Web-Scale Academic Name Disambiguation: the WhoIsWho Benchmark, Leaderboard, and Toolkit

Bo Chen
Tsinghua University
cb21@mails.tsinghua.edu.cn

Jing Zhang*
Renmin University of China
zhang-jing@ruc.edu.cn

Fanjin Zhang
Tsinghua University
zfj17@mails.tsinghua.edu.cn

Tianyi Han
Zhipu.AI
tianyi.han@aminer.cn

Yuqing Cheng
Zhipu.AI
yuqing.cheng@aminer.cn

Xiaoyan Li
Zhipu.AI
xinyan.li@aminer.cn

Yuxiao Dong
Tsinghua University
yuxiaod@tsinghua.edu.cn

Jie Tang*
Tsinghua University
jietang@tsinghua.edu.cn

ABSTRACT

Name disambiguation—a fundamental problem in online academic systems–is now facing greater challenges with the increasing growth of research papers. For example, on AMiner, an online academic search platform, about 10% of names own more than 100 authors. Such real-world challenging cases have not been effectively addressed by existing researches due to the small-scale or low-quality datasets that they have used. The development of effective algorithms is further hampered by a variety of tasks and evaluation protocols designed on top of diverse datasets. To this end, we present WhoIsWho owning, a large-scale benchmark with over 1,000,000 papers built using an interactive annotation process, a regular leaderboard with comprehensive tasks, and an easy-to-use toolkit encapsulating the entire pipeline as well as the most powerful features and baseline models for tackling the tasks. Our developed strong baseline has already been deployed online in the AMiner system to enable daily arXiv paper assignments[1] [2].

CCS CONCEPTS

• **Information systems → Data management systems**; **Information integration**; **Entity resolution**;

KEYWORDS

name disambiguation, benchmark

ACM Reference Format:
Bo Chen, Jing Zhang, Fanjin Zhang, Tianyi Han, Yuqing Cheng, Xiaoyan Li, Yuxiao Dong, and Jie Tang. 2023. Web-Scale Academic Name Disambiguation: the WhoIsWho Benchmark, Leaderboard, and Toolkit. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '23), August 6–10, 2023, Long Beach, CA, USA.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3580305.3599930

1 INTRODUCTION

Name disambiguation, aiming to clarify who is who, is one of the fundamental problems in online academic systems such as Google Scholar[3], Semantic Scholar[4], and AMiner[5]. The past decades have witnessed a huge proliferation of research papers in all fields of science. For example, Google Scholar, Bing Academic Search, and AMiner have all indexed about 300 million papers [10, 33, 36]. As a result, the author name ambiguity problem—the same authors with different name variants, or the different authors with the exact same name or homonyms—has become increasingly sophisticated in modern digital libraries. For example, as of January 2023, there were over 10,000 authors with the name "Yang Yang" on AMiner. Three of them are displayed in Figure 2. Since all three authors are computer scientists, there are intricate connections between their papers. Paper P_5, which belongs to "Yang Yang(THU)", was mistakenly assigned to "Yang Yang(UND)", because both "Yang Yang" coauthored with "Yizhou Sun", leading to the appearance of reliable co-author and co-keyword relationships between P_5 and the correct paper P_4 of "Yang Yang(UND)". Furthermore, "Yang Yang(THU)" and "Yang Yang(ZJU)" are the same person but are separated into two different authors due to organization shifts after graduation. This real-world example demonstrates the great challenges of name disambiguation in online academic systems, which, however, can not be addressed by existing efforts [3, 18, 20, 21, 32, 35, 38, 47–49], because of the small-scaled low-quality benchmark and non-uniform task designs with evaluation settings.

In particular, even though several name disambiguation benchmarks, such as PubMed [40, 45], MAG [46], DBLP [14], etc. [17, 37], have been directly harvested from existing digital libraries, inevitably spurious information and assignment mistakes, as shown in Figure 2, are detrimental to build effective algorithms [4, 44]. In light of this, others attempt to manually annotate a small amount of high-quality data from the online noisy data in order to reduce

[*] Jing Zhang and Jie Tang are the corresponding authors.
[1] The public leaderboard is available at http://whoiswho.biendata.xyz/. The toolkit is at https://github.com/THUDM/WhoIsWho. The online demo of daily arXiv paper assignments is at https://na-demo.aminer.cn/arxivpaper.
[2] This work has been accepted by KDD 2023 ADS Track.

KDD '23, August 6–10, 2023, Long Beach, CA, USA
© 2023 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0103-0/23/08.
https://doi.org/10.1145/3580305.3599930

[3] https://scholar.google.com/
[4] https://www.semanticscholar.org/
[5] https://aminer.org

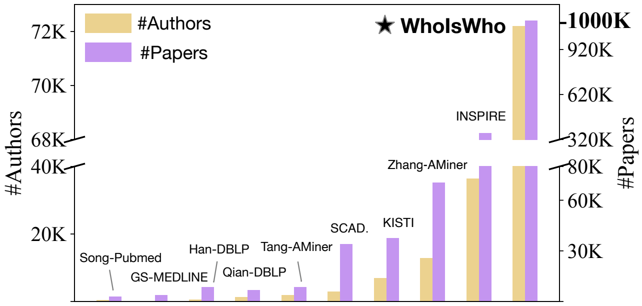


Figure 1: The sizes of the prevailing name disambiguation benchmarks. Among these, WhoIsWho is the largest one with 1,000+ names, 70,000+ authors, and 1,000,000+ papers.

the negative impact of these noises [11, 30, 34]. However, as illustrated in Figure 1, the majority of them lack an adequate number of instances. Additionally, on top of these benchmarks, previous efforts have defined a variety of tasks and evaluation protocols, preventing us from fairly comparing different methods to promote the development of the name disambiguation community.

Present Work. We present WhoIsWho, a benchmark, a leaderboard, together with a toolkit for web-scale academic name disambiguation. Specifically, WhoIsWho has the following characteristics:

• **Interactive large-scale benchmark construction.** To create a challenging benchmark, we devise an interactive annotation process to label paper-author affiliations under a single name with high ambiguity with the aid of the developed visualization tool. 10+ professional annotators were employed to conduct the annotation task with each of them spending about 24 working months. To date, we have released a large-scale, high-quality, and challenging benchmark that contains over 1,000 names, 70,000 authors, and 1,000,000 papers. Figure 1 shows the WhoIsWho benchmark is orders-of-magnitude larger than existing manually-labeled datasets.

• **Contest leaderboard with comprehensive tasks.** To fairly compare various name disambiguation methods, we sponsor contests with two tracks: The first is *From-scratch Name Disambiguation* (SND) aiming at grouping papers by the same author together in order to fulfill the need to create an original academic system from scratch. The other is *Real-time Name Disambiguation* (RND), also known as incremental name disambiguation, which targets at assigning newly-arrived papers to the existing clarified authors. The RND task is crucial to maintain a regular assignment of papers on existing online academic systems owning a substantial amount of clarified author profiles. Beyond these, we additionally define *Incorrect Assignment Detection* (IND), which attempts to remedy online paper-author affiliation errors in order to guarantee the reliability of academic systems. To date, three-round contests have been held on the first two tasks, attracting more than 3,000 researchers. Furthermore, we host a regular leaderboard to keep track of recent advances. The contest for the IND task is under active preparation.

• **Easy-to-use toolkit.** To facilitate researchers to quickly get started in the name disambiguation area, we summarize our research findings and organize an end-to-end pipeline to standardize the entire name disambiguation process, including data loading, feature creation, model construction, and evaluation, We thoroughly investigate the contest winner methods, assemble the most effective features and models, and encapsulate them into the toolkit. The end users are free to directly invoke the baselines and encapsulated features to develop their own algorithms.

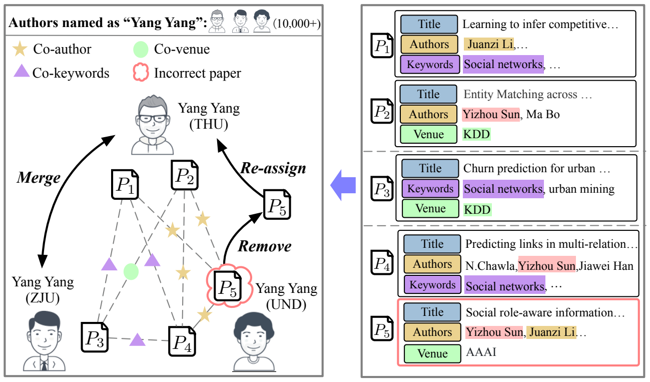

Authors named as "Yang Yang":
(10,000+)
Co-author
Co-venue
Co-keywords
Incorrect paper
Yang Yang (THU)
Merge
Re-assign
Remove
Yang Yang (ZJU)
Yang Yang (UND)
P_1
P_2
P_3
P_4
P_5
Title Learning to infer competitive…
Authors Juanzi Li,…
Keywords Social networks, …
Title Entity Matching across …
Authors Yizhou Sun, Ma Bo
Venue KDD
Title Churn prediction for urban …
Keywords Social networks, urban mining
Venue KDD
Title Predicting links in multi-relation…
Authors N.Chawla,Yizhou Sun,Jiawei Han
Keywords Social networks, …
Title Social role-aware information…
Authors Yizhou Sun, Juanzi Li…
Venue AAAI

Figure 2: Illustration of the challenges for annotating authors with the name "Yang Yang". Paper p_5 is incorrectly assigned because of the coauthorship with the same third person. Two authors are mistakenly separated due to the organization shift.

We provide in-depth analyses of the features adopted in methods of contest winners, finding that blending the multi-modal features, i.e., the semantic features involving paper attributes and the relational features created by co-author, co-organization, and co-venue links, contributes the most to the performance of name disambiguation methods. On top of these discoveries, we provide simple yet effective baselines (RND/SND-all) that perform on par with the top contest methods. Particularly, RND-all has been deployed on AMiner for daily arXiv paper assignment.

To sum up, WhoIsWho is an ongoing, community-driven, open-source project. We intend to update the leaderboard as well as offer new datasets and methods over time. We also encourage contributions at oagwhoiswho@gmail.com.

2 WHOISWHO BENCHMARK

This section first introduces the interactive annotation process for constructing the large-scale high-quality benchmark and then presents the intrinsic distributions of the benchmark.

2.1 Interactive Benchmark Construction

We formalize the interactive benchmark construction pipeline into two sub-modules: data collection and data annotation.

2.1.1 Dataset Collection. Practically, we collect the raw data from AMiner [33]. To acquire name disambiguation data with less noise and also higher ambiguity, we adopt the following rules,

Select authors by H-index. For each author in AMiner, we compute the H-index [12], a metric used to measure the impact of experts, and then we keep the authors with the higher H-index scores. If authors are more well-known, it is assumed that their profiles contain less noise, because they may have already clarified themselves on the academic platform. Concretely, we filtered out authors with an H-index less than 5 by sorting them in descending order based on their H-index values. This threshold is a widely

Table 1: Data annotation pipeline. For operations performed via three annotators, major voting is applied to solve conflicts. Anno. is the abbreviation of annotators.

Steps	#Anno.	Operations
Clean (Roughly)	1	**1**. Delete papers not belonging to the concerned author; **2**. Split over-merged author profiles into multiple authors.
Validate	3	**1**. Same as the "Clean" step to deal with more difficult incorrect papers.
Add	3	**1**. Add unassigned papers to certain authors.
Merge	3	**1**. Merge separate author profiles into a single author.

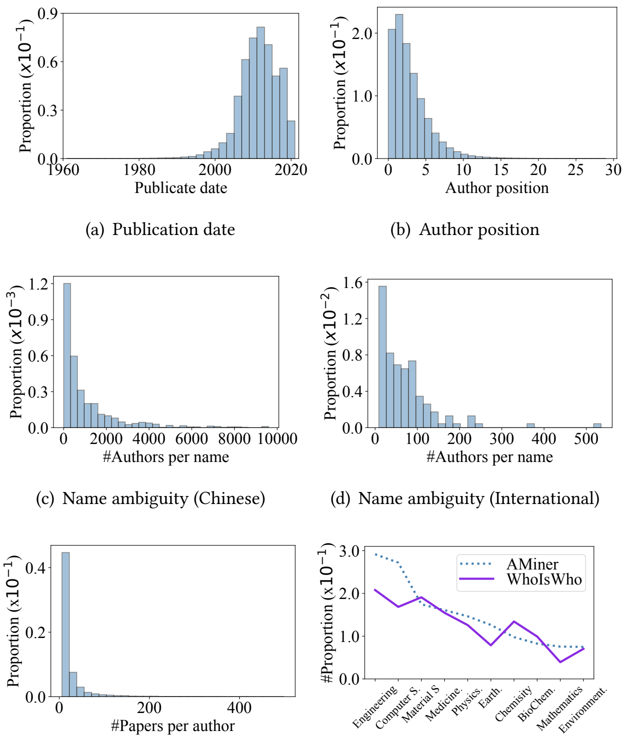
accepted criterion in the literature for identifying authors with significant impact in their research field.

Choose names with high ambiguity. We count the number of authors with the same name in AMiner. The term "same name" refers to the name-blocking ways to unify names, such as moving the last name to the first or preserving all name initials but the last name [2, 14]. For example, the variants of "Jing Zhang" include "Zhang Jing", "J Zhang" and "Z Jing". A name is more ambiguous if it is used by more authors. We filter names with fewer authors than a threshold to make WhoIsWho challenging[6].

After obtaining names with high ambiguity with the corresponding authors for each name, we collect papers for each author. Specifically, we collect the title, author names, organizations of all authors, keywords, abstract, publication year, and venue (conference or journal) as attributes of papers. Additionally, there are a large number of papers that have yet to be assigned to any authors. To increase the challenge of the benchmark, we also gather these papers, denoted as unassigned papers, whose authors share the same name as these in the benchmark, which may be assigned to the authors in the benchmark during the data annotation pipeline.

2.1.2 Dataset Annotation. Figure 2 demonstrates some real-world hard cases of name disambiguation, which are quite challenging for annotators to label because of the intricate relationships between papers. In light of this, we design an interactive annotation tool[7] adapted from [29] to not only provide detailed information about papers and authors but also to offer various practical atomic operations to help annotators in performing arbitrary actions. A toy example is shown in Figure 12. The tool allows annotators to annotate interactively because each time an action is taken, the author profiles are updated and displayed to the annotators.

With the help of the tool, we establish four standardized annotation steps (detailed in Table 1) to ensure the manual labeling process can be conducted in a reasonable manner. Overall, the annotators are authorized to remove incorrect papers, add unassigned papers, split an author into two authors, and merge two authors. Specifically, the first "**Clean**" step allows annotators to remove or split obviously incorrect papers from the concerned author. Such papers cover different topics with the concerned author. Then, the "**Validate**" step allows annotators to conduct the same "Clean" function on incorrect papers that are hard to identify. Such papers cover relevant topics to the concerned author. After that, the "**Add**" step enables annotators to add unassigned papers to associated authors. Finally, the "**Merge**" step allows annotators to blend the papers of two authors into a single author. Since the last three steps are more challenging than the first step, three annotators are requested to annotate the same name with their results aggregated by majority voting. Notably, annotators label all the papers of authors under the same name together each time. To prevent them from simply removing arbitrary papers, annotators must retain at least 80% of the papers for each author.



(a) Publication date (b) Author position

(c) Name ambiguity (Chinese) (d) Name ambiguity (International)

(e) Paper number (f) Domains (Top-10)

Figure 3: Statistics of WhoIsWho benchmark

In summary, on one hand, the devised interactive annotation process, which provides abundant facts among papers, fully supports annotators to label the dataset effectively. On the other hand, each paper is examined by at least 10 skilled annotators, which further guarantees the quality of WhoIsWho.

2.2 Statistics of WhoIsWho Benchmark

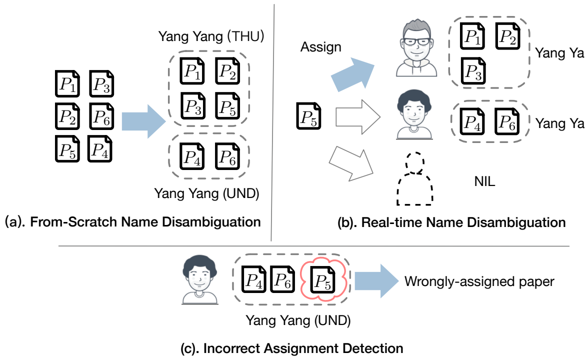
We present the holistic analysis to demonstrate the superiority of the WhoIsWho benchmark in multi-facets, as illustrated in Figure 3.

Accuracy of the Annotated Authorship. We first check the accuracy of the manually-labeled authorship. To achieve this, we randomly sample 1,000 papers from the benchmark and manually verify which papers belong to which authors. Each paper is verified by three skilled annotators via major voting. The resultant accuracy is 99.6% with only four assignment errors, indicating that the benchmark offers a large number of high-quantity instances.

Publication Date Distribution. Figure 3(a) illustrates the distribution of paper publication date. Few scientific documents were

[6] We set the threshold as 6 in WhoIsWho.

[7] https://www.aminer.cn/billboard/id:5e42777f530c70f19522863e



(a). From-Scratch Name Disambiguation

(b). Real-time Name Disambiguation

(c). Incorrect Assignment Detection

Figure 4: Three name disambiguation tasks.

recorded before the year 2000 since managing digital libraries was still a relatively new technique at that time. As the internet develops rapidly after 2020, the number of digital records increases more quickly. However, there are fewer records around 2022 than there were around 2010, suggesting that the online name disambiguation system may not be able to assign the latest papers in time.

Author Position Distribution. Several datasets focus on disambiguating the author on a particular position in the paper. For example, Song-PubMed [30] is created for disambiguating the first author, which introduces biased information to name disambiguation methods toward certain specific author positions. On the contrary, the WhoIsWho benchmark takes all author positions equally into account, as shown by the rational long-tail curve in Figure 3(b).

Name Ambiguity Distribution. Author names of different ethnic groups typically have varying degrees of ambiguity. Chinese authors, for example, are more difficult to disambiguate than other nationalities [9, 15, 16]. Figure 3(c) and 3(d) illustrate the distribution of the clarified author profiles per Chinese and international name respectively in AMiner, indicating Chinese names are more ambiguous than international ones. As we focus on constructing a benchmark with high ambiguity that facilitates name disambiguation methods, we collect more Chinese names, covering about 87% of author names in our dataset, than international names.

Paper Number Distribution. We also present the distribution of the number of papers per author in the benchmark, as shown in Figure 3(e). The long-tail distributions indicate that most of the cases have a manageable quantity and only a few famous scientists own hundreds of publications.

Domain Distribution. Compared with several datasets that merely cover biased domains, such as datasets based on PubMed [30, 45] focus on the field of medical science, WhoIsWho has great coverage of general disciplines. To confirm this, we randomly sample 100,000 papers and then adopt the taxonomy rank of SCImago Journal Rank (SJR)[8] from Scopus to obtain paper domains. The top-10 highest frequency domains are shown in Figure 3(f), which implies the benchmark not only covers a variety of domains but also is a representative of the overall distribution in AMiner.

3 WHOISWHO TASKS & CONTESTS

In this section, we first present three name disambiguation tasks with standardized evaluation protocols. Then we review three-round historical contests and a regular leaderboard built on defined tasks with different released versions.

3.1 Task Formations and Evaluation Protocols

Here we formalize the three tasks i.e, from-scratch name disambiguation, real-time name disambiguation, and incorrect assignment detection, with evaluation metrics, as shown in Figure 4.

Definition 1. ***Paper***. *A paper p is associated with multiple fields of attributes, i.e., $p = \{x_1, \cdots, x_F\}$, where $x_f \in p$ represents the f-th attribute. F is the number of attributes.*

Definition 2. ***Author***. *An author a is comprised of a set of papers, i.e., $a = \{p_1, \cdots, p_n\}$, where each paper $p_i = \{x_1, \cdots, x_F\}$ and n is the number of papers authored by a.*

Definition 3. ***Candidate Papers***. *Given a person name denoted by na, $\mathcal{P}^{na} = \{p_1^{na}, \ldots, p_N^{na}\}$ is a set of candidate papers written by any author with the name na.*

Definition 4. ***Candidate Authors***. *Given a person name denoted by na, $\mathcal{A}^{na} = \{a_1^{na}, \ldots, a_M^{na}\}$ is a set of candidate authors with the same name na. The term "same name" refers to the ways to unify names using name blocking techniques [2, 14].*

3.1.1 *From-scratch Name Disambiguation.*

At the beginning of building digital libraries, we need to partition a large number of published papers into groups, each of which represents papers that belong to a single person. To achieve this, we formalize from-scratch name disambiguation as a clustering problem.

Problem 1. ***From-scratch Name Disambiguation (SND)***. *Given a set of candidate papers $\mathcal{P}^{na}$, SND aims at finding a function Φ to partition $\mathcal{P}^{na}$ into a set of disjoint clusters C^{na}, i.e.,*

$$\Phi(\mathcal{P}^{na}) \rightarrow C^{na}, \text{where } C^{na} = \{C_1^{na}, C_2^{na}, \cdots, C_K^{na}\},$$

where each cluster consists of papers owned by the same author, i.e., $\mathbb{I}(p_i^{na}) = \mathbb{I}(p_j^{na}), \forall (p_i^{na}, p_j^{na}) \in C_k^{na} \times C_k^{na}$, and different clusters contain papers from different authors, i.e., $\mathbb{I}(p_i^a) \neq \mathbb{I}(p_j^a), \forall (p_i^a, p_j^a) \in C_k^a \times C_{k'}^a, k \neq k'$. $\mathbb{I}(p_i^{na})$ is the author identification of the paper p_i^{na}.

Evaluation Protocol. We adopt the macro pairwise-F1 to evaluate the performance of related SNA methods, which is widely adopted by many SND methods [18, 28, 31, 47, 49].

3.1.2 *Real-time Name Disambiguation.*

Assigning new papers to existing authors is crucial for online digital libraries at the current stage. For instance, AMiner receives over 500,000 new papers each month. To this end, we formalize the real-time name disambiguation as a classification problem.

Problem 2. ***Real-time Name Disambiguation (RND).*** *Given a paper p^{na}, i.e, the paper with its author name na to be disambiguated, and the set of candidate authors $\mathcal{A}^{na}$, the right author a^* can be either a real author in $\mathcal{A}^{na}$ or a non-existing author profile, i.e., NIL. We target at learning a function to assign the paper p^{na} to a^*, i.e.,*

$$\Psi(p^{na}, A^{na}) \rightarrow a^*$$

[8] http://www.scimagojr.com

Note that NIL situations are found frequently in online academic platforms. Assuming that undergraduate students publish their first paper at a conference or journal, but the current database has not yet established their author profile, it is infeasible to assign the paper to any authors. In light of this, we have incorporated the NULL scenarios in the RND task. Formal efforts [3] also take into account the NIL situation, however, they create synthesized NIL labels rather than incorporating the actual NIL cases. To our best knowledge, we are the first to consider the NIL situation in the WhoIsWho benchmark with manually-labeled real NIL cases.

Evaluation Protocol. We propose the weighted-F1 to evaluate the methods that solve the RND problem. For an author a to be disambiguated, we calculate the metrics as follows:

$$\begin{aligned}
Precision^a &= \frac{\#PapersCorrectlyAssignedToTheAuthor}{\#PapersAssingedToTheAuthor},\\
Recall^a &= \frac{\#PapersCorrectlyAssignedToTheAuthor}{\#PapersOfTheAuthorTobeAssinged},\\
Weight^a &= \frac{\#PapersOfTheAuthorTobeAssigned}{\#TotalPapersTobeAssigned},
\end{aligned}$$

where precision measures the correctness of papers predicted to a, and recall measures how many papers from a's actual papers could be correctly assigned to a. Then we calculate the F1 score by the precision and recall for each author. After that, we average the F1 score by the weight of each author which is determined by the percentage of their papers that will be assigned. We adopt the weighted average strategy to alleviate the negative effects of some extreme cases, like authors who only have one paper.

3.1.3 *Incorrect Assignment Detection.*

As inevitable cumulative errors brought via the methods of SND and RND greatly affect the efficacy of subsequent assignments, Incorrect Assignment Detection is a vital task to detect and remove wrongly-assigned papers.

Problem 3. ***Incorrect Assignment Detection (IND).*** *Given a conflated author entity* $a^* = \{p_i, \cdots, p_j, p_a, \cdots, p_b, \cdots, p_m, \cdots, p_n\}$ *comprising multiple papers from* K *different authors* $\{a_1, \cdots, a_K\}$, *where* $a_1 = \{p_i, \cdots, p_j\}, a_2 = \{p_a, \cdots, p_b\}$, *and* $a_K = \{p_m, \cdots, p_n\}$. *Assuming* a_1 *covers the highest percentage of papers within* a^*, *we set* $a^* = a_1$. *Consequently, the papers owned by* $\{a_2, \cdots, a_K\}$ *are defined as incorrectly-assigned papers to be detected.*

Evaluation Protocol. We leverage Area Under ROC Curve (AUC), broadly adopted in anomaly detection [22] and Mean Average Precision (MAP), which pays more attention to the rankings of incorrect cases, as the evaluation metrics.

3.1.4 *Discussion.*

The proposed three name disambiguation tasks shed light on the life cycle of concerned name disambiguation problems in online digital libraries. Specifically, the SND task reflects the requirements of building digital libraries at the early stage; the RND task corresponds to the urgent needs of current online platforms; and the IND task is devoted to correcting the accumulated errors of name disambiguation algorithms, which is critical to maintaining the reliability of the name disambiguation system. In addition, the three tasks can serve as the backbone of any other complex name disambiguation tasks. We believe name disambiguation methods, which perform better on these tasks, are powerful enough to handle the majority of name disambiguation situations. Although Zhang and Tang [44] have already proposed similar types of tasks, we improve them by 1) taking the NIL issue into account and formalizing the RND problem into a more general classification problem instead of a ranking problem, 2) standardizing the evaluation protocol of the three tasks, and 3) arranging contests for the first two tasks to prompt their accomplishments.

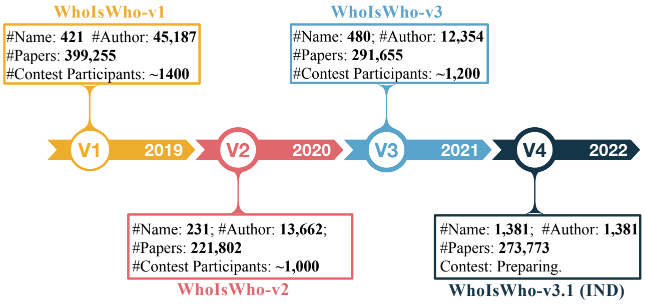

WhoIsWho-v1
#Name: 421 #Author: 45,187
#Papers: 399,255
#Contest Participants: ~1400
WhoIsWho-v3
#Name: 480; #Author: 12,354
#Papers: 291,655
#Contest Participants: ~1,200
V1
2019
V2
2020
V3
2021
V4
2022
#Name: 231; #Author: 13,662;
#Papers: 221,802
#Contest Participants: ~1,000
WhoIsWho-v2
#Name: 1,381; #Author: 1,381
#Papers: 273,773
Contest: Preparing.
WhoIsWho-v3.1 (IND)

Figure 5: The released time of WhoIsWho benchmark and launched contests.

3.2 Historical Contests & Regular Leaderboard

From 2019 to 2022, WhoIsWho periodically released three versions of benchmarks. To promote the development of the community, we sponsored three rounds of name disambiguation contests on BienData[9]. The timeline of released benchmarks and corresponding contests is depicted in Figure 5. To date, more than 3,000 people in the world, have downloaded the WhoIsWho benchmark more than 10,000 times. WhoIsWho has already become one of the most well-known and representative benchmarks of the name disambiguation community. In addition, to assist researchers who are interested in resolving name disambiguation problems at any time, we maintain a regular leaderboard with the contest based on the most recent benchmarks released by WhoIsWho.

In the following part, we briefly revisit the methodologies proposed by contest winners, based on which we conduct an in-depth empirical analysis to probe key factors that may have a significant impact on the performance of name disambiguation methods.

3.2.1 *Methodologies of the Contest Winner.*

We revisit the approaches of contest winners in the first two tasks of SND and RND since they have the best performance to date. How to measure the fine-grained similarities between papers and authors is vital to finding a solution to both tasks. Thus, to measure these similarities, we need to build the interaction between authors and papers, which needs to be primarily explored. In the following part, we skip over some technical details and focus on the strategies to quantify connections between papers and authors.

From-scratch Name Disambiguation. The SND task aims to group the papers written by the same author. The contest winner divides the similarities across papers into two categories.

[9] https://www.biendata.xyz/

Semantic Aspect. The contest winner views the paper's title, venue, organization of authors, year, and keywords as the semantic features, based on which they measure the topical similarities between papers. Specifically, they first learn word2vec [23] embeddings based on the semantic features of all the papers in the WhoIsWho benchmark. Then they project the semantic features of a paper into corresponding word embeddings and average them as the paper embedding. Finally, they calculate the soft semantic similarities between papers based on these semantic embeddings.

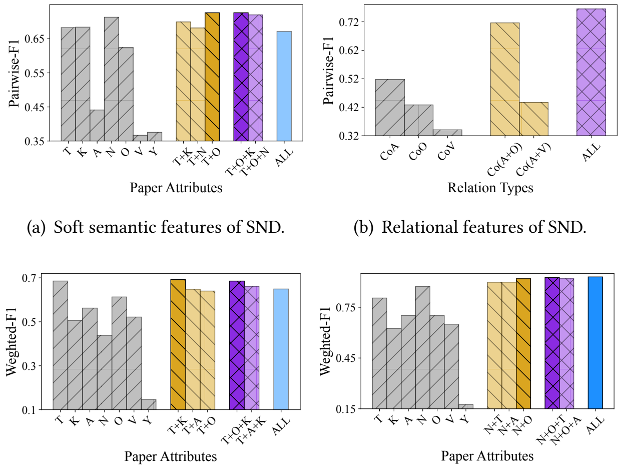
Relational Aspect. The contest winner takes author names and organizations as the relational features of papers. For example, the concurrence of the same author name in two papers reflects their relationships. Specifically, they construct a relational graph by considering papers as nodes and the connections between papers as edges. If two papers have identical coauthors' names, the edge of co-authors is added. When two papers have the same organization for the concerned author, the edge of co-organization is added. After that, they employ the metapath2vec [6] to obtain relational embeddings of papers. Finally, they calculate the relational similarity score between papers based on these relational embeddings.

Furthermore, the contest winner combines the two multi-modal similarities to estimate the final similarities between papers and then uses DBSCAN [7] to obtain the clustering results.

Real-time Name Disambiguation. The RND task focuses on measuring connections between the paper and a collection of papers from each candidate author. The contest winner captures more precise semantic features between unassigned papers and candidate authors than the SND task as follows.

Semantic Aspect. Besides the soft semantic features, i.e., those measured via embedding techniques, they also consider the ad-hoc semantic features, i.e., those measured via hand-crafted features. In terms of the soft semantic features, they identify similarities between the target paper and each paper of the candidate author, just like SND does. Then they adopt aggregation functions to obtain overall similarities between the target paper and all papers of the candidate author. As for the ad-hoc semantic features, they propose 36-dimensional hand-crafted features to explicitly capture the semantic correlations between the target paper and the candidate author. The complete features are listed in Table 6. Finally, they concatenate the soft semantic features and the ad-hoc semantic features to create the final similarity features. Then they adopt ensemble methods to acquire the classification results.

Being aware that the contest winner's methods disregarded the characterization of relationship properties. We make the following hypotheses: 1) Unlike the SND task, which only requires building a relational graph of papers from one name once, the RND task needs to build time-consuming graphs between unassigned papers and corresponding candidate authors with each unassigned paper once. 2) Some ad-hoc features can somewhat capture relational correlations. For example, the coauthor-occurrence feature, which counts the number of coauthors between the target paper and a candidate author, can be viewed as the coauthor edge weight on virtual paper-author graphs. Nevertheless, how to model the relational correlations in the RND task is still under-explored.



(a) Soft semantic features of SND.

(b) Relational features of SND.

(c) Soft semantic features of RND.

(d) Ad-hoc semantic features of RND.

Figure 6: Feature importance on the SND and RND tasks.

Incorrect Assignment Detection. The IND task targets at detecting accumulated incorrect papers, which is important to guarantee the reliability of academic systems. However, there is no available IND benchmark in the current stage. To this end, we have released V3.1 data consisting of 1,000+ authors and 200,000+ papers dedicated to the IND task. To our best knowledge, we are the first to specify and release the corresponding IND benchmark. Furthermore, we are planning a contest based on the released WhoIsWho-v3.1 benchmark for the IND task in a few months.

3.2.2 Discussion. In summary, we observe a crucial insight of establishing a good approach to comprehensively measure the correlations among papers is to intertwine multi-modal features, i.e., semantic and relational features. The contest results show that methods capturing both two aspects of features produce impressive results. Although the contest for the third task IND has not been held, we assume a similar result may be drawn for the IND task, as they also depend on evaluating the agreements among papers.

4 EMPIRICAL FACTOR ANALYSIS

We conduct in-depth ablation studies to understand the effect of various factors on name disambiguation performance. To ensure fair comparisons, we only modify the factors of interest, leaving others unaltered. We adopt metrics defined in WhoIsWho tasks for evaluations. For each experiment, we run 5 trials and report the mean results at the WhoIsWho-v3 validation set.

4.1 Semantic Feature Importance

We study the effects of accessible paper attributes, i.e., title (T), keywords (K), abstract (A), venue/journal (V), year (Y), author names (N), and organizations of authors (O), on the SND and RND tasks.

From-scratch Name Disambiguation. To perform the soft semantic feature analysis, we adopt a similar implementation pipeline with the contest winner method while exploring different attributes.

Results. The results are shown in Fig. 6(a). The fields of title, keywords, author name, and organization play a more significant effect

Table 2: Performance (%) of different feature modalities (semantic or relational) and their combinations.

Tasks	Semantic Feature		Relational Feature		All
	Soft	**Ad-hoc**	**Relation**	**Ego**	
SND	72.64	-	76.52	-	**88.46**
RND	76.55	93.01	-	72.92	**93.40**

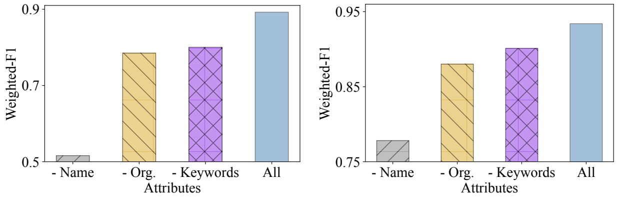
on disambiguation than others. The field of abstract contains much redundant words and noises. The venue and year also fail to access the similarities among papers. **(1) Combining consistent attributes might better express semantic correlations.** We combine these four effective single features, as shown in the yellow bars. The title + name even perform worse than its constituent single attribute. We speculate that compared to title, keywords, and organization which have semantic correlations among papers, the author's name has more linguistic qualities. Thus, combining two disparity attributes result in performance degradation. The performance of the title improves when it is paired with keywords or organization, suggesting that a consistent attributes combination may better express semantic correlations. **(2) Combining title, keywords, and organization performs the best**. Finally, the combination of title, keywords, and organization, represented by the blue bars, performs better than mixing all the attributes together, represented by the blue bar. This suggests that adding more attributes without calibrating may result in noise and lower performance.

Real-time Name Disambiguation. We also adopt the RND contest winner method's implementation pipeline. In addition to the soft semantic feature analysis, we also explore how various paper attributes affect the performance of name disambiguation methods using hand-crafted features listed in Table 6.

Results. The results are shown in Figure 6(c) and 6(d). **(1) The soft semantic features share a similar trend on both tasks.** Regarding the soft semantic features, Figure 6(c) and 6(a) show that both tasks share a common trend: 1) the attributes of title, keywords, and organization perform well and 2) the combination of title, keywords, and organization performs better than just mixing all the considered features. This is expected because both tasks measure the agreements between papers and authors via the same soft semantic feature modality. In terms of the ad-hoc semantic features, shown in Figure 6(d), the author name is the most effective factor to determine the performance of algorithms. **(2) Mixing all attributes performs best.** Surprisingly, the blue bar, which represents the performance of combining all features, outperforms other combination patterns, suggesting that despite falling into the semantic feature category, the ad-hoc feature characterization frameworks have different underlying biases than the soft one.

4.2 Relational Feature Importance

Empirically, the fields of the author name and venue show a greater relational dependency between papers. Moreover, the field of organization has both relational and semantic characteristics. Therefore, we build three relational edges between papers: CoAuthor, where two papers have a relationship only if they share the same author name; CoOrg, where two papers have a relationship only if they share the same affiliations[10]; CoVenue, where two papers have a relationship only if they are published in the same venue or journal.



(a) SND.

(b) RND.

Figure 7: Realistic cases analysis.

From-scratch Name Disambiguation. We also follow the implementation pipeline of the contest winner method to obtain the relational paper embeddings in the built rational graphs, while exploring the effects of different relational edges.

Results. Fig. 6(b) presents the performance of using different relation types. The grey bars, which show that CoAuthor performs the best among the single relational types, suggest that the author name has more important relational information than the semantic information. CoVenue performs the worst because massive papers from various domains may be published in the same venue/journal. Combining all three features yields the best results when taking into account the mixed outcomes, represented by the yellow and purple bars, which is consistent with empirical findings from Section 4.1 that consistent attribute combinations can improve performance.

4.3 Feature Modality Importance

We explore how the semantic and relational features affect the effectiveness of disambiguation. We conduct a thorough examination about the combination patterns of multi-modal features to see which ones perform the best. For the SND task, we leverage the paper attributes of title, keywords, and organization as the soft semantic features. For the relational features, we adopt three relation types, i.e., CoAuthor, CoOrg, and CoVenue. For the RND task, in addition to the soft and ad-hoc semantic features used in Section 4.1, we build the heterogeneous ego-graph for each pair of the target paper and a candidate author in order to add relational features.

Results. Table 2 shows the performance of single feature modalities and their combinations. **(1) Mixing multi-modal features performs best.** We observe the single modality, i.e., semantic or relational features, underperforms their combination patterns, i.e., SND-all and RND-all, indicating that the semantic and relational features are complementary to one another. However, for the RND task, the ad-hoc semantic features alone can compete with their combinations. The relational features make marginal improvements. That explains why the best contest approach in this task doesn't take advantage of relational features. Therefore, how to effectively incorporate relational aspects is still an open question.

4.4 Overall Evaluation

In this section, we compare the proposed SND-all and RND-all frameworks with existing state-of-the-art name disambiguation

[10] We only take the organization of the author to be disambiguated into consideration.

Table 3: Performance of from-scratch name disambiguation (%).

Model	Pairwise-Precision	Pairwise-Recall	Pairwise-F1
G/L-Emb	50.77	84.64	63.48
LAND	61.20	61.12	61.12
IUAD	58.82	65.22	61.63
Contest Winner	82.72	**96.59**	89.14
SND-all	**83.06**	96.35	**89.22**

Table 4: Performance of real-time name disambiguation (%).

Model	Weighted-Precision	Weighted-Recall	Weighted-F1
IUAD	75.53	90.49	82.34
CONNA	90.54	89.22	89.64
CONNA+Ad-hoc.	90.23	92.64	91.14
Contest Winner	92.09	**94.95**	93.49
RND-all	**92.14**	94.94	**93.52**

methods of the SND and RND tasks[11], respectively. The experimental results are performed on the WhoIsWho-v3 test set.

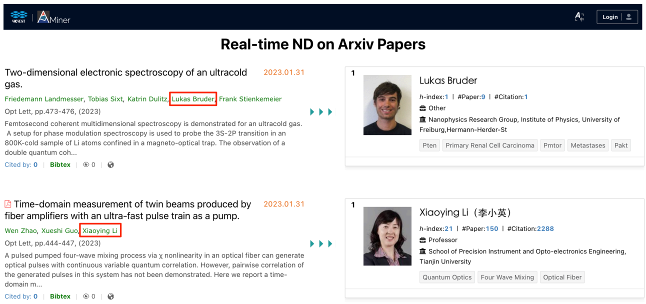
Compared Baselines. Besides the methods of the contest winner, we also compare other prevailing methods,

From-scratch Name Disambiguation. **G/L-Emb** [47] learns paper embeddings on a global paper-paper network and then fine-tunes the embeddings on a local paper-paper network built for each name by graph auto-encoding. **LAND** [28] constructs the heterogeneous knowledge graphs (KGs) with papers and authors and leverages KGs embedding techniques to obtain the node embeddings, based on which it performs clustering methods. **IUAD** [18] determines the authorship of papers via reconstructing the collaboration network where nodes are authors and edges are the coauthor relationships. **SND-all** is our proposed strong baseline based on the empirical studies in Section 4.3. It mixes soft semantic features with heterogeneous relational graph features to perform the SND task.

Real-time Name Disambiguation. We adopt the following baselines, **IUAD** [18] is also employed to perform the RND task via reconstructing the collaboration network between newly-arrived papers and existing authors. **CONNA** [3] is an interaction-based model. The basic interactions are built between the token embeddings of two attributes, then different attributes matrices are aggregated as the paper-level interactions, and finally, the paper-level matrices are aggregated as author-level interactions, and **CONNA+Ad-hoc.** is also a combination methodology that incorporates hand-crafted features into CONNA framework introduced in [3]. For fair comparisons, we leverage features used in Table 6. **RND-all** is also our proposed method based on the findings in Section 4.3. It adopts the soft and ad-hoc semantic features used in Section 4.1. It also builds heterogeneous ego-graphs as relational features. The two features are combined to make predictions.

Other prevailing methods, such as Louppe et al. [20], Zhang et al. [41], Camel [43], etc, are empirically proven to be less powerful than the adopted baselines, and thus are ignored in the experiments.

[11] We only consider the baselines with the released code.



Real-time ND on Arxiv Papers

Two-dimensional electronic spectroscopy of an ultracold gas. 2023.01.31
Friedemann Landmesser, Tobias Sixt, Katrin Dulitz, Lukas Bruder, Frank Stienkemeier
Opt Lett, pp.473-476, (2023)
Femtosecond coherent multidimensional spectroscopy is demonstrated for an ultracold gas. A setup for phase modulation spectroscopy is used to probe the 3S-2P transition in an 800K-cold sample of Li atoms confined in a magneto-optical trap. The observation of a double quantum coh...
Cited by: 0 | Bibtex | 0 |

1 Lukas Bruder
h-index:1 | #Paper:9 | #Citation:1
Other
Nanophysics Research Group, Institute of Physics, University of Freiburg,Hermann-Herder-St
Pten | Primary Renal Cell Carcinoma | Pmtor | Metastases | Pakt

Time-domain measurement of twin beams produced by fiber amplifiers with an ultra-fast pulse train as a pump. 2023.01.31
Wen Zhao, Xueshi Guo, Xiaoying Li
Opt Lett, pp.444-447, (2023)
A pulsed pumped four-wave mixing process via χ nonlinearity in an optical fiber can generate optical pulses with continuous variable quantum correlation. However, pairwise correlation of the generated pulses in this system has not been demonstrated. Here we report a time-domain m...
Cited by: 0 | Bibtex | 0 |

1 Xiaoying Li（李小英）
h-index:21 | #Paper:150 | #Citation:2288
Professor
School of Precision Instrument and Opto-electronics Engineering, Tianjin University
Quantum Optics | Four Wave Mixing | Optical Fiber

Figure 8: A demo about disambiguating daily papers from arXiv.org.

Results. Table 3 and Table 4 demonstrate the performance of various name disambiguation methods on the two tasks. The proposed SND-all, RND-all, and the contest winner significantly outperform other baselines by 25.74~28.10% pairwise-F1 and 2.35~11.80% weighted-F1 respectively. The significant performance gap between our proposed method and baselines proposed in recent research sheds light on the capability of prevailing name disambiguation methods is still far from satisfactory, which also reflects the significance of the WhoIsWho benchmark. Moreover, our proposed simple yet effective methods slightly outperform the contest winner method, suggesting that our empirical factor analysis successfully captures the essential components that enhance the effectiveness of name disambiguation methods.

4.5 Performance in Realistic Cases

Papers in the WhoIsWho benchmark always contain rich information since annotators prefer to work on papers owning abundant attributes that provide helpful evidence to support their decisions. Unfortunately, online digital libraries always contain a lot of papers with sparse attributes, meaning that papers with multiple attributes are absent. Taking AMiner for example, almost half of the newly-arrived papers lack the attributes of organizations. To understand the online name disambiguation scenarios on these papers, we perform SND-all and RND-all on these sparse-attributes cases.

Results. The results are shown in Figure 7. Among these, the papers without author names perform worst, dropping 36.72% pairwise-F1 and 15.58% weighted-F1. The absence of the attribute of organizations or keywords also significantly degenerates the online performance of name disambiguation algorithms on both tasks by dropping 8.97~10.51% pairwise-F1 and 4.28~6.40% weighted-F1. The results indicate that the online name disambiguation scenario is even more sophisticated than what we show on WhoIsWho. We will update datasets with sparse attributes to encourage more real-world online name disambiguation scenarios in the future.

5 WHOISWHO TOOLKIT

By automating data loading, feature creation, model construction, and evaluation processes, the WhoIsWho toolkit is easy for researchers to use and let them develop new name disambiguation approaches. The overview of the toolkit pipeline is illustrated in Figure 9. The toolkit is fully compatible with PyTorch and its associated deep learning libraries, such as Hugging face [39]. Additionally, the toolkit offers library-agnostic dataset objects that can be used

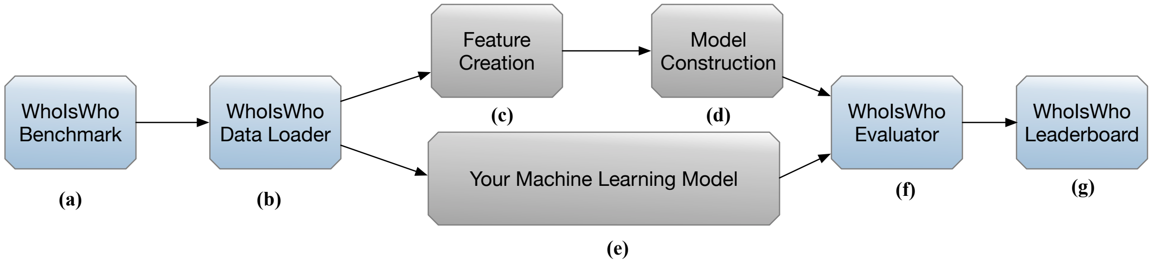


Figure 9: Overview of the WhoIsWho toolkit pipeline. (a) WhoIsWho provides the large-scale benchmark with high ambiguity and large quantity. **(b)** The WhoIsWho toolkit automates dataset processing and splitting. That is, the data loader automatically loads arbitrary versions of datasets, and further split the datasets in a standardized manner. **(c)** WhoIsWho toolkit provides flexible modules for feature creation including semantic features characterization and relational graph construction, based on that **(d)** researchers can adopt models pre-defined in the toolkit library for training and prediction. Moreover, **(e)** researchers can build their own feature processing process and develop ML models. **(f)** WhoIsWho evaluates the model in a task-dependent manner and outputs the model performance on the validation set. Finally, **(g)** WhoIsWho provides public leaderboards to keep track of recent advances.

by any other Python deep learning frameworks such as Tensorflow [1]. To keep things simple, we concentrate on building a basic RND method using PyTorch shown in Listing 1. More details refer to https://github.com/THUDM/WhoIsWho.

Disambiguating Arxiv Papers. We deploy the RND-all method implemented by our toolkit on AMiner to disambiguate daily papers from arXiv.org on-the-fly. A demo page is depicted in Figure 8. The details refer to Section A.8. We manually check the latest 100 disambiguation results reflecting that 90% assignments are accurate.

6 RELATED WORK

Here, we recall the prevailing name disambiguation datasets and the state-of-the-art name disambiguation algorithms.

Name Disambiguation Datasets. The size of datasets heavily influences the performance of name disambiguation algorithms. To address the problem, the community has created a large number of name disambiguation datasets recently. Among them, several efforts directly harvest datasets from existing digital libraries, including PubMed [40, 45], DBLP [14], etc. [17, 37, 46]. However, the assignment mistakes, as shown in Figure 2, hamper the development of effective algorithms [4, 44]. Others attempt to manually label a small amount of data based on noisy data from existing databases to reduce data noises [11, 13, 20, 24, 26, 30, 32, 34, 38, 47]. Most of them, however, do not have sufficient instances, as shown in Figure 1. The detailed data statistics refer to Table 5. Some of them have restricted scopes, for example, SCAD-zbMATH [24] is customized for a mathematical domain. The fragile inductive bias affects the performance and generalization of name disambiguation methods that are trained on these datasets. Subramanian et al. [31] build a unified dataset via aggregating several small scales of datasets. However, the quality of constituents has not been checked.

Practical Tasks & Algorithms. Most efforts focus on the SND task. Generally, they operate via three steps: blocking, paper similarity matching, and clustering. Backes [2] discusses the name-blocking step. Several works lay emphasis on paper similarity matching and clustering steps. Early attempts designed hand-crafted similarity metrics [5] to measure paper similarities. Then, researchers discover that constructing paper similarity graphs excels at learning high-order similarity [6, 8, 13, 27, 47]. As for clustering steps, the clustering methods such as hierarchical agglomerative clustering and DBSCAN are adopted. Among them, DBSCAN is preferred by practitioners as there is no need to specify the cluster number.

The RND task, which aims to assign newly-arrived papers to existing authors, is a more practicable scenario for online academic systems, Besides adopted baselines, Qian et al. [26] predict the likelihood of a paper being written by a specific author via the attributes of coauthor and keyword. Pooja et al. [25] utilize dynamic graph embedding to model evolving graphs. Several works [18, 42] further employ a probabilistic model for online paper assignments.

Inevitable cumulative errors will greatly affect the efficacy of name disambiguation algorithms. Thus, the IND task is vital to guarantee the reliability of academic systems. Unfortunately, the issue has not received much attention [4].

Previous methods are usually evaluated on diverse small-scale datasets, which hamper the development of the community. Thus, a large-scale benchmark, a regular leaderboard with comprehensive tasks, together with an easy-to-use toolkit for web-scale academic name disambiguation should be concerned.

7 CONCLUSIONS

This paper delivers WhoIsWho including a benchmark, a leaderboard, and a toolkit for web-scale academic name disambiguation. Specifically, the large-scale benchmark with high ambiguity enables the devising of robust algorithms. Sponsored contests with two tracks promote the advances of the name disambiguation community. A regular leaderboard is publicly available to keep track of recent advances. An easy-to-use toolkit is designed to allow end users to rapidly build their own algorithm and publish their results on a regular leaderboard that records recent advances. In summary, WhoIsWho is an ongoing, community-driven, open-source project. We also encourage contributions from the community.

Acknowledgments. This work was supported by Technology and Innovation Major Project of the Ministry of Science and Technology of China under Grant 2020AAA0108400 and 2020AAA0108402, the NSF of China for Distinguished Young Scholars (No. 61825602), NSF of China (No. 62076245, 62276148), CCF-Zhipu202306, and the Public Computing Cloud at Renmin University of China.

REFERENCES

[1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. 2016. Tensorflow: a system for large-scale machine learning.. In *Osdi*, Vol. 16. Savannah, GA, USA, 265–283.

[2] Tobias Backes. 2018. The impact of name-matching and blocking on author disambiguation. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. 803–812.

[3] Bo Chen, Jing Zhang, Jie Tang, Lingfan Cai, Zhaoyu Wang, Shu Zhao, Hong Chen, and Cuiping Li. 2020. CONNA: Addressing Name Disambiguation on The Fly. *IEEE Transactions on Knowledge and Data Engineering* (2020).

[4] Bo Chen, Jing Zhang, Xiaokang Zhang, Yuxiao Dong, Jian Song, Peng Zhang, Kaibo Xu, Evgeny Kharlamov, and Jie Tang. 2022. GCCAD: Graph Contrastive Learning for Anomaly Detection. *TKDE'22* (2022).

[5] Ricardo G Cota, Anderson A Ferreira, Cristiano Nascimento, Marcos André Gonçalves, and Alberto HF Laender. 2010. An unsupervised heuristic-based hierarchical method for name disambiguation in bibliographic citations. *Journal of the American Society for Information Science and Technology* 61, 9 (2010), 1853–1870.

[6] Yuxiao Dong, Nitesh V Chawla, and Ananthram Swami. 2017. metapath2vec: Scalable representation learning for heterogeneous networks. In *KDD'17*. 135–144.

[7] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise.. In *kdd*, Vol. 96. 226–231.

[8] Xiaoming Fan, Jianyong Wang, Xu Pu, Lizhu Zhou, and Bing Lv. 2011. On graph-based name disambiguation. *Journal of Data and Information Quality* 2, 2 (2011), 1–23.

[9] Janaína Gomide, Hugo Kling, and Daniel Figueiredo. 2017. Name usage pattern in the synonym ambiguity problem in bibliographic data. *Scientometrics* 112, 2 (2017), 747–766.

[10] Michael Gusenbauer. 2019. Google Scholar to overshadow them all? Comparing the sizes of 12 academic search engines and bibliographic databases. *Scientometrics* 118, 1 (2019), 177–214.

[11] Hui Han, Hongyuan Zha, and C Lee Giles. 2005. Name disambiguation in author citations using a k-way spectral clustering method. In *Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries*. 334–343.

[12] Jorge E Hirsch. 2005. An index to quantify an individual's scientific research output. *PNAS'05* 102, 46 (2005), 16569–16572.

[13] In-Su Kang, Pyung Kim, Seungwoo Lee, Hanmin Jung, and Beom-Jong You. 2011. Construction of a large-scale test set for author disambiguation. *Information Processing & Management* 47, 3 (2011), 452–465.

[14] Jinseok Kim. 2018. Evaluating author name disambiguation for digital libraries: A case of DBLP. *Scientometrics* 116, 3 (2018), 1867–1886.

[15] Jinseok Kim and Jana Diesner. 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 (2016), 1446–1461.

[16] Jinseok Kim, Jinmo Kim, and Jason Owen-Smith. 2019. Generating automatically labeled data for author name disambiguation: An iterative clustering method. *Scientometrics* 118, 1 (2019), 253–280.

[17] Jinseok Kim and Jason Owen-Smith. 2021. ORCID-linked labeled data for evaluating author name disambiguation at scale. *Scientometrics* 126, 3 (2021), 2057–2083.

[18] Na Li, Renyu Zhu, Xiaoxu Zhou, Xiangnan He, Wenyuan Cai, Ming Gao, and Aoying Zhou. 2021. On disambiguating authors: Collaboration network reconstruction in a bottom-up manner. In *2021 IEEE 37th International Conference on Data Engineering (ICDE)*. IEEE, 888–899.

[19] Xiao Liu, Da Yin, Jingnan Zheng, Xingjian Zhang, Peng Zhang, Hongxia Yang, Yuxiao Dong, and Jie Tang. 2022. Oag-bert: Towards a unified backbone language model for academic knowledge services. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 3418–3428.

[20] Gilles Louppe, Hussein T Al-Natsheh, Mateusz Suszhanik, and Eamonn James Maguire. 2016. Ethnicity sensitive author disambiguation using semi-supervised learning. In *KESW'16*. 272–287.

[21] Dongsheng Luo, Shuai Ma, Yaowei Yan, Chunming Hu, Xiang Zhang, and Jinpeng Huai. 2020. A collective approach to scholar name disambiguation. *IEEE Transactions on Knowledge and Data Engineering* 34, 5 (2020).

[22] Xiaoxiao Ma, Jia Wu, Shan Xue, Jian Yang, Chuan Zhou, Quan Z Sheng, Hui Xiong, and Leman Akoglu. 2021. A comprehensive survey on graph anomaly detection with deep learning. *IEEE Transactions on Knowledge and Data Engineering* (2021).

[23] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[24] Mark-Christoph Müller, Florian Reitz, and Nicolas Roy. 2017. Data sets for author name disambiguation: an empirical analysis and a new resource. *Scientometrics* 111, 3 (2017), 1467–1500.

[25] KM Pooja, Samrat Mondal, and Joydeep Chandra. 2022. Online author name disambiguation in evolving digital library. *Neurocomputing* 493 (2022), 1–14.

[26] Yanan Qian, Qinghua Zheng, Tetsuya Sakai, Junting Ye, and Jun Liu. 2015. Dynamic author name disambiguation for growing digital libraries. *Information Retrieval Journal* 18, 5 (2015), 379–412.

[27] Ziyue Qiao, Yi Du, Yanjie Fu, Pengfei Wang, and Yuanchun Zhou. 2019. Unsupervised author disambiguation using heterogeneous graph convolutional network embedding. In *IEEE Big Data'19*. IEEE, 910–919.

[28] Cristian Santini, Genet Asefa Gesese, Silvio Peroni, Aldo Gangemi, Harald Sack, and Mehwish Alam. 2022. A knowledge graph embeddings based approach for author name disambiguation using literals. *Scientometrics* 127, 8 (2022), 4887–4912.

[29] Qiaomu Shen, Tongshuang Wu, Haiyan Yang, Yanhong Wu, Huamin Qu, and Weiwei Cui. 2016. Nameclarifier: A visual analytics system for author name disambiguation. *IEEE transactions on visualization and computer graphics* 23, 1 (2016), 141–150.

[30] Min Song, Erin Hea-Jin Kim, and Ha Jin Kim. 2015. Exploring author name disambiguation on PubMed-scale. *Journal of informetrics* 9, 4 (2015), 924–941.

[31] Shivashankar Subramanian, Daniel King, Doug Downey, and Sergey Feldman. 2021. S2and: A benchmark and evaluation system for author name disambiguation. In *JCDL'21*. IEEE, 170–179.

[32] Jie Tang, Alvis CM Fong, Bo Wang, and Jing Zhang. 2012. A unified probabilistic framework for name disambiguation in digital library. *IEEE Transactions on Knowledge and Data Engineering* 24, 6 (2012), 975–987.

[33] Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. 2008. Arnetminer: extraction and mining of academic social networks. In *SIGKDD'08*. 990–998.

[34] Dina Vishnyakova, Raul Rodriguez-Esteban, and Fabio Rinaldi. 2019. A new approach and gold standard toward author disambiguation in MEDLINE. *Journal of the American Medical Informatics Association* 26, 10 (2019), 1037–1045.

[35] Haiwen Wang, Ruijie Wan, Chuan Wen, Shuhao Li, Yuting Jia, Weinan Zhang, and Xinbing Wang. 2020. Author name disambiguation on heterogeneous information network with adversarial representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 238–245.

[36] Kuansan Wang, Zhihong Shen, Chiyuan Huang, Chieh-Han Wu, Yuxiao Dong, and Anshul Kanakia. 2020. Microsoft academic graph: When experts are not enough. *Quantitative Science Studies* 1, 1 (2020), 396–413.

[37] Ruijie Wang, Yuchen Yan, Jialu Wang, Yuting Jia, Ye Zhang, Weinan Zhang, and Xinbing Wang. 2018. Acekg: A large-scale knowledge graph for academic data mining. In *Proceedings of the 27th ACM international conference on information and knowledge management*. 1487–1490.

[38] Xuezhi Wang, Jie Tang, Hong Cheng, and S Yu Philip. 2011. Adana: Active name disambiguation. In *ICDM'11*. 794–803.

[39] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *EMNLP'20*. Association for Computational Linguistics, Online, 38–45. https://www.aclweb.org/anthology/2020.emnlp-demos.6

[40] Tong Zeng and Daniel E Acuna. 2020. Large-scale author name disambiguation using approximate network structures. In *International Conference on Computational Social Science*.

[41] Baichuan Zhang and Mohammad Al Hasan. 2017. Name disambiguation in anonymized graphs using network embedding. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. 1239–1248.

[42] Baichuan Zhang, Murat Dundar, Vachik Dave, and Mohammad Hasan. 2019. Dirichlet process Gaussian mixture for active online name disambiguation by particle filter. In *JCDL'19*. IEEE, 269–278.

[43] Chuxu Zhang, Chao Huang, Lu Yu, Xiangliang Zhang, and Nitesh V Chawla. 2018. Camel: Content-Aware and Meta-path Augmented Metric Learning for Author Identification. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*. 709–718.

[44] Jing Zhang and Jie Tang. 2021. Name disambiguation in AMiner. *Science China-information sciences* 64, 4 (2021), 10–1007.

[45] Li Zhang, Yong Huang, Qikai Cheng, and Wei Lu. 2020. Mining Author Identifiers for PubMed by Linking to Open Bibliographic Databases. In *2020 IEEE 20th International Conference on Software Quality, Reliability and Security Companion (QRS-C)*. IEEE, 209–212.

[46] Li Zhang, Wei Lu, and Jinqing Yang. 2021. *arXiv preprint arXiv:2104.01821* (2021).

[47] Yutao Zhang, Fanjin Zhang, Peiran Yao, and Jie Tang. 2018. Name Disambiguation in AMiner: Clustering, Maintenance, and Human in the Loop.. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1002–1011.

[48] Zhenyu Zhang, Bowen Yu, Tingwen Liu, and Dong Wang. 2020. Strong Baselines for Author Name Disambiguation with and Without Neural Networks. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 369–381.

[49] Xin Zheng, Pengyu Zhang, Yanjie Cui, Rong Du, and Yong Zhang. 2021. Dual-Channel Heterogeneous Graph Network for Author Name Disambiguation. *Information* 12, 9 (2021), 383.

```
# Module-1: Data Loading
from whoiswho.dataset import LoadData, SplitDataRND
# Load specific versions of dataset.
train = LoadData(name="v3", type="train", partition=None)
# Split data into unassigned papers and candidate authors
unassigns, candidates = SplitDataRND(train, split="time",
      ratio=0.2)

# Modules-2: Feature Creation
from whoiswho.featureGenerator import AdHocFeatures
# Extract default n-dimensional ad-hoc features.
pos_feats, neg_feats = AdHocFeatures(unassigns,
     candidates, feature_mode="default", negatives=3)

# Module-3: Model Construction
from whoiswho.loadmodel import ClassficationModels
# build a basic classfication model.
predictor=ClassficationModels(type="MLP", ensemble=False)
# Automatic training
from whoiswho.training import AutoTrainRND
predictor = AutoTrainRND(inputs = (pos_feats, neg_feats),
      predictor, epoch=1, bs=1, early_stop=None)

# Modules-4: Evaluation on the validation data
# Load validation data
unassigns, candidates, gt = LoadData("v3", type="Valid",
     task="RND")
# Assign unassigned papers
assign_res = predictor.predict(unassigns, candidates)
# Evaluate the RND results
from whoiswho.evaluation import RNDeval
weighted_Precision, weighted_Recall, weighted_F1 =
     RNDeval(assign_res, gt)
```

Listing 1: Basic RND algorithm

A APPENDICES

A.1 WhoIsWho Toolkit Pipeline.

Figure 9 demonstrates the overview of the WhoIsWho toolkit pipeline. A toy example of building basic RND algorithms is shown in Listing 1.

A.2 Data Organizations in WhoIsWho Benchmark.

To date, WhoIsWho released three versions of datasets, i.e., WhoIsWho-v1, -v2, and -v3, with one specified dataset, WhoIsWho-v3.1, for the IND task. Among them, v1,v2,v3 datasets have the same organizations with the SND and RND tasks. Here, we briefly review the data organizations.

A.2.1 WhoIsWho-v1/v2/v3.

The datasets are organized into the format of a two-level dictionary, i.e., names-authors-papers as shown in Listing 2. The key of the first-level dictionary is author names and the value is author profiles with the "same name". The term "same name" refers to the ways to unify names using name blocking techniques [2, 14], such as moving the last name to the first or preserving all name initials but the last name. For example, the variants of "Jing Zhang" are "Zhang Jing", "J. Zhang", and "Z. Jing". The author profiles are also organized as a dictionary with the key being author IDs and the value being paper IDs of the author. For each paper, we collect the title, author names, organizations of all the authors, keywords, abstract, publication year, and venue (conference or journal) as its attributes. A toy example of the paper with ID "9PgiwDo7" is shown in Listing 5.

A.2.2 WhoIsWho-v3.1 for the IND task.

WhoIsWho-v3.1 is organized as a one-level dictionary, as the IND task aims to detect and remove the error papers within each author. The key of the dictionary is author IDs, and the value is the papers belonging to the author, i.e., the normal data, and the manually detected error papers, i.e., the outliers. A demo case is present in Listing 3.

A.3 Data Split of WhoIsWho Contest

We claim the process of splitting the WhoIsWho datasets into training, validation, and test sets for the contest of three tasks, i.e., SND, RND, and IND, respectively.

From-scratch Name Disambiguation. The SND task targets at partitioning papers of the author's name into different groups. Each group contains papers from the same author while papers in different groups belong to different authors. Thus, we first split the datasets into training, validation, and test set via the level of author names following specific ratios. Then for the validation and test sets, we delete the authorships between authors and papers in each name as shown in Listing 4. Researchers should correctly cluster papers belonging to the same author into the same group.

Real-time Name Disambiguation. The RND task aims at assigning newly-arrived papers to existing authors. Thus, firstly, we also split the datasets into training, validation, and test sets via the level of author names following specific ratios. Then for the validation and test sets, we sort papers within each author via the published year in ascending order. To simulate the real RND scenario, we treat the latest papers as the new-arrived unassigned papers and the remains as existing author profiles in each author, as shown in Listing 6. We also add several NIL papers, i.e., papers that can not be assigned to any existing author profiles, to the unassigned papers. Researchers need not only correctly assign papers to the right author, but also to distinguish NIL papers.

Incorrect Assignment Detection. The IND task is designed for detecting and removing the error papers within each author. Concretely, we construct the dataset of the IND task as follows: 1) as illustrated in Table 1, the overall data annotation pipeline includes a 'Clean' step, during which annotators remove or split obviously incorrect papers from the concerned author. The subsequent 'Validate' step allows annotators to perform the same 'Clean' function on incorrectly assigned papers that are more difficult to identify. These two stages provide a sufficient number of incorrectly assigned papers to be detected in the IND task. 2) Some authors manually maintain their profiles, such as adding new papers or removing papers that do not belong to them. We also collect the removed papers as targets for detection in the IND task. Thereby, we split the training, validation, and test sets via the author groups.

```
{
    "guanhua_du": { # Author name,
        "zsOOUZxZ": [ # Author IDs,
            "QDMcmF8V", # Paper IDs,
            "9PgiwDo7"
        ]
    },
    "bin_yu": {
        "HoH18DsE": [
            "VMYs96sn",
            "YT4XzThC",
            "S3RARClD",
            "wM8dXlKT"
        ],
        "WYZVZfO0": [
            "OTvYjfnt",
            "EzzruFin"
        ],
        "9BMlVP0u": [
            "brlzUqnH",
            "HWAfXDPx",
            "RZPhHOMm",
            "vebukM2n"
        ]
    }
}
```

Listing 2: Data organizations of WhoIsWho-v1/v2/v3

```
{
    "HoH18DsE": { # Author IDs,
        "name": "xxx", # Name of the author,
        "normal_data":[ # Papers belong to the author,
        "VMYs96sn",
        "YT4XzThC",
        "S3RARClD",
        "wM8dXlKT"
        ],
        "outliers":[ # Papers wrongly assigned to the
 author,
        "OTvYjfnt",
        "EzzruFin"
        ]
    },
}
```

Listing 3: Data organizations of WhoIsWho-v3.1

```
{
    "bin_yu": {
        [
            "VMYs96sn",
            "YT4XzThC",
            "S3RARClD",
            "wM8dXlKT"
            "OTvYjfnt",
            "EzzruFin",
            "brlzUqnH",
            "HWAfXDPx",
            "RZPhHOMm",
            "vebukM2n"
        ]
    }
}
```

Listing 4: Data organizations of the validation/test set in the RND task

```
{
    "9PgiwDo7": {
        "id": "9PgiwDo7",
        "title": "Constrained phase transformation of
 prestrained TiNi fibers embedded in metal matrix
 smart composite",
        "abstract": "The reverse martensitic
 transformation of TiNi fibers embedded in a metal
  matrix smart composite has been studied. Results
  show that under the influence of temperature and
  recovery stress, the reverse martensitic
 transformation of TiNi fibers can be divided into
  two parts with different kinetic characteristics
 : the reverse transformation of self-
 accommodating martensite and that of oriented
 martensite. The relationship between martensitic
 fraction and temperature was calculated.",
        "keywords": [
            "null"
        ],
        "authors": [
            {
                "name": "Yanjun Zheng",
                "org": "Dalian University of
 Technology(Dalian University of Technology,Dalian
  Univ. of Technol.),Dalian,China"
            },
            {
                "name": "Lishan Cui",
                "org": "China University of Petroleum
 - Beijing(China University of Petroleum,
 University of Petroleum),Beijing,China"
            },
            {
                "name": "Dan Zhu",
                "org": "China University of Petroleum
 - Beijing(China University of Petroleum,
 University of Petroleum),Beijing,China"
            },
            {
                "name": "Dazhi Yang",
                "org": "Dalian University of
 Technology(Dalian University of Technology,Dalian
  Univ. of Technol.),Dalian,China"
            }
        ],
        "venue": "Materials Letters",
        "year": 2000
    }
}
```

Listing 5: Data organizations of WhoIsWho-v1/v2/v3

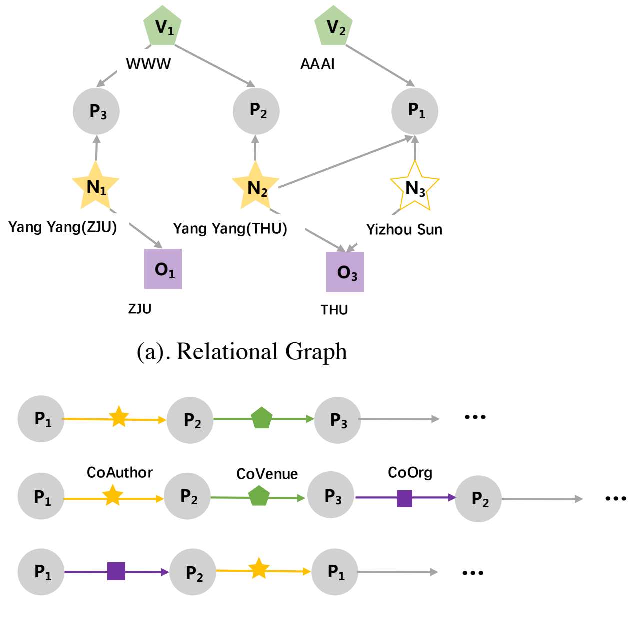
```
# Existing author profiles.
{
    "HoH18DsE": [
        "VMYs96sn",
        "YT4XzThC",
        "S3RARClD",
        "wM8dXlKT"
    ],
    "9BMlVP0u": [
        "brlzUqnH",
        "HWAfXDPx",
        "RZPhHOMm",
    ]
}

# Newly-arrived unassigned papers.
{
    [
    "vebukM2n-1", # "-1" means the 2nd (Author index
     begins at 0) author should be disambiguated.
    "RZPhHOMm-3"
    ]
}
```

Listing 6: Data organizations of WhoIsWho-v1/v2/v3

A.4 Running Environment

We implement all the experiments model by PyTorch and run the code on an Enterprise Linux Server with 40 Intel(R) Xeon(R) CPU cores (E5-2640 v4 @ 2.40GHz and 252G memory) and 1 NVIDIA Tesla V100 GPU core (32G memory).



(a). Relational Graph

(b). Meta-paths

Figure 10: Heterogeneous relational graph and meta-paths adopted in the SND-all method.

A.5 Implementation Details of the SND-all

Here, we introduce the technical details of the SND-all method, consists of the soft semantic features, the heterogeneous relational graph features, and the combination patterns.

A.5.1 Soft Semantic features. We extract all paper attributes from the WhoIsWho benchmark, including title, abstract, venue, keywords, year, and author's organization, to train the word embedding model using word2vec implemented by genism with embedding dimensions set to 100. The hyper-parameters are defined as follows: min_count=2, window=5, negative=5. Then, for each paper, we average the word embedding of the title, keywords, and the organization of the author being clarified as the semantic feature as the final paper embeddings. Specifically, the fields of title and keywords are directly picked from the paper attributes. While for the organization, we extract the organizations of the target author to be disambiguated.

A.5.2 Relational Semantic Features. For each target name, we build a heterogeneous network among papers with three relation types CoAuthor, CoOrg, and CoVenue. The adopted meta-paths are shown in Figure 10. The weights of the heterogeneous edge are determined by the number of co-occurrences, i.e., the number of co-authors for CoAuthor, the number of co-occurring words in organization for CoOrg, and also the number of co-occurring words in venue/journal for CoVenue. We set a random probability as 0.1 that the paper walks to neighboring papers via the CoVenue edge to reduce the impact of common word noises in the venue. The walk length is set to 20, and the number of random walks for each node is set to 5. Then, we save the collection of paths for each paper and train them using word2vec. The isolated papers that lack neighbors are preserved for post-match processing.

A.5.3 Feature Combinations. Finally, we calculate pairwise distances among papers via the semantic and relational embeddings of papers, respectively. Then we simply add the two similarity matrices to form the combined similarity matrix. To obtain the clustering results, we adopt DBSCAN to generate clusters with db_eps as 0.2 and db_min as 4. For outliers that can not be assigned to any groups, we conduct a rule-based method for post-matching. Specifically, we conduct character matching on names of co-authors as CoA, titles, and keywords of two papers as CoW, and we adopt tanimoto distance for calculating the similarity of organizations and venues as CoO and CoV. CoA, CoW, CoO, and CoV are added with the weights of {1.5, 0.33, 1.0, 1.0} respectively. Then we assign the papers to the group if the score is above a pre-defined threshold, i.e., 1.5 in our method.

A.6 Implementation Details of RND-all

Here, we introduce the technical details of the RND-all method, which consists of the soft semantic features, the ad-hoc semantic features, the ego graphs, and the combination patterns.

A.6.1 Soft Semantic Features. Similarly, we obtain the soft semantic embeddings of papers like the SND-all does while leveraging OAGBERT [19], based on which we get the soft semantic similarities between the target paper and each paper of the candidate author.

A.6.2 Ad-hoc Semantic Features. Table 6 presents the detailed definitions of 36-dimensional hand-crafted features between each unassigned paper and candidate author pair.

A.6.3 Relational Ego Graph Features. We collect historical bibliographic data to construct heterogeneous ego-graphs for target papers and candidate authors, with authors, papers, and organizations as nodes. As shown in Figure 11, For the concerned center node of the author, we first use OAGBERT to get the paper embedding, then we average all the paper embeddings belong the author as the author embeddings. After that, we obtain the updated paper or author embeddings via training the graph attention networks (GAT), based on which we get the relational similarities between the center paper and the center authors.

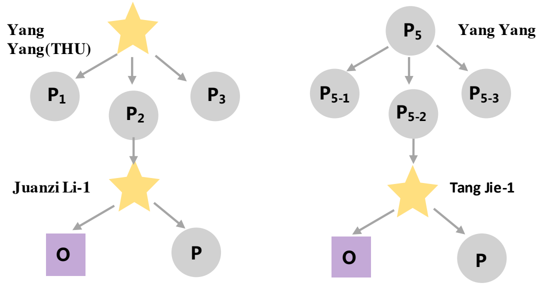
A.6.4 Feature Combinations. For the soft semantic features and relational ego graph features, we adopt an RBF kernel function used in [3] to extract 41 dimensional aggregated features based on the similarity scores between the target paper and the candidate author. For the ad-hoc semantic features, we directly adopt 36-dimensional hand-crafted features. Finally, we simply concatenate aggregated features from the three feature modalities to obtain the 118-dimensional features between the target paper and the candidate author, and adopt ensemble GBDT models to make predictions.

A.7 Implementation Details of baseline methods

Here we elaborate on the details of name disambiguation baselines used in the paper.

Table 5: . Statistics of prevailing manually-labeled name disambiguation datasets. Fewer names with more authors and papers indicates the dataset with higher ambiguity.

Datasets	#Names	#Authors	#Papers	Source
Song-PubMed	36	385	2,875	PubMed (Biomedicine)
GS-MEDLINE	-	-	3,756	PubMed (Biomedicine)
Han-DBLP	14	479	8,453	DBLP (Computer Science)
Qian-DBLP	680	1,201	6,783	DBLP (Several Domains)
Tang-AMiner	110	1782	8386	AMiner (General Domains)
SCAD-zbMATH	2919	2946	33,810	zbMATH (MATH)
Zhang-AMiner	100	12,798	70,258	AMiner (General Domains)
INSPIRE	**12,458**	36,340	360,066	INSPIRE (Physics)
WhoIsWho	2,495	**72,609**	**1,102,249**	AMiner (General Domains)



(a). Ego graph of the author (b). Ego graph of the paper

Figure 11: The built ego graph on the target paper and the candidate author in the RND-all method.

A.7.1 Baselines of the SND task. **G/L-emb**. Is a method for constructing paper-paper graphs using co-authorship connections. The process begins by generating initial paper embeddings through a weighted average of Word2Vec embeddings for all tokens within a paper. Next, the method fine-tunes these embeddings by first learning on a global paper-paper network and then adapting them to a local paper-paper network, specific to each author, using graph auto-encoding. Lastly, a hierarchical agglomerative clustering algorithm (HAC) is employed to segregate these papers into distinct groups. **LAND**. Is a method for constructing heterogeneous scholarly knowledge graphs (KGs) that encompass multiple entities, including papers, authors, venues, affiliations, and more. These KGs also feature various relations, such as co-authorship, and publication venues. Entity input embeddings are initialized using BERT models, and KG embedding techniques are then applied to derive paper and author embeddings. Finally, hierarchical agglomerative clustering (HAC) methods are utilized to group these entities based on the embeddings. **IUAD**. Is a method that constructs collaboration networks by treating papers as nodes and creating edges between two papers if they share the same author. It employs probabilistic generative methods to determine whether two papers in the collaboration network belong to a single author, with the goal of reconstructing the complete collaboration network accurately. Once the probabilistic models are trained, they are utilized to perform SND (Supervised Name Disambiguation) tasks, enhancing the overall disambiguation process. **SND-all**. Is a proposed baseline method based on empirical studies from contest-winning approaches. It starts by estimating semantic correlations among papers to be disambiguated, using title, keywords, and organizations as soft semantic features. For each paper, SND-all projects these features into corresponding word embeddings via Word2Vec and averages them to create paper embeddings. Cosine similarities are then calculated between papers based on these semantic embeddings. Additionally, SND-all constructs a heterogeneous network featuring three relational edges among papers: co-author, co-organization, and co-venue. The method employs metapath2vec to generate relational embeddings of papers, and computes relational similarity scores based on these embeddings. Lastly, SND-all combines both multi-modal similarities to calculate overall similarities among papers. DBSCAN is used to derive the final clustering results based on these similarities. **Contest Winner Method**: Follows a similar pipeline to SND-all. However, it uses all paper attributes to estimate semantic correlations among papers, which has proven to be less effective than utilizing a few informative attributes, such as title, keywords, and organizations, as implemented in the SND-all method. Furthermore, when estimating relational similarities among papers, the Contest Winner Method only considers co-author and co-organization relational edges, offering a more limited perspective compared to SND-all.

A.7.2 Baselines of the RND task. **IUAD**. Is also employed for the RND task by reconstructing the collaboration network between newly-arrived papers and existing authors. **CONNA**. Is a bottom-up, interaction-based model. To determine similarities between the target paper and candidate authors, it first constructs basic interactions between token embeddings of corresponding attributes for the target paper and each paper of the candidate author. Then, different attribute matrices are aggregated to form paper-level interactions. Specifically, CONNA uses all paper attributes, treating author names as one field and all other attributes as another field. Finally, Learning-to-Rank techniques are used to score the agreements between the target paper and all candidate authors based on the aggregated paper-level matrices. **CONNA+Ad-hoc**. Is a combination methodology that integrates hand-crafted features into the CONNA framework. For fair comparisons, the same ad-hoc features used in RND-all are leveraged. **RND-all**. Is another proposed method based on ablation study findings. It utilizes the soft and ad-hoc semantic features described in Section 4.1. Additionally, to incorporate relational features, RND-all constructs heterogeneous ego-graphs for the target paper and all candidate authors, as shown in Figure 11 of the appendix. A Graph Attention Network (GAT) is employed to obtain fused center node embeddings for the target paper and candidate authors. The relational correlations between the target paper and each candidate author are determined by calculating cosine similarities between the center node embeddings.

Table 6

The detailed definitions of 36-dimensional hand-crafted features. p: target paper, a: target author in p, c: candidate person.

No.	Feature description
1	TF-IDF score of a's coauthors in c
2	TF-IDF score of a's coauthors in c multiplied by co-occurrence times in c
3	Ratio of a's coauthors in p's author names
4	Ratio of a's coauthors in c's author names
5	TF-IDF score of a's title common part in c
6	TF-IDF score of a's title common part in c multiplied by co-occurrence times in c
7	Ratio of a's title common part in a's title
8	Ratio of a's title common part in c's titles
9	Max Jaccard similarity between a's title and c's titles
10	Mean Jaccard similarity between a's title and c's titles
11	Max Jaro–Winkler similarity between a's title and c's titles
12	Mean Jaro–Winkler similarity between a's title and c's titles
13	TF-IDF score of a's venue common part in c
14	TF-IDF score of a's venue common part in c multiplied by co-occurrence times in c
15	Ratio of a's venue common part in a's venue
16	Ratio of a's venue common part in c's venues
17	Max Jaccard similarity between a's venue and c's venues
18	Mean Jaccard similarity between a's venue and c's venues
19	Max Jaro–Winkler similarity between a's venue and c's venues
20	Mean Jaro–Winkler similarity between a's venue and c's venues
21	TF-IDF score of a's organization common part in c
22	TF-IDF score of a's organization common part in c multiplied by co-occurrence times in c
23	Ratio of a's organization common part in a's organization
24	Ratio of a's organization common part in c's organizations
25	Max Jaccard similarity between a's organization and c's organizations
26	Mean Jaccard similarity between a's organization and c's organizations
27	Max Jaro–Winkler similarity between a's organization and c's organiza-tions
28	Mean Jaro–Winkler similarity between a's organization and c's organiza-tions
29	TF-IDF score of a's keywords common part in c
30	TF-IDF score of a's keywords common part in c multiplied by co-occurrence times in c
31	Ratio of a's keywords common part in a's keywords
32	Ratio of a's keywords common part in c's keywords
33	Max Jaccard similarity between a's keywords and c's keywords
34	Mean Jaccard similarity between a's keywords and c's keywords
35	Max Jaro–Winkler similarity between a's keywords and c's keywords
36	Mean Jaro–Winkler similarity between a's keywords and c's keywords

Lastly, RND-all combines the soft semantic features, ad-hoc semantic features, and proposed relational features to make predictions. **Contest Winner Method**. Estimates the agreement between the target paper and each candidate author using only soft semantic features and ad-hoc semantic features. It also employs all paper attributes when constructing semantic features, which has been proven less effective than adopting a few informative attributes, as used in the RND-all method.

A.8 Online deployment of disambiguating daily papers from arXiv.

We have deployed the proposed RND-all method on AMiner to disambiguate daily papers from arXiv.org. Practically, for each name in the paper to be disambiguated, instead of the adopted name blocking strategy, i.e., moving the last name to the first or preserving all name initials but the last name, we adopt Elastic-Search[12] to perform the online fuzzy search. Finally, we apply RND-all to estimate the similarity between each candidate author and the target paper. To solve NIL cases that there are no right authors, we pre-defined a threshold and return the candidate with the highest score exceeding the threshold as the right author on AMiner.

A.9 Dataset Statistics.

The detailed data statistics are shown in Figure 5.

A.10 Interactive Annotation Tool

Figure 12 depicts the framework of the designed interactive annotation tool, which consists of two main parts, i.e., the annotation panel and the information panel.

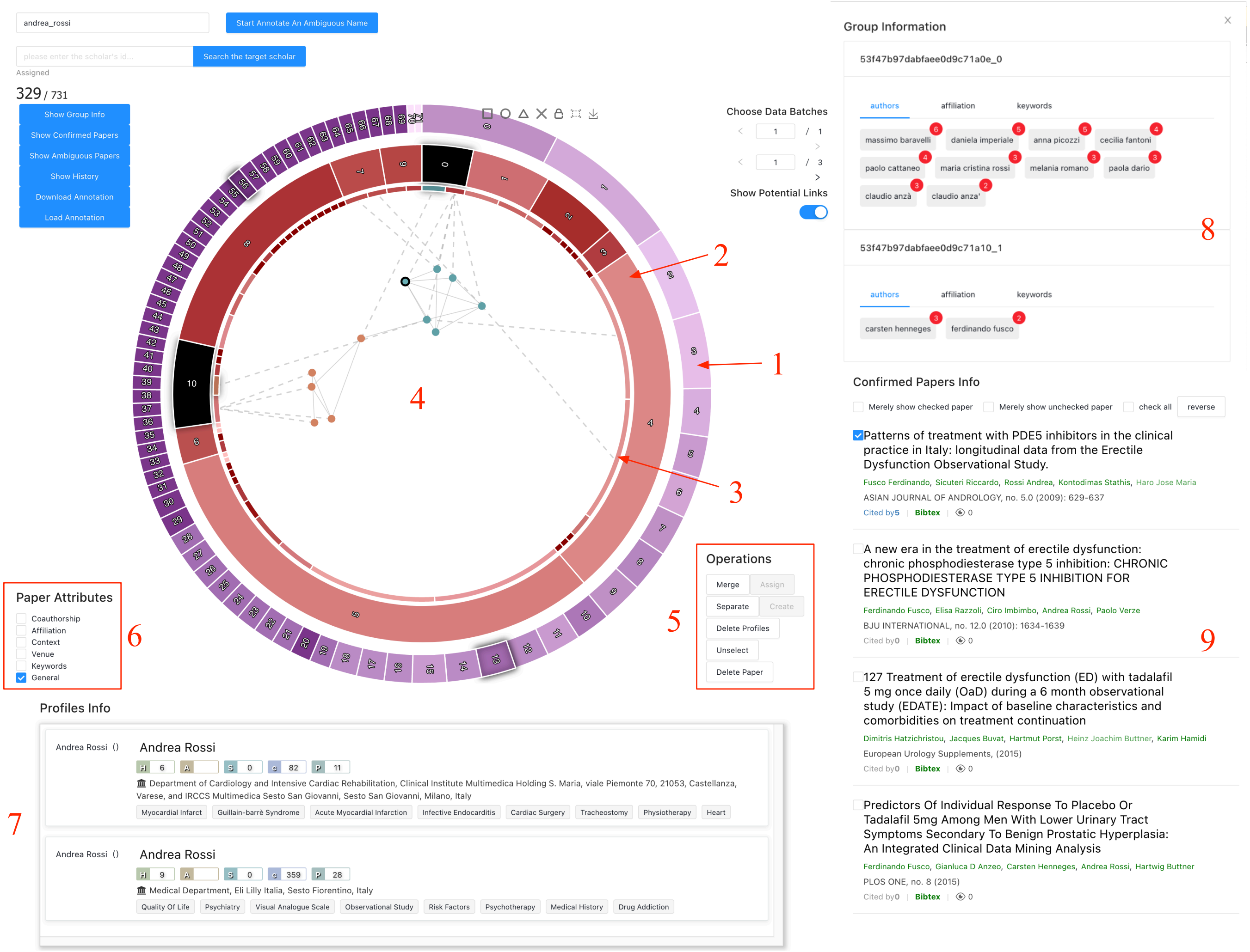
Annotation Panel. The first 3 regions construct the ring with three stack layers. Specifically, the outer layer, i.e., region "1", shows the collected unassigned papers with the target author named "Andrea Rossi". Each block in the middle layer, i.e., region "2", represents the author named "Andrea Rossi". To facilitate annotators to grasp the global relationships among papers, we adopt the clustering methods to partition papers within each author into several groups, as shown in the region "3". Each group contains papers with similar attributes, such as those are published in the same venue, coauthored by the same authors, etc. By clicking the authors in the region "2", we can see the inter-connections, i.e., the dotted line, and intra-connections, i.e., the solid line among papers. Annotators can freely select different attributes via the region "6". Then, annotators can perform operations via region "5".

Information Panel. Regions "7", "8", and "9" provide comprehensive information about the selected papers and authors, which support annotators to conduct accurate operations. Among these, the region "7" presents the profile comparisons among selected authors. Then, region "8" shows more detailed comparisons between selected authors, such as coauthors, affiliations, and keywords. Finally, region "9" supply the complete information of selected papers.

Overall, the interactive annotation tool not only provides convenient atomic operations to improve the efficiency of annotators, but also prepares comprehensive information to support them make decisions. With the help of the effective visualization tool, we plan to annotate and update more datasets to WhoIsWho in the future.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009

[12] https://www.elastic.co



Andrea Rossi ()

Andrea Rossi

H 6 | A | S 0 | c 82 | P 11

🏛 Department of Cardiology and Intensive Cardiac Rehabilitation, Clinical Institute Multimedica Holding S. Maria, viale Piemonte 70, 21053, Castellanza, Varese, and IRCCS Multimedica Sesto San Giovanni, Sesto San Giovanni, Milano, Italy

Myocardial Infarct | Guillain-barrè Syndrome | Acute Myocardial Infarction | Infective Endocarditis | Cardiac Surgery | Tracheostomy | Physiotherapy | Heart

Andrea Rossi ()

Andrea Rossi

H 9 | A | S 0 | c 359 | P 28

🏛 Medical Department, Eli Lilly Italia, Sesto Fiorentino, Italy

Quality Of Life | Psychiatry | Visual Analogue Scale | Observational Study | Risk Factors | Psychotherapy | Medical History | Drug Addiction

53f47b97dabfaee0d9c71a0e_0

massimo baravelli 6 | daniela imperiale 5 | anna picozzi 5 | cecilia fantoni 4 | paolo cattaneo 4 | maria cristina rossi 3 | melania romano 3 | paola dario 3 | claudio anzà 3 | claudio anza' 2

53f47b97dabfaee0d9c71a10_1

carsten henneges 3 | ferdinando fusco 2

☑ Patterns of treatment with PDE5 inhibitors in the clinical practice in Italy: longitudinal data from the Erectile Dysfunction Observational Study.

Fusco Ferdinando, Sicuteri Riccardo, Rossi Andrea, Kontodimas Stathis, Haro Jose Maria

ASIAN JOURNAL OF ANDROLOGY, no. 5.0 (2009): 629-637

Cited by5 | Bibtex | 👁 0

□ A new era in the treatment of erectile dysfunction: chronic phosphodiesterase type 5 inhibition: CHRONIC PHOSPHODIESTERASE TYPE 5 INHIBITION FOR ERECTILE DYSFUNCTION

Ferdinando Fusco, Elisa Razzoli, Ciro Imbimbo, Andrea Rossi, Paolo Verze

BJU INTERNATIONAL, no. 12.0 (2010): 1634-1639

Cited by0 | Bibtex | 👁 0

□ 127 Treatment of erectile dysfunction (ED) with tadalafil 5 mg once daily (OaD) during a 6 month observational study (EDATE): Impact of baseline characteristics and comorbidities on treatment continuation

Dimitris Hatzichristou, Jacques Buvat, Hartmut Porst, Heinz Joachim Buttner, Karim Hamidi

European Urology Supplements, (2015)

Cited by0 | Bibtex | 👁 0

□ Predictors Of Individual Response To Placebo Or Tadalafil 5mg Among Men With Lower Urinary Tract Symptoms Secondary To Benign Prostatic Hyperplasia: An Integrated Clinical Data Mining Analysis

Ferdinando Fusco, Gianluca D Anzeo, Carsten Henneges, Andrea Rossi, Hartwig Buttner

PLOS ONE, no. 8 (2015)

Cited by0 | Bibtex | 👁 0

Figure 12: A toy annotation example for annotating authors with author name "Andrea Rossi".