*, *1*(2), 1–38.

To, H. Q., Nguyen, K. V., Nguyen, N. L.-T., & Nguyen, A. G.-T. (2020). Gender prediction based on vietnamese names with machine learning techniques. In *NLPIR 2020: 4th International Conference on Natural Language Processing and Information Retrieval, Seoul, Republic of Korea, December 18–20, 2020* (pp. 55–60). ACM. https://doi.org/10.1145/3443279.3443309

Torvik, V. I., & Agarwal, S. (2016). Ethnea—An instance-based ethnicity classifier based on geo-coded author names in a large-scale bibliographic database. In *International Symposium on Science of Science*. Library of Congress.

Torvik, V. I., & Smalheiser, N. R. (2009). Author name disambiguation in MEDLINE. *ACM Transactions on Knowledge Discovery from Data*, *3*(3), 1–29. https://doi.org/10.1145/1552303.1552304

Treeratpituk, P., & Giles, C. L. (2012). Name-ethnicity classification and ethnicity-sensitive name matching. In J. Hoffmann & B. Selman (Eds.), *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence, July 22–26, 2012, Toronto, Ontario, Canada*. AAAI Press http://www.aaai.org/ocs/index.php/AAAI/AAAI12/paper/view/5180

Vishnyakova, D., Rodriguez-Esteban, R., & Rinaldi, F. (2019). A new approach and gold standard toward author disambiguation in MEDLINE. *Journal of the American Medical Informatics Association*, *26*(10), 1037–1045. Oxford University Press. https://doi.org/10.1093/jamia/ocz028

Vishnyakova, R., Esteban, K., & Ozol, F. R. (2016). Author name disambiguation in MEDLINE based on journal descriptors and semantic types. In *Fifth Workshop on Building and Evaluating Resources for Biomedical Text Mining, December* (pp. 134–142). https://doi.org/10.5167/uzh-132256

Wais, K. (2016). Gender prediction methods based on first names with genderizeR. *The R Journal*, *8*, 17–37. https://doi.org/10.32614/RJ-2016-002

Wang, K., Shen, Z., Huang, C., Wu, C.-H., Dong, Y., & Kanakia, A. (2020). Microsoft academic graph: When experts are not enough. *Quantitative Science Studies*, *1*(1), 396–413.

Wang, X., Tang, J., Cheng, H., & Yu, P. S. (2011). ADANA: Active name disambiguation. In D. J. Cook, J. Pei, W. Wang, O. R. Zaïane, & X. Wu (Eds.), *11th IEEE International Conference on Data Mining, ICDM 2011, Vancouver, BC, Canada, December 11–14, 2011* (pp. 794–803). IEEE Computer Society. https://doi.org/10.1109/ICDM.2011.19

Wu, J., & Ding, X.-H. (2013). Author name disambiguation in scientific collaboration and mobility cases. *Scientometrics*, *96*(3), 683–697.

Wu, J., Sefid, A., Ge, A. C., & Giles, C. L. (2017). A supervised learning approach to entity matching between scholarly big datasets. In C. Óscar, K. Janowicz, G. Rizzo, I. Tiddi, & D. Garijo (Eds.), *Proceedings of the Knowledge Capture Conference, K-CAP 2017, Austin, TX, December 4–6, 2017* (pp. 41:1–41:4). ACM. https://doi.org/10.1145/3148011.3154470

Xiao, Z., Zhang, Y., Chen, B., Liu, X., & Tang, J. (2020). *A framework for constructing a huge name disambiguation dataset: Algorithms, visualization and human collaboration*. CoRR. https://arxiv.org/abs/2007.02086

Zeng, T., & Acuna, D. E. (2020). Large-scale author name disambiguation using approximate network structures. In *International Conference on Computational Social Science*. MIT.

Zhang, J., Wu, X., & Sheng, V. S. (2016). Learning from crowdsourced labeled data: A survey. *Artificial Intelligence Review*, *46*(4), 543–576.

Zhang, L., Huang, Y., Cheng, Q., & Lu, W. (2020). Mining author identifiers for PubMed by linking to open bibliographic databases. In *Proceedings—Companion of the 2020 IEEE 20th International Conference on Software Quality, Reliability, and Security, QRS-C 2020* (pp. 209–212). IEEE. https://doi.org/10.1109/QRS-C51114.2020.00043

Zhang, L., Huang, Y., Yang, J., & Lu, W. (2021). Aggregating large-scale databases for PubMed author name disambiguation. *Journal of the American Medical Informatics Association*, *28*, 1–9. https://doi.org/10.1093/jamia/ocab095

Zhang, Y., Zhang, F., Yao, P., & Tang, J. (2018). Name disambiguation in AMiner: Clustering, maintenance, and human in the loop. In Y. Guo & F. Farooq (Eds.), *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19–23, 2018* (pp. 1002–1011). ACM. https://doi.org/10.1145/3219819.3219859

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Zhang, L., Lu, W., & Yang, J. (2023). LAGOS-AND: A large gold standard dataset for scholarly author name disambiguation. *Journal of the Association for Information Science and Technology*, *74*(2), 168–185. https://doi.org/10.1002/asi.24720