Author Name Disambiguation using a Categorical Distribution Similarity

Shaohua Li, Gao Cong, and Chunyan Miao

Nanyang Technological University
shaohua@gmail.com, {gaocong, ascymiao}@ntu.edu.sg

Abstract. Author name ambiguity has been a long-standing problem which impairs the accuracy of publication retrieval and bibliometric methods. Most of the existing disambiguation methods are built on similarity measures, e.g., "Jaccard Coefficient", between two sets of papers to be disambiguated, each set represented by a set of categorical features, e.g., coauthors and published venues[1]. Such measures perform bad when the two sets are small, which is typical in Author Name Disambiguation. In this paper, we propose a novel categorical set similarity measure. We model an author's preference, e.g., to venues, using a categorical distribution, and derive a likelihood ratio to estimate the likelihood that the two sets are drawn from the same distribution. This likelihood ratio is used as the similarity measure to decide whether two sets belong to the same author. This measure is mathematically principled and verified to perform well even when the cardinalities of the two compared sets are small. Additionally, we propose a new method to estimate the number of distinct authors for a given name based on the name statistics extracted from a digital library. Experiment shows that our method significantly outperforms a baseline method, a widely used benchmark method, and a real system.

Keywords: Name Disambiguation, Categorical Sampling Likelihood Ratio

1 Introduction

Bibliometrics is an important methodology to assess the output and impact of researchers and institutions. Ambiguous names which correspond to many authors are a long-standing headache for bibliometric assessors and users of digital libraries. For example, in DBLP, there are at least 8 authors named *Rakesh Kumar*, and their publications are mixed in the retrieved citations. The ambiguity on Chinese names is more severe, as many Chinese share a few family names such as *Wang*, *Li*, and *Zhang*. An extreme example is *Wei Wang*. According to our labeling, it corresponds to over 200 authors in DBLP! As more and more researchers become active, the ambiguity problem will only become graver.

Author Name Disambiguation refers to splitting the bibliographic records by different authors with the same name into different clusters, so that each cluster belongs to one author and each author's works are gathered in one cluster.

[1] Venues here refer to the journal or conference, such as *J. ACM* or *SIGIR*.

For each paper, we consider 3 features: *coauthors*, *published venue* and *title*, by following the setting used in previous work [5,3,12]. Under this setting, our proposed method can be general and applicable to the existing bibliography databases, e.g., DBLP, since they contain information on the three features for each paper. Each feature serves as a body of evidence used to decide whether two homonymous authors are the same person. *Coauthors* and *venues* are two important features that have categorical values. During disambiguation, we need measure the similarity between two clusters of papers. Naturally the feature values in each cluster form a set of categorical data, and thus a categorical set similarity measure is an important foundation of a disambiguation algorithm.

Given two sets of categorical data, previous methods of name disambiguation use set similarity measures, such as *Jaccard Coefficient* ([2,12]) or *cosine similarity* ([8]), which often fail when the sets are unbalanced in cardinality, or when the frequencies of the elements in each set have distinctive patterns (to be explained in Section 4). We exploit the property that categorical sets from the same author follow similar distributions, and propose a generative probabilistic model to estimate the similarity of two sets. We name this novel similarity measure as **Categorical Sampling Likelihood Ratio (CSLR)**.

In addition, the ambiguity (number of distinct people) of a disambiguated name needs to be estimated to guide the disambiguation process. We exploit the property that the different parts of a person name in a given culture are chosen roughly independently, and derive a simple statistical method to estimate the ambiguity, based only on the name statistics in a digital library. The estimated ambiguity is shown to be reasonably close to the actual value for Chinese names.

We evaluate our system on two test sets extracted from the January 2011 dump of DBLP. Experiments show that our method significantly outperform one baseline method (by 2-12%), a representative previous method DISTINCT (by 4-13%) and a well-known system Arnetminer [9] (`http://arnetminer.org/`) (by 6-17%) in terms of macro-average F1 scores.

The rest of this paper is organized as follows. In Section 2, we review related work. In Section 3, we define basic notations used in this paper, and state the objective of Author Name Disambiguation. In Section 4, we establish the novel set similarity measure CSLR. In Section 5, we outline our clustering system based on CSLR. In Section 6, we describe the *name ambiguity estimation* method. In Section 7, we report experimental results. Finally, we conclude in Section 8. In addition, all proofs are in the full version of this paper ([6]). The source code and data set are available at `http://github.com/askerlee/namedis`.

2 Related Work

A pioneering work [5] on Author Name Disambiguation presents two supervised learning approaches, using Naive Bayes and SVM, respectively. For each name to be disambiguated, a specific classifier is trained. Therefore, hand-labeled papers for each name are needed. This overhead is unaffordable in practice.

The method DISTINCT [12] uses SVM to learn the weights of features. The training data for SVM is generated automatically. The title is considered the

unigram "bag-of-words" (BoW). Each cluster of papers has a few features, and the similarity between feature value sets of two clusters is calculated using Jaccard Coefficient. As another similarity measure, the connection strength between clusters is measured by a random walk probability. The two similarity measures are combined and form the similarity used in the agglomerative clustering.

The work [2] formulats the Name Disambiguation problem as a hypergraph, where each author is one node. Relationships among authors, such as the coauthorship of a few authors, are represented as hyperedges. The similarity between two clusters is measured by comparing their "neighboring sets" (other clusters they connect with), using Jaccard Coefficient or Adamic/Adar Similarity.

Torvik et al. ([10]) develops a disambiguation system on MEDLINE. First a training set is automatically generated, and the likelihood ratio of each feature value as its evidential strength is estimated from the training set. Evidence provided by different feature values is aggregated under the Naive Bayes assumption, and the probability that two papers belong to the same author is estimated. Finally, a maximum likelihood agglomerative clustering is conducted.

Recently, Tang et al. ([8,11]) presents two closely-related methods based on Pairwise Factor Graph models. The authorship is modeled as edges between observation variables (papers) and hidden variables (author labels). Features of each paper, and relationships such as CoPubVenue and CoAuthor, have impact on the probability of each assignment of labels. The similarity between two clusters is encoded in different "factors" (edge potentials) on different features. The clustering process tries different author label assignments and finds the one with maximal probability. Moreover, [11] improves the disambiguation resultes based on user feedback, and is being used online in Arnetminer for disambiguation (`http://arnetminer.org/disambiguation`).

In addition to the title, co-authorship and venue information, authors' homepages ([11]), and results returned by a search engine ([7]) are also used for disambiguation. However, such information is not always available.

3 Problem Formulation

In a digital library, each author name e may correspond to one or more authors $\{a_1, a_2, \cdots, a_{\kappa(e)}\}$. Each a_i is called e's **namesake**. The number of namesakes $\kappa(e)$ is the **ambiguity** of name e. The estimated ambiguity is denoted by $\hat{\kappa}(e)$. The name e being disambiguated is called the **focus name**. Each paper d has a set of authors $A_d = \{a_1, a_2, \cdots, a_m\}$. Suppose a_i has name e. The rest authors (if any) $A_d \setminus \{a_i\}$ are the **coauthors** with regard to paper d, denoted by $\mathrm{co}(d)$.

We represent a collection of categorical data as a **multiset**. In contrast to the traditional set, here each element x in set S has a frequency value $\mathrm{freq}_S(x)$. $\mathrm{freq}_S(x)$ could be a real number after scaling. The **cardinality** of a multiset S, denoted by $|S|$, is the sum of frequencies of all its elements: $|S| = \sum_{x \in S} \mathrm{freq}_S(x)$. A multiset S is often represented as a list of pairs as $\{x_1 : f_1, \cdots, x_m : f_m\}$, where $f_i = \mathrm{freq}_S(x_i)$. Often we simply refer to a multiset as a **set** when the meaning is clear from context.

Given a set of papers $C = \{d_1, d_2, \cdots, d_n\}$ written by author a, the **coauthor set** of C is the union of coauthors[2] of all d_i, i.e., $\text{co}(C) = \cup_{i=1}^{n} \text{co}(d_i)$. Each coauthor $b_i \in \text{co}(C)$ has a frequency $\text{freq}_{\text{co}(C)}(b_i)$, which is the count of papers in C having b_i as a coauthor.

Likewise, we refer to the multiset of publication venues for the set of papers C as the **venue set** of C, denoted by $V(C)$. Each venue $v_i \in V(C)$ has a frequency $\text{freq}_{V(C)}(v_i)$, which is the number of papers in C published in v_i.

Problem Statement Given a focus name e and a set of papers authored by name e: $\mathcal{P}(e) = \{d_1, d_2, \cdots, d_n\}$, the problem of **name disambiguation** is to partition $\mathcal{P}(e)$ into different clusters $\{C_1, \cdots, C_{\kappa(e)}\}$, so that all papers in C_i are authored by person a_i and all the papers in $\mathcal{P}(e)$ by a_i are in C_i.

Before we present the proposed method for name disambiguation in Section 5, we first present the proposed similarity measure in Section 4, which lays the foundation of our method.

Notation	Description
e	An ambiguous name
$\kappa(e)$	Ambiguity of name e
a_i	An author (with no ambiguity)
C	A cluster of papers that belong to the same author
$\text{co}(C)$	Coauthor multiset of C: the union of coauthors of all papers in C
$V(C)$	Venue multiset of C: the union of venues of all papers in C
$\text{freq}_S(x)$	Frequency of an element x in a multiset S
S	A multiset, where each element $x \in S$ has a frequency
$\|S\|$	Cardinality of a multiset, i.e., the sum of frequencies of all elements
$\boldsymbol{p} = (p_0, p_1, \cdots, p_m)$	A parameter vector of a categorical distribution
B	Base Set (the larger one of two compared multisets S_1 and S_2)
BCD, $\mathcal{B}$	Base Categorical Distribution where B is drawn
A	Sampled Set (the smaller one of two multisets S_1 and S_2)
$\tilde{A}$	Conflated sampled set (all "unseen" outcomes become UNSEEN)
A'	Tolerated sampled set (by reducing some UNSEEN counts from $\tilde{A}$)
$\text{Cat}(\boldsymbol{p})$	A categorical distribution with the parameter vector $\boldsymbol{p}$
$\Pr(S\|\boldsymbol{p})$	Probability of drawing set S from $\text{Cat}(\boldsymbol{p})$
$S \sim \mathcal{D}$	The case of drawing S from distribution $\mathcal{D}$
$\Lambda(A, B)$	Categorical Sampling Likelihood Ratio (CSLR) between A and B

Table 1. Notation table

4 Categorical Sampling Likelihood Ratio – A Categorical Set Similarity Measure

In Section 4.1, we use a categorical distribution to model the preference of each author, introduce the intuition behind Categorical Sampling Likelihood Ratio

[2] As different coauthors with the same name are literally indistinguishable, the *coauthor* here may correspond to more than one actual author.

(CSLR), and formulate CSLR as the ratio of two likelihoods. In Section 4.2, we present methods to approximate the two likelihoods. Section 4.3 presents the proposed CSLR.

For ease of discussion, we present CSLR in the context of two venue sets, each representing a set of papers by an author. The comparison between two coauthor sets can be computed similarly.

4.1 Modeling using the Categorical Distribution and Motivation

Each author has preferences to the publication venue, and such preferences can be represented as a categorical distribution, namely the *Preference Distribution*. The frequency that the author published in a venue reflects the preference of this author to the venue. Consider a cluster of papers C belonging to author a. The venue of each paper in C is an observation of the preference distribution, and the whole venue set $V(C)$ forms a **sample** of that distribution. Suppose there are m possible outcomes (i.e., venues) in this distribution, denoted by $x_i,\ i = 1, \cdots, m$. Each x_i has a probability p_i drawn from this distributuion. We denote all the outcome probabilities as a vector: $\boldsymbol{p} = (p_1, \cdots, p_m)$. A categorical distribution with a parameter vector $\boldsymbol{p}$ is denoted by $\mathrm{Cat}(\boldsymbol{p})$. Therefore author a's preference distribution is $\mathrm{Cat}(\boldsymbol{p})$.

Different authors usually have distinctive preference distributions. Hence we can estimate the possibility that two clusters belong to the same author, by comparing the two distributions from which these venue sets are drawn. Such a problem is traditionally known as the *two-sample problem* ([4]).

The biggest challenge of the two-sample problem in Author Name Disambiguation is: during the clustering, a cluster of papers are often a small fragment of the complete set of papers by that author, and therefore the venue set is a **small sample** and often only a **partial observation** of the preference distribution. It is difficult to compare two distributions based only on two partial observations. Traditional categorical set/distribution similarity measures, such as *Jaccard Coefficient*: $J(A, B) = \frac{|A \cap B|}{|A \cup B|}$, its variant *Adamic/Adar Similarity*, *cosine similarity*, or *Kullback-Leibler divergence*, perform well when the sets A and B are large and good approximations of the underlying distributions, but do not fit well with Author Name Disambiguation. We take Jaccard Coefficient to illustrate the problems of these measures:

1. Sets A and B often have unbalanced cardinalities, and $J(A, B)$ is sensitive to the relative set cardinalities. In the extreme case that $A \subset B$, intuitively A, B are probably drawn from the same distribution (A is a smaller sample); however $J(A, B) = \frac{|A|}{|B|}$ varies drastically with the cardinality of either set;
2. The evidential strength of each shared element is usually regarded as the same, regardless of their relative importance. But some elements are more discriminative than others. For example, suppose x is the most frequent elemnt in B, but absent in A. Then it is strong evidence that A and B follow different distributions, and are dissimilar. But if x appears once in B and absent in A, it is only weak evidence. Note adding weights to elements

does not help much, e.g., *Adamic/Adar Similarity*, the weighted version of Jaccard Coefficient, is shown to perform worse than Jaccard Coefficient ([2]).

To this end, we propose a new measure. Assume two multisets A, B have arisen under one of the two hypotheses H_0 and H_1. The null hypothesis H_0 here is: A and B are drawn from different distributions (and thus belong to different authors). The alternative hypothesis H_1 is: A and B are drawn from the same distribution (and thus belong to the same author). We want to see how likely one hypothesis holds relative to the other. The more likely H_1 is relative to H_0, the more similar are A and B.

Formally, we estimate both $\Pr(H_1|B,\, A)$ and $\Pr(H_0|B,\, A)$. We compare these two posterior probabilities and get a likelihood ratio $\Lambda = \dfrac{\Pr(H_1|B,\, A)}{\Pr(H_0|B,\, A)}$. We use the likelihood ratio as the similarity between A and B.

We assume a flat prior on the two hypotheses: $\Pr(H_0) = \Pr(H_1) = 0.5$. By applying Bayes' theorem (the proof can be found in [6]), we get

Theorem 1.

$$\Lambda = \frac{\Pr(H_1|B,\, A)}{\Pr(H_0|B,\, A)} = \frac{\Pr(A|B,\, H_1)}{\Pr(A|B,\, H_0)}.$$

To compute the likelihood ratio, we need to compute the two probabilities that A is seen, given B and one of the hypotheses, H_0 and H_1.

4.2 Calculating the Two Likelihoods

Computing $\Pr(\boldsymbol{A|B,\, H_1})$ Consider two authors a_1 and a_2, whose preference distributions are $\mathrm{Cat}(\boldsymbol{p_1})$ and $\mathrm{Cat}(\boldsymbol{p_2})$, respectively, and whose venue sets are A and B, respectively.

We proceed to estimate $\Pr(A|B,\, H_1)$. First, suppose hypothesis H_1 holds. Then $\boldsymbol{p_1} = \boldsymbol{p_2}$. This implies, given B and H_1, A is drawn from $\mathrm{Cat}(\boldsymbol{p_2})$. Let $\Pr(A|\boldsymbol{p_2})$ be the probability that A is drawn from $\mathrm{Cat}(\boldsymbol{p_2})$. Then $\Pr(A|B,\, H_1) = \Pr(A|\boldsymbol{p_2})$.

We estimate $\boldsymbol{p_1}, \boldsymbol{p_2}$ from A and B and get $\hat{\boldsymbol{p}}_1, \hat{\boldsymbol{p}}_2$, respectively. Then

$$\Pr(A|B,\, H_1) = \Pr(A|\boldsymbol{p_2}) \approx \Pr(A|\hat{\boldsymbol{p}}_2).$$

Note in Theorem 1, A and B are symmetric and exchangeable. Empiracally a larger sample tends to better reflect the actual distribution $\mathrm{Cat}(\boldsymbol{p_i})$. Without loss of generality, suppose $|B| \geq |A|$. Then $\mathrm{Cat}(\hat{\boldsymbol{p}}_2)$ is probably a better estimation of $\mathrm{Cat}(\boldsymbol{p}_2)$ than $\mathrm{Cat}(\hat{\boldsymbol{p}}_1)$ as an estimation of $\mathrm{Cat}(\boldsymbol{p}_1)$. The likelihood $\Pr(A|\hat{\boldsymbol{p}}_2)$ would likely be more accurate than $\Pr(B|\hat{\boldsymbol{p}}_1)$. So we choose B as the conditioning set, namely the *Base Set*, from which we estimate a *Base Categorical Distribution* (BCD) $\mathcal{B}$, and the smaller set A as the conditioned set, namely the *Sampled Set*. If $|A| > |B|$, we simply exchange A and B.

Let us denote the base set as $B = \{x_1 : f_1,\, x_2 : f_2,\, \cdots,\, x_n : f_n\}$, and the sampled set as $A = \{y_1 : g_1,\, y_2 : g_2,\, \cdots,\, y_m : g_m\}$, where x_i, y_j are outcomes (venues), and $f_i = \mathrm{freq}_B(x_i)$, $g_j = \mathrm{freq}_A(y_j)$. We can estimate $\mathcal{B}$ from B using Maximum Likelihood Estimation (MLE): $\hat{p}_i = \frac{f_i}{\sum_i f_i}$.

Considering that B may not cover all the outcomes in $\mathcal{B}$, we should tolerate outcomes in A but not in B. We introduce a "wildcard" outcome: UNSEEN (denoted by x_0, drawn with a small probability p_0). Any outcome in A but not in B is treated as UNSEEN, without discrimination. We adopt the widely used *Jeffreys prior*([1]) to assign a pseudocount $\delta = 0.5$ to UNSEEN and all the observed outcomes in B. The smoothed estimator gives the following parameters:

$$\hat{p}_0 = \frac{\delta}{\delta(n+1) + \sum_i f_i}, \qquad \hat{p}_i = \frac{f_i + \delta}{\delta(n+1) + \sum_i f_i}, \text{ for } i = 1, \cdots, n. \quad (1)$$

The estimated $\mathcal{B}$ is $\hat{\mathcal{B}} = \text{Cat}(\hat{\boldsymbol{p}}_2) = \text{Cat}(\hat{p}_0, \hat{p}_1, \cdots, \hat{p}_n)$.

Before calculating the probability that A is drawn from $\hat{\mathcal{B}}$, we partition A into two sets – the "seen" outcomes A_s and the "unseen" ones A_u, and conflate A_u into UNSEEN:

1) $A_s = A \cap B$. Suppose $A_s = \{y_1 : g_1, \ldots, y_t : g_t\}$. We align (relabel) the elements in B with A_s, so that $x_i = y_i$, for $i = 1, ..., t$ (the remaining outcomes in B are labeled as $x_{t+1}, \cdots, x_n$ arbitrarily). Then outcome y_i is drawn with probability $\hat{p}_i$ from $\hat{\mathcal{B}}$;
2) $A_u = A \setminus B$ is the unseen outcomes. Suppose $A_u = \{y_{t+1} : g_{t+1}, \ldots, y_m : g_m\}$. All elements in A_u are "conflated" to UNSEEN (x_0). Let the frequency of x_0 be g_0, then $g_0 = |A_u| = \sum_{i=t+1}^{m} g_i$.

We denote the conflated set as $\tilde{A}$. We have $\tilde{A} = \{x_0 : g_0, y_1 : g_1, \ldots, y_t : g_t\}$. Note the conflation does not change the cardinality of the set, i.e., $|\tilde{A}| = |A|$. Then the probability that drawing A from distribution $\mathcal{B}$, denoted by $A \sim \mathcal{B}$, is approximated by the probability that $\tilde{A} \sim \hat{\mathcal{B}}$:

$$\Pr(A|B, H_1) \approx \Pr(\tilde{A}|\hat{\boldsymbol{p}}_2) = \binom{|A|}{g_0, g_1, \cdots, g_t} \hat{p}_0^{g_0} \prod_{i=1}^{t} \hat{p}_i^{g_i}, \quad (2)$$

where $\binom{|A|}{g_0, g_1, \cdots, g_t}$ is the multinomial coefficient, counting the total number of sequences with the same frequencies of outcomes as in A.

Toleration of Preference Divergence: Converting from A to A' The preference distribution of an author often evolves slowly with time. Thus an author has different preference distributions at different periods; however typically these categorical distributions share many common outcomes, and the probabilities of shared outcomes are still close. Thus the difference between the preference distributions of the same author at different times is usually much smaller than the difference between the distributions of different authors.

Consider two sets A and B, both belonging to author a, are drawn from slightly different preference distributions $\text{Cat}(\boldsymbol{p_1})$ and $\text{Cat}(\boldsymbol{p_2})$, respectively, where the parameter vectors $\boldsymbol{p_1}$ and $\boldsymbol{p_2}$ are similar but not identical. Let B be the base set, and $\hat{\mathcal{B}}$ is the estimated BCD. When we calculate the probability that $A \sim \hat{\mathcal{B}}$, A may contain a few "unseen" outcome occurrences with respect to $\hat{\mathcal{B}}$, as well as a lot of "seen" outcome occurrences. These UNSEEN occurrences are all assigned a tiny probability $\hat{p}_0$, and contribute $c \cdot \hat{p}_0^{g_0}$ (c is a small factor

in the multinomial coefficient) in (2), which reduces the probability drastically (although the majority of outcome occurrences are "seen"), wrongly indicating that A and B unlikely belong to the same author. The "culprit" of this undesirable result is the few "unseen" outcomes. In other words, the direct likelihood estimation is too stringent and intolerant to deviation from $\hat{\mathcal{B}}$.

To allow for preference divergence, before we calculate the likelihood, we reduce some count of UNSEEN, proportional to the cardinality of A. This strategy is called *toleration*. The kept outcome occurrences form a new *Tolerated Set* A'.

To perform toleration on set A, first we conflate the "unseen" outcomes in A and get $\tilde{A}$. Parameter θ_t controls the UNSEEN count to be reduced relative to A's cardinality, i.e., UNSEEN frequency g_0 will be reduced by $\theta_t|A|$. If UNSEEN frequency $g_0 < \theta_t|A|$, then the new frequency $g_0' = 0$. We set $\theta_t = \frac{1}{3}$. We denote the tolerated set as $A' = \{x_0 : h_0, y_1 : h_1, \cdots, y_r : h_r\}$, where $h_0 = g_0'$, and $h_i = \text{freq}_A(y_i)$, for $\forall i > 0$. The probability in (2) becomes $\Pr(A'|\hat{\boldsymbol{p}}_2)$:

$$\Pr(A|B,\, H_1) \approx \Pr(A'|\hat{\boldsymbol{p}}_2) = \binom{|A'|}{h_0,\, h_1,\, \cdots,\, h_r} \hat{p}_0^{h_0} \prod_{i=1}^{r} \hat{p}_i^{h_i}. \tag{3}$$

Computing $\Pr(A'|B,\, H_0)$ In the following, the sampled set in our likelihood estimation is the tolerated set A'. We will estimate $\Pr(A'|B,\, H_0)$ first.

The hypothesis H_0 states that A' and B are drawn from different categorical distributions, i.e., A' is drawn from a distribution other than $\text{Cat}(\boldsymbol{p}_2)$. Since any randomly-chosen categorical distribution is probably dissimilar to $\text{Cat}(\boldsymbol{p}_2)$, we can approximate $\Pr(A'|B,\, H_0)$ by $\Pr(A')$, i.e., the probability that A' is drawn from a categorical distribution $\text{Cat}(\boldsymbol{p})$, where we have no information about $\boldsymbol{p}$.

We limit the sample space of any possible categorical distribution $\text{Cat}(\boldsymbol{p})$ to the set of outcomes in $\mathcal{B}$: $\{x_1, \cdots, x_n\}$. Naturally, we assume a flat Dirichlet $\text{Dir}(\mathbf{1}_n)$ as the prior distribution of $\boldsymbol{p}$, where $\mathbf{1}_n = (1, \cdots, 1)$ is n dimensional.

Suppose $A' = \{x_0 : h_0, y_1 : h_1, \cdots, y_{r-1} : h_{r-1}, y_r : h_r\}$, then we can represent A' by the frequency vector of its elements: $\boldsymbol{h} = (h_0, h_1, \cdots, h_r, h_{r+1}, \cdots, h_n)$, where $h_{r+1} = \cdots = h_n = 0$. Then we have the following Theorem.

Theorem 2.

$$\Pr(A'|B,\, H_0) \approx \Pr(A') = \int_{\boldsymbol{p}} \Pr(\boldsymbol{h}|\boldsymbol{p}) \Pr(\boldsymbol{p}; \mathbf{1}_n) d\boldsymbol{p} = \frac{1}{\binom{|A'|+n}{n}}, \tag{4}$$

where $\Pr(\boldsymbol{p}; \mathbf{1}_n)$ *denotes the probability of drawing* $\boldsymbol{p}$ *from* $\text{Dir}(\mathbf{1}_n)$.

The proof can be found in [6]. Theorem 2 reveals an interesting fact: $\Pr(A')$ is only determined by $|A'|$, A's cardinality, and n, the number of categories in B, but irrelevant to the histogram of outcome frequencies in A'.

4.3 Categorical Sampling Likelihood Ratio (CSLR)

As we have obtained two approximations of the two likelihoods in Eq. (3) and Theorem 2, we combine them and get the approximation of Λ:

$$\Lambda \approx \frac{\Pr(A'|\hat{\boldsymbol{p}}_2)}{\Pr(A')} = \binom{|A'|}{h_0,\, h_1,\, \cdots,\, h_r} \binom{|A'|+n}{n} \hat{p}_0^{h_0} \prod_{i=1}^{r} \hat{p}_i^{h_i}. \tag{5}$$

We name Λ as *Categorical Sampling Likelihood Ratio* (CSLR). It is directly used as the similarity between two categorical sets, such as venue sets and coauthor sets. For two sets A and B, we denote their CSLR as $\Lambda(A, B)$.

5 Clustering Framework

5.1 Overview of the Clustering Procedure

We use Agglomerative Clustering as the basic framework. It starts with each paper being a cluster, and at each step we find the most similar (the similarity measures will be defined later) pairs of clusters, and merge them, until the maximal similarity falls below certain threshold, or the cluster number is smaller than the estimated ambiguity of the disambiguated name. The whole clustering process divides into two stages:

1. Merge based on the evidence from shared coauthors;
2. Merge based on the combined similarity defined on the title sets and venue sets of each pair of clusters.

The reasons for developing the two-stage clustering are twofold: First, coauthors generally provide stronger evidence than other features, based on which the generated cluster usually comprises of papers of the same author, but the papers of an author may distribute among multiple clusters ([3]); Second, the venue and title features are relatively weak evidence, based on which we can further merge clusters from the same author.

5.2 Stage 1: Merging by Shared Coauthors

The existing work ([5,12,10,2,3]) usually takes shared coauthors as a crucial feature. They usually treat all authors equally, and combine two clusters if they have shared coauthors. However, we observe that the strength of the evidence provided by a shared coauthor varies from one to another. If a coauthor collaborates with many people, it is likely that the coauthor collaborate with different people with the same focus name. Especially when the focus name to be disambiguated has high ambiguity, the chance of different people sharing the same coauthor names would be high. Hence, we propose to distinguish those weak evidential coauthors from the strong evidential coauthors and treat them differently. For example, consider to disambiguate "Wei Wang". Coauthors *Jiawei Han* and *Jian Pei* both collaborate with different "Wei Wang". We observe that both *Jiawei Han* and *Jian Pei* have over 200 collaborators, and thus they should be treated as weak evidential coauthors when disambiguating "Wei Wang".

We proceed to present a statistical approach to estimating the probability that a coauthor b works with only one namesake of a given name e. Given that a coauthor b is shared by two clusters C_1 and C_2, the alternative hypothesis H_1 says C_1 and C_2 belong to the same author. If $\Pr(H_1|b)$ is large enough ($\geq \theta_{co}$), then b is regarded as *strong evidential*, and we merge C_1 and C_2. Otherwise b is *weak evidential*. Here θ_{co} is the decision threshold. We choose $\theta_{co} = 0.95$.

Let e be the disambiguated focus name. Suppose that the coauthor b randomly chooses n authors from the whole author set $\mathbb{A}$[3] to collaborate with, and among the n collaborators at least one person a_1 has name e. The total count of authors is denoted by $M = |\mathbb{A}|$. We assume the choice of collaboration follows a uniform distribution $\mathcal{U}$ over $\mathbb{A}$. Thus the n collaborators are viewed as n independent trials from $\mathcal{U}$, where each author $a_i \in \mathbb{A}$ has probability $1/M$ to be chosen[4]. Since one trial is reserved for a_1, only $n-1$ trials are really random. Suppose we have known e's ambiguity $\kappa(e)$. Then in each trial, choosing another author with name e has probability $\frac{\kappa(e)-1}{M-1} \approx \frac{\kappa(e)-1}{M}$.

The probability that no other collaborator of b has name e is:

$$\Pr(H_1^*|b) = (\frac{M-\kappa(e)}{M})^{n-1} \approx 1 - \frac{(n-1)\kappa(e)}{M}, \tag{6}$$

considering $\kappa(e) \ll M$. H_1^* means that for any pair of clusters C_1 and C_2, H_1 holds. So $H_1^* \implies H_1$, and $\Pr(H_1^*|b) \leq \Pr(H_1|b)$.

But we do not know n, the actual number of collaborators of b. We only know b has collaborated with $|\operatorname{co}(b)|$ **names**. So $n \geq |\operatorname{co}(b)|$. We can obtain n's expectation $\mathrm{E}[n]$ as n's estimation:

$$\mathrm{E}[n] \approx \frac{M(|\operatorname{co}(b)|-1)}{M - \sum_{e_i \in \operatorname{co}(b)} (\kappa(e_i)-1)}, \tag{7}$$

where $\kappa(e_i)$ is approximated by $\hat{\kappa}(e)$ in Section 6, and $M \approx \sum_{e \in \mathbb{A}} \hat{\kappa}(e)$.

Strong evidential coauthors require $\Pr(H_1|b) \geq \theta_{co}$. Combining this with Eq. (6), we obtain

$$n \leq \frac{(1-\theta_{co})M}{\kappa(e)} + 1. \tag{8}$$

The right-hand value of Eq. (8) is a threshold value to partition authors into two groups: one contains authors who have fewer coauthors than the threshold, and thus provide strong evidence; the other contains authors who have more coauthors than the threshold and thus offer weak evidence.

Given two clusters C_1 and C_2, if there is one shared strong evidential coauthors, then we see enough evidence supporting H_1, and then we merge them. Otherwise all shared coauthors are weak evidential. We use CSLR to see how likely the two coauthor sets are drawn from the same distribution. If $\Lambda(\operatorname{co}(C_1), \operatorname{co}(C_2)) > 1$, we merge C_1 and C_2.

5.3 Stage 2: Merging by Venue Set and Title Set

Consider a pair of clusters C_1 and C_2 with venue sets V_1, V_2, and title sets T_1, T_2. We denote the Venue Set Similarity by $\operatorname{sim}_V(V_1, V_2)$, and Title Set Similarity by

[3] $\mathbb{A}$ includes all authors in the DBLP dump.

[4] The n trials is without replacement. The probability is approximated by trials with replacement. This approximation is good, since $n \ll M$.

$\text{sim}_T(T_1, T_2)$. These two similarity measures are heterogeneous metrics, and we multiply them to compute the combined similarity:

$$\text{sim}(C_1, C_2) = \text{sim}_V(V_1, V_2) \cdot \text{sim}_T(T_1, T_2). \tag{9}$$

As the ambiguity $\kappa(e)$ of an author e increases, there are more and more authors working in the same subfields and publishing in the same venues. Therefore the clustering threshold in this stage, denoted by θ_c, should increase monotonically with $\kappa(e)$. We set θ_c as a linear function of $\hat{\kappa}(e)$:

$$\theta_c(\hat{\kappa}(e)) = 0.2 \cdot \max(1, \frac{1}{5}\hat{\kappa}(e)) \tag{10}$$

Due to space limitations, the technical details of using CSLR to compute the similarity $\text{sim}_V(V_1, V_2)$ and using BoW to compute $\text{sim}_T(T_1, T_2)$ are omitted here, and can be found in the full version of this paper ([6]).

Next we briefly introduce the idea of computing the two similarities.

Venue Set Expansion and Similarity We use CSLR to compare two venue sets. But CSLR treats different outcomes as disparate and their correlations are not considered. Often two venue sets do not share common venues, but the venues are correlated, such as "TKDE" in one set, and "CIKM" in the other. They still favor (to certain degree) the hypothesis that the two clusters are from the same author. In this case, CSLR returns a very small likelihood ratio.

To remedy this problem, before computing CSLR, we expand each venue set with correlated venues first. Now a venue set {TKDE: 2, CIKM: 3} could become {TKDE: 2, CIKM: 3, ICDM: 1, KDD: 0.5}, and the CSLR value between it and another set {ICDM: 3, KDD: 1} will become reasonably large.

The idea is to predict the frequencies of absent but correlated venues of a set, based on observed venues, and then add the predicted $\{venue : frequency\}$ pairs into that set. The correlated venues are mined using *linear regression* on the 1.5 million DBLP papers.

We denote the expanded venue set of V_i as $\tilde{V}_i$, then $\text{sim}_V(V_1, V_2) = \Lambda(\tilde{V}_1, \tilde{V}_2)$.

Title Set Similarity based on Unigram BoW We adopt the traditional unigram BoW model to represent two title sets and calculate their similarity. The similarity $\text{sim}_T(T_1, T_2)$ is the weighted sum of shared unigrams[5]. The weighting scheme is a variant of TF*IDF, which regards all the titles of an author as a single document when calculating the Inverse Document Frequency (IDF).

6 Name Ambiguity Estimation

We present a statistical method to estimate the ambiguity $\kappa(e)$ of each focus name e. The estimation $\hat{\kappa}(e)$ is used in (8) and (10). In addition, it plays two

[5] Words in the titles are so sparse and diverse that even if two title sets belong to the same author, the two corresponding sets of words are usually not drawn from the same distribution, and thus CSLR does not fit in here.

other roles: First, it is one of the stop criteria of the clustering. Once we reach $\hat{\kappa}(e)$ clusters, we should stop merging. Note the clustering may stop before the number of clusters becomes $\hat{\kappa}(e)$ due to other criteria. Second, if $\hat{\kappa}(e)$ is much less than 1, it means name e is rare, and it is highly possible that only one person has this name, regardless how many papers is authored by e. For example, in our dataset, 448 papers have author name *Jiawei Han*. We assert that all of them are by the same person, given that *Jiawei Han*'s estimated ambiguity is 0.29.

Our method is inspired by the "Ambiguity Estimate" intuition in [2]. Our estimation only needs the names statistics in a digital library.

In the digital library names in a given culture usually have a fixed number of parts. For example in DBLP, a Chinese name usually has 2 parts (e.g., "Xiaofeng" and "Wang" for name "Xiaofeng Wang"). Suppose that these parts were chosen roughly independently with each other. Thus we can estimate the probability of each option of each part, and then the probability of a full name is the joint probability of its parts.

We formulate the case of 3-part names as an example. Suppose a name e in a given culture consists of a given name $G(e)$, a middle name $M(e)$ and a family name $F(e)$, i.e., $e = G(e)+M(e)+F(e)$, where "+" means string concatenation.

For any name e in this culture, we assume $G(e)$, $M(e)$ and $F(e)$ are drawn independently from 3 categorical distributions Cat_G, Cat_M and Cat_E, respectively. Then $\Pr(e) = \Pr(G(e)) \Pr(M(e)) \Pr(F(e))$.

The parameters of Cat_G, Cat_M and Cat_E are estimated using MLE. Take Cat_G as an example. Let $\mathbb{E}$ be the set of all names in this culture, and $\mathbb{G}$ be the set of all given names in this culture,

$$\forall g \in \mathbb{G}, \quad \Pr(G(e) = g) \approx \frac{\sum_{e \in \mathbb{E}, G(e)=g} \kappa(e)}{\sum_{\forall e \in \mathbb{E}} \kappa(e)}. \tag{11}$$

Noticing $\sum_{\forall e \in E} \kappa(e)$ is the total number of different authors in this culture, the MLE of the instances (i.e., ambiguity) of name e in the DBLP author set is:

$$\hat{\kappa}(e) = \Pr(G(e)) \Pr(M(e)) \Pr(F(e)) \sum_{\forall e \in E} \kappa(e). \tag{12}$$

We do not know $\kappa(e)$, and thus we use $\hat{\kappa}(e)$ in place of $\kappa(e)$, and evaluate (11) and (12) iteratively, until $\hat{\kappa}(e)$ converges. It is possible that $\hat{\kappa}(e) < 1$ (a rare name), so during the iteration, we round $\hat{\kappa}(e)$ to 1 if $\hat{\kappa}(e) < 1$. Specifically,

1. Initially, $\forall e, \hat{\kappa}_0(e) = 1$;
2. In the $(i+1)$-th iteration, we plug $\max(\hat{\kappa}_i(e), 1)$ for $\kappa(e)$ into (11) and (12), evaluate them and get $\hat{\kappa}_{i+1}(e)$. Repeat this step until $|\sum_{\forall e} \hat{\kappa}_{i+1}(e) - \sum_{\forall e} \hat{\kappa}_i(e)| \leq \epsilon_m$, where ϵ_m is a small number to measure the convergence.

When the estimation converges at the n-th iteration, we round $\hat{\kappa}_n(e)$ up to 1 and get $\hat{\kappa}(e)$. If we want to check the rarity of a name, we use $\hat{\kappa}_n(e)$ directly.

Note the name-part independence assumption holds only among names in a given culture. Given names from one culture and family names from another culture are usually anti-correlated, for example "Jacob Li" is a very rare combination. So Ambiguity Estimation should be conducted culture-wise. For names in

a culture which are too few in the digital library to form a large enough sample, external demographic data could be incorporated to get better estimation.

Table 2. Statistics of Data Set 1*

Name e	**#Pubs**	$\kappa(e)$	$\hat{\kappa}(e)$
Hui Fang	9	3	1.62
Ajay Gupta	16	4	n/a
Joseph Hellerstein	151	2	n/a
Rakesh Kumar	36	2	n/a
Michael Wagner	29	5	n/a
Bing Liu	89	6	6.91
Jim Smith	19	3	n/a
Lei Wang	55	13 (31)	22.34
Wei Wang	140	14 (57)	49.43
Bin Yu	44	5 (11)	8.7

Table 3. Statistics of Data Set 2

Name e	**#Pubs**	$\kappa(e)$	$\hat{\kappa}(e)$
Hui Fang	45	8	6.8
Ajay Gupta	25	8	n/a
Joseph Hellerstein	234	2	n/a
Rakesh Kumar	104	8	n/a
Michael Wagner	61	16	n/a
Bing Liu	192	23	21.0
Jim Smith	54	5	n/a
Lei Wang	400	144	104.6
Wei Wang	833	216	254.2
Bin Yu	102	18	17.3

* [12] removed authors who have only one paper from their data set. So for the last three names in Table 2, [12] reported much smaller ambiguities than the real values, which are given in the parentheses.

7 Experimental Results

7.1 Experimental Setting

Data Set Two test sets are used. For fairness of comparison, both use the same set of names as in [12]. Papers written by these names in DBLP are extracted for disambiguation.

Set 1 is the same dataset as that used in [12]. Its statistics are listed in Table 2. This data set was extracted from a 2006 dump of DBLP.

Set 2 is extracted from a January 2011 dump of DBLP. Each name corresponds to many more papers (and bigger ambiguity, as more authors with these names publish) in Set 2 than Set 1. Their statistics are in Table 3. All these papers were hand-labeled and available at the URL given in Section 1.

As a part of our experiments, we test Ambiguity Estimation on Chinese author names, and list the results on names in the test set in Tables 2 and 3. Set 1 was built at the beginning of year 2006 ([12]), so we use the DBLP statistics before 2006 to estimate these ambiguities. Set 2 contains all authors and papers in DBLP till January 2011, and we use the whole DBLP statistics to estimate these ambiguities. The actual ambiguities $\kappa(e)$ are obtained by hand-labeling.

For Chinese names, our method gives a reasonable estimation: $\hat{\kappa}(e) \in (0.5\kappa(e), 1.5\kappa(e))$. We have not estimated the ambiguities of names in other cultures. But usually their ambiguities are small (below 30) and we set all of them to 2. Experiments show such inaccuracy does not impair the performance of our system noticeably.

Evaluation As in [12,8], we use *Pairwise Precision*, *Pairwise Recall*, and *Pairwise F1* scores to evaluate the performance of our method and other methods. Specifically, any two papers that are annotated with the same label in the ground truth are called a correct pair, and any two papers that are predicted with the same label (if they are grouped in the same cluster, we also call they have the same label) by a system but are labeled differently in the ground truth are called a wrong pair. Note the counting is for pairs of papers with the same label (either predicted or labeled) only. Thereafter, we define the three scores:

$$\text{Prec} = \frac{\#\,\text{PairsCorrectlyPredicted}}{\#\,\text{TotalPairsPredicted}} \quad \text{Rec} = \frac{\#\,\text{PairsCorrectlyPredicted}}{\#\,\text{TotalCorrectPairs}}$$

$$\text{F1} = \frac{2 \times \text{Prec} \times \text{Rec}}{\text{Prec} + \text{Rec}}$$

Experimental Details We evaluated one baseline, denoted by Jac, which uses Jaccard Coefficient for coauthor/venue sets, the unigram BoW based similarity for title sets, and Eq. (10) as its clustering threshold. The optimal Jaccard Coefficient thresholds for coauthor sets and venue sets are different. We tested Jac with different thresholds, and chose the thresholds for coauthor sets and venue sets that produce the highest macro-average F1 scores, respectively. The best thresholds are 0.03 for coauthor sets, and 0.04 for venue sets.

We compared our method with two representative methods: DISTINCT ([12]) and Arnetminer ([11]). We acquired the original source code of DISTINCT. DISTINCT uses randomly generated training sets, and in different runs its performance varies greatly. Moreover, DISTINCT does not have a mechanism to determine a clustering threshold for a given name. Instead it tries 12 different thresholds between $[0, 0.02]$. For each name, different thresholds lead to disparate performance. So we ran DISTINCT 10 times and averaged its scores at each threshold, and then took the threshold that gives the highest macro-average F1 score over all names, as the chosen threshold (0.002 for Set 1, 0.001 for Set 2). Additionally, we crawled the disambiguation pages of these 10 names from `http://arnetminer.org/` on March 12, 2012, and extracted the disambiguation results. These results are generated by the up-to-date work of [11]. As Arnetminer contains papers newer than the release date of our DBLP dump, we discarded papers that are not in our data sets.

We refer to our own method as CSLR. It has 3 important parameters: θ_t, which controls the degree of toleration; θ_{co}, which controls the decision threshold between strong/weak-evidential coauthors; and $\theta_c(\hat{\kappa}(e))$, which controls when to stop the second-stage clustering. They are tuned on a development set of 5 names: *Tao Peng*, *Peng Cheng*, *David Jensen*, *Xiaodong Wang*, and *Gang Wu*.

7.2 Experimental Results and Discussion

The results for all methods are shown in Table 4 and 5. For each method, the most important measure, the macro-average F1 score over all names, is underlined. On both sets, CSLR significantly outperforms all the other methods.

Table 4. Comparison of Performance on Set 1

Name	Jac			Arnetminer			DISTINCT			Our (CSLR)		
	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
Hui Fang	**100.0**	**100.0**	**100.0**	55.6	**100.0**	71.4	85.6	**100.0**	88.7	**100.0**	**100.0**	**100.0**
Ajay Gupta	**100.0**	93.1	96.4	**100.0**	**100.0**	**100.0**	67.7	94.5	78.8	**100.0**	93.1	96.4
Joseph Hellerstein	50.7	83.9	63.2	97.4	**97.4**	**97.4**	92.4	80.6	84.6	**100.0**	69.7	82.1
Rakesh Kumar	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**
Michael Wagner	**100.0**	64.0	78.1	**100.0**	33.7	50.5	90.1	**96.2**	**92.9**	**100.0**	64.0	78.1
Bing Liu	**99.8**	84.5	**91.5**	86.2	79.8	82.9	86.5	82.0	83.6	91.8	**87.0**	89.4
Jim Smith	**100.0**	83.1	90.8	**100.0**	84.5	91.6	95.6	**91.7**	**93.3**	**100.0**	87.3	93.2
Lei Wang	**100.0**	71.2	**83.2**	59.4	**94.2**	72.9	42.5	75.0	51.8	**100.0**	63.3	77.6
Wei Wang	**60.5**	83.7	**70.2**	28.1	98.5	43.8	31.0	**98.8**	47.1	59.3	72.4	65.2
Bin Yu	70.7	64.7	67.6	87.8	**95.3**	**91.4**	77.1	89.2	81.3	**98.8**	68.5	80.9
Avg. (macro-F1)	88.2	82.8	84.1	81.5	88.4	80.2	76.9	**90.8**	80.2	**95.0**	80.5	**86.3**

Table 5. Comparison of Performance on Set 2

Name	Jac			Arnetminer			DISTINCT			Our (CSLR)		
	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
Hui Fang	**100.0**	68.8	81.5	59.1	63.7	61.3	81.3	**97.9**	88.0	**100.0**	78.9	**88.2**
Ajay Gupta	**96.0**	47.0	63.1	60.0	65.4	62.6	65.3	**87.9**	**74.2**	**96.0**	39.6	56.1
Joseph Hellerstein	52.8	80.5	63.7	94.5	**95.9**	**95.2**	92.3	89.5	90.0	**100.0**	79.6	88.6
Rakesh Kumar	**100.0**	89.0	94.2	98.4	89.3	93.7	89.9	96.0	92.5	99.9	**97.8**	**98.8**
Michael Wagner	**92.8**	59.4	72.4	55.6	36.7	44.2	67.4	**98.2**	**79.1**	88.1	64.6	74.6
Bing Liu	97.8	67.0	79.5	75.7	67.2	71.2	83.0	**84.7**	83.3	**98.1**	74.7	**84.8**
Jim Smith	**100.0**	44.1	61.2	88.6	45.1	59.7	94.8	**87.8**	**90.0**	**100.0**	48.8	65.6
Lei Wang	30.0	79.8	43.6	18.1	83.1	29.8	29.3	85.9	42.4	**78.1**	**87.6**	**82.6**
Wei Wang	40.2	77.0	52.8	9.7	**88.2**	17.5	25.8	84.2	38.9	**81.0**	71.8	**76.1**
Bin Yu	70.6	42.8	53.3	72.4	**62.2**	**66.9**	54.0	62.0	57.0	**88.0**	49.1	63.0
Avg. (macro-F1)	78.0	65.5	66.5	63.2	69.7	60.2	68.3	**87.4**	73.5	**92.9**	69.2	**77.8**

On Set 1 DISTINCT has a lower macro-average F1 score than that reported in [12]. We think it is partly due to the random nature of DISTINCT when it chooses a random training set to trains the feature weights. But since we have run DISTINCT for consecutive 10 times, we think the average scores truly reflect its performance in practice without ground truth to select the best trained weights.

On Set 2 Arnetminer has a sudden performance drop compared to its performance on Set 1. One important "culprit" is its precision on *Wei Wang* is extremely low. As we can see in the actual disambiguation result online at `http://arnetminer.org/`, 727 papers are credited to the professor at UNC, among which we believe only < 200 papers are authored by her. The reason might be Arnetminer merges clusters based on a few weak evidential coauthors.

The baseline Jac performs well on Set 1. This may ascribe to two factors: 1) It uses the optimal Jaccard Coefficient thresholds, which are impossible to obtain in practice without ground truth; 2) It uses the same estimated name ambiguity to set the clustering threshold. However, Jac's performance plunges on Set 2 where the ambiguity of each name is larger. This contrast suggests the adverse effect of the inaccuracy of Jaccard Coefficient intensifies as the ambiguity grows.

Compared to other methods, our system has slightly lower recall, but much higher precision. We think a major reason is that CSLR returns a high similarity only when two clusters follow similar distributuions. Sometimes clusters of papers

by the same author are drastically different (e.g., very few shared venues and shared terms in titles), and it is difficult even for a human to decide whether they belong to the same author. From a user's perspective, it is often more frustrating to see papers of different authors are mixed up (low precision), than to see papers of the same author are split into smaller clusters (low recall).

8 Conclusions and Future Work

In this paper, we present a novel categorical set similarity measure named CSLR for two sets which both follow categorical distributions. It is applied in Author Name Disambiguation to measure the similarity between two venue sets or coauthor sets. It is verified to be better than the widely used *Jaccard Coefficient*. We have also proposed a novel method to estimate the distinct author number of each name, which gives reasonable estimation. Our experiments show that our system clearly outperforms other methods of comparison.

We envision broad applications of CSLR since it is a general categorical set similarity measure. In scenarios such as Social Networks and Natural Language Processing, an entity often has a set of contextual features. Often these features have categorical values, and two entities are similar iff these sets follow similar categorical distributions. Some previous work used Jaccard Coefficient etc. as the similarity measures. We expect CSLR will perform better than them.

References

1. A. Agresti. *Categorical data analysis*. Wiley series in probability and statistics. Wiley-Interscience, 2002.
2. I. Bhattacharya and L. Getoor. Collective entity resolution in relational data. *ACM Trans. Knowl. Discov. Data*, 1, March 2007.
3. R. G. Cota, A. A. Ferreira, C. Nascimento, M. A. Gonalves, and A. H. F. Laender. An unsupervised heuristic-based hierarchical method for name disambiguation in bibliographic citations. *J. Am. Soc. Inf. Sci. Technol.*, 61(9):1853–1870, 2010.
4. A. Gretton, K. Borgwardt, M. Rasch, B. Schlkopf, and A. Smola. A kernel method for the two sample problem. In *NIPS 19*, pages 513–520. MIT Press, 2007.
5. H. Han, L. Giles, H. Zha, C. Li, and K. Tsioutsiouliklis. Two supervised learning approaches for name disambiguation in author citations. JCDL '04. ACM, 2004.
6. S. Li, G. Cong, and C. Miao. Supplementary material to author name disambiguation using a categorical distribution similarity. `http://git.io/namedis`.
7. 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. JCDL '09. ACM, 2009.
8. J. Tang, A. C. Fong, B. Wang, and J. Zhang. A unified probabilistic framework for name disambiguation in digital library. *IEEE TKDE*, 99(PrePrints), 2011.
9. J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su. Arnetminer: extraction and mining of academic social networks. In *KDD '08*. ACM, 2008.
10. V. I. Torvik and N. R. Smalheiser. Author name disambiguation in medline. *ACM Trans. Knowl. Discov. Data*, 3:11:1–11:29, July 2009.
11. X. Wang, J. Tang, H. Cheng, and P. S. Yu. Adana: Active name disambiguation. In *ICDM '2011*, 2011.
12. X. Yin, J. Han, and P. S. Yu. Object distinction: Distinguishing objects with identical names by link analysis. In *ICDE '07*, 2007.