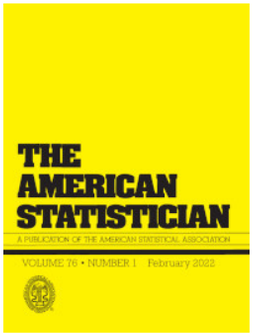

The American Statistician

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/utas20

A Practical Approach to Proper Inference with Linked Data

Andee Kaplan, Brenda Betancourt & Rebecca C. Steorts

To cite this article: Andee Kaplan, Brenda Betancourt & Rebecca C. Steorts (2022): A Practical Approach to Proper Inference with Linked Data, The American Statistician, DOI: 10.1080/00031305.2022.2041482

To link to this article: https://doi.org/10.1080/00031305.2022.2041482

THE AMERICAN STATISTICIAN
2022, VOL. 00, NO. 0, 1–10: Statistical Practice
https://doi.org/10.1080/00031305.2022.2041482

A Practical Approach to Proper Inference with Linked Data

Andee Kaplan[a], Brenda Betancourt[b], and Rebecca C. Steorts[c]

[a]Department of Statistics, Colorado State University, Fort Collins, CO; [b]Department of Statistics, University of Florida, Gainesville, FL; [c]Departments of Statistical Science and Computer Science, Duke University, Durham, NC

ABSTRACT

Entity resolution (ER), comprising record linkage and deduplication, is the process of merging noisy databases in the absence of unique identifiers to remove duplicate entities. One major challenge of analysis with linked data is identifying a representative record among determined matches to pass to an inferential or predictive task, referred to as the *downstream task*. Additionally, incorporating uncertainty from ER in the downstream task is critical to ensure proper inference. To bridge the gap between ER and the downstream task in an analysis pipeline, we propose five methods to choose a representative (or *canonical*) record from linked data, referred to as *canonicalization*. Our methods are scalable in the number of records, appropriate in general data scenarios, and provide natural error propagation via a Bayesian canonicalization stage. The proposed methodology is evaluated on three simulated datasets and one application – determining the relationship between demographic information and party affiliation in voter registration data from the North Carolina State Board of Elections. We first perform Bayesian ER and evaluate our proposed methods for canonicalization before considering the downstream tasks of linear and logistic regression. Bayesian canonicalization methods are empirically shown to improve downstream inference in both settings through prediction and coverage.

ARTICLE HISTORY
Received February 2021
Accepted February 2022

KEYWORDS
Bayesian methods; Canonicalization; Entity resolution; Error propagation; Record linkage

1. Introduction

In many practical problems analysts seek to remove duplicate records across multiple noisy databases, a process known as entity resolution (ER). This is just one important task in "data cleaning" or "data integration," where the outputs are then used for inferential and predictive analyses in areas of application such as government statistics, human rights, economics, precision medicine, and others (Christen 2012; Ilyas and Chu 2019; Christophides et al. 2020; Binette and Steorts 2021; Papadakis et al. 2021). This is just one stage of an *analysis pipeline*, typically consisting of four data cleaning stages prior to the inferential task (Christen 2012; Hogan et al. 2013; Herzog, Scheuren, and Winkler 2007; Abel et al. 2016; Christen 2019; O'Hare, Jurek-Loughrey, and de Campos 2019; Vidhya and Geetha 2019; Papadakis et al. 2021),

$$\text{schema alignment} \rightarrow \text{blocking} \rightarrow \text{ER} \rightarrow \text{canonicalization} \rightarrow \text{analysis}. \quad (1)$$

Equation (1) shows a potential pipeline, including the cleaning stages of (a) schema alignment, (b) blocking, (c) ER, and (d) canonicalization. The first stage, known as "schema alignment," entails identifying available information across multiple data sources for the purpose of joining these together. In the second stage, "blocking," the goal is to place similar records together as a computational tool to speed finding matches. In the third stage, known as ER, one identifies duplicate records and outputs clusters of records that belong to the same partition. Similar records that are considered to represent the same latent entity are grouped together. The ER result can be represented as links between records that belong to the same cluster (the *linkage structure*) or, equivalently, a partition of records into clusters. The resulting dataset that is created as a result of ER is generally referred to as *linked data*. The fourth stage, canonicalization, is the subject of this work. Canonicalization is an optional stage, depending on the inferential goal of an analysis. The partition of records that is the output of ER is used to create a representative record for each cluster. This process creates a representative dataset, which is used as a set of inputs for the inferential analysis.

While removing duplicate entities through ER and canonicalization is worthy as a stand-alone data-cleaning goal, most analyses are primarily motivated by the performance of inferential or predictive analyses on the linked data (e.g., regression or classification). Going forward, we will denote any such postlinkage analysis as the *downstream task*. In an ideal scenario, all records would be correctly clustered by the linkage structure that results from an ER model. In practice, however, ER error is a common occurrence and methods to propagate this error throughout subsequent analysis stages are essential. The primary contribution of this work is the development of a scalable, multi-stage approach, where the ER error is propagated through the downstream task via a stage that we refer to as *posterior canonicalization*.

CONTACT Andee Kaplan andee.kaplan@colostate.edu Department of Statistics, Colorado State University, Fort Collins, CO.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/TAS.

Our multi-stage approach handles the ER and *canonicalization* stages separately from the downstream task and involves using the resulting clusters from the ER model to construct a dataset (without duplicate information) comprised of the most representative or *canonical* set of records. Undoubtedly, the quality of the linkage directly affects the canonicalization step and subsequent downstream tasks. Therefore, the construction of this canonical dataset plays a key role in performing inferential downstream tasks—such as regression analysis—with the linked data.

1.1. Motivating Example

Our work is motivated by voter registration data published by the North Carolina State Board of Elections (NCSBE). The NCSBE provides regular updates of their database, releasing snapshots of their registration database for archival purposes (North Carolina State Board of Elections 2019). Each snapshot contains identifying information about voters such as first name, middle name, last name, full address, city, state, date of birth, state of birth, and phone number. Unlike curated versions of this data (Christen 2014), the NCSBE snapshots contain voter party registration status, which can be useful for inferential or predictive tasks. In addition, given the updating snapshot structure of the NCSBE, individuals are duplicated between (and potentially within) each dataset. For example, an individual can be duplicated when they move or when they perform a name change. We consider five snapshots taken between 2018 and 2019, and refer to this collection as the North Carolina Voter Data (NCVD).

Table 1 illustrates an individual who is believed to be duplicated in the NCVD five times. One goal of the analysis is to predict political affiliation using demographic information, such as race, sex and age. Unfortunately, it's not clear which record in Table 1 most accurately reflects voter "Mark Baker." We will attempt to overcome this issue using canonicalization. The process of canonicalization involves finding the "best" representation of this voter's data fields in order to predict party affiliation. Multiple definitions of "best" lead to multiple methods for performing canonicalization and we detail some of these choices in Section 2.

Given that the NCSBE does not release the details of the process by which they remove duplicates or how they choose the most representative record of an individual, this application motivates our methodology. We will return to this example in Section 4.

1.2. Related Work

The most common approach for ER is the Fellegi–Sunter model based upon pairwise comparisons of common data fields that are used to estimate conditional probabilities of matches and nonmatches (Fellegi and Sunter 1969; Sadinle and Fienberg 2013). A drawback of these and other related probabilistic approaches (e.g., Larsen and Rubin 2001; Hof, Ravelli, and To 2017) is the lack of natural uncertainty quantification in the linkage structure, limiting natural error propagation to the downstream task. This particular limitation has led to many recent developments of Bayesian ER approaches (Tancredi and Liseo 2011; Zhao et al. 2012; Gutman, Afendulis, and Zaslavsky 2013; Sadinle 2014; Steorts 2015; Steorts, Hall, and Fienberg 2016; Zanella et al. 2016; Sadinle 2017; McVeigh, Spahn, and Murray 2020; Marchant et al. 2021).

While the canonicalization methods that we explore in this work are not tied to a particular ER model, we employ the approach of Marchant et al. (2021) to obtain posterior samples of the linkage structure (see Sections 3 and 4). In this model, ER is framed as a bipartite matching problem that links records to (unknown) latent entities. Although most Bayesian ER methods are flexible and can propagate error, they are known to suffer from scalability issues in realistically sized databases ($n \gtrsim 10^4$ records). Scalability has been addressed previously by splitting the data using a variable thought to be relatively clean (*blocking*, see Christen 2012; Steorts et al. 2014) and applying a Bayesian ER model to each subset separately (Tancredi and Liseo 2011; Murray 2015; Steorts, Hall, and Fienberg 2016; Sadinle 2014, 2017). However, since the blocking and ER task are not modeled jointly, the error cannot be quantified exactly. These limitations are addressed in Marchant et al. (2021), which is a scalable model that propagates uncertainty from blocking and ER simultaneously.

The earliest proposals of canonicalization were deterministic, rule-based methods, which were application-specific and fast to implement (Cohen and Sagiv 2005). Other existing literature involving probabilistic approaches commonly assumes the availability of training data in order to select the canonical dataset. This assumption has lead to several optimization and semisupervised methods for finding the most representative or canonical records (Yan and Ozsu 1999; Bohannon et al. 2005; Culotta et al. 2007). Compared to the existing literature, the canonicalization methods that we explore in this work do not rely on any training data (which can be expensive or difficult to obtain), making them fully unsupervised. Additionally, one of our proposed methods fully exploits the uncertainty quantification properties of the Bayesian framework to propagate the ER error to the downstream task (see Section 2). For a review of canonicalization and data fusion techniques, see Bleiholder and Naumann (2009).

Recent work on the relationship between ER and downstream tasks commonly makes assumptions that remove the need for canonicalization. In fact, some downstream tasks have no need for a canonicalization stage because they use only functions of the linkage structure, for example, population size estimation (Tancredi and Liseo 2011; Tancredi, Steorts, and Liseo 2020). For those downstream tasks that do require the linked data, the literature can be classified into two main frameworks—single- and two-stage approaches.

Single-stage approaches build one joint model for the ER and the downstream task, while two-stage approaches treat each model separately. Single-stage approaches for regression and classification have been limited mainly to linking two databases and do not easily generalize beyond this framework (Gutman, Afendulis, and Zaslavsky 2013; Hof, Ravelli, and To 2017; Dalzell and Reiter 2018; Steorts, Tancredi, and Liseo 2018). Moreover, they require knowledge of the model specification up front such that if an additional downstream task is required (after the single-stage joint approach has been fitted), the linkage

would need to be repeated in a new joint model for valid inference. Because ER is the most computationally costly piece of the joint model, this can be an argument against using a single-stage approach.

Performing ER and the downstream task in a two-stage approach allows for specification of downstream models post-linkage, which can expand the computational feasibility of the methods at the cost of no longer having the downstream model inform the linkage. Most of the two-stage literature joins two databases and often assumes that the error from the ER task occurs only in the response variable (Lahiri and Larsen 2005; Kim and Chambers 2012; Goldstein, Harron, and Wade 2012; Hof and Zwinderman 2012; Chambers et al. 2019). In a more general setting where more than two databases are considered or where duplication and linkage errors can affect both response or explanatory variables, there is a dearth of literature. Our proposed work on canonicalization is intended to bridge the gap between ER and the downstream task in this more general setting.

Our Contributions: The main contribution of this work is to provide a practical approach to downstream inference with linked data while maintaining principled error propagation under mild conditions. We propose a scalable, multi-stage approach, where the ER error is propagated to the downstream task via a Bayesian canonicalization stage. A central advantage of the proposed methodology for canonicalization is its generality—it is applicable in the most general data scenario, one in which we have any number of databases to link and duplication can occur in all downstream variables. Our proposed canonicalization methods can be used following any ER method that produces a partition of the records, and prior to any downstream task. With this work, our goal is to make a step toward the formalization of a process that historically has been an ad hoc procedure. We aim to provide guidance for analysts and statisticians who are currently working with linked data on how they can achieve appropriate inference when working with less than perfect and noisy datasets.

The remainder of the article is structured as follows. Section 2 defines a *canonical record* and *canonical dataset*, proposes five unsupervised methods of canonicalization, and describes the computational cost of our recommended method. Section 3 and 4 illustrate the proposed multi-stage approach on simulated data and the NCVD, respectively. Section 5 provides a discussion, including advice for practitioners and directions for future work. We provide further details and a reproducible code base in the supplementary materials.

2. Canonicalization

We provide general definitions of a *canonical record* and *canonical dataset* before detailing five methods for canonicalization.

2.1. Notation

Consider a collection of T databases indexed by i, each with n_i records (rows) indexed by j, and p fields (columns) indexed by ℓ. For instance, the NCVD data is a collection of $T = 5$ databases where each record is comprised of data fields such as first and last name, race, sex, age and political party (see Table 1). Marchant et al. (2021) assumes each record (i, j) links to a single entity, denoted by λ_{ij}, from a fixed population of entities indexed by $e \in \{1, \ldots, E\}$. For example, the five records in Table 1 presumably link to the same individual voter. Denote the value of the ℓ-th field for record (i, j) by $x_{ij\ell}$ and the collection of all field values for all records to be linked is then denoted $\boldsymbol{x} = \{x_{ij\ell} : i = 1, \ldots, T; j = 1, \ldots, n_i; \ell = 1, \ldots, p\}$. We assume that the fields ℓ can be split into those that are used for ER, $\boldsymbol{x}^a = \{x_{ij\ell} : i = 1, \ldots, T; j = 1, \ldots, n_i; \ell \in 1, \ldots, p$ such that field ℓ is aligned in all databases$\}$, and those that are used for the downstream task, $\boldsymbol{x}^d$. Thus, $\boldsymbol{x} = \{\boldsymbol{x}^a, \boldsymbol{x}^d\}$. In the NCVD application (see Section 4), $\boldsymbol{x}^a$ corresponds to the fields of first and last name, sex, and race, while $\boldsymbol{x}^d$ comprises age, ethnicity, and party affiliation.

Table 1. Five records that represent the same voter in the NCVR dataset.

Last	First	Race	Sex	Age	Party
BAKER	MARK	WHITE	MALE	42	REP
BAKER	MARK	WHITE	MALE	41	DEM
BAKER	MARK	WHITE	MALE	41	REP
BAKER	MARK	WHITE	MALE	40	DEM
BAKER	MARK	WHITE	MALE	40	REP

The observations of the aligned fields $x^a_{ij\ell}$ are assumed to be noisy observations of the true field value $z_{\lambda_{ij}\ell}$ of the linked entity. The linkage structure is defined as $\boldsymbol{\Lambda} = \{\lambda_{ij} : i = 1 \ldots T; j = 1 \ldots n_i\}$, where $\lambda_{ij} = e$ means that record (i, j) is linked to the e latent entity. If two records (i, j) and (i', j') refer to the same entity, then $\lambda_{ij} = \lambda_{i'j'}$ must hold. Under this framework, the linkage structure $\boldsymbol{\Lambda}$ induces a partition of the data into clusters where the number of clusters in the partition corresponds to the number of unique values of e observed in the linkage structure. Let $C \equiv C(\boldsymbol{\Lambda})$ be the set of clusters of records resulting from the linkage structure $\boldsymbol{\Lambda}$, assumed to fully partition the data $\boldsymbol{x}$. For simplicity of notation we will often drop the functional notation for C.

Note that when ER is performed with all the variables used for the downstream task, the latent values, $z_{\lambda_{ij}\ell}$, can be used as a representation of the truth in a downstream task. This requires all variables used in the downstream task (e.g., explanatory and response variables $\boldsymbol{x}^d$) to be also used as linkage variables (i.e., be included in $\boldsymbol{x}^a$), otherwise the latent values, $z_{\lambda_{ij}\ell}$, would not be available for those variables. However, it is a fundamental assumption in ER models that all fields in the linkage model are independent, which is not the case for the downstream variables. The effect of breaking this independence assumption has not been investigated in the literature. Furthermore, it may not be advisable to use downstream variables as linkage variables due to known noisiness or unreliability in the data (e.g., age or party affiliation in Table 1). Additionally, since the use of canonicalization does not require all downstream variables to be known prior to the linkage, the ability to perform exploratory analyses is not limited (Tukey 1977). Thus, canonicalization remains a necessary step in many analysis pipelines.

2.2. Definitions

The definitions of a canonical record and canonical dataset are presented below.

Definition 1. For each cluster $c \in C$, the canonical record $\boldsymbol{r}_c$, is defined as a function of the cluster and some known parameters $\boldsymbol{\theta}$, $r_c = \psi_{\boldsymbol{\theta}}(c, \boldsymbol{x}_c)$, where $\boldsymbol{x}_c = \{x_{ij\ell} : (i,j) \in c; \ell = 1, \ldots, p\}$ denotes the record values in cluster c.

Definition 2. A canonical dataset $\boldsymbol{r}$, is obtained by applying the canonicalization function $\psi_{\boldsymbol{\theta}}$ over each cluster that results from $\boldsymbol{\Lambda}$, that is $\boldsymbol{r} = \{\boldsymbol{r}_c : c \in C\}$.

Definitions 1 and 2 include the record selection canonicalization methods proposed by Culotta et al. (2007) for the case where $\psi_{\boldsymbol{\theta}}$ is the minimum average edit distance between strings as a function of fixed or learned costs. However, our definition is more general as it includes categorical, ordinal and numerical fields, as well as unsupervised functions $\psi_{\boldsymbol{\theta}}$.

2.3. Point-Estimate Canonicalization

In this section, we assume that an ER task has been performed as to provide a point estimate $\hat{\boldsymbol{\Lambda}}$ of the linkage structure $\boldsymbol{\Lambda}$. In practice, the point estimate could result from any ER model that results in a partitioning of records. We propose three unsupervised methods for performing canonicalization based on $\hat{\boldsymbol{\Lambda}}$.

First, we define *random canonicalization*, which will serve as a baseline for comparison. For each cluster $c \in C$, choose the canonical record $\boldsymbol{r}_c$ randomly,

$$\boldsymbol{r}_c = \{\boldsymbol{x}_{ck} : k \sim \text{Categorical}(\{(i,j) \in c\}, \boldsymbol{\theta})\},$$

where $\boldsymbol{x}_{ck}$ denotes the kth record in $\boldsymbol{x}_c$ and $\boldsymbol{\theta}$ represents the vector of selection probabilities for the records (i,j) in cluster c.

Composite canonicalization is defined as an aggregate record that includes information from each linked record in the cluster,

$$\boldsymbol{r}_c = \{r_{c\ell} : \ell = 1, \ldots, p\}, \quad r_{c\ell} = \bar{f}_\ell(\{x_{ij\ell} : (i,j) \in c\}),$$

where $\bar{f}_\ell$ is a an aggregation function for each column, $\ell = 1, \ldots, p$. The form of aggregation depends on the column type and can be weighted by some prior knowledge of the data sources when available. Composite canonicalization alters the original data values for records with duplicates (unless they are exact duplicates), which can heavily affect inference results for some downstream tasks.

Minimax canonicalization is a point-estimate canonicalization method designed to choose a canonical record that "most closely captures" the underlying true unknown (latent) entity. We propose selecting the record whose farthest neighbor within the cluster is closest, where "closeness" is measured by a pairwise record distance function, denoted by $d(\cdot, \cdot)$. We define the canonical record as the record $\boldsymbol{r}_c$ within each cluster c such that

$$\boldsymbol{r}_c = \arg\min_{(i,j) \in c} \max_{(i',j') \in c} d(\boldsymbol{x}_{ij}, \boldsymbol{x}_{i'j'}), \quad c \in C. \tag{2}$$

The result is a set of representative records, one for each latent entity, that is central to the other records in each cluster. Many distance functions can be used depending on the context of the problem and have a critical role in determining the resulting canonical dataset. One potential distance function is provided in Appendix A.1 in the supplementary materials. Ties in maximum record distance within the cluster can be handled multiple ways, depending on the computational constraints and anticipated number of ties. One option is to select the record that has the closest farthest neighbor (minimax record distance) when compared to all other records in the dataset (not within the cluster) that match on categorical variable levels. Another is to simply break the tie randomly. If tied records are identical, we select the record with the lowest index.

We note that Bayesian models for ER are commonly sensitive to choice of hyperparameters, indicating a need for knowledge of the data collection process (Steorts 2015; Steorts, Hall, and Fienberg 2016; Sadinle 2017; Aleshin-Guendel and Sadinle 2021) and the choice of distance function in the minimax canonicalization method is similarly critical to its success.

2.4. Posterior Canonicalization

We now propose two alternative approaches that use the marginal posterior distribution of the linkage structure, $P(\boldsymbol{\Lambda}|\boldsymbol{x})$, to inform canonicalization using the probability that each record in the data is canonical, $P((i,j) \in \boldsymbol{r}|\boldsymbol{x})$. To estimate this probability, let $C^{(m)}$ represent the partition of the data into clusters for iteration $m \in \{1, \ldots, M\}$ of the Markov chain Monte Carlo (MCMC) samples from $P(\boldsymbol{\Lambda}|\boldsymbol{x})$. We apply the following procedure,

1. For each m and $c \in C^{(m)}$, find the canonical records, $\boldsymbol{r}_c^{(m)}$ based on minimax canonicalization defined in Equation (2).
2. Compute the *posterior canonicalization (PC) weights* that approximate the posterior probability of each record (i,j) being selected as a canonical record,

$$\hat{p}_{ij} \equiv \frac{1}{M} \sum_{m=1}^{M} \mathbb{I}((i,j) \in \boldsymbol{r}^{(m)}) \approx P((i,j) \in \boldsymbol{r}|\boldsymbol{x}), \tag{3}$$

where, $\boldsymbol{r}^{(m)} = \{\boldsymbol{r}_c^{(m)} : c \in C^{(m)}, m = 1, \ldots, M\}$.

Note that the linkage uncertainty is captured by the (possibly) different partitions, $C^{(m)}$, obtained from the MCMC samples of $P(\boldsymbol{\Lambda}|\boldsymbol{x})$. The PC weights described in Equation (3) provide a vehicle through which to pass linkage uncertainty to the downstream task.

To incorporate the ER uncertainty in a downstream task via the PC weights $\hat{\boldsymbol{p}} = \{\hat{p}_{ij} : i = 1, \ldots, T; j = 1, \ldots, n_i\}$, we take advantage of the multitude of methods available for incorporating survey weights into downstream analyses (e.g., Little 1991; Pfeffermann 1993). For illustration, assume the downstream task is a linear model of the form

$$Y|\boldsymbol{\beta}, \Sigma_y, \boldsymbol{X} \sim \text{MVN}(\boldsymbol{X}^T\boldsymbol{\beta}, \Sigma_y), \tag{4}$$

where the response variable is represented as Y and explanatory variables X. The *PC weighted* method incorporates the uncertainty from both entity resolution and canonicalization as the weights in weighted linear regression, where

$$\Sigma_y = Q_y\sigma^2 \quad \text{and}$$
$$Q_y^{-1} = \text{diag}(w_{11}, \ldots, w_{1n_1}, \ldots, w_{T1}, \ldots, w_{Tn_T}),$$

and $w_{ij} = \hat{p}_{ij}$ defined to be the PC weight for record (i,j) from Equation (3). Note this is not a formal canonicalization method

as defined in Definition 2 given that all the original records are passed to the downstream task along with their respective PC weights, but it does allow for proper accounting of uncertainty.

To construct a canonical dataset based on the PC weights, we use thresholding. The *PC threshold canonicalization* method can be implemented similarly for the downstream task described in Equation (4), where now

$$w_{ij} = \boldsymbol{I}\{\hat{p}_{ij} \geq \tau_{\text{PC}}\}$$

for each record (i, j). A value of $\tau_{\text{PC}} = 0.5$ is a natural choice because it results in canonical records having posterior probability greater than 0.5 of being representative. We describe one way to validate the choice of threshold τ_{PC} in Section 4. Note that the PC threshold method can result in multiple (or zero) records being selected as canonical for a cluster.

2.5. Computational Complexity

The posterior canonicalization approach allows us to deliver a method that is still feasible even for large datasets. To see this, let $|c|$ denote the size of each cluster $c \in C^{(m)}$, for $m = 1, \ldots, M$ where M is the number of posterior samples of the linkage structure available and $C^{(m)}$ is the resulting cluster from $\boldsymbol{\Lambda}^{(m)} \sim P(\boldsymbol{\Lambda}|\boldsymbol{x})$. ER has been shown to exhibit the properties of *micro-clustering*, where the cluster sizes grow sub-linearly as the size of the data increases (Zanella et al. 2016; Betancourt, Zanella, and Steorts 2020). Given the expected sub-linear growth of the cluster sizes, it is common to assume that

$$\frac{1}{n} \max_{\substack{c \in C^{(m)} \\ 1 \leq m \leq M}} |c| \to 0 \text{ as } n \to \infty, \tag{5}$$

for $n = \sum_{i=1}^{T} n_i$, the total number of records in all T databases. Thus, the computational complexity of our proposed posterior canonicalization methodology (with randomly broken ties) is $\left(\max_{c,m} |c|\right)^2$, subquadratic with respect to n. For proof of this assertion, see Appendix A.2 in the supplementary materials. We make no statement about the computational feasibility of other methods, rather we simply note that the computational complexity of the posterior canonicalization method is not prohibitive of larger analyses.

3. Simulation Study

Given that the true underlying relationship between variables is not available in any real data application, we construct simulated datasets that contain this information as well as the true records. The aim of this simulation study is to provide an overview of the performance of different canonicalization methods and their effect in a downstream task. We frame the simulation study to perform the downstream task of linear regression. The simulated datasets are generated using the GeCO tool (Tran, Vatsalan, and Christen 2013) and three different levels of noise in the relationship between predictors and response variables through Gaussian noise, where $\sigma = 1, 2, 5$. Each of the three datasets contain a total of 500 records, 30% duplication, and the maximum number of duplicates of each record is 5. Each dataset contains the following fields: first name, last name, birth date, sex, education level, income (in 1000's), and blood pressure (bp). The bp variable was generated with a known (noisy) relationship to sex and income, and our goal is to assess how the fitted model is altered based on the canonical dataset passed after the ER task. Additionally, we generated three sets of test records, of 500 records each, following the same data generation mechanism. We consider two data scenarios—the most general case, in which all variables in the downstream task are subject to ER error and a more common case in the existing literature, where only the explanatory variables are subject to this error. For full details on the data generation process, refer to Section B.1 in the supplemental materials.

First, we perform ER on the GeCO datasets using the Bayesian ER model of Marchant et al. (2021). In this simulated dataset, we have a true known unique identifier, and thus, we are able to ascertain the true performance of the ER task. To evaluate ER performance we compute the pairwise precision and recall (Christen 2012), which correspond to the proportion of predicted links that are correctly estimated and the proportion of true links that are correctly estimated, respectively. In this case, the pairwise precision and recall for the point-wise linkage are 0.97 and 0.88, respectively.[1] To obtain a point estimate for the linkage structure, we use the shared most probable maximal matching set (SMPMMS) from Steorts (2015). Alternative point estimates were considered, including decision theoretical approaches for optimal Bayesian estimation based on multiple loss functions (Binder's, the Normalized Information Distance, and the Variation of Information) for the space of partitions (Lau and Green 2007; Wade and Ghahramani 2018; Rastelli and Friel 2018; Rastelli 2021). The adjusted Rand indices when compared to the SMPMMS were all greater than 0.9722, indicating robustness.

To assess the performance of canonicalization, we evaluate the distributional closeness of the canonical dataset generated from canonicalization to the true records using an empirical Kullback–Leibler (KL) divergence metric (Wang, Kulkarni, and Verdú 2005; Silva and Narayanan 2007). Values closer to zero indicate closer distributions. These results are based on simulating 100 datasets. As expected, the empirical KL divergence values are slightly higher for the scenario where all the downstream variables are susceptible to linkage error ([0.0036, 0.0085]) compared to the error-free response setting ([0.0036, 0.0071]). For all levels of noise in the two error scenarios, the closest distributions to the truth are achieved through the minimax method ([0.0036, 0.0045]) with PC threshold method ([0.0038, 0.0047]) and PC weighted ([0.0041, 0.0049]) very close behind. We expect the composite method to perform poorly in terms of distributional closeness since this method will alter all records with a duplicate, and this is indeed the case ([0.0057, 0.0085]).

Finally, we examine the performance of the canonicalization methods through the downstream task of linear regression via three metrics: (a) bias in the fitted coefficients, (b) coverage of the credible intervals, and (c) mean square error (MSE) for

[1] Note that while there are three GeCO datasets, they only differ in the noise level for the relationship between explanatory and response variables. Thus, the noise differences do not affect the record linkage, which is performed just once with the linkage variables.

test records from each of the models fitted with the canonical datasets. We fit the models with a Bayesian specification in `stan` (Stan Development Team 2016, 2020), with Gaussian prior distributions for the parameters (Gelman et al. 2008, 2013) and parameters centered and scaled to be weakly informative. See Appendix B.3 in the supplementary materials for more detail.

The linear regression model is specified as

$$\begin{aligned} Y_i|\boldsymbol{\beta}, \sigma, \boldsymbol{X}_i &\overset{\text{ind}}{\sim} \text{N}(\boldsymbol{X}_i^T\boldsymbol{\beta}, \sigma^2) \\ \beta_j|b_j &\overset{\text{ind}}{\sim} \text{N}(0, b_j) \\ \sigma|a &\sim \text{Exponential}(a). \end{aligned} \tag{6}$$

In this simulation, we are interested in assessing the effect of canonicalization method on inference in the relationship between blood pressure and sex and income. Due to the focus on effect of canonicalization method, rather than the effect of model misspecification in the downstream task, we have generated the data from a known (and correctly specified) model for the purposes of being able to evaluate bias and coverage after linkage for each of the proposed canonicalization methods.

Table 2 displays the MSE, bias, and coverage of the 90% credible interval of the regression coefficient for the income variable. See Tables 3–5 in Appendix B in the supplementary materials for results of the other model coefficients. The MSE, bias, and coverage results are promising for minimax, PC Threshold, and PC Weighted. However, PC Weighted is consistently the optimal choice (or close to it) in all three metrics, with the best predictive performance when compared to the other canonicalization methods. The coverage and bias behavior of the PC weighted method is consistent for the other model terms, again see Tables 3–5 in Appendix B in the supplementary materials for evidence. The inferential and predictive results highlight the advantage of error propagation from the ER phase of the PC weighted method for downstream tasks.

4. Application to NCVD

We now return to the motivational example of the NCVD from Section 1.1. Due to the size of the NCSBE snapshot dataset (each snapshot contains between 8 million and 29 million voter registration records, resulting in more than 500 million total records), we limit our investigation to five snapshot datasets recorded on April 30, 2019, January 1, 2019, November 6, 2018, May 8, 2018, and January 1, 2018. Specifically, we consider Caswell County, which is located in North Central North Carolina and has nearly even membership among Democrats and Republicans, as well as a diverse population in terms of gender, age, and race (North Carolina State Board of Elections 2020). This results in 54,716 records to be linked. For further detail on how we curated the NCVD dataset, see Appendix C.1 in the supplementary materials.

Discrepancies with respect to the NCSBE voter identifiers have been previously noted, leading others to question whether the assignment of voter identifiers in each snapshot correspond to the same individual voter (Wortman 2019). Due to the suspected issues with the state-assigned voter identities, it is impossible to assess the accuracy of our ER or canonicalization procedures. We first perform ER using the model of Marchant et al. (2021) and apply our proposed canonicalization methods before investigating how well we can predict party affiliation in a downstream classification task.

We provide the pairwise precision and recall, 0.979 and 0.787, respectively, for the point-estimate $\hat{\boldsymbol{\Lambda}}$ of the linkage structure when compared to the NCSBE voter identifiers (which may or may not themselves be accurate). This indicates that our method of de-duplication is linking more entities than the NCSBEs procedure. Appendix C.3 in the supplementary materials provides trace-plots and hyperparameter values. The resulting 95% credible interval for the number of unique voters in the dataset is [14,394, 14,590] and the point estimate is $\hat{n} = 14{,}484$.

Table 2. Mean and standard deviation (in parenthesis) for MSE, bias, and coverage of the 90% credible interval for income for regression based on five canonicalization methods and the true dataset for levels of noise $\sigma = 1, 2, 5$.

	Errors in all downstream variables			Errors in explanatory variables only		
Method	MSE	Bias	Coverage	MSE	Bias	Coverage
$\sigma = 1$						
Random	37.55(10.67)	0.0862(0.06)	0.31	38.72(12.02)	0.08262(0.06)	0.28
Composite	24.95(8.67)	0.04961(0.04)	0.51	23.93(8.19)	0.05907(0.05)	0.39
Minimax	7.28(3.14)	0.00083(0.01)	0.97	**6.44 (2.65)**	**−0.00256 (0.01)**	0.98
PC weighted	**7.16 (2.17)**	0.00971(0.01)	**0.92**	6.54(2.00)	0.00663(0.01)	**0.97**
PC threshold	7.27(3.08)	**0.00021 (0.01)**	0.97	6.47(2.60)	−0.00259(0.01)	0.98
True	2.23(0.08)	0.00086(0.01)	0.92	2.24(0.08)	−4e-05 (0.01)	0.91
$\sigma = 2$						
Random	45.51(10.26)	0.08976(0.05)	0.23	43.95(11.10)	0.08272(0.05)	0.31
Composite	31.74(8.98)	0.0552(0.05)	0.48	32.12(9.16)	0.06139(0.05)	0.5
Minimax	12.6(3.04)	−0.00315(0.02)	0.92	13.25(2.94)	0.00121(0.01)	0.95
PC weighted	**11.83 (2.09)**	0.00746(0.02)	**0.9**	**12.11 (2.14)**	0.01101(0.02)	0.87
PC threshold	12.52(3.05)	**−0.00269 (0.01)**	0.95	13.1(2.86)	**0.00076 (0.02)**	**0.92**
True	8.36(0.35)	0.00048(0.01)	0.9	8.34(0.31)	0.00106(0.01)	0.92
$\sigma = 5$						
Random	89.32(11.75)	0.09106(0.06)	0.4	89.81(12.74)	0.09(0.07)	0.38
Composite	74.82(9.57)	0.0536(0.06)	0.67	74.12(9.52)	0.05039(0.05)	0.64
Minimax	55.61(3.80)	**−0.004 (0.03)**	0.91	55.61(3.62)	**−0.00336 (0.03)**	0.95
PC weighted	**47.64 (2.63)**	0.00443(0.03)	0.87	**47.61 (2.64)**	0.0053(0.03)	0.86
PC threshold	55.67(3.88)	−0.00468(0.03)	**0.9**	55.44(3.64)	−0.00361(0.03)	**0.9**
True	51.09(2.19)	−0.00233(0.03)	0.92	51 (1.85)	−0.00324(0.03)	0.93

NOTE: Results are based on 100 simulated datasets. Best performing method for each data scenario and metric are bolded.

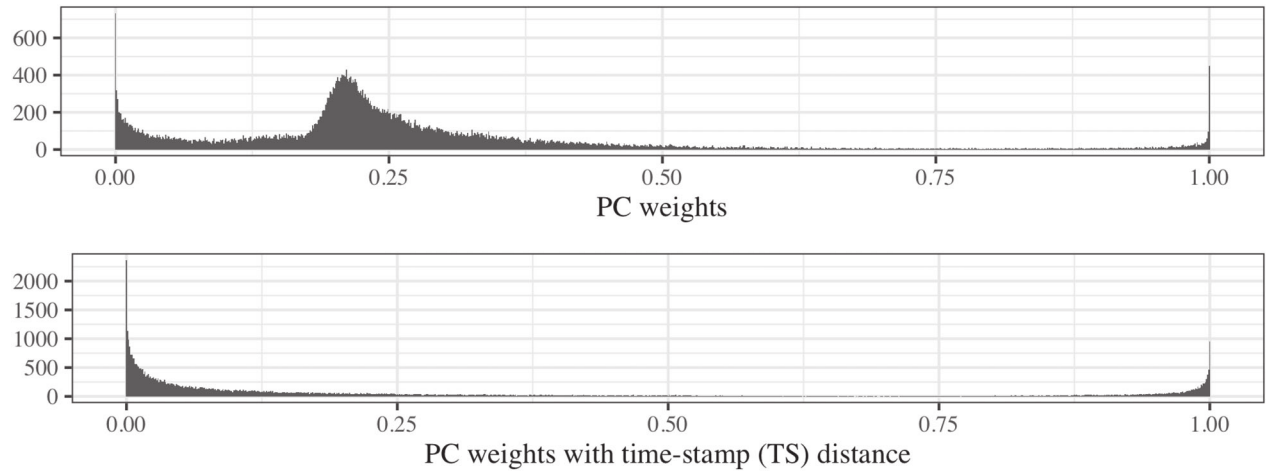

Figure 1. (Top) Distribution of PC weights for Caswell NCVD. (Bottom) Distribution of PC weights using a distance function that includes time-stamps.

According to the NCSBE, there are 14,740 registered voters in Caswell County, which is in agreement with the comparative overlinking evidenced in the precision and recall.

Next, we perform canonicalization using all methods in Section 2. The top of Figure 1 displays the distribution of the resulting posterior canonicalization weights. Due to the large size of the data, we have broken ties randomly. Many records have weights around 0.2, suggesting the presence of many ties, which may lead to the exclusion of voters from the canonical dataset. In fact, the choice of $\tau_{\text{PC}} = 0.5$ results in only 4919 canonical records using the PC threshold method, which is far less than the point estimate of the number of unique individuals in the data (14,484).

We overcome this issue by using time-stamps in the distance function, which are available in this dataset. This choice reflects our belief that the most recent record is likely the most accurate for this application. The bottom of Figure 1 displays the distribution of PC weights for a distance function that includes time-stamp information. Many records have PC weights close to 0 or 1, making the decision of excluding or including records in the canonical dataset straightforward. Incorporating the time-stamp data in the distance function has eliminated the need for a robust tie-breaking procedure, allowing us to use the more computationally efficient version of posterior canonicalization. The choice of $\tau_{PC} = 0.5$ results in 11,217 canonical records using the PC threshold method, which is closer to the point estimate of the number of unique individuals in the dataset (14,484).

We compute the empirical KL divergence between the canonical datasets of each of the proposed methods to the only notion of truth that we have—the dataset released for Caswell county by the NCSBE. KL values close to zero indicate distributional closeness between datasets. We reiterate that it is not known how the NCSBE performs ER or canonicalization, and thus, the quality of this dataset is near impossible to ascertain. We find that the PC threshold Time-stamp (TS) method shows the lowest empirical KL divergence (0.03) when compared to the records released by the NCSBE. The PC weighted TS also provides a very low empirical KL divergence (0.04). On the other hand, the PC threshold method without temporal information provides the highest value (0.1). Certainly, the chosen value of τ_{PC} affects the result, but this behavior appears to be closely related to the large presence of record distance ties in this relatively clean dataset. We emphasize this point, as it further highlights the importance of choosing a record distance function that has strong discriminatory power to distinguish between records and places the (believed) truth at the center of a cluster. From this outcome, we conclude that the PC methods that incorporate time-stamp information produce results that are most consistent with the deduplication approach undertaken by the NCSBE.

Table 3. Five records that represent the same voter with their respective PC weights both with and without time-stamp information.

Last	First	Race	Sex	Age	Party	PC Weight	PC Weight TS
BAKER	MARK	WHITE	MALE	42	REP	0.295	**0.793**
BAKER	MARK	WHITE	MALE	41	DEM	0.325	0.463
BAKER	MARK	WHITE	MALE	41	REP	0.306	0.326
BAKER	MARK	WHITE	MALE	40	DEM	0.124	0.05
BAKER	MARK	WHITE	MALE	40	REP	0.232	0.053

NOTE: PC weights above $\tau_{PC} = 0.5$ are bolded.

Returning to the illustrative example of Table 1, we have appended two columns with the PC weights (both with and without time-stamp information) for each record in Table 1 in Table 3. PC weights above $\tau_{\text{PC}} = 0.5$ are bolded. Based on these weights, the canonical record that would be selected is the first entry in Table 3 using the PC threshold TS method. Based on the PC threshold (no time-stamp information), none of these records would be selected and this individual would be left out of future analyses.

We next consider the downstream task on Caswell County, where our goal is to be able to model the relationship between party affiliation and the following demographic variables (fields): sex, age, race, and ethnicity. We consider a logistic regression model fit using `stan` (Stan Development Team 2016) with Gaussian prior distributions for the parameters (Gelman et al. 2008, 2013) and parameters centered and scaled to be weakly informative (see Appendix C.2 in the supplementary materials),

$$Y_i|\boldsymbol{\beta}, \boldsymbol{X}_i \overset{\text{ind}}{\sim} \text{Bernoulli}\left(\frac{\exp\{\boldsymbol{X}_i^T\boldsymbol{\beta}\}}{1+\exp\{\boldsymbol{X}_i^T\boldsymbol{\beta}\}}\right)$$

$$\beta_i|b_i \overset{\text{ind}}{\sim} \text{N}(0, b_i). \tag{7}$$

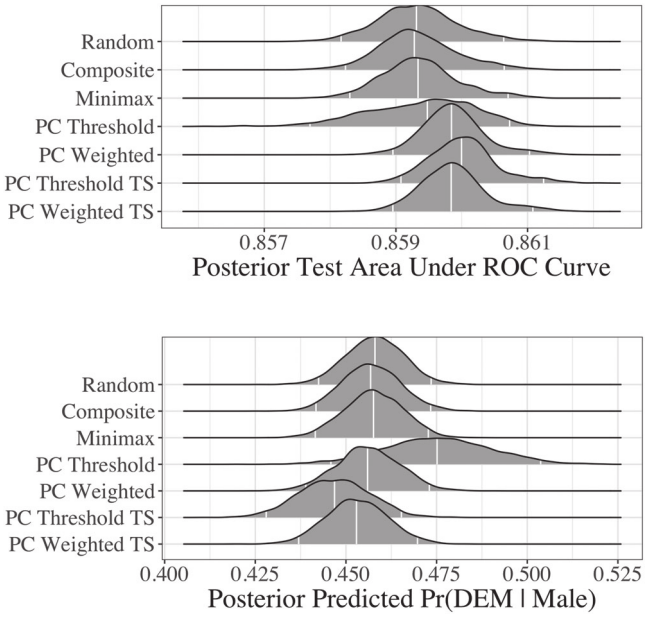

Figure 2. (Left) Posterior test area under ROC curve for the logistic regression models fit after performing each type of canonicalization. PC weighted, PC weighted TS, and PC threshold TS show the best predictive results. (Right) Posterior predicted Pr(DEM|Male) holding all other features at a typical value in the test dataset. There is a difference in the inferential relationship between outcome and predictors for models fit on point-wise canonical datasets versus posterior canonical datasets.

After completing both ER and canonicalization, we assess the performance of the downstream task by obtaining out-of-sample predictions of party affiliation from a recent snapshot of the NCSBE (May 14, 2019) that was not included in either the ER task or the canonicalization tasks (test dataset). Figure 2 shows two methods for assessing the affect of canonicalization on the downstream task. On the left, we assess the predictive performance of the model fit on each canonical dataset using the posterior test AUC. Based on these distributions, PC weighted, PC weighted TS, and PC threshold TS show the best predictive results. On the right of Figure 2 we see the posterior predicted *Pr*(DEM|Male), holding all other fields fixed at a typical value in the test dataset (King, Tomz, and Wittenberg 2000). There is a clear difference in this relationship for models fit on point-wise canonical datasets and posterior canonical datasets. This result in conjunction with the improved inference evidenced in Section 3, indicates improved inferential performance via the natural error propagation from the posterior canonicalization methods.

5. Discussion

In this article, we have presented a practical approach to proper inference with linked data via canonicalization. This approach allows error to propagate naturally into downstream analyses, such as prediction of voter affiliation in Caswell County or regression of blood pressure on sex and income. We have proposed several methods to find canonical records, including those based on point estimates and posterior distributions of linkage. Additionally, we have empirically shown the benefits of error propagation with posterior canonicalization through an inferential downstream task in simulated experiments, as well as a real data example.

In general, we recommend the use of posterior canonicalization when two conditions are met—Bayesian record linkage is used and a discriminatory distance function is available that places the true records central to their clusters. When Bayesian record linkage is not used, then we recommend minimax canonicalization with a robust tie-breaking procedure and when a discriminatory distance function is not available we recommend composite canonicalization. Finally, in the case where there is no reason to believe that the true records can be placed central within their clusters, then we are left with only random canonicalization. In this case, the use of a very informative distribution for records within cluster may be helpful. See Appendix D in the supplementary materials for a potential decision making process to determine which canonicalization method to use.

The key advantages of our proposed methodology for canonicalization are generality and computational efficiency. The methods are applicable in general data scenarios with multiple databases where duplication can occur in all downstream variables at a relatively low computational cost. Canonicalization can be a crucial step that facilitates the transition between the ER stage and the subsequent downstream tasks in general applications with linked data. Future areas of research include investigating tradeoffs in choice of ER model under simulated and real data and determining automated methods for choosing τ_{PC}. It is also of interest to consider more downstream tasks, such as generalized linear models, small area estimation, and alternative classification methods.

Supplementary Materials

The supplement contains additional details concerning data for the simulation study and NCVD application, additional results, diagnostics, and hyperparameters for the fitted models, as well as further advice for practitioners. The R code for reproducing the results is also available.

Funding

This work was partially supported by NSF-1652431, NSF-1534412, NSF-2051911, and the Laboratory for Analytic Sciences.

References

Abel, D., Melanie, A., Blache, A., and Saïdi, A. (2016), "Measuring the Quality of a Probabilistic Linkage through Clerical-Reviews," in *Statistics Canada Symposium 2016*. [1]

Aleshin-Guendel, S., and Sadinle, M. (2021), "Multifile Partitioning for Record Linkage and Duplicate Detection," *Journal of the American Statistical Association*, 1–25, DOI: 10.1080/01621459.2021.2013242. [4]

Betancourt, B., Zanella, G., and Steorts, R. C. (2020), "Random Partition Models for Microclustering Tasks," *Journal of the American Statistical Association*, 1–13, DOI: 10.1080/01621459.2020.1841647. [5]

Binette, O., and Steorts, R. C. (2021), "Some of Entity Resolution," arXiv:2008.04443. [1]

Bleiholder, J., and Naumann, F. (2009), "Data Fusion," *ACM Computing Surveys*, 41, 1–41. [2]

Bohannon, P., Fan, W., Flaster, M., and Rastogi, R. (2005), "A Cost-based Model and Effective Heuristic for Repairing Constraints by Value Modification," in *Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data* (pp. 143–154), ACM. [2]

Chambers, R., Salvati, N., Fabrizi, E., and da Silva, A. D. (2019), "Domain Estimation Under Informative Linkage," *Statistical Theory and Related Fields*, 3, 90–102. [3]

Christen, P. (2012), *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*, Data-Centric Systems and Applications, Berlin: Springer-Verlag. [1,2,5]

——— (2014), "Preparation of a Real Temporal Voter Data Set for Record Linkage and Duplicate Detection Research," Technical Report, Australian National University. [2]

——— (2019), "Data Linkage: The Big Picture," *Harvard Data Science Review*, 1(2). [1]

Christophides, V., Efthymiou, V., Palpanas, T., Papadakis, G., and Stefanidis, K. (2020), "An Overview of End-to-End Entity Resolution for Big Data," *ACM Computing Surveys*, 53, 1–42. [1]

Cohen, S., and Sagiv, Y. (2005), "An Incremental Algorithm for Computing Ranked Full Disjunctions," in *Proceedings of the 24th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems* (pp. 98–107), ACM. [2]

Culotta, A., Wick, M., Hall, R., Marzilli, M., and McCallum, A. (2007), "Canonicalization of Database Records Using Adaptive Similarity Measures," in *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 201–209), ACM. [2,4]

Dalzell, N. M., and Reiter, J. P. (2018), Regression Modeling and File Matching Using Possibly Erroneous Matching Variables," *Journal of Computational and Graphical Statistics*, 27, 728–738. [2]

Fellegi, I. P., and Sunter, A. B. (1969), "A Theory for Record Linkage," *Journal of the American Statistical Association*, 64, 1183–1210. [2]

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013), "*Bayesian Data Analysis*," Texts in Statistical Science (3rd ed.), New York: Chapman and Hall/CRC. [6,7]

Gelman, A., Jakulin, A., Pittau, M. G., and Su, Y.-S. (2008), "A Weakly Informative Default Prior Distribution for Logistic and Other Regression Models," *The Annals of Applied Statistics*, 2, 1360–1383. [6,7]

Goldstein, H., Harron, K., and Wade, A. (2012), "The Analysis of Record-linked Data Using Multiple Imputation with Data Value Priors," *Statistics in Medicine*, 31, 3481–3493. [3]

Gutman, R., Afendulis, C. C., and Zaslavsky, A. M. (2013), "A Bayesian Procedure for File Linking to Analyze End-of-life Medical Costs," *Journal of the American Statistical Association*, 108, 34–47. [2]

Herzog, T., Scheuren, F., and Winkler, W. (2007), *Data Quality and Record Linkage Techniques*, New York: Springer. [1]

Hof, M. H., Ravelli, A. C., and To, A. H. Z. (2017),"A Probabilistic Record Linkage Model for Survival Data," *Journal of the American Statistical Association*, 112, 1504–1515. [2]

Hof, M. H. P., and Zwinderman, A. H. (2012), "Methods for Analyzing Data from Probabilistic Linkage Strategies Based on Partially Identifying Variables," *Statistics in Medicine*, 31, 4231–4242. [3]

Hogan, H., Cantwell, P. J., Devine, J., Mule, V. T., and Velkoff, V. (2013), "Quality and the 2010 Census," *Population Research and Policy Review*, 32, 637–662. [1]

Ilyas, I. F., and Chu, X. (2019), *Data Cleaning*, New York: Association for Computing Machinery. [1]

Kim, G., and Chambers, R. (2012), "Regression Analysis Under Incomplete Linkage," *Computational Statistics and Data Analysis*, 56, 2756–2770. [3]

King, G., Tomz, M., and Wittenberg, J. (2000), "Making the Most of Statistical Analyses: Improving Interpretation and Presentation," *American Journal of Political Science*, 44, 347–361. [8]

Lahiri, P., and Larsen, M. (2005),"Regression Analysis With Linked Data," *Journal of the American Statistical Association*, 100, 222–230. [3]

Larsen, M. D., and Rubin, D. B. (2001), "Iterative Automated Record Linkage Using Mixture Models," *Journal of the American Statistical Association*, 96, 32–41. [2]

Lau, J. W., and Green, P. J. (2007), Bayesian Model-based Clustering Procedures," *Journal of Computational and Graphical Statistics*, 16, 526–558. [5]

Little, R. J. (1991), Inference with Survey Weights," *Journal of Official Statistics*, 7, 405–424. [4]

Marchant, N. G., Kaplan, A., Elazar, D. N., Rubinstein, B. I., and Steorts, R. C. (2021), "d-blink: Distributed End-to-end Bayesian Entity Resolution," *Journal of Computational and Graphical Statistics*, 30, 406–421. [2,3,5,6]

McVeigh, B. S., Spahn, B. T., and Murray, J. S. (2020), "Scaling Bayesian Probabilistic Record Linkage with Post-Hoc Blocking: An Application to the California Great Registers," arXiv:1905.05337. [2]

Murray, J. S. (2015), "Probabilistic Record Linkage and Deduplication after Indexing, Blocking, and Filtering," *Journal of Privacy and Confidentiality*, 7, 3–24. [2]

North Carolina State Board of Elections (2019), North Carolina State Board of Elections, Accessed December 24, 2019. [2]

——— (2020), Voter Registration Statistics - Caswell County, Accessed February 14, 2020. [6]

O'Hare, K., Jurek-Loughrey, A., and de Campos, C. (2019), *A Review of Unsupervised and Semi-supervised Blocking Methods for Record Linkage* (pp. 79–105), Cham: Springer. [1]

Papadakis, G., Ioannou, E., Thanos, E., and Palpanas, T. (2021), "The Four Generations of Entity Resolution," *Synthesis Lectures on Data Management*, 16, 1–170. [1]

Pfeffermann, D. (1993), "The Role of Sampling Weights When Modeling Survey Data," *International Statistical Review*, 61, 317–337. [4]

Rastelli, R. (2021), *GreedyEPL: Greedy Expected Posterior Loss*. R package version 1.2. [5]

Rastelli, R., and Friel, N. (2018), "Optimal Bayesian Estimators for Latent Variable Cluster Models," *Statistics and Computing*, 28, 1169–1186. [5]

Sadinle, M. (2014), "Detecting Duplicates in a Homicide Registry Using a Bayesian Partitioning Approach," *The Annals of Applied Statistics*, 8, 2404–2434. [2]

——— (2017), "Bayesian Estimation of Bipartite Matchings for Record Linkage," *Journal of the American Statistical Association*, 112, 600–612. [2,4]

Sadinle, M., and Fienberg, S. E. (2013), "A Generalized Fellegi-Sunter Framework for Multiple Record Linkage With Application to Homicide Record Systems," *Journal of the American Statistical Association*, 108, 385–397. [2]

Silva, J., and Narayanan, S. (2007), "Universal Consistency of Data-driven Partitions for Divergence Estimation," in *IEEE International Symposium on Information Theory, 2007* (pp. 2021–2025), IEEE. [5]

Stan Development Team. (2016), *rstanarm: Bayesian applied regression modeling via Stan*. R package version 2.13.1. [6,7]

——— (2020), *RStan: the R interface to Stan*. R package version 2.21.2. [6]

Steorts, R. C. (2015), "Entity Resolution with Empirically Motivated Priors," *Bayesian Analysis*, 10, 849–875. [2,4,5]

Steorts, R. C., Hall, R., and Fienberg, S. E. (2016), "A Bayesian Approach to Graphical Record Linkage and Deduplication," *Journal of the American Statistical Association*, 111, 1660–1672. [2,4]

Steorts, R. C., Tancredi, A., and Liseo, B. (2018), "Generalized Bayesian Record Linkage and Regression with Exact Error Propagation," in *International Conference on Privacy in Statistical Databases* (pp. 297–313), Springer. [2]

Steorts, R. C., Ventura, S. L., Sadinle, M., and Fienberg, S. E. (2014), "A Comparison of Blocking Methods for Record Linkage," in *Privacy in Statistical Databases*, Lecture Notes in Computer Science, eds. J. Domingo-Ferrer and F. Montes (pp. 253–268), Cham: Springer. [2]

Tancredi, A., and Liseo, B. (2011), "A Hierarchical Bayesian Approach to Record Linkage and Population Size Problems," *The Annals of Applied Statistics*, 5, 1553–1585. [2]

Tancredi, A., Steorts, R., and Liseo, B. (2020), "A Unified Framework for De-Duplication and Population Size Estimation" (with discussion), *Bayesian Analysis*, 15, 633–682. [2]

Tran, K.-N., Vatsalan, D., and Christen, P. (2013), "GeCo: An Online Personal Data Generator and Corruptor," in *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management* (pp. 2473–2476), ACM. [5]

Tukey, J. W. (1977), *Exploratory Data Analysis*, Boston, MA: Addison-Wesley. [3]

Vidhya, K. A., and Geetha, T. V. (2019), "Entity Resolution and Blocking: A Review," *2019 IEEE 9th International Conference on Advanced Computing (IACC)*, 2019, pp. 133–140. [1]

Wade, S., and Ghahramani, Z. (2018), "Bayesian Cluster Analysis: Point Estimation and Credible Balls" (with discussion), *Bayesian Analysis*, 13, 559–626. [5]

Wang, Q., Kulkarni, S. R., and Verdú, S. (2005), "Divergence Estimation of Continuous Distributions Based on Data-dependent Partitions," *IEEE Transactions on Information Theory*, 51, 3064–3074. [5]

Wortman, J. P. H. (2019), *Record Linkage Methods with Applications to Causal Inference and Election Voting Data*, PhD thesis, Duke University. [6]

Yan, L. L., and Ozsu, M. T. (1999), "Conflict Tolerant Queries in AURORA," in *Proceedings of the 4th IFCIS International Conference on Cooperative Information Systems* (pp. 279–290), IEEE. [2]

Zanella, G., Betancourt, B., Wallach, H., Miller, J., Zaidi, A., and Steorts, R. C. (2016), "Flexible Models for Microclustering with Application to Entity Resolution," in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16 (pp. 1425–1433). [2,5]

Zhao, B., Rubinstein, B. I. P., Gemmell, J., and Han, J. (2012), "A Bayesian Approach to Discovering Truth from Conflicting Sources for Data Integration," *Proceedings of the VLDB Endowment*, 5, 550–561. [2]