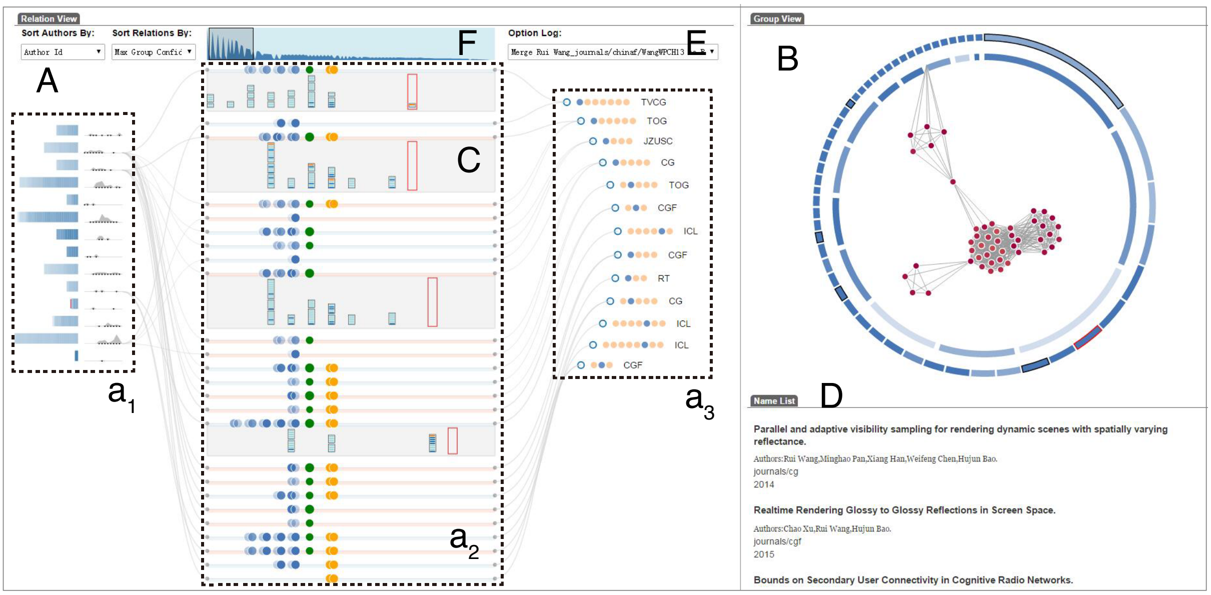
NameClarifier: A Visual Analytics System for Author Name Disambiguation

Qiaomu Shen, Tongshuang Wu, *Student Member, IEEE*, Haiyan Yang, Yanhong Wu, *Student Member, IEEE*, Huamin Qu, *Member, IEEE*, and Weiwei Cui, *Member, IEEE*



Sort Authors By: Author Id

Sort Relations By: Max Group Confic

Option Log: Merge Rui Wang_journals/chinaf/WangWPCH13_...

Parallel and adaptive visibility sampling for rendering dynamic scenes with spatially varying reflectance.

Authors:Rui Wang,Minghao Pan,Xiang Han,Weifeng Chen,Hujun Bao.
journals/cg
2014

Realtime Rendering Glossy to Glossy Reflections in Screen Space.

Authors:Chao Xu,Rui Wang,Hujun Bao.
journals/cgf
2015

Bounds on Secondary User Connectivity in Cognitive Radio Networks.

Fig. 1. Interface for the NameClarifier, which contains the following: (A) a relation view that contrasts papers containing ambiguous author names with confirmed authors, resulting in easier classification of ambiguous names; (B) a group view that supports the relation view by assessing whether the ambiguous names have been correctly and comprehensively classified; (C) a temporal view that verifies whether a specific paper fits into a confirmed author's publication trajectory; and (D) a list containing all papers with ambiguous author names so that the users can always refer back to the original metadata.

Abstract— In this paper, we present a novel visual analytics system called NameClarifier to interactively disambiguate author names in publications by keeping humans in the loop. Specifically, NameClarifier quantifies and visualizes the similarities between ambiguous names and those that have been confirmed in digital libraries. The similarities are calculated using three key factors, namely, co-authorships, publication venues, and temporal information. Our system estimates all possible allocations, and then provides visual cues to users to help them validate every ambiguous case. By looping users in the disambiguation process, our system can achieve more reliable results than general data mining models for highly ambiguous cases. In addition, once an ambiguous case is resolved, the result is instantly added back to our system and serves as additional cues for all the remaining unidentified names. In this way, we open up the black box in traditional disambiguation processes, and help intuitively and comprehensively explain why the corresponding classifications should hold. We conducted two use cases and an expert review to demonstrate the effectiveness of NameClarifier.

Index Terms—Name disambiguation, analytical reasoning

1 Introduction

Name ambiguity, which refers to the many-to-many mapping relationships between persons and their names [21], is a common problem in the context of bibliographic citation records. In many cases, names may not be sufficient to distinguish a person from another because of the two types of name ambiguity (i.e., homonyms and synonyms) [38]. Homonyms refer to cases wherein the same name is shared by multiple persons [21]. This issue is particularly common in Chinese names, because many of them have similar spellings even when they consist of different Chinese characters. For example, DBLP, a well-acknowledged bibliographic database for computer science journals and proceedings, contains at least 15 different researchers whose names are all "Rui Wang"[1] and 17 ones whose names are "Kai Chen"[2]. In contrast, a synonym refers to a person with different names. For example, "Rui Wang" is often replaced by its surname and the first name initial, i.e., "R. Wang", on many occasions.

Making distinctions between persons with the same name is an important prerequisite ensuring the effectiveness of person search [21], the high quality of services and content in digital libraries [14], and any application related to networks of scientific publications [25]. For this reason, current major bibliographic databases (e.g., DBLP, PubMed, and Medline) and the research community have devoted a great deal of

- *Qiaomu Shen, Tongshuang Wu, Haiyan Yang, Yanhong Wu, and Huamin Qu are with the Hong Kong University of Science and Technology. E-mail: {qshen, twuac, hyangac, ywubk, huamin}@ust.hk.*
- *Weiwei Cui is with Microsoft Research and is the corresponding author. E-mail: weiwei.cui@microsoft.com.*

Manuscript received 31 Mar. 2016; accepted 1 Aug. 2016. Date of publication 15 Aug. 2016; date of current version 23 Oct. 2016.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TVCG.2016.2598465

[1] http://dblp.uni-trier.de/search/author?q=Rui%20Wang

[2] http://dblp.uni-trier.de/search/author?q=Kai%20Chen

effort on name disambiguation. Existing solutions generally take advantage of name initials [31], co-authorship [13,21], self-citation [25], etc. Although some of them have achieved fairly convincing accuracies or have been deployed to real world systems, few existing mining algorithms have offered comprehensive solutions to this problem [21]. One shared characteristic in these proposed algorithms is that they aim at solving name ambiguities with a universal model. However, in reality, many ambiguous cases are tricky and cannot be easily solved for diverse reasons that are difficult to model exhaustively. For instance, in some complicated cases, ambiguous papers have common co-authors who also have ambiguous names, and researchers with the same name may work in the same research area or even appear in the same paper [14]. These cases are very difficult, if not impossible, to capture with a general model designed for common cases. Moreover, bibliographic data in different disciplines often have varied characteristics, which bring additional challenges in designing a universal algorithm. Ferrira et al. [14] discovered that humanities or medicine papers may have different publication patterns from computer science papers (e.g., publications with a sole author or with many co-authors).

To address the aforementioned overgeneralization problem, this study proposes a new visual analytics system called NameClarifier to customize the disambiguation on a case-by-case basis. On the basis of name similarity, co-author relationship, venue similarity, and temporal information, we compare unidentified author names with the ones that are already identified in digital libraries by default. Our deliberately designed visual cues guide users to adjust their attention to different factors, and thus allow them to interactively disambiguate these cases individually and progressively, or restore misclassified cases. Our contributions are described as follows:

- We approach the name disambiguation problem from the visualization perspective. By linking mining algorithms with visual perceptions [22, 27], we turn the traditional black-box solution into a white-box procedure, which is capable of articulating underlying reasons for each ambiguous case through visual cues, thereby offering better insights into the disambiguation results.
- The system provides guidance instead of classification results for ambiguous cases; thus we can easily adjust it for various datasets.
- Our two-fold disambiguation process iteratively refines initial results. Through interaction, not only can users move forward and evaluate new ambiguous cases, they can also take a step back and tune the identified author groups to avoid inaccuracies due to taking misclassified cases as evidence. The system also instantly updates visual feedback after every interaction, so that the interact-and-update cycle can advance the overall process smoothly.

2 Related Work

In this section, we give a brief overview of both the mining algorithms and visual designs for name disambiguation problems.

2.1 Mining Algorithms for Name Disambiguation

The core issue in name disambiguation is to determine whether two similar names in archival records refer to the same person [38]. This issue has long been well recognized as early as in the late 1960s [15]. Early attempts in this field mostly include manual disambiguation [29]. However, the rapid growth of the number of researchers in large-scale digital libraries makes manual checking methods unpractical [36,38]. Therefore, numerous advanced methods have been proposed to identify authorship automatically [14,34].

Most of the existing works pre-select various strong features from huge bibliographic databases and compute their similarities to identify publication records authored by the same person. The simplest way is to rely on author names, because it is the most commonly available feature shared by all records. For instance, Milojević [31] tested the usefulness of initial-based disambiguation on synthetic datasets. Although this simple method works well for specific cases, most scholars who only use initial-based disambiguation have acknowledged that these methods could mishandle ambiguities [23]. Thus, to further improve the accuracy, more researchers developed automatic algorithms by fusing the name with other attributes, which either exist in digital libraries by default or are extracted additionally. These attributes typically include titles [41], self-citations [25], shared references [38], the characteristics of author names [31], etc. Among them, co-authorship has been proven to be the most accessible and influential one [21]. Some studies [19, 32, 40] even included additional web information. For example, Pereira et al. [32] resolved disambiguation with curricula vitae and Web pages containing publication records of the ambiguous authors. Yang et al. [40] focused on topic similarities and web correlations. Although more attributes may increase the accuracy, few of them are generally applicable. This is because bibliographic metadata and online information are intrinsically sparse. The attributes required by a model are not always available and have varying usefulness, thus resulting in inconsistent performances in various real cases.

Our work shares the use of metadata with previous studies. However, to balance the sparsity of data and the correctness, we guide users to make case-by-case comparisons. In this sense, Gurney et al. [16] implemented a related approach which calculates similarities dynamically, so to compare records on completely different metadata. However, their supervised method requires manually labeled training data for better performance. We borrow its idea of customized evaluations, but save the effort of labeling by investigating the problem through a hybrid of unsupervised learning and visual feedback.

2.2 Name Disambiguation Visualization

Only a few works directly relate to our system. Strotmann et al. [36] presented a visual design that exactly aims at name disambiguation by taking advantage of co-authorships and publication venues. While their co-author network proved their algorithm's efficiency, it cannot facilitate the process. Besides, Bilgic et al. designed D-Dupe [1, 20] for entity resolution [6]. It combines data matching algorithms with network visualization, enabling users to resolve aliases with an entity's relational context. While targeting at a closely related problem, D-Dupe primarily differs from our work in two aspects. First, D-Dupe mainly compares authors using researcher-oriented attributes in bibliographic collections, such as affiliation and country. However, most digital libraries, including DBLP, do not provide such metadata by default. Thus, D-Dupe cannot directly work for these databases. Meanwhile, NameClarifier takes advantage of more common publication-oriented metadata, such as publication venue and year. As demonstrated in our use cases and confirmed by domain experts, these attributes are critical for effective name disambiguation. Second, despite the diverse attributes it uses to compute similarities, D-Dupe only visualizes co-author relationships to facilitate manual disambiguation, which is not enough. This only visual display can be fallible when the co-authorship of two researchers greatly overlap. In contrast to its heavy dependency on the co-authorship, our design tightly connects the co-authorship, venue and temporal information. The multifaceted visual attributes help users compare between authors more comprehensively and resolve ambiguous names more confidently.

In a broader sense, our work is also related to text [8, 10, 12], network [9, 30], multivariate visualization [4, 5], and visual analytics of digital libraries [24,28,35]. For instance, an effective approach to presenting inter- and intra-record relationships with composite features is through glyph-based designs [2, 3, 39]. NameClarifier also use glyphs to pack multiple metadata together. To achieve relative comparisons between paper relations (see R.5 in Section 3.2), our glyphs are simple enough to be displayed simultaneously without causing confusions. We take advantage of visual paths for comparisons and spatial arrangements of such glyphs. Dörk et al. [11] exposed faceted relations as visual paths, and Hoque and Carenini proposed ConVis [17] to compare different facets of blog conversations with similar approaches. However, while the visual paths can effectively help explore relations, both of them only present the relations in a very abstract way, either with single textural titles or visual summaries of blog conversations. Given that our system requires more detailed visual comparisons of multiple ambiguous cases, applying them directly is inadequate. Nevertheless, their work motivates us to design our relation view to compare two authors explicitly, as described in Section 7.1.

3 Design Consideration

3.1 Task Analysis

Because of the diversity of name ambiguity cases, most of the existing methods' performances are limited by overly generalized models. Such diversity is essentially caused by the uniqueness of every attribute, which we refer to as "attribute uncertainties". For instance, publication venues have uncertainties in the research areas they cover. IJCV is only for computer vision, whereas TVCG covers computer graphics and visualization. Venues covering a narrower area can pinpoint one's research interest more precisely. The uncertainty of co-author names also matters: shared co-authors themselves may suffer from the name ambiguity problem inherently if they have popular names, which could link unrelated authors together and thereby introduce additional noise. Such uncertainties make it impossible to have a universally well-performed rule that can instantly cater to every case.

To solve this, we build a visual system that counts in the specialty of every case, presents them to our users, and collects users' feedback to iteratively refine disambiguation results. Specifically, we break it down into three analytical tasks to first generate an initial result (T.1) and then perform two-fold refinements (T.2 and T.3) as follows:

T.1. Estimate the identifications of ambiguous names. This is the fundamental task of our system. For each ambiguous name (or essentially a paper with an ambiguous author name that needs to be classified), we first estimate its similarities with the ones that have been confirmed in a digital library. Such estimation is essential, because it roughly provides an initial disambiguation guidance, based on which uncertainties can be further taken care of.

T.2. Allow iterative refinements of disambiguation. As we have previously mentioned, the literature follows the "black-box" convention, meaning all the classifications are performed simultaneously with the inner processes hidden from users. However, because of papers' diverse characteristics, fixed settings and unified thresholds maximizing the global performance may, in fact, cause some boundary cases to be missed out. To handle different boundary cases, our system needs to support iterative refinements of disambiguation results. The iteration also provides more insights into how and why the result is improved.

T.3. Verify the automatically classified names. Digital libraries often maintain a list of classified author names in publications, which is achieved by automatic algorithms. However, over- and under-classifications may happen for these names, where multiple distinct authors are merged into one or an author is identified as many. Given that our preliminary estimations heavily rely on the previously confirmed names in digital libraries, the uncertainties within the confirmed names are also worth investigating. To improve the accuracy of our estimations, we also need to explore and quantify the qualities of the automatically classified names, and rectify false positives introduced by misclassified names in the data.

3.2 Design Rationales

In response to the aforementioned tasks, we compile the following design requirements to guide our name disambiguation process:

R.1. Quantify similarities between ambiguous names and confirmed ones. Quantifying similarities is the basic and foremost requirement, on the basis of which we can estimate the identifications of ambiguous names (T.1). Such quantification should support progressive exploration to achieve case-by-case disambiguation (T.2).

R.2. Support multiple attributes. The previous literature [16] has shown that the comparison of confirmed authors to ambiguous names is unreliable with only a single attribute. Thus, we need to consider multiple attributes simultaneously to make the comparison more solid.

R.3. Encode the mapping of attributes to similarities. To support the manual evaluation of each ambiguous name (T.2), we need to elaborate the comparisons of every attribute. In other words, we need to build visual cues that can present which attributes lift the similarity, and which have a negative effect on it. By transforming the black-box solution into a white-box one in this way, users can understand the different aspects of the comparison profoundly. This helps them consider re-allocating attribute weights implicitly when identifying ambiguous cases based on such visual feedback.

R.4. Encode qualities of automatically classified names. As mentioned in T.3, the qualities of the previously confirmed names are important factors for subsequent comparisons and assignments. To help users identify and rectify the cases of over- or under- classifications as early as possible, our system must intuitively encode the (dis-)similarities between the names that are classified as the same person and between the same names that are classified as different people.

R.5. Enable multi-case comparisons. Whereas the absolute quantification (R.1) helps generate the initial result, and the aforementioned encodings provide a general sense of refinement direction (R.3 and R.4), relative comparisons among ambiguous and confirmed names are more crucial for guiding the disambiguation. Comparing one ambiguous case with multiple confirmed cases helps assign this particular case, and comparing across multiple ambiguous cases facilitates the selection of the best possible assignment currently available to arrive at the next stage of disambiguation. Therefore, to provide a clearer guidance, the visual form must support relative comparisons intuitively.

R.6. Support interactive refinements with prompt visual feedback. To progressively improve disambiguation results, we need to allow smooth interactions that help users assign ambiguous names to confirmed authors, or eliminate noise from the latter. Our system must promptly update the visual display in response to every interaction, such that users can intuitively sense how certain modifications universally affect disambiguation results.

4 System Overview

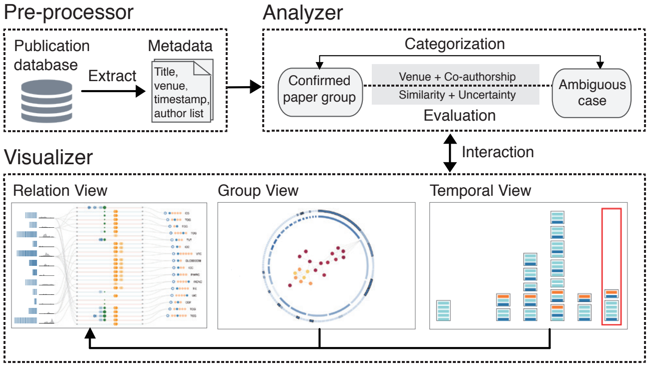


Fig. 2. System overview: NameClarifier consists of three major modules, namely, Pre-processor, Analyzer, and VisualizerPre-processor (Section 5), Analyzer (Section 6), and Visualizer (Section 7).

Fig. 2 illustrates the pipeline of our system. Given an author name *NM*, the input to NameClarifier is a collection of publications with the name *NM* listed as an author from digital libraries. In such a collection, some of the names have already been classified to certain authors, while the true identities of the other names remain unknown. Our Pre-processor extracts metadata for each publication record, and then sends them to the Analyzer, which organizes and compares the collected publication records. Specifically, the Analyzer first categorizes all the papers that contain a confirmed identity of name *NM* into groups based on their identities (*Person*1, *Person*2, etc.), and lists the remaining publications as ambiguous cases waiting to be classified. Thereafter, the Analyzer compares and estimates the similarities between the confirmed groups and the ambiguous cases using their publication venues and co-authorships with respect to publication time.

The Visualizer then transforms the output from the Analyzer into a comprehensible visualization consisting of three linked views, namely, the relation view, the group view, and the temporal view. The relation view is the essential channel for contrasting confirmed cases with ambiguous cases based on co-authorship and venue information. It presents such inter-comparisons compactly to help classify the papers with ambiguous author names into the existing confirmed groups (T.1). Unlike the relation view, the group view and the temporal view serve as verification methods (T.2 and T.3). The group view helps "verify persons". It rectifies the confirmed paper groups by (1) identifying additional groups from all the ambiguous records, and (2) correcting false positives that are already in the confirmed groups. The temporal view, on the other hand, "verifies papers". It compares a specific ambiguous paper to the publication trajectory of its potential identities discovered in the relation view, and helps confirm the allocation using the temporal patterns of publications. Various interactions, such as filtering and grouping, are implemented to help users assess name ambiguity with these three views interchangeably and, therefore, understand it more rationally. The Visualizer is closely connected to the Analyzer in that the evaluation metrics are re-calculated dynamically after each interaction, in which the confirmed groups and ambiguous lists are refined. This interact-and-update cycle facilitates the completion of iterative refinements of the disambiguation.

5 Data Pre-processing

Our data are taken from DBLP, one of the most acknowledged and representative bibliographic databases in this field. During preprocessing, each record in the database is summarized into several descriptive attributes: title (for identifying the paper), publication venue (i.e., conference/journal name), timestamp (i.e., publication year), and author list. Whereas many other attributes can be extracted from the raw publication (e.g., keywords), we explicitly select these attributes for two reasons. First, these are the most commonly available metadata in most bibliographic databases. Thus, using them can make NameClarifier much more compatible with various databases. Second, according to Tang et al. [37], the two strong indicators for name disambiguation are similar contents of papers (e.g., publication venue) and co-authors. In addition, we also observe that the publication timestamp plays an important role in determining similarities of papers [16]: consecutively published papers by the same person tend to share similar research areas, venues, etc., whereas greater differences may occur if there are temporal gaps in between.

6 Data Analysis

The Analyzer contains two parts, namely, the categorization of known and unknown authors, and similarity evaluations between papers.

	(a) Allocation Likelihood
cm_i	**Co-author matching** The likelihood of two papers written by the same researchers.
vm_i	**Venue matching** If two papers are published at the same venue.
AL	**Allocation likelihood** The likelihood of allocating an ambiguous paper to a confirmed author group.
	(b) Confidence Measurements
	Venue Confidence
vr	**The venue research interest** The level of research area concentration of a venue.
•	
ad	**The author's addiction** The author loyalties to certain venues.
+	
vs	**The similarity of publication venues** The research area overlap between two venues.
↓	
vc	**Venue Confidence** How certain we are to use a venue as an evidence.
	Co-author Confidence
DC	**Connection closeness** Directly connected: close collaborations, Indirectly connected: "friends of friends".
•	
cf	**Collaboration frequency** The number of papers co-authored between two authors.
+	
gq	**Group quality** The confidence of a co-author group: co-authors with a same name indeed refer to an identical person?
↓	
cc	**Co-author Confidence** How certain we are to use a co-author as an evidence.

Table 1. Summary of notations introduced in the evaluation method.

6.1 Categorization Method

Given an author name, the Analyzer first extracts all the publications that contain the name from the dataset, and then splits the extracted papers into two categories: a confirmed list and an ambiguous list.

The **confirmed list** contains a list of indexed authors accompanied by their corresponding paper sets that are already identified by DBLP's default method. For instance, for the name "Wei Chen", DBLP would index "Wei Chen 0001" for one group of papers written by one researcher named "Wei Chen" and "Wei Chen 0002" for a different researcher whose name is also "Wei Chen". Notice that every record in the list, which is later referred as a **confirmed group**, is essentially a researcher and all his/her confirmed publications.

The **ambiguous list** collects all the unidentified papers. While every confirmed author may be associated with multiple papers, records in this list are all one-to-one mappings between papers and ambiguous names. In the rest of this paper, we use the phrases **ambiguous author** and **ambiguous paper** interchangeably to refer to the same thing.

Thus, our ultimate goal is to put every ambiguous paper into one of the confirmed groups and rectify any misclassified ones in the confirmed groups through a series of similarity evaluations. To simplify descriptions in the following sections, we denote a confirmed group by G, which contains a confirmed author A_G and a set of publications $P_G = \{p_1, ..., p_n\}$. An ambiguous author is denoted by A, and the associated paper is p_A because of the one-to-one mapping.

6.2 Evaluation Method

The evaluation step contrasts between ambiguous papers and confirmed groups using the attributes summarized from the Pre-processor. First, we define an initial comparison method named **allocation likelihood**, which scores the likelihood of allocating an ambiguous paper to a confirmed group with respect to shared publication venues and co-authors (R.1 and R.2). This score is used as a supportive hint, but not a decisive evidence, of the ambiguous paper. Next, we define and quantify a series of metrics to take care of the attribute uncertainties. We denote these metrics as "confidences" because they essentially measure how trustable the attributes are. Both the comparison score and the uncertainty measurements are fed into the Visualizer, so to provide all-rounded visual cues and help users process the disambiguation.

6.2.1 Allocation Likelihood

To compute the allocation likelihood AL of an ambiguous author A to a confirmed group G, we first consider the similarity between p_A and $p_i \in P_G$: The greater the similarity , the more possible it is that p_A is in G. Given the attributes collected by the Pre-processor, their estimated similarity is the weighted sum with respect to two parts:

(1) **Co-author matching** $cm_i = |C(p_i) \cap C(p_A)|/|C(p_i) \cup C(p_A)|$, where $C(p)$ denotes the co-authors of the paper p, uses the Jacquard Index [18] to define the overlap ratio of their co-author lists. High cm_i means p_A and p_i are likely to be written by the same research team.

(2) **Venue matching** $vm_i = \mathrm{sgn}(vs(v_A, v_i) - s)$ describes whether p_A and p_i are published at venues of similar research focus. The function vs, which evaluates the similarity between the publication venues of p_A and p_i, denoted by v_A and v_i, respectively, is introduced below in Section 6.2.2. The threshold for vs is set to be $s = 0.1$, which effectively highlights the $5,000$ most similar venues out of $4,900,000$ pairs of venues that have non-zero vs scores in DBLP (i.e., top 0.01%). We declare here an exact interest match as long as vs passes s, and the uncertainties are considered later.

The final AL takes the average of all the paper similarities:

$$AL(A, G) = \frac{1}{n} \sum_{i=1}^{n} (\alpha_c \cdot cm_i + \alpha_v \cdot vm_i),$$

where α_c and α_v are the weights for cm and vm, respectively, with their default values being $\alpha_c = 0.8$ and $\alpha_v = 0.2$.

In addition, as we also consider the possibility that a confirmed paper may be misclassified, for every paper $p_i \in P_G$, we take it out, treat it as an ambiguous paper, and compute its allocation likelihood AL_i against the remaining papers in the group $P_G \backslash \{p_i\}$. This collection of allocation likelihoods $AL_G = \{AL_1, \ldots, AL_n\}$ is also used to evaluate the likelihood $AL(A,G)$. If $AL(A,G)$ is larger than some $AL_i \in AL_G$, then we know A is more likely to belong to G than some papers that are already in the group.

6.2.2 Confidence Measurements

We supply $AL(A,G)$ with multiple dimensions of confidences measured for venues and co-authorships, which assess the certainty of using the specific attributes as evidence.

Venue Confidence The confidence dimensions for venues are designed based on an empirical assumption: authors who (1) *frequently* publish papers in (2) *similar venues*, such as VIS and EuroVis, tend to share similar (3) *research interests*; moreover, it is unlikely to have multiple researchers of the same name in one field. Thus, we find the following factors that may affect our confidence of venues:

(1) **Author's addiction** to a venue v, denoted by $ad(v)$, is defined as the proportion of the author's confirmed papers that are published in venue v. It mainly describes an author's loyalty to a certain venue. Thus, the chance that an ambiguous paper that is published at venue v belongs to a confirmed group G is much higher if $ad(v)$ is high for G.

(2) **Similarity between venues** v_a and v_b, denoted by $vs(v_a, v_b)$, mainly concerns the authors who publish papers in both venues, which indicate how likely the research areas of these two venues overlap with each other. Similar to the co-author matching metric, we define $vs(v_a, v_b) = |A(v_a) \cap A(v_b)| / |A(v_a) \cup A(v_b)|$, where $A(v)$ indicates the set of researchers who have publications in the venue v.

(3) **Venue research interest**, denoted by $\mathbb{1}_{vr}(v)$, defines the level of research area concentration of a venue v. If certain venues focus on fewer topics, we are more confident that researchers publishing here indeed share the same research interests. Empirically, a venue similar to multiple other venues is more likely to accept papers with diverse research interests. Therefore, we first selected all the venues v_a in DBLP whose vm with at least one other venue v_b is larger than 0.1 (i.e., it is at least similar to another venue). We then manually checked the extracted 869 venues, and blocked those very inclusive ones who claim to accept papers on multiple topics ($\mathbb{1}_{vr}(v) = 0$). All other venues are treated as "concentrated" ones ($\mathbb{1}_{vr}(v) = 1$).

Note that while vr works as a filter, both ad and vs reflect our confidences about the venue factor of AL. When comparing the venue v_A of p_A with the venue v_i of $p_i \in P_G$, we can declare that, to a certain extent, we are confident in the similarity of p_i and p_A as long as $ad(v_i)$ is large (i.e., the confirmed author A_G is a major contributor to v_i, though v_A may not relate to v_i), or $vs(v_i, v_A)$ is large (i.e., p_A is highly related to some research area that A_G once contributed to, though maybe only occasionally). Thus, we model the **venue confidence** about the similarity between p_A and p_i as

$$vc(p_i, p_A) = \mathbb{1}_{vr}(v_i) \cdot (ad(v_i) + vs(v_i, v_A)).$$

Co-author Confidence Given an ambiguous paper p_A and the paper set P_G of a confirmed group G, we collect two co-author sets $C_A = C(p_A) \backslash \{A\}$ and $C_G = \cup_{p \in P_G} C(p) \backslash \{A_G\}$, respectively, where $C(p)$ refers to the co-authors of p. To derive the confidence about the co-author factor of AL, we consider two cases in terms of **connection closeness**: directly connected (DC) and indirectly connected (IC).

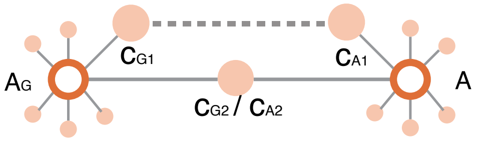


Fig. 3. An example for DC and IC. For a confirmed author A_G and an ambiguous author A, suppose we have $c_{G1},\ c_{G2} \in C_G$ and $c_{A1},\ c_{A2} \in C_A$. If c_{G1} has collaborated with c_{A1} in a paper without A_G or A, c_{G1} is an IC to A. If c_{G2} and c_{A2} are in fact the same person, then c_{G2} is an DC to A.

DC refers to the overlapped co-authors in C_A and C_G, i.e., $C_A \cap C_G$. These DC cases may strongly indicate that the ambiguous author A should be classified to G.

IC refers to those co-authors in $C_A \backslash C_G$ who have collaborated with some co-authors in C_G. Though less suggestive, it still helps to evaluate the confidence [13]. There are, of course, more indirect connections who are "friends of friends". We do not consider those cases because we believe IC is fuzzy enough for visual analysis in our experiment.

For DC, an informative measure is **collaboration frequency** $cf(c)$, which describes the number of papers co-authored by the confirmed author A_G and an overlapped author $c \in C_A \cap C_G$. For example, $cf(c) = 1$ means they only co-author one paper in P_G. Thus, the bigger the $cf(c)$ value is, the stronger an evidence the co-author c is for disambiguation. This measure is meaningless for IC because there are no collaborations between A_G and any author in $C_A \backslash C_G$.

Note that measuring the connection closeness and the collaboration frequency is based on a belief that co-authors of the same name in p_A and P_G always refer to the same person. However, these co-authors are essentially also identities with ambiguity. Thus, we introduce **group quality** $gq(c)$, which describes our confidence about the group of same author names discovered in $P_A \cup \{p_A\}$ referring to the same person. Similar to cf, gq is also only defined for DC cases, because IC cases could include too much noise, which renders their gq meaningless.

Intuitively, gq measures the density of co-author and venue graphs, denoted by $G_c = (V_c, E_c)$ and $G_v = (V_v, E_v)$, respectively, which are generated from the associated papers of a co-author group. Nodes in G_c represent researchers who at least appear once in these papers, and edges in G_c represent co-author relationships found in DBLP. Similarly, nodes in G_v are venues collected from these papers, and edges in G_v indicate the adjacent venues match (Section 6.2.1). It is apparent that the higher the density of these two graphs, the closer this group is connected in terms of collaborations and research interests, which are two key factors for measuring the identity of this co-author group. Mathematically, the density for an arbitrary graph is defined as its number of edges normalized by the number of edges of the corresponding fully connected graph [7]. Thus, the group quality gq is defined as a weighted sum of the density scores of G_c and G_v. The weights α_c and α_v are taken from Section 6.2.1 to globally weigh our interest on co-authorship and venue:

$$gq = \alpha_c \cdot \frac{2|E_c|}{|V_c|^2 - |V_c|} + \alpha_v \cdot \frac{2|E_v|}{|V_v|^2 - |V_v|}.$$

Similar to venue confidence vc, the final **co-author confidence** $cc(c)$ about using co-author c as an evidence for disambiguation is

$$cc(c) = \mathbb{1}_{DC}(c) \cdot (cf(c) \ + \ gq(c)).$$

7 Visual Design

In designing our visualization techniques, we follow the design rationales discussed in Section 3.2, to present the otherwise too abstract quantitative comparisons between confirmed and ambiguous authors.

7.1 Relation View

The relation view presents the many-to-many comparisons between confirmed and ambiguous authors. Inspired by ConVis [17], we design a three column visualization to assist intuitive comparisons. Once an author name is selected, the first column displays the confirmed list (a_1 in Fig. 1), and the third one is for all the ambiguous authors (a_3). These two columns are connected with a **comparison link** column in the center (a_2). Notice that we simultaneously display multiple ambiguous authors and comparison links.This is because besides the one-to-one comparison for individual confirmed-ambiguous pairs, we also want to learn the relative contrast among various pairs (R.5).

7.1.1 First Column: Confirmed List

The first column displays all confirmed groups in rows. Every row (Fig. 4(a)) encodes the attributes of all papers in a group G with a rectangle (B in Fig. 4) and a small line chart (C in Fig. 4). The rectangle consists of multiple thick bars that have the same width (D in Fig. 4), each representing a paper $p_i \in P_G$. These bars are sorted and filled with colors based on the corresponding allocation likelihoods in AL_G to help reveal the accuracy of the group (R.4): if the bar saturations are all high, then the confirmed group is very likely to have no false positives. The line chart summarizes the **temporal distributions** of the publications, so to put the time information into consideration.

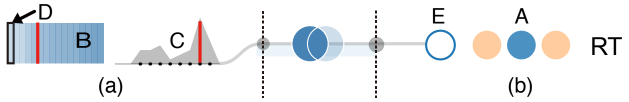


Fig. 4. Three-column design for the relation view. (a) denotes a confirmed group, and (b) an ambiguous author.

7.1.2 Third Column: Ambiguous List

The rows of the third column show the ambiguous groups. Given that these authors and their corresponding papers essentially have one-to-one relationships, we directly encode the papers' information (Fig. 4(b)). As shown in Fig. 4(b), for an ambiguous paper, we arrange a series of filled circles from left to right to represent the coauthor list. The ambiguous author currently under investigation is highlighted with blue (A) to indicate his/her position in the list. For further clarity, we place an empty circle (E) to the leftmost to denote the connection of this ambiguous paper with the confirmed ones.

7.1.3 Second Column: Comparison Link

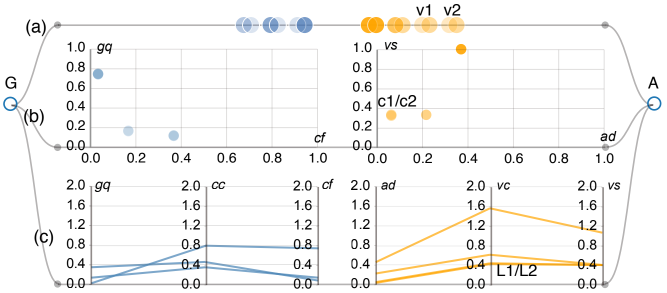


Fig. 5. Three ways to visually compare a confirmed group G with an ambiguous paper A: (a) the Venn glyph we design, (b) the side-by-side scatterplots, with the left representing co-authors and the right for venues, and (c) the side-by-side parallel coordinates. $v1$, $c1$, $L1$ all represent one venue, so as $v2$, $c2$ and $L2$.

The second column contains all the comparison links for contrasting the other two columns. Every link connects two rows, and shows the comprehensive similarity evaluation of the corresponding confirmed group G and ambiguous author A by representing both (1) the allocation likelihood AL for assigning A to G, and (2) our confidence in the evidence proving such likelihood.

For AL, because it only serves as hints toward the exploration sequences for disambiguation, but not as strong evidence to peremptorily identify an ambiguous author, we directly reflect it on the ordering of the comparison links. In addition, because there are too many possible links, we place a line chart at the top of this column (F in Fig. 1) to summarize the distribution of AL scores. Users can brush it to select and display a small set of comparison links for detailed explorations.

To help users comprehensively evaluate how trustful the estimated AL values are, we need to integrate all the measures modeling venue and co-author confidences in Table. 1(b). Two straightforward ways to do so is to visualize venue and co-author confidences as two side-by-side parallel coordinates (Fig. 5(b)) or scatterplots (Fig. 5(c)). However, the visual elements of similar measures can easily overlap in these designs. For instance, in Fig. 5, it is very difficult to distinguish c_1 and c_2 or L_1 and L_2. Moreover, these designs take too much space to be compared with each other efficiently (R.5). We need a design that can balance the data encodings and the space efficiency.

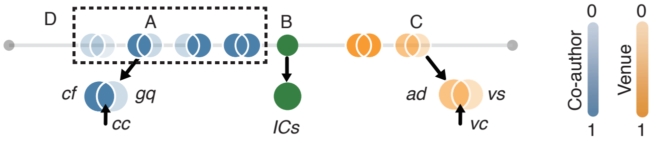


Fig. 6. Design of the comparison link. Each link contains multiple glyphs representing confidences.

Thus, we use small two-set Venn Diagrams [33] to present both the individual attributes (i.e., ad, vs, cf, and gq) and their combinations (i.e., vc and cc) compactly (Fig. 5(a)). For a pair of confirmed group G and ambiguous author A, their directly overlapped co-authors (i.e., $C_A \cap C_G$) and venues with concentrated research interest (i.e., $\{v : \mathbb{1}_{vr}(v) = 1,\ v \text{ is the venue of } p \in P_G\}$) are encoded using Venn glyphs (Fig. 6) on a comparison link. We use colors to distinguish these two types of glyphs: blue represents co-authors, and orange represents the venues. For the co-author glyphs, the saturation of the left circle encodes cf, and that of the right circles encodes gq. For venue glyphs, the saturation of the left indicates ad, and the right one is for vs. The higher the measurements, the more saturated the corresponding circles would be. Therefore, the intersection aggregating the saturations of both circles intuitively encodes cc or vc. It will be very conspicuous as long as at least one circle is highly saturated. The distance between circle centers is kept consistent so the size of the intersection does not affect the visual perception of saturations. The Venn glyph encoding effectively reveal different attribute combinations. For example, D in Fig. 6 shows four co-author glyphs. Both circles in the leftmost glyph are lightly colored, indicating that the corresponding co-author is a weak evidence. Meanwhile, the three other glyphs have at least one highly saturated circle, and thus saturated intersections. Therefore, they all convey confidences towards the co-authors.

Aside from the above, we further collect all the IC cases into one set and represent them with a single green circle (B in Fig. 6). The size of this always-fully-saturated circle encodes the number of IC authors to show the extensiveness of this group. This is because the useful metrics, ad and gq, are not defined for IC. Hence, we can ensure IC cases do not dominate the visual display facilitating the final judgment.

7.2 Temporal View

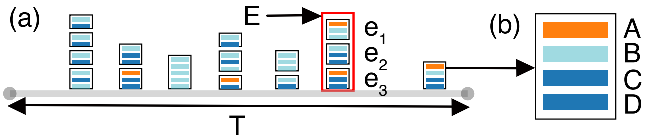


Fig. 7. Design of the temporal view; (a) is the overall layout, and (b) the specific design for paper rectangles in the confirmed group.

While the relation view provides an overview of comparisons, it is still not sufficient. Because authors' profiles and research interests may change over time, those confirmed papers only work as strong indicators if they are published around the ambiguous one temporally. For instance, if a confirmed author has no papers around the time when the ambiguous one is published, the paper might belong to a different author even though this confirmed author does share multiple co-authors with the ambiguous one. Therefore, we support every comparison link with a **temporal view** by showing detailed temporal distributions of the confirmed papers.

Specifically, when users find a comparison link interesting and expands it, a temporal view (Fig. 7(a)) appears for the corresponding confirmed and ambiguous authors. Each paper in the confirmed group is represented by a rectangle (Fig. 7(b)), which embodies the detailed comparison between this confirmed paper and the ambiguous one. Such a rectangle contains multiple segments that denote strict overlaps for venues and co-authors between the two papers. The first segment encodes the publication venue, and the rest represent co-authors of this confirmed paper. If it is published at the same venue as the ambiguous paper, the first line is colored in orange (A in Fig. 7(b)). If any of its co-authors also appear in the ambiguous paper, the corresponding segments are colored in blue (C and D). Otherwise, the segments are colored in very light cyan just to denote their existence (B). These color encodings align with those in the comparison link.

To further show the temporal features, we stack these individual paper rectangles vertically together based on their publication years on a horizontal timeline (T in Fig. 7(a)). We outline the year when the ambiguous paper is published with a red border (E). For example, we see three papers published in the same year as the ambiguous one in Fig. 7(a). Two of them (e_1 and e_3) are on the same venue as the ambiguous one, and two (e_2 and e_3) share two co-authors with it separately. This way, we can easily visualize whether the ambiguous paper is published around the same time when the confirmed author is active at the same venue, or with the same group of collaborators.

7.3 Group View

Both the aforementioned two views strictly take the pre-identified authors as premises. This means identifying new or misclassified authors is not possible. To cover this shortage, we further design a **group view** to "verify the persons", thus facilitating the evaluation of whether all the researchers with the same name have been thoroughly classified.

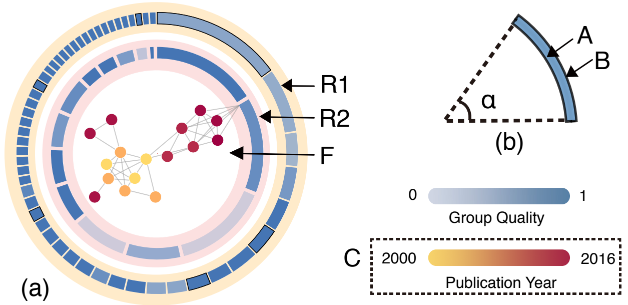


Fig. 8. (a) The layout for the group view. R_1 and R_2 present paper collections that potentially belong to the same person, and F shows the inner structures of one collection. (b) An enlarged arc taken from R_1.

Concretely, the group view does two things. First, it divides the input papers into subsets of two types:

- **Confirmed subsets**, each representing a confirmed author identified by digital libraries or users; and
- **Ambiguous subsets**, each containing a number of ambiguous papers that may belong to the same author. To obtain initial ambiguous subsets, we roughly allocate all the ambiguous papers into the subsets, such that papers share co-author similarities only with those in the same subsets.

Then, the group view guides users to explore their inner relationships, enabling them to recognize new authors or rectify misclassified ones.

Therefore, it is designed to be an enclosure structure. All the candidate subsets are encoded using two concentric rings (Fig. 8(a)), with the outer layer R_1 gathering all the ambiguous subsets, and the inner one R_2 representing all the confirmed ones. Each arc in these two ring layers represent one subset. The central angle of an arc (α in Fig 8(b)) encodes the total number of papers in the corresponding subset, and the saturation (A in Fig 8(b)) conveys our confidence (R.4) that this subset indeed refers to a person (i.e., its group quality gq).

Once an arc is selected, the relationship of its papers is displayed as a force layout in the center (F in Fig. 8(a)). Every paper is encoded with a node. The nodes are colored with respect to their publication years to emphasize the temporal differences (C in Fig. 8). If two papers share the same co-authors, we add an edge to the corresponding nodes to simulate a co-author graph [21, 23]. Thus, the denser the graph, the more likely it is that those papers belong to one person.

Moreover, given that papers in an ambiguous subset may belong to a confirmed author, we further link the ambiguous subsets with the confirmed ones if they have overlapped co-authors. We highlight such ambiguous subsets by drawing a dark stroke (B in Fig. 8(b)) around the corresponding outer arcs. When showing the co-author graph for an ambiguous subset, all the related confirmed arcs are connected. That is, if an ambiguous node shares co-authors with papers in a confirmed subset, the corresponding arc will emit an edge toward the node and exerts an attraction force on it. In addition, nodes in the confirmed arc can also be added to the the co-author graph. If so, the original ambiguous paper nodes will be enclosed by a dark stroke (C in Fig. 13(b)). In this way, not only can we recognize new authors from ambiguous subsets, but we can also double-check whether the potential new authors can merge with the existing confirmed ones.

7.4 Interaction

NameClarifier supports various interactions, which are described below, to facilitate the effective performance of name disambiguation.

Browse the three views interchangeably. Our relation, graph, and temporal view are closely connected, with the latter two supporting the first one. Users can display a temporal view with a single click on the corresponding comparison link, and then the temporal view can support the disambiguation with additional details. As for verifying authors, when users hover on a specific confirmed/ambiguous author record in the relation view, the corresponding author arc or paper node in the group view is highlighted with red borders, and vice versa. Similarly, when users select an ambiguous author arc in the group view, all the involved ambiguous papers are brought to the top in the relation view to help users simultaneously browse them side-by-side.

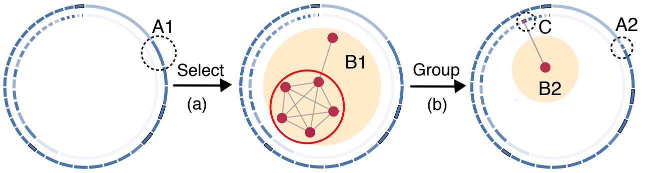


Fig. 9. Interactions in the group view. (a) Users can select an arc A1 to display its inner structures B1. (b) Circling the closely connected paper nodes in B1 creates a new confirmed arc C. The original A1 is then updated into A2, whose inner structure displays as B2.

Query visual feedback on demand. We allow users to query the data they find interesting. Users can brush on the *AL* distribution bar (F in Fig. 1) to select a specific group of ambiguous papers, so to avoid visual clutter when the name is too popular. They can also select a specific author arc in the group view to evaluate the corresponding co-author graph (Fig. 9(a)). Moreover, sorting the relation view helps prioritize the best possible case to disambiguate next. In this way, users can easily access the disambiguation in a targeted manner.

Iteratively disambiguate the data collection. Users can directly identify ambiguous authors. In the relation view, if users find an ambiguous paper to be significantly similar to a confirmed group, they can merge the ambiguous one into the confirmed one by selecting the corresponding comparison link. Once such ambiguous papers are all assigned, users can move to the group view and identify new confirmed authors by exploring co-author graphs (Fig. 9(b)). They can also eliminate noises from a confirmed group by dragging out paper nodes from the co-author graph of a confirmed arc. In particular, merging ambiguous papers, identifying new confirmed authors, and eliminating noises all trigger the Analyzer. By changing the confirmed author list, these operations enable the Analyzer to completely re-calculate the allocation likelihood, as well as the confidences for the ambiguous papers that are connected to the affected confirmed groups. These updated attributes are then reflected on the visualization to support the next round of disambiguation (R.6).

Track back the performed disambiguation processes. Because users only make judgements based on the most recent visual feedback, they may misclassify ambiguous papers. To rectify such misclassification, we provide a track log to help them reverse their disambiguation process (E in Fig. 1).

8 USE CASES

In our experiment, we have applied our system to DBLP dataset and successfully disambiguated numerous popular names. Here, we take two Chinese names as examples because many duplicated spellings make the name disambiguation problem particularly tricky for Asian names. The first case "Wei Chen" mainly demonstrates the basic functionalities of our system, whereas the second case "Rui Wang" evaluates the usefulness of NameClarifier in dealing with difficult cases.

8.1 The Case of "Wei Chen"

We use 1,170 publications with "Wei Chen" appearing in the author list to test whether NameClarifier can effectively guide the disambiguation. In this dataset, 25 distinct researchers have been identified and indexed in DBLP, from "Wei Chen 0001" to "Wei Chen 0025".

We notice that the first comparison links for many ambiguous papers have at least one highly saturated Venn glyph and a large green circle. This means that the paper has a very high allocation likelihood *AL* (observed from the ordering) to the corresponding confirmed author group with fairly good co-author/venue confidences. These cases can often be easily disambiguated. For instance, in Fig. 10(a), an ambiguous author A_1 yields two comparison links L_1 and L_2 to two confirmed groups G_1 and G_2, respectively. They both have two orange glyphs, but L_1 has two more blue glyphs representing more directly connected co-authors. In particular, glyph c_1 is especially eye-catching because of two significantly saturated circles. It shows that a co-author of A_1 who does not suffer from ambiguity greatly has collaborated frequently with the confirmed author of G_1. Hovering on both links, we observe that the red strokes in G_1 and G_2 are also informative. The red stroke in the middle of the blue rectangle of G_1 means A_1 is more likely to belong to G_1 than many papers that are already in G_1, and the stroke on the timeline shows that the paper falls in a peak period during which the confirmed author publishes frequently. We open the temporal view and observe that the red border in the view for L_1 is indeed surrounded by rectangles with many orange or blue strokes, but this is not the case for L_2. Thus, we confirm A_1 fits well into the publication trajectory of G_1, and place A_1 into G_1.

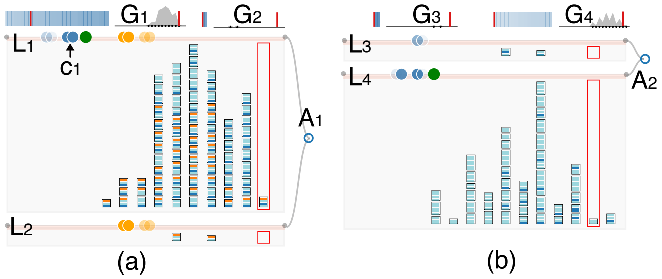


Fig. 10. Disambiguating two "Wei Chen" cases. A_1 and A_2 are two ambiguous papers, and $G_i, i \in \{1,2,3,4\}$ are confirmed groups. L_i are the comparison links connecting with the corresponding G_i. (a) A_1 is easily allocated to G_1 because L_1 leads in both AL and confidence. (b) Putting A_2 into G_4 is trickier, since L_4 has lower AL but higher confidence.

In certain instances, AL and confidence do not align with each other. For instance, in Fig. 10(b), the system places comparison link L_3 on the top of L_4, which indicates that $AL(A_2, G_3) > AL(A_2, G_4)$. However, L_4 is much more attractive in terms of the number of Venn glyphs and their saturations. In the temporal view, we observe that AL for G_3 is scored higher because the group size of G_3 is relatively small. Undoubtedly, $AL(A_2, G_3)$ is significantly increased because the only two papers in G_3 both have similar co-authors with the ambiguous one. However, the comparison link can alert us to verify the result further.

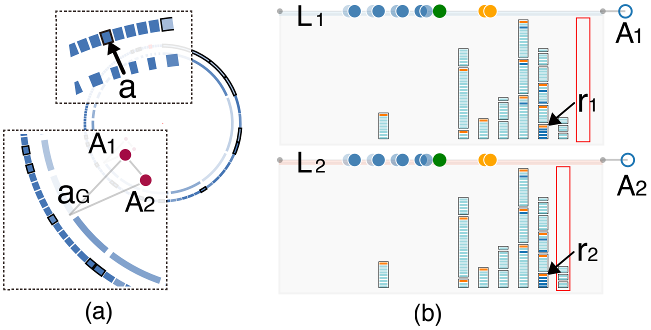


Fig. 11. Locating papers with the group view: we find from the outer layer two papers that are highly similar. Selecting them helps us notice their promising confidence.

The cases with clear evidence are all resolved after several rounds of exploration in the relation view. Then, when the low confidences reflected on the comparison links block us from browsing further, we switch to the group view to look for additional clues. We immediately notice several small ambiguous arcs that have high saturations and are enclosed by dark strokes. These arcs contain a small number of ambiguous papers that are closely connected to certain confirmed groups. For instance, in Fig. 11(a), we expand ambiguous arc a and find it contains two inter-connected ambiguous papers A_1 and A_2, which are also connected to confirmed arc a_G. We refer back to their comparison links L_1 and L_2 in the relation view (Fig. 11(b)). The saturated Venn glyphs indicate we are very confident in L_1 and L_2, although their AL is not high enough to be ranked at the top. Browsing their temporal views, we observe that it s due to the low co-author overlap (i.e., there are only a small number of blue strokes in the entire view). We further notice that these highlighted blue strokes are all stacked around the year right before the red border, including highly similar paper rectangles r_1 and r_2. This means that the overlapped collaborations have just been established, and that the ambiguous paper is possibly a follow-up paper carried out by the same team. Therefore, we confidently classify this paper into the corresponding confirmed author group. Although the automated quantification does not work well in this case, our temporal view quickly provides additional guidances.

8.2 The Case of "Rui Wang"

Distinguishing two recognized researchers named "Rui Wang" in DBLP, Rui Wang 0003 and 0004, is much trickier. This is because (1) their research area, both being graphics and visualization, greatly overlaps, (2) they have similar co-authors on various papers, and (3) they have even collaborated with each other on a few papers. We load this dataset, which contains 560 paper records (179 confirmed and 381 ambiguous ones), into NameClarifier to see if it can handle such tricky cases. A total of 15 researchers have been confirmed.

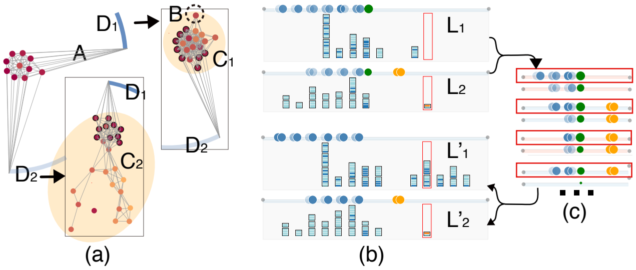


Fig. 12. Disambiguating papers of Rui Wang. (a) Papers in the ambiguous arc A are strongly and loosely connected to papers in the confirmed arcs D_1 and D_2, respectively. (b) and (c) Iterative disambiguation in the relation view improves L_1 and L_2 into more distinguishable L_1' and L_2'.

We find the great overlap is reflected in the comparison links and temporal views. For instance, L_1 and L_2 are a pair of comparison links for one paper (Fig. 12(b)). They cannot be easily distinguished because L_1 has more co-authors, whereas L_2 shows another paper at a similar venue exactly when this paper is published. We locate these papers in the group view (Fig 12(a)) and find they are all located in one ambiguous author arc that contains a highly connected co-author graph (A). This cluster is connected to two confirmed author arcs D_1 and D_2. We extend these two arcs into the co-author graph A separately to evaluate the relationships among D_1, D_2, and papers in A. We can see that papers in D_1 merge with papers in A to form a tight cluster C_1. By contrast, the merged result C_2, formed by paper in A and D_2, only contains nodes that are loosely connected. We suspect all these papers in A belong to D_1. Thus, we start with those nodes that are at the farthest from D_2 (e.g., B in Fig. 12(a)), and classify these ambiguous papers one by one. Through several rounds of iteration in the relation view (Fig. 12(c)), the comparison links and temporal views in L_1 and L_2 are updated into L_1' and L_2' respectively (Fig. 12(b)). We easily affirm L_1' is more prominent in this case. This proves that our iterative update can smoothly deal with those otherwise indistinguishable cases.

After distinguishing these two authors and disambiguating other confident cases in the relative view, we move to the graph view to identify new confirmed authors. We start with the longest ambiguous arc A. From its co-author graph (Fig. 13(a)), we first notice a significant compact cluster B. We group the paper nodes in B into a new confirmed group (A_{G2} in Fig. 13(b)). A new confirmed author row is added to the relation view, and we determine from its dark blue rectangles (Fig. 13(c)) that papers in B indeed belong to one researcher. Then, we explore the relationships between the only related confirmed arc (A_{G1} in Fig. 13(a)) and the cluster of ambiguous papers (C in Fig. 13(a)) that connects to it. We merge the papers of A_{G1} with the remaining co-author graph (Fig. 13(b)) and see how these nodes influence each other. We find a paper (p in Fig. 13(b)), which originally belongs to A_{G1}, is dragged close to cluster C. This strange behavior makes us suspect that this paper p by Liu et al. [26] was misclassified by DBLP, which was confirmed later during our research. Thus, NameClarifier can effectively refine the predefined confirmed groups.

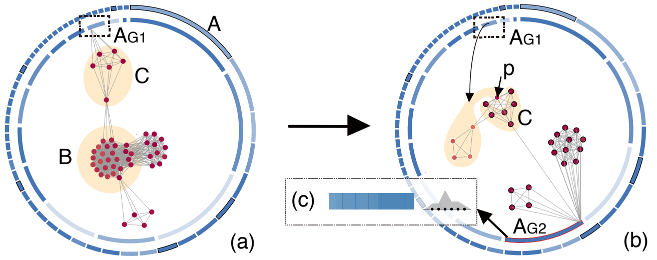


Fig. 13. Rectifying a wrongly classified case p by identifying a new researcher from cluster B and observing the co-author graph.

The subsequent exploration reveals another interesting case, in which the temporal information plays an important role. Fig. 14 shows an independent ambiguous arc whose co-author graph is divided into two parts in terms of time and collaboration closeness denoted as C_1 and C_2. To check whether these all belong to the same author, we first group the more compact cluster C_1into as a new confirmed group A_G. Then, we evaluate the relationship between A_G and its linked paper p in the temporal view (Fig. 14(b)). We observe a large temporal gap T between the red border denoting the publication year t_p for p and A_G's main publication period t_{AG}. Moreover, blue and orange strokes for co-authors and venues are rare. Thus, we confirm that the four papers in C_2 are not published by the same researcher in C_1.

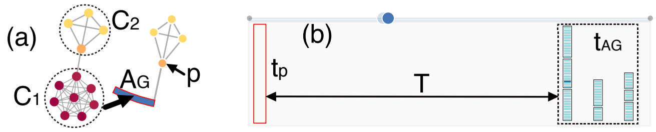


Fig. 14. A distinguishable case through temporal information.

9 Expert Review

To evaluate the effectiveness of NameClarifier, we conducted one-on-one interviews with two experts who are responsible for maintaining the publication records of the university.

Background. Expert A has worked on literature retrieval and scholar database design for sixteen years, and expert B on scholar database management for five years. In a prior discussion, both experts agreed that name disambiguation was a big concern in their work. They merely used automatic methods because of their inaccurate performance. Currently, they devote a large amount of manpower to manually compare ambiguous names with all the papers that have been identified. They know nothing about NameClarifier before.

Process. In each interview, we first introduced the background and visual encodings of NameClarifier, and demonstrated how the system worked using the "Wei Chen" case (Section 8.1). Then we asked the experts to freely explore ambiguous names with NameClarifier on their own, during which we answered their questions and observed their behaviors. Finally, we collected their feedback on their use experience with the system. One such section lasted about 90 minutes.

Feedback. Overall, both experts felt that they can easily understand our system and resolve ambiguous cases according to the visual hint, which improved the efficiency significantly. Expert A especially emphasized that the system can effectively help him find the relationships between ambiguous names and confirmed authors, and thereby narrow down the scopes he needed to search. The experts credited this convenience to the well selected attributes, which were exactly what they commonly used to address the same problem. They further suggested us present as much information as possible. For example, titles and keywords in the publication could be very helpful if we use them together with co-authors and venues.

As for the designs of individual views, Expert A particularly liked the temporal view. He said this view provided an intuitive overview on the concrete matching between papers, which can replace their current manual process. When using it together with the relation view, his confidence was greatly improved. Expert B agreed that he would always look for additional supports in the temporal view however significant the comparison links were. He further suggested that we simplify the temporal view and make it more intuitive by deleting unnecessary details (e.g., cyan lines in those totally unmatched paper rectangles). Nevertheless, he was impressed by the group view because it was helpful in organizing ambiguous names. He also confirmed that the relation view was very useful for comparing the strengths of the different relations. However, Expert B did find the encodings of this view complicated. Though he understood the metaphors during the introduction, it was difficult to remember how the visual encoding related to attributes. Therefore, he transferred the detailed encodings into more intuitive "darker is better" instructions in practice.

Because expert A liked our system very much, he also discussed the possibility of using NameClarifer in the university library to improve their efficiency on name disambiguation. He hoped that we can expand the system to support more available attributes like email and affiliation, so that the system can better fit smaller scholar databases. He also mentioned that the combination of the group view and the relation review may relate to how skillful the users are. Thus, he suggested us add mis-operation warning functions to further guide the users.

10 Discussion and Conclusion

In this paper, we propose NameClarifier, an interactive visual system for name disambiguation. NameClarifier contains three linked views that transfer multi-faceted comparisons and intrinsic uncertainties into visual feedback, thus handing back the decision-making to the users. Two use cases and an expert review prove that our system effectively guides users to iteratively enrich and rectify the confirmed group evidence, providing support for both apparent and tricky cases. While NameClarifier is specifically designed for name disambiguation, it is also valuable for much broader purposes. For instance, the three-column relation view can help resolve various cases that involve many-to-many comparisons and joint structures of multiple attributes. It can also work for general entity resolutions or interactive visual labeling in machine learning problems.

Although useful and effective, NameClarifier has some design limitations. First, while we integrate various metrics into our visualization to describe the paper similarities, NameClarifier still heavily relies on human beings' subjective judgments. Subjective assessment is, on the one hand, undoubtedly important for measuring tricky cases. However, on the other hand, the lack of deterministic mathematical verification can potentially lead to dilemmatic decisions when different users have inverse preferences on cases with low confidences. Moreover, because there is no ground truth collection in which every paper's author is known, we cannot verify our correctness. Thus, achieving a confirmed, complete, and correct disambiguation still requires more effort. This can be alleviated by combining our system with some well-defined mining algorithms, so the disambiguation can be double-checked by the human and the algorithm. Another limitation is the scalability of our system. When an ambiguous author name is too popular, the relation view can only display a very small portion of possible comparisons, which limits our understanding on the whole dataset. We plan to further adopt non-linear scaling or fisheye interactions, so that the system can help pinpoint specific visual shapes while maintaining a macro-perception of the whole dataset.

In the future, we will first collect an artificial dataset to verify our system's accuracy. We will also test some other attributes for measuring author similarities so as to pinpoint the best combination. Furthermore, we will investigate more data mining methods that could be incorporated into NameClarifier so to achieve a more symbiotic relationship between subjective assessments and objective measurements.

Acknowledgments

We gratefully thank the experts for their feedback during the expert review, and the anonymous reviewers for their valuable and constructive comments. The project was supported by RGC GRF 618313.

References

[1] M. Bilgic, L. Licamele, L. Getoor, and B. Shneiderman. D-dupe: An interactive tool for entity resolution in social networks. In *Visual Analytics Science And Technology, 2006 IEEE Symposium On*, pages 43–50. IEEE, 2006.

[2] R. Borgo, J. Kehrer, D. H. Chung, E. Maguire, R. S. Laramee, H. Hauser, M. Ward, and M. Chen. Glyph-based visualization: Foundations, design guidelines, techniques and applications. *Eurographics State of the Art Reports*, pages 39–63, 2013.

[3] N. Cao, D. Gotz, J. Sun, and H. Qu. Dicon: Interactive visual analysis of multidimensional clusters. *Visualization and Computer Graphics, IEEE Transactions on*, 17(12):2581–2590, 2011.

[4] N. Cao, Y.-R. Lin, L. Li, and H. Tong. g-Miner: Interactive visual group mining on multivariate graphs. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 279–288. ACM, 2015.

[5] W. W.-Y. Chan. A survey on multivariate data visualization. *Department of Computer Science and Engineering. Hong Kong University of Science and Technology*, 8(6):1–29, 2006.

[6] P. Christen. *Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection*. Springer Science & Business Media, 2012.

[7] T. F. Coleman and J. J. Moré. Estimation of sparse jacobian matrices and graph coloring blems. *SIAM journal on Numerical Analysis*, 20(1):187–209, 1983.

[8] C. Collins. Docuburst: Document content visualization using language structure. In *Proceedings of IEEE Symposium on Information Visualization, Poster Compendium*, 2006.

[9] T. Crnovrsanin, I. Liao, Y. Wu, and K.-L. Ma. Visual recommendations for network navigation. In *Computer Graphics Forum*, volume 30, pages 1081–1090. Wiley Online Library, 2011.

[10] W. Cui, Y. Wu, S. Liu, F. Wei, M. X. Zhou, and H. Qu. Context preserving dynamic word cloud visualization. In *2010 IEEE Pacific Visualization Symposium (PacificVis)*, pages 121–128. IEEE, 2010.

[11] M. Dörk, N. H. Riche, G. Ramos, and S. Dumais. Pivotpaths: Strolling through faceted information spaces. *Visualization and Computer Graphics, IEEE Transactions on*, 18(12):2709–2718, 2012.

[12] A. Endert, P. Fiaux, and C. North. Semantic interaction for visual text analytics. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 473–482. ACM, 2012.

[13] X. Fan, J. Wang, X. Pu, L. Zhou, and B. Lv. On graph-based name disambiguation. *Journal of Data and Information Quality (JDIQ)*, 2(2):10, 2011.

[14] A. A. Ferreira, M. A. Gonçalves, and A. H. Laender. A brief survey of automatic methods for author name disambiguation. *ACM Sigmod Record*, 41(2):15–26, 2012.

[15] E. Garfield. British quest for uniqueness versus American egocentrism. *Nature*, 223:763, 1969.

[16] T. Gurney, E. Horlings, and P. Van Den Besselaar. Author disambiguation using multi-aspect similarity indicators. *Scientometrics*, 91(2):435–449, 2012.

[17] E. Hoque and G. Carenini. Convis: A visual text analytic system for exploring blog conversations. In *Computer Graphics Forum*, volume 33, pages 221–230. Wiley Online Library, 2014.

[18] P. Jaccard. *Etude comparative de la distribution florale dans une portion des Alpes et du Jura*. Impr. Corbaz, 1901.

[19] P. H. Kanani, A. McCallum, and C. Pal. Improving author coreference by resource-bounded information gathering from the web. In *International Joint Conferences on Artificial Intelligence (IJCAI)*, pages 429–434, 2007.

[20] H. Kang, L. Getoor, B. Shneiderman, M. Bilgic, and L. Licamele. Interactive entity resolution in relational data: A visual analytic tool and its evaluation. *Visualization and Computer Graphics, IEEE Transactions on*, 14(5):999–1014, 2008.

[21] I.-S. Kang, S.-H. Na, S. Lee, H. Jung, P. Kim, W.-K. Sung, and J.-H. Lee. On co-authorship for author disambiguation. *Information Processing & Management*, 45(1):84–97, 2009.

[22] D. A. Keim, F. Mansmann, and J. Thomas. Visual analytics: how much visualization and how much analytics? *ACM SIGKDD Explorations Newsletter*, 11(2):5–8, 2010.

[23] J. Kim, H. Kim, and J. Diesner. The impact of name ambiguity on properties of coauthorship networks. *Journal of Information Science Theory and Practice*, 2(2):6–15, 2014.

[24] B. Lee, M. Czerwinski, G. Robertson, and B. B. Bederson. Understanding research trends in conferences using paperlens. In *CHI'05 extended abstracts on Human factors in computing systems*, pages 1969–1972. ACM, 2005.

[25] M. Levin, S. Krawczyk, S. Bethard, and D. Jurafsky. Citation-based bootstrapping for large-scale author disambiguation. *Journal of the American Society for Information Science and Technology*, 63(5):1030–1047, 2012.

[26] D. Liu, E. Liu, Z. Zhang, R. Wang, Y. Ren, Y. Liu, I.-H. Ho, X. Yin, and F. Liu. Secondary network connectivity of ad hoc cognitive radio networks. *Communications Letters, IEEE*, 18(12):2177–2180, 2014.

[27] S. Liu, W. Cui, Y. Wu, and M. Liu. A survey on information visualization: recent advances and challenges. *The Visual Computer*, 30(12):1373–1393, 2014.

[28] J. Matejka, T. Grossman, and G. Fitzmaurice. Citeology: visualizing paper genealogy. In *CHI'12 Extended Abstracts on Human Factors in Computing Systems*, pages 181–190. ACM, 2012.

[29] R. L. Maxwell. *Maxwell's guide to authority work*. American Library Association, 2002.

[30] M. J. McGuffin. Simple algorithms for network visualization: A tutorial. *Tsinghua Science and Technology*, 17(4):383–398, 2012.

[31] S. Milojević. Accuracy of simple, initials-based methods for author name disambiguation. *Journal of Informetrics*, 7(4):767–773, 2013.

[32] D. A. Pereira, B. Ribeiro-Neto, N. Ziviani, A. H. Laender, M. A. Gonçalves, and A. A. Ferreira. Using Web information for author name disambiguation. In *Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries*, pages 49–58. ACM, 2009.

[33] F. Ruskey and M. Weston. A survey of venn diagrams. *Electronic Journal of Combinatorics*, 4:3, 1997.

[34] N. R. Smalheiser and V. I. Torvik. Author name disambiguation. *Annual review of information science and technology*, 43(1):1–43, 2009.

[35] J. Stasko, J. Choo, Y. Han, M. Hu, H. Pileggi, R. Sadanaand, and C. D. Stolper. Citevis: Exploring conference paper citation data visually. *Posters of IEEE InfoVis*, 2013.

[36] A. Strotmann, D. Zhao, and T. Bubela. Author name disambiguation for collaboration network analysis and visualization. *Proceedings of the American Society for Information Science and Technology*, 46(1):1–20, 2009.

[37] J. Tang, A. C. M. Fong, B. Wang, and J. Zhang. A unified probabilistic framework for name disambiguation in digital library. *Knowledge and Data Engineering, IEEE Transactions on*, 24(6):975–987, 2012.

[38] L. Tang and J. Walsh. Bibliometric fingerprints: name disambiguation based on approximate structure equivalence of cognitive maps. *Scientometrics*, 84(3):763–784, 2010.

[39] M. O. Ward. Multivariate data glyphs: Principles and practice. In *Handbook of data visualization*, pages 179–198. Springer, 2008.

[40] K.-H. Yang, H.-T. Peng, J.-Y. Jiang, H.-M. Lee, and J.-M. Ho. Author name disambiguation for citations using topic and web correlation. In *Research and advanced technology for digital libraries*, pages 185–196. Springer, 2008.

[41] J. Zhu, X. Zhou, and G. P. C. Fung. A term-based driven clustering approach for name disambiguation. In *Advances in Data and Web Management*, pages 320–331. Springer, 2009.