Active Learning for Probabilistic Record Linkage.*

Ted Enamorado†

September 25*th*, 2018

Abstract

Integrating information from multiple sources plays a key role in social science research. However, when a unique identifier that unambiguously links records is not available, merging datasets can be a difficult and error-prone endeavor. Probabilistic record linkage (PRL) aims to solve this problem by providing a framework in which common variables between datasets are used as potential identifiers, with the goal of producing a probabilistic estimate for the unobserved matching status across records. In this paper, I propose an active learning algorithm for PRL, which efficiently incorporates *human judgment* into the process and significantly improves PRL's performance at the cost of manually labelling a small number of records. Using data from local politicians in Brazil, where a unique identifier is available for validation, I find that the proposed method bolsters the overall accuracy of the merging process. In addition, I examine data from a recent vote validation study conducted for the ANES, and I show that the proposed method can recover estimates that are indistinguishable from those obtained from a more extensive, expensive, and time-consuming clerical review.

Key Words: Active learning, EM algorithm, precision, recall, record linkage

*I want to thank Kosuke Imai, Matias Iaryczower, John B. Londregan, Marc Ratkovic, and Leonard Wantchekon for their invaluable encouragement and advice. I also wish to thank Winston Chou, Naoki Egami, Ben Fifield, Adeline Lo, Gabriel López-Moctezuma, Fabi Pineda, Soichiro Yamauchi, and Yang-Yang Zhou for helpful comments and suggestions. Special thanks to Matt DeBell for all his help and technical assistance during my visit to the ANES offices at Stanford University.

†Ph.D. Candidate, Department of Politics, Princeton University, Princeton NJ 08544. Email: tede@princeton.edu, URL: http://www.tedenamorado.com

1 Introduction

Modern social science research often relies on bringing together information from different sources to advance our understanding about questions of interest. From studies that seek to explain the differences between self-reported and actual behavior (Ansolabehere and Hersh 2012, Barbera 2015, Meredith and Morse 2015, Berent, Krosnick, and Lupia 2016, Hill and Huber 2017, Jackman and Spahn 2018, Bonica 2018); the effects of the national news media on mass public and elite behavior (DellaVigna and Kaplan 2007, Hopkins and Ladd 2014, Arceneaux et al. 2016, Martin and Yurukoglu 2017); historical accounts about the electorate (Acharya, Blackwell, and Sen 2016, Spahn 2017, Hall, Huff, and Kuriwaki 2018); the impacts of lobbying activities on trade (Bombardini and Trebbi 2012, Bertrand, Bombardini, and Trebbi 2014, Kim 2017); to studies on clientelism and redistributive politics (De La O 2013, Zucco 2013, 2015, Rueda 2016), etc; scholars have spent considerable amounts of time and effort assembling detailed datasets from multiple sources to conduct sound empirical analyses.

When merging data, the main difficulty faced by the researcher is that oftentimes a unique identifier that unambiguously links records across two datasets, such as the social security number, does not exist. Under this scenario the true match status of all the pairwise comparisons across two datasets is unknown and merging data is prone to mis-classifications; in particular, we might fail to find true matches in the data (false negatives) or classify as matches observations that do not refer to the same entity (false positives). This problem is more pronounced when the data are noisy, either due to missing information or typographical errors.

Since the work of Fellegi and Sunter (1969), who formalized the notion of probabilistic record linkage (PRL), a growing literature in statistics, computer science, and more recently in the social sciences, has aimed to solve this problem via a principled framework that uses variables in common between datasets as potential identifiers. The goal is to produce a probabilistic estimate for the latent matching status across pairs of records. The advantages of such an approach are that it is devised specifically as a mechanism to control for possible error rates and to account for any remaining uncertainty into subsequent empirical analyses.

As recently noted by many authors (see e.g., McVeigh and Murray 2017, Sadinle 2017, Enamorado, Fifield, and Imai 2018, and references therein), using PRL to merge two datasets that lack a unique identifier is a good strategy, especially in situations where the amount of overlap between two datasets is large – even in the presence of moderate amounts of noise in the data. However, as originally noted by Winkler (2002), in many common situations, PRL struggles to

accurately match records e.g., when the overlap between datasets is small, when only a few common potential identifiers exist, when the amount of noise in the data is large, when the modeling assumptions are violated, etc.

A common approach to detect problems with PRL is to perform a detailed ex-post clerical review or to use heuristics like random spot checking. Yet, these methods often provide insufficient guidance about which observations are more informative or how to utilize these qualitative assessments about the matching status of a particular set of records to improve the accuracy of our estimates. Hence, the critical question is: can we do better at detecting and addressing problems with PRL? and if so, how?

In this paper, I propose a robust approach for probabilistic record linkage via active learning. The proposed method efficiently incorporates *human judgment* into the merging process by sampling the most informative pairs of observations for manual labeling, rather than requiring the analyst to hand-code every pair of observations in the clerical review region or to use ad-hoc procedures to reduce the size of clerical review task. Thus, through the use of sampling, the researcher can avoid clerical reviews that are orders of magnitude larger while improving the overall accuracy of the results. In addition, through its iterative nature, the model behind the proposed method borrows strength from the small set of manually labeled cases to improve the precision of quantities on interest by directly incorporating them into the estimation stage. Finally, it does not impose major bottlenecks in terms of computing-time as it is fully incorporated into the computational improvements to PRL recently introduced by Enamorado, Fifield, and Imai (2018) and implemented in fastLink an R-package for PRL (Enamorado, Fifield, and Imai 2017).

I demonstrate the proposed method using two empirical examples. In the first application, I validate the model by merging two datasets on local-level candidates in Brazil for the 2012 and 2016 municipal elections. Each dataset contains more than 450,000 observations, and are ideal to test the performance of probabilistic record linkage algorithm as they contain a unique identifier: the *Cadastro de Pessoas Físicas* (CPF), which is the Brazilian individual taxpayer registry identification number. More importantly, while the CPF is perfectly recorded for each individual, the other common variables between datasets are not as they are manually entered into the database, and therefore may be subject to misspellings and other errors.

The matches identified by a PRL model without labeled data are of high quality; however, the estimated rates of false positives and false negatives obtained are at least 6 and 8 percentage points away from the ground truth, respectively. Such a discrepancy might lead a researcher

to conclude, for example, that the rate of politicians that switch their party affiliation (a common practice among Brazilian politicians) is 58%, while 53% actually did. Using my proposed methodology, I find that 53% local politicians in fact joined a different party. To reach this level of precision, it was necessary to manually label just 4% of the usual clerical review proposed by other PRL approaches.

In a second application, I revisit a recent vote validation study conducted by Enamorado and Imai (2018) for the 2016 American National Election Study (ANES). In that study, ANES respondents were matched with a voter file of more than 180 million observations. Enamorado and Imai (2018) conducted a lengthy and detailed clerical review, in which the matching status of 4,271 respondents was carefully evaluated. Using unlabeled data alone, PRL produces a validated turnout rate among ANES respondents that is 5 percentage points larger than the actual turnout rate in the population of interest. In contrast, by manually labeling fewer than 15% of the number of cases included in the original clerical review, the proposed methodology recovers a validated turnout rate that is within the margin of error of the true population-level turnout rate and which is virtually identical to the rate found after a more comprehensive clerical review.

The paper is organized as follows. First, I briefly describe the canonical model of probabilistic record linkage and discuss its shortcomings. Second, I introduce the proposed methodology. Third, I present the results from the two empirical applications. The first application serves as a validation exercise, while the second application highlights the advantages of the proposed methodology. Finally, some concluding remarks and avenues for future extensions are discussed.

2 Probabilistic Record Linkage

In this section, I start by describing how to construct an agreement pattern, perhaps the most important concept in the probabilistic record linkage literature. Then, I introduce the canonical model of probabilistic record linkage and describe how this model classifies pairs of records as matches or as non-matches. Finally, I discuss the problems and challenges a researcher might face when conducting a merge using the Fellegi-Sunter framework.

2.1 Representing Comparisons as Agreement Patterns

Suppose that we wish to merge two data sets, $\mathcal{A}$ and $\mathcal{B}$, with sample sizes $N_{\mathcal{A}}$ and $N_{\mathcal{B}}$, respectively. The problem is that the true matching status for all the $N_{\mathcal{A}} \times N_{\mathcal{B}}$ distinct pairs is

unknown. Therefore, the best strategy to conduct a merge under those circumstances is to use the variables which are common to both datasets as potential identifiers. Unfortunately, in practice, most data are noisy, which means that if the variables in common are recorded with error (e.g., misspellings, missing information, etc.), an exact match on these fields would classify many true matches as non-matches. We use the function $\gamma_k(i,j)$ to represent the level of within-pair similarity for the kth variable between the ith observation of data set $\mathcal{A}$ and the jth observation of data set $\mathcal{B}$. Thus, for each pairwise comparison, an agreement pattern $\boldsymbol{\gamma}(i,j) = \{\gamma_1(i,j), \gamma_2(i,j), \ldots, \gamma_k(i,j), \ldots, \gamma_K(i,j)\}$ represents a sequence of similarity levels across all the K variables used to link files. Intuitively, the higher the level of agreement across fields as recorded in $\boldsymbol{\gamma}(i,j)$, the more likely a pair of records is to be a match.

To construct each $\gamma_k(i,j)$, we first need to calculate a measure of distance $S_k(i,j)$ between the observed values for variable k that the ith and jth observations take. Consequently, the smaller the value of $S_k(i,j)$ the closer are the values being compared. In the case of string-valued variables, there are three prominent options for $S_k(\cdot)$: Levenshtein, Jaro, and Jaro-Winkler; all involving character-wise comparisons of two strings (see Cohen, Ravikumar, and Fienberg 2003; Yancey 2005 for a detailed description of each measure). For numeric-valued variables, $S_k(\cdot)$ can be represented by an $L1$ (absolute value of the difference) or an $L2$ norm (euclidean distance) as measures of distance between two values.

Let the number of agreement levels in the kth variable be denoted by L_k, then, for example, if $L_k = 2$ we have that:

$$\gamma_k(i,j) = \begin{cases} 1 & \text{if } S_k(i,j) \leq \tau \qquad \text{“identical (or nearly so)”} \\ 0 & \text{otherwise} \qquad\qquad\qquad\qquad\quad \text{“different”} \end{cases}$$

where τ is a threshold value set at the discretion of the researcher – see Jaro (1989) and Winkler (1990) for examples of threshold values commonly used by the U.S. Census Bureau.[1]

Of course, not all data at our disposal is perfectly recorded, and missing values are common features of social science data. The latter means that some components of $\boldsymbol{\gamma}(i,j)$ might be missing. That is why we define a missingness vector of length K, denoted by $\delta(i,j)$, where its kth element $\delta_k(i,j)$ is equal to 1 if at least one record in the pair (i,j) has a missing value in the kth variable and is equal to 0 otherwise.

Figure 1 presents an illustrative example on how agreement patterns are constructed. The

[1]Note that comparisons can be classified using more than two levels e.g., using three agreement levels, we can classify comparisons into: `identical (or nearly so)`, `similar`, and `different`.

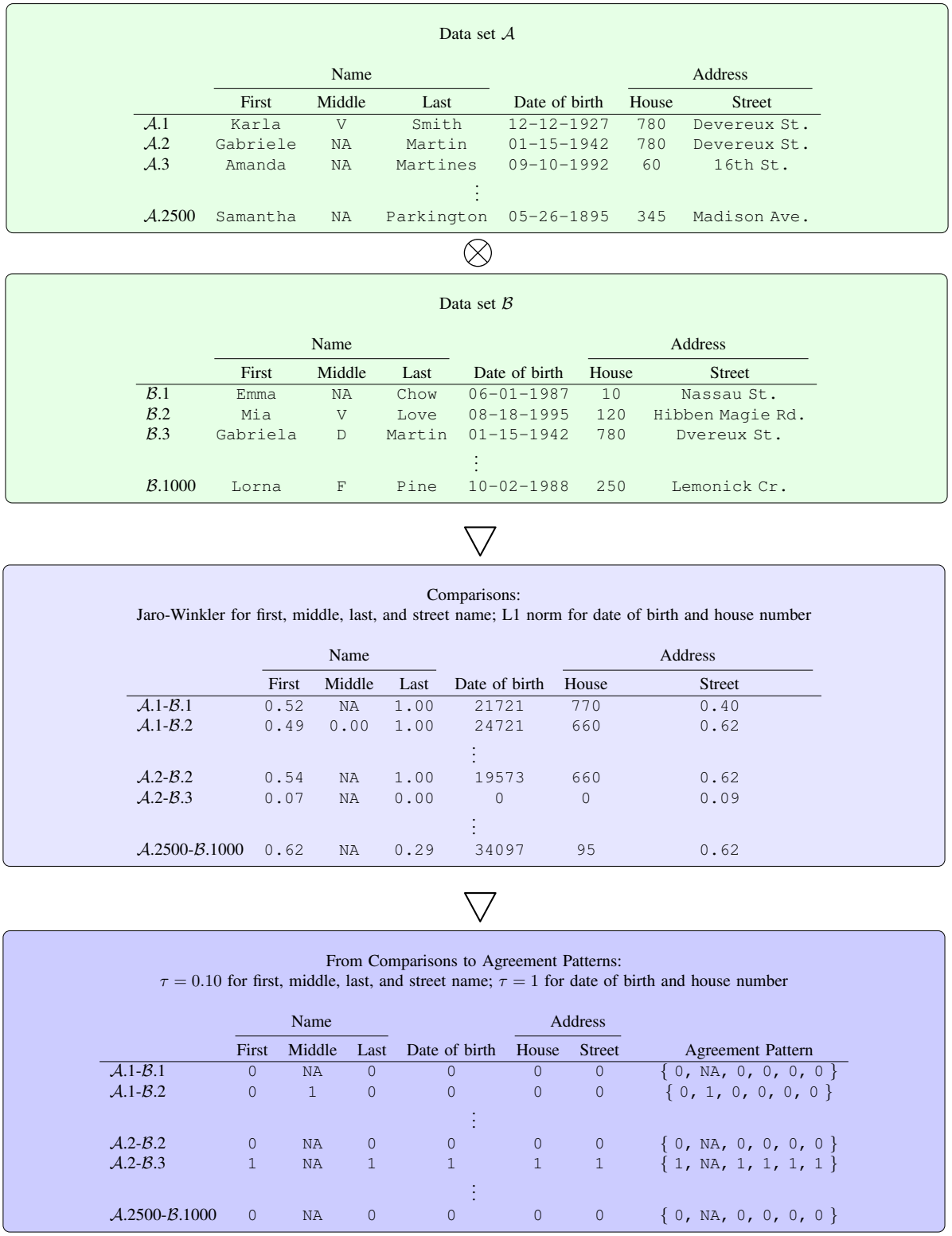
top panels (in green) represent two artificial data sets, $\mathcal{A}$ and $\mathcal{B}$, of 2500 and 1000 records, respectively. As noted above, the first step to obtain each $\gamma(i, j)$ is to compare the values of a pair of records for a given variable. In this example, such an operation translates into 2.5 million comparisons per variable.

The third panel of figure 1 (in light blue) presents examples of pairwise distances across six variables: first, middle, and last name, house number and street name. String-valued variables are compared using the Jaro-Winkler string distance – a value of 0 represents that the two values being compared are the same and a value of 1 means that they are different. In the case of numeric-valued variables, the absolute value of the difference (L1 norm) is used. For example, if we compare observations $\mathcal{A}.1$ and $\mathcal{B}.2$, we obtain a renormalized Jaro-Winkler score of 0.49 for first name, 0 for middle name, 1 for last name, and 0.62 for street name; in addition, those two observations are 24,721 days apart in terms of age and their house numbers differ in 660 units.

For all the variables, let $L_k = 2$ and $\tau = 0.10$ for string-valued variables and $\tau = 1$ for numeric-valued variables. Then, for example, we get an agreement pattern for observations $\mathcal{A}.1$ and $\mathcal{B}.2$ equal to $\gamma(\mathcal{A}.1, \mathcal{B}.2) = \{$`0, 1, 0, 0, 0, 0`$\}$ i.e., only the middle names are identical for those observations – see the last panel (blue) of figure 1. Note that a comparison involving at least one missing value is indicated by `NA`, and in our notation this is described by the vector $\delta(i, j)$. For example, $\delta(\mathcal{A}.2, \mathcal{B}.3) = \{$`0, 1, 0, 0, 0, 0`$\}$ indicates the middle name comparison is not possible due to a missing value when comparing records $\mathcal{A}.2$ and $\mathcal{B}.3$.

Without a doubt, constructing agreement patterns is the most computational expensive step of any record linkage task as the number of comparisons grows quadratically with the size of the datasets. To reduce the number of comparisons, a common technique used in the related literature is blocking i.e., make comparisons only for observations that share the same value in a variable, treating as non-matches observations that differ on that variable. For example, a researcher may only make comparisons for observations that share the same gender.[2] Without loss of generality, the proposed methodology in this paper can be applied with or without blocking. To facilitate the exposition (notation-wise), no blocking scheme is assumed.

[2]Christen (2012) and Steorts et al. (2014) are two excellent reviews of blocking techniques.



Data set $\mathcal{A}$

	Name				Address	
	First	Middle	Last	Date of birth	House	Street
$\mathcal{A}$.1	Karla	V	Smith	12-12-1927	780	Devereux St.
$\mathcal{A}$.2	Gabriele	NA	Martin	01-15-1942	780	Devereux St.
$\mathcal{A}$.3	Amanda	NA	Martines	09-10-1992	60	16th St.
			⋮			
$\mathcal{A}$.2500	Samantha	NA	Parkington	05-26-1895	345	Madison Ave.

⊗

Data set $\mathcal{B}$

	Name				Address	
	First	Middle	Last	Date of birth	House	Street
$\mathcal{B}$.1	Emma	NA	Chow	06-01-1987	10	Nassau St.
$\mathcal{B}$.2	Mia	V	Love	08-18-1995	120	Hibben Magie Rd.
$\mathcal{B}$.3	Gabriela	D	Martin	01-15-1942	780	Dvereux St.
			⋮			
$\mathcal{B}$.1000	Lorna	F	Pine	10-02-1988	250	Lemonick Cr.

▽

Comparisons:
Jaro-Winkler for first, middle, last, and street name; L1 norm for date of birth and house number

	Name				Address	
	First	Middle	Last	Date of birth	House	Street
$\mathcal{A}$.1-$\mathcal{B}$.1	0.52	NA	1.00	21721	770	0.40
$\mathcal{A}$.1-$\mathcal{B}$.2	0.49	0.00	1.00	24721	660	0.62
				⋮		
$\mathcal{A}$.2-$\mathcal{B}$.2	0.54	NA	1.00	19573	660	0.62
$\mathcal{A}$.2-$\mathcal{B}$.3	0.07	NA	0.00	0	0	0.09
				⋮		
$\mathcal{A}$.2500-$\mathcal{B}$.1000	0.62	NA	0.29	34097	95	0.62

▽

From Comparisons to Agreement Patterns:
$\tau = 0.10$ for first, middle, last, and street name; $\tau = 1$ for date of birth and house number

	Name				Address		
	First	Middle	Last	Date of birth	House	Street	Agreement Pattern
$\mathcal{A}$.1-$\mathcal{B}$.1	0	NA	0	0	0	0	{ 0, NA, 0, 0, 0, 0 }
$\mathcal{A}$.1-$\mathcal{B}$.2	0	1	0	0	0	0	{ 0, 1, 0, 0, 0, 0 }
				⋮			
$\mathcal{A}$.2-$\mathcal{B}$.2	0	NA	0	0	0	0	{ 0, NA, 0, 0, 0, 0 }
$\mathcal{A}$.2-$\mathcal{B}$.3	1	NA	1	1	1	1	{ 1, NA, 1, 1, 1, 1 }
				⋮			
$\mathcal{A}$.2500-$\mathcal{B}$.1000	0	NA	0	0	0	0	{ 0, NA, 0, 0, 0, 0 }

Figure 1: An Illustrative Example on How to Construct Agreement Patterns. The top panels (in green) show two artificial data sets, $\mathcal{A}$ and $\mathcal{B}$, with 2500 and 1000 records, respectively. The third panel shows how the comparisons across values for the different variables are made. The bottom panel shows how we can move from comparisons to agreement values, and consequently to agreement patterns.

2.2 The Fellegi-Sunter Model

Fellegi and Sunter (1969) formalized the intuition behind the ideas of Newcombe et al. (1959) and Newcombe and Kennedy (1962), and proposed what is today the workhorse model of probabilistic record linkage. The Fellegi-Sunter model is a two-class mixture model, where the unobserved variable $M(i,j)$ indicates whether the ith record in the data set $\mathcal{A}$ and the jth record in the data set $\mathcal{B}$ is a match or not.

The model has the following structure:

$$\gamma_k(i,j) \mid M(i,j) = m \overset{\text{indep.}}{\sim} \text{Discrete}(\boldsymbol{\pi}_{km}) \tag{1}$$

$$M(i,j) \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(\lambda) \tag{2}$$

where $\boldsymbol{\pi}_{km}$ is a vector of length L_k, containing the probability of each agreement level for the kth variable given that the pair is a match ($m = 1$) or a non-match ($m = 0$), and λ represents the probability of a match across all pairwise comparisons. Through the parameter $\boldsymbol{\pi}_{k0}$, the model allows for the possibility that two records can have identical values for some variables even when they do not represent a match.

The model is based on three key independence assumptions: 1. the latent matching status $M(i,j)$ is assumed to be independently and identically distributed; 2. conditional independence of the agreement levels across merging variables; and 3. missing at random (MAR). This last assumption was recently introduced in the record linkage literature by Sadinle (2017) and Enamorado, Fifield, and Imai (2018), to avoid ad-hoc decisions about how to represent comparisons involving a missing value.[3]

Under these assumptions, the observed-data likelihood function of the model defined in equations (1) and (2) is given by,

$$\mathcal{L}_{obs}(\lambda, \boldsymbol{\pi} \mid \boldsymbol{\delta}, \boldsymbol{\gamma}) \propto \prod_{i=1}^{N_{\mathcal{A}}} \prod_{j=1}^{N_{\mathcal{B}}} \left\{ \sum_{m=0}^{1} \lambda^m (1-\lambda)^{1-m} \prod_{k=1}^{K} \left(\prod_{l=0}^{L_k - 1} \pi_{kml}^{\mathbf{1}\{\gamma_k(i,j)=l\}} \right)^{1-\delta_k(i,j)} \right\}$$

where π_{kml} represents the lth element of probability vector $\boldsymbol{\pi}_{km}$, i.e., $\pi_{kml} = \Pr(\gamma_k(i,j) = l \mid M(i,j) = m)$.

Following the work of Winkler (1988), the model parameters are estimated using the Expectation-Maximization (EM) algorithm (Dempster, Laird, and Rubin 1977).[4] Note that there is no guar-

[3] For example, a common practice in the record linkage literature has been to categorize comparisons involving a missing value as `different` values (see e.g., Sariyar, Borg, and Pommerening 2012b).

[4] See appendix A.1.1 for more details on how the model parameters are obtained.

antee that the Fellegi-Sunter model will lead to one-to-one matching assignment. If the researcher is interested in imposing such a restriction, the ex-post approach of Jaro (1989) can be used.

2.3 Fellegi-Sunter Decision Rules

Fellegi and Sunter (1969) propose an optimal classification plan, where the goal is to separate pairs of records into three groups: matches, non-matches, and cases where model had problems in terms of classification. It is in this last group, known as the clerical review region, where the matching status of each pair of records is adjudicated by *human judgment* in an ex-post manner.

Figure 2 presents the intuition behind the Fellegi-Sunter decision rules on how to separate records into groups. The idea is to select two thresholds so that the false positive and false negative rate are controlled, in other words, keep them below a certain pre-determined value by the researcher. Formally, to separate records into one of the aforementioned classes, the first step Fellegi and Sunter (1969) take is to rank observations according to their likelihood of being a match. To do so, they introduce the following weights for each agreement pattern:

$$W_h = \log\left(\frac{\Pr(\boldsymbol{\gamma}_h \mid M_h = 1)}{\Pr(\boldsymbol{\gamma}_h \mid M_h = 0)}\right) \tag{3}$$

where $h \in \{1, \ldots, H\}$ indexes unique instances of an agreement pattern and H represents the total number of distinct agreement patterns we observe in the data. Note that larger Fellegi-Sunter weights imply more support in favor of the hypothesis that the pairs of records (i, j) with agreement pattern $\boldsymbol{\gamma}_h$ are a match.

The next step in the process is to order pairs and obtain an upper threshold W_1 from:

$$c_1 = \sum_{h:W_h \geq W_1} \Pr(\boldsymbol{\gamma}_h \mid M_h = 0) \tag{4}$$

and, similarly, a lower threshold W_2 from:

$$c_2 = \sum_{h:W_h \leq W_2} \Pr(\boldsymbol{\gamma}_h \mid M_h = 1) \tag{5}$$

where c_1 and c_2 represent values for the false positive and false negative rates that the researcher feels comfortable with e.g., $c_1 = c_2 \leq 0.01$. Fellegi and Sunter (1969) proved that the the area between W_1 and W_2, the clerical review region, is optimal in the sense that there is no other decision rule, for the same level of error tolerance, that will include fewer observations.

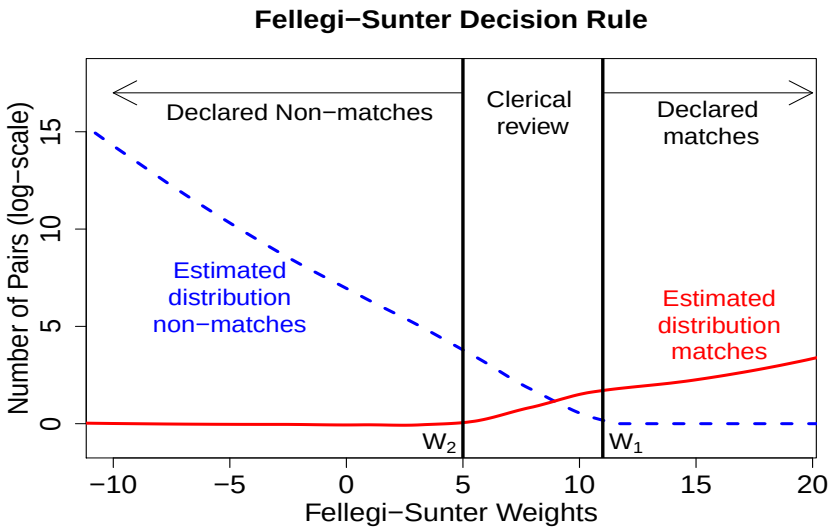


Figure 2: Illustration of the Fellegi-Sunter Decision Rules. Pairwise comparisons with agreement patters to the left (right) of W_2 (W_1) are classified as non-matches (matches). Adjudication of the matching status of pairwise comparisons between W_2 and W_1 is left for clerical review. Note that W_2 (W_1) is selected such that the area below the red solid (blue dashed) line which represents the estimated distribution of matches (non-matches) is small.

2.4 A Mismatch Between True and Estimated Error Rates

Oftentimes, due to the large number of cases in the clerical review region, researchers avoid a case-by-case manual labeling process; pairs of records are either divided into matches and non-matches, or as recently proposed in Enamorado, Fifield, and Imai (2018) the probability of being a match can be used to control for the uncertainty related to the merging process in subsequent empirical analysis. As noted by Winkler (2002), these approaches are appropriate so long as the overlap between datasets is large, the set of matches and non-matches are well separated, typographical error rates are low or at most moderate, and there are enough fields in common between datasets to overcome faulty data in an specific field. Moreover, as noted by Sadinle (2017), there is always the hope that model assumptions are met (at least in expectation), so that the mixture classes correspond to the set of matches and non-matches.

In situations where those conditions are not met, authors like Thibaudeau (1993), Winkler (1993), Thibaudeau (1993), Belin and Rubin (1995), Larsen and Rubin (2001), Winkler (2006), Winkler and Yancey (2006), Herzog, Scheuren, and Winkler (2010) Murray (2016), Herzog, Scheuren, and Winkler (2010), Sadinle (2017), and Enamorado, Fifield, and Imai (2018) have found that the Fellegi-Sunter model returns estimated error rates (false positives and false negatives) that are not a good approximation of the truth.

Figure 3 illustrates the mismatch between true and estimated error rates. The panel on the

left shows that while the estimated false positive rate (the area under the dashed blue line to the right of W_1) is less than 0.01, the true false positive rate is significantly larger (the area under the blue solid line to the right of W_1); such a mismatch between an estimate and the truth false positive rate is illustrated by the blue shaded region. Similarly, on the right panel of figure 3 the mismatch between true and estimated false negative rates is illustrated (red shaded area). Moreover, these differences between true and estimated error rates are not greatly improved when relaxing some of the model assumptions made in the Fellegi-Sunter framework (Larsen and Rubin 2001, Sadinle 2017, and Enamorado, Fifield, and Imai 2018). Thus, the question is: can we improve upon the performance of the Fellegi-Sunter model while still retain some of its advantages, such as its simplicity and scalability? The next section lays out a solution that applies selective human judgment to improve the performance of the Fellegi-Sunter model in these scenarios.

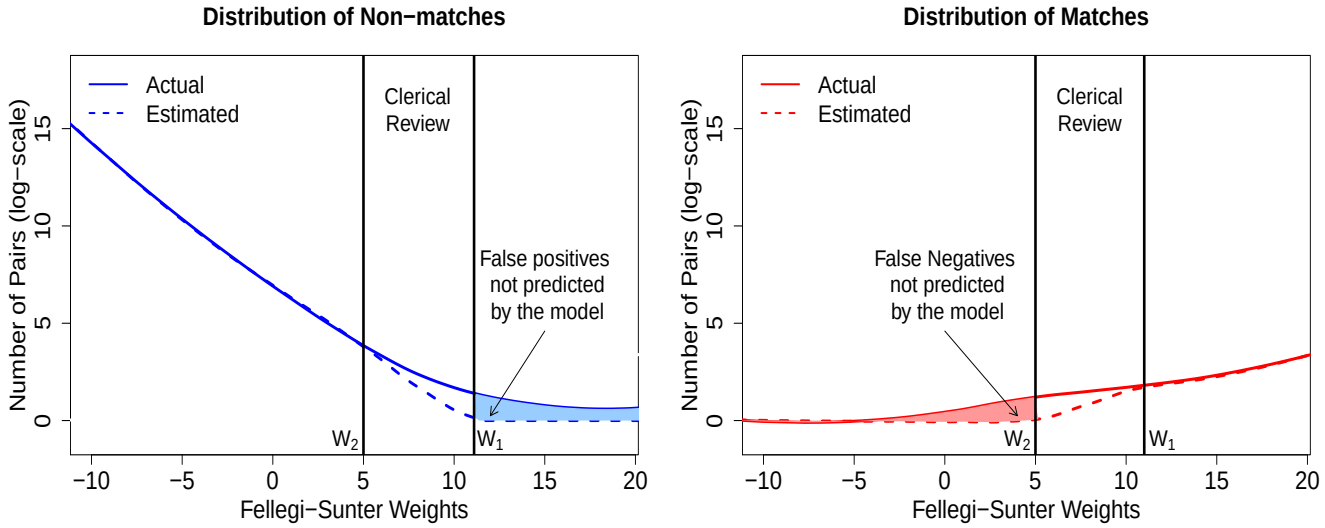


Figure 3: Illustration of the Problem: a Mismatch Between True and Estimated Error Rates. The shaded area in both panels represents the size of bias of the estimated false positive (left panel) and false negative (right panel) error rates if compared to their true counterparts.

3 Active Learning for Probabilistic Record Linkage

In this section, I start by presenting a brief description of different machine learning approaches for classification tasks. In addition, I situate the proposed method in the related literature, with an emphasis on how it is designed to directly overcome the drawbacks of previous approaches that aim to incorporate human judgment into probabilistic record linkage. The latter is fol-

lowed by a detailed description of the proposed methodology.

3.1 Background

In the machine learning literature, unsupervised learning is a task where only unlabeled observations are used to learn the inherent structure of the data. In classification problems, this translates into learning the unobserved group assignment of different observations. The Fellegi-Sunter model is an example of an unsupervised learning algorithm. For each pairwise comparison, we observe its corresponding agreement pattern ($\gamma(i,j)$), and use this information to predict (learn) its unobserved matching status ($M(i,j)$).

In contrast to unsupervised learning, supervised learning uses labeled data only i.e., observations for which both the predictors and the outcome (labels) are observed. The main advantage of supervised learning over unsupervised learning is that we can actually test how effective a method is, as we have at our disposal objective measures of what constitutes success e.g., how far is our prediction from the truth (Hastie, Tibshirani, and Friedman 2009). In record linkage, supervised learning corresponds to a situation where for each pair of records we observe both its agreement pattern and its true matching status.

Semi-supervised learning aims to encompass the ideas behind supervised and unsupervised learning. In other words, in semi-supervised learning, to train a model we make use labeled and unlabeled data. Still, the problem is that due to the scarcity of labeled data, the amount of unlabeled data at our disposal is orders of magnitude larger. Therefore, for the model to extract signal from the labeled data and not be informed by unlabeled data alone, it is key to increase its relative importance in a principled way. An additional hurdle is that only in exceptional cases a researcher will have access to labeled data. Thus, for semi-supervised learning to be incorporated into any record linkage framework, it is important to find efficient ways to obtain the most informative labeled data. Larsen and Rubin (2001) and Winkler (2002), introduce semi-supervised learning to improve the performance of the Fellegi-Sunter model. They show that by including labeled data in a probabilisitic record linkage framework the mismatch between true and estimated error rates can be greatly reduced.

In particular, Winkler (2002) building on the work of Nigam et al. (2000) proposed a model that combines unlabeled and (pre-existing) labeled data to learn the best set of decision rules to divide pairwise comparisons into either matches or non-matches. One of the main advantages of this method is that it balances how much the labeled and unlabeled data contribute to the likelihood of observing an agreement pattern. Thus, the amount of labeled data does not need

to be large but a representative sample of the data of interest. Using this approach, Winkler (2002) finds that the Fellegi-Sunter model, under the conditional independence assumptions discussed in section 2.2, is quite accurate in terms of producing error rates that are statistically indistinguishable from the truth. However, as recognized by Nigam et al. (2000) and Winkler (2002), the main disadvantage of this method is that the set of labeled cases need to exist before training a model. In addition, labeled cases need to be a representative sample of the datasets being merged. For example, Sariyar, Borg, and Pommerening (2012a) note that because of the low frequency of some comparisons, building a representative sample for labeling is a difficult and prohibitively time-consuming task.

Larsen and Rubin (2001), on the other hand, propose an iterative procedure that, as a first step, selects among multiple Fellegi-Sunter style candidate models.[5] Once a model is selected, the second step is to define a clerical review region using a similar procedure as the one described in section 2.3. After those cases are manually labeled, the third step is to fit the model again using labeled and unlabeled data. Finally, these steps are repeated until the number of observation labeled as matches is minimal. Similarly to Winkler (2002), Larsen and Rubin (2001) find that when labeled data is included to recover the parameters of a Fellegi-Sunter model, the estimated false positive and false negative rates are close to the true error rates.

A major caveat with the approach of Larsen and Rubin (2001) is that the number of cases for clerical review, at each iteration, tends to be incredibly large. The reason, as noted above, is that for a semi-supervised model to learn from labeled data and not be guided by unlabeled data alone, labeled observations need to be abundant and at least represent a reasonable fraction of the unlabeled cases (Winkler 2002). Of course, one way to achieve such a goal is to label as many cases as possible. For example, Larsen and Rubin (2001) merged datasets of roughly 10 thousand records each, and had to manually label between two and ten thousand pairs of record to improve the performance of the model. In practice, labelling more than a few thousand cases may take more than a couple days, even for an experienced clerk (Winkler 1995). Therefore, scalability concerns about this approach arise.

3.2 The Proposed Method

To overcome the aforementioned challenges, I propose an active learning approach for probabilisitic record linkage. Active learning is a special form of semi-supervised learning that does

[5]For example, this step involves the process of selecting between one Fellegi-Sunter model assuming conditional independence across the merging variables and another model relaxing that specific assumption.

	Semi-supervised approaches to Probabilistic Record Linkage within the Fellegi-Sunter framework		
	Larsen and Rubin 2001	Winkler 2002	This paper
No Pre-existing labeled data required	✓	✗	✓
Reasonably sized clerical review tasks	✗	✗	✓
Weights labeled and unlabeled data differently	✗	✓	✓
Iterative process	✓	✗	✓

Table 1: Comparison of the Different Semi-supervised Approaches for Probabilistic Record Linkage. Four dimensions are examined: pre-existence of labeled data, size of the clerical review task, how the likelihood is informed from labeled and unlabeled data, and the iterative nature of the process. The approach proposed in this paper overcomes the problems faced by Larsen and Rubin 2001 and Winkler 2002 by resorting to active learning.

not require any pre-existing labeled data. In active learning, labels are obtained when a classifier (the model) interactively queries an oracle (a human) about a sample of cases where it has problems to adjudicate labels (cases difficult to classify). Therefore, a classifier can be initially trained using unlabeled data only and subsequently incorporate *human judgment* into the process by manual labeling for those cases difficult to classify.[6] Thus, in probabilistic record linkage problems, active learning directly addresses the practical issues with the methods of Larsen and Rubin (2001) and Winkler (2002), as no pre-existing labeled data is needed and, by requiring a small sample of those cases that are difficult to classify, clerical review tasks can be dramatically reduced.

As noted in table 1, this paper incorporates the virtues of the approaches of Larsen and Rubin (2001) and Winkler (2002), and addresses their main drawbacks using active learning. Additionally, it lays out a sampling scheme specifically designed to tackle the mismatch between true and estimated error rates. It differs from other active learning approaches to record linkage (Sarawagi and Bhamidipaty 2002, Bilenko 2006, Bellare et al. 2012, and Sariyar, Borg, and Pommerening 2012a) as it is embedded a probabilistic record linkage framework. Furthermore, the proposed method directly learns from both labeled and unlabeled data – not only from labeled data as it is the case in the above-mentioned works. Finally, as recognized by Bilenko

[6]See Settles (2010) for a great introduction to active learning.

2006, while those active learning approaches to record linkage are quite appealing, scalability to large datasets is still a concern because the number comparisons needed grows quadratically with the size of the datasets in question. Under the recent computational improvements proposed by Enamorado, Fifield, and Imai (2018), the Fellegi-Sunter model becomes a promising tool in an active learning framework for probabilistic record linkage.

3.3 The Algorithm

The basic idea behind the proposed method is to efficiently incorporate *human judgment*, in the form of manually labeled cases, directly into the Fellegi-Sunter model. Due to the iterative nature of active learning, it is best to explain the proposed method through each one of its steps. This is done as follows:

`Step 1.` **Initialization:** as in most record linkage tasks, we do not observe the true matching status for any pairwise comparison in the cross-product ($N_{\mathcal{A}} \times N_{\mathcal{B}}$) of two datasets $\mathcal{A}$ and $\mathcal{B}$. Thus, in this step, we fit the Fellegi-sunter model with unlabeled data only and under the assumptions described in section 2.2. As in the Fellegi-Sunter framework, we learn the threshold W_1 and W_2 (illustrated in figure 2) using the estimated parameters of the model. These thresholds will serve to guide the selection of cases to be included in the clerical review process.

`Step 2.` **Sampling informative cases:** in active learning, uncertainty sampling is one the most popular ways to query cases for which the learning algorithm is the least certain about (Lewis and Catlett 1994). For example, in probabilistic record linkage, uncertainty sampling will prioritize labeling efforts in cases where the probability of being a match $\xi_{ij} = \Pr(M(i,j) = 1|\boldsymbol{\gamma}(i,j), \delta(i,j)) \approx 0.50$.[7]

In uncertainty sampling, the prioritization of queries happens in a stream-based fashion i.e., an observation that is difficult to be classified is manually labeled first, the model is fitted to the data again, and then another observation is requested for labeling. Only when the cases predicted as the most difficult for classification are labeled, observations that are farther away in that regard are considered for labeling. In record linkage applications, since labeling is an expensive and time-consuming task, a stream-based strategy is not

[7]Formally, given the observed comparison data, $\boldsymbol{\gamma}(i,j)$, to select the least confident case, we choose $\boldsymbol{\gamma}(i,j)^* = \arg\max_{\boldsymbol{\gamma}(i,j)}(1 - \hat{\xi}_{ij})$, where $\hat{\xi}_{ij} = \max\{\xi_{ij}, 1 - \xi_{ij}\}$. As a consequence observations with $\xi_{ij} = 0.50$ are the ones with the smallest margin between classes, and therefore the most uncertain ones in terms of matching status.

ideal as it would take a lot of time and effort (in terms of labeling) to reach cases that are blatant errors that the model predicts as accurate classifications as it is illustrated in figure 3.

To solve this problem, in this paper, I will follow a pool-based approach to select pairwise comparisons to be queried. The latter means that many observations will be chosen for labelling simultaneously. To that end, let W_c denote the Fellegi-Sunter weight associated with observation(s) that have a match probability $\xi_h = c$, where e.g., $c = 0.50$. Then, our measure of uncertainty sampling w_h can be written as:

$$w_h(W_h, W_c) = K\Big(W_h - W_c\Big) \tag{6}$$

where h indexes observed agreement patterns and $K(\cdot)$ is a kernel function centered around W_c e.g., a Gaussian kernel. Thus, if we select N_L observations for manual labeling, the number of randomly sampled cases per unique agreement pattern will be given by:

$$N_{L_h} = \begin{cases} N_L \times \sum\limits_{\{h':W_{h'} \leq W_h\}} w_h & \text{if } W_h \leq W_c \\ N_L \times \sum\limits_{\{h':W_{h'} \geq W_h\}} w_h & \text{otherwise} \end{cases} \tag{7}$$

where $\sum_{h=1}^{H} N_{L_h} = N_L$.

Note that if we were to follow the Fellegi-Sunter decisions rules to define a clerical review region, all the observations in the blue shaded area (left panel of figure 4) would be checked by clerks. However, the sampling scheme proposed in this paper will focus its attention in pairwise comparisons around W_c. In other words, observations in the green shaded region (right panel figure 4) are more likely to be selected, but by virtue of random sampling, a few observations to the left (right) of W_2 (W_1) will be selected as well.[8]

`Step 3.` **Clerical review:** after N_L pairwise comparisons have been selected for clerical review, the researcher will have to manually adjudicate the matching status for each of them (pseudo truth). In practice three options are given to classify the pair of records: match, non-match, and inconclusive. In the first two cases, a pair of records is considered labeled data, while in the last it is still considered unlabeled.

[8] In the empirical applications (section 4), it is shown that sampling between 10 and 20 observation per iteration works well in practice.

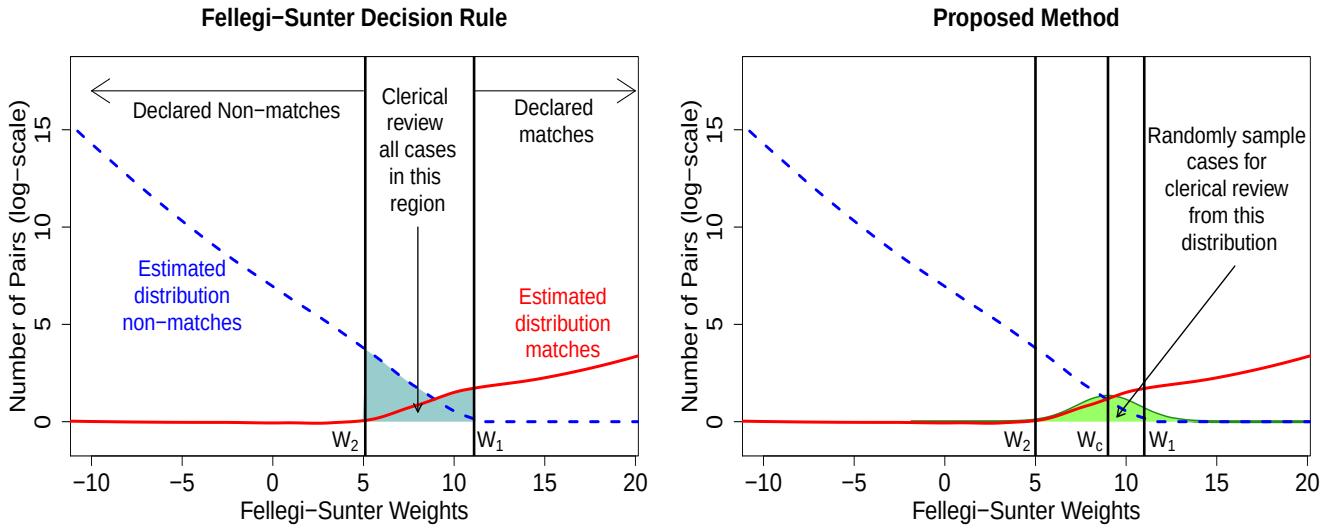


Figure 4: Intuition Behind the Sampling Scheme for Clerical Review Cases. Instead of embarking in a time consuming clerical review tasks as the one the Fellegi-Sunter decision rules would recommend (blue shaded region), the proposed method samples N_L observations inside and outside the Fellegi-Sunter clerical review region (green shaded area) according to the uncertainty sampling weights w_h centered around W_c.

`Step 4.` **Incorporate newly labeled cases into the model:** to incorporate labeled data into the Fellegi-Model, as Winkler (2002), we will follow the EM-Ω approach of Nigam et al. (2000) for semi-supervised learning. The idea is to find the best set of parameters that fit both the labeled data and unlabeled data within the Fellegi-Sunter model. The problem is that in record linkage applications, as it is the case for most semi-supervised learning tasks as well, the amount of labeled data overwhelms the information one can extract from the labeled data in a likelihood framework. Hence, a balancing weight $\Omega \in [0,1]$ whose purpose is to down-weight the importance of the unlabeled data is needed. Note that for smaller (larger) values of Ω, the less (more) information the unlabeled data brings to the model.

Let $T(i,j) = t$ denote the label status of a pair of records, that is, if $t = 1$ the matching status for the pair (i,j) has been labeled by a clerk, while if $t = 0$ it means that the matching status is still unobserved (unlabeled). Then, the complete data log-likelihood for this modified version of Fellegi-Sunter model can be written as:

$$\log \mathcal{L}_c(\lambda, \boldsymbol{\pi} \mid \boldsymbol{\delta}, \boldsymbol{\gamma}, T, \Omega) \propto$$

$$\underbrace{\sum_{i=1}^{N_A}\sum_{j=1}^{N_B}\mathbf{1}\{T(i,j)=1\}\sum_{m=0}^{1}\log\left\{\lambda^m(1-\lambda)^{1-m}\prod_{k=1}^{K}\left(\prod_{l=0}^{L_k-1}\pi_{kml}^{\mathbf{1}\{\gamma_k(i,j)=l\}}\right)^{1-\delta_k(i,j)}\right\}}_{\text{Labeled Data}}$$
$$+\Omega\underbrace{\sum_{i=1}^{N_A}\sum_{j=1}^{N_B}\mathbf{1}\{T(i,j)=0\}\sum_{m=0}^{1}\log\left\{\lambda^m(1-\lambda)^{1-m}\prod_{k=1}^{K}\left(\prod_{l=0}^{L_k-1}\pi_{kml}^{\mathbf{1}\{\gamma_k(i,j)=l\}}\right)^{1-\delta_k(i,j)}\right\}}_{\text{Unlabeled Data}} \tag{8}$$

This is a non-concave problem, for which a local maximum can still be found with the EM algorithm. Actually, under the conditional independence assumptions made in section 2.2, closed-form solutions exists for all the parameters of the model (see Appendix A.1.2).

`Step 5.` **Stopping criteria:** if the difference between the parameters of the models is larger than some pre-defined e.g., 0.0001, the process stops. If not, we cycle through `Step 2` to `5` until the stopping criteria is met.

As it will be demonstrated in the next section via two empirical applications, adopting an active learning approach and the use of the sampling scheme proposed in this paper results in a more robust probabilistic record linkage process.

4 Empirical Applications

As noted throughout this paper, the lack of a unique identifier that links datasets is a common problem faced by researchers when merging datasets. Still, additional challenges might appear as well e.g., the potential identifiers for the merge are noisy, the number of potential identifiers is small, the information in those identifiers is not rich enough to separate matches from not matches, etc. These are all situations where previous probabilistic record linkage models perform poorly. In this section, I test the robustness and demonstrate the advantages of the proposed methodology in two empirical applications. The first application validates the method as the true matching status for each pair of records is known ex-ante. In the second application, I will revisit a turnout validation study that involved one of the most extensive, detailed, and time-consuming clerical reviews in the social sciences.

4.1 A Validation Study: Merging Data on Local-level Politicians in Brazil

Instead of making use of synthetic data to test the performance of the model, I validate the model by merging two datasets where the true matching status is know. To do this, I merge data on local-level candidates in Brazil for the 2012 and 2016 municipal elections. They were obtained from the *Tribunal Superior Eleitoral* (TSE), the Brazilian office in charge of electoral matters. Each dataset contains information for more than 450,000 local politicians spread across the 5,570 Brazilian municipalities. In addition, these datasets contain more than 20 variables with information about the office for which a candidate was running, her party affiliation, the coalition she represented (if any), whether they were elected or not, and other information.[9]

Among all the variables, only six can help identify an individual across the two datasets: names, date of birth, marital status, the municipality and state where they were candidates, and the *Cadastro de Pessoas Físicas* (CPF), which is the Brazilian individual taxpayer registry identification number. Indeed, what makes these datasets ideal to validate the proposed method is that they include the CPF which is a unique identifier for each candidate. Furthermore, while the CPF is perfectly recorded for each individual, the other potential identifiers between datasets are not as they are manually entered into the database making them subject to misspellings and other types of errors. In addition, some of these fields might have changed over the course of four years e.g., marital status and names.

Merging two datasets with more than 450,000 records each would result in more than 200 billion comparisons. To reduce the number of comparisons, a blocking scheme is necessary. Specifically, for each local-level politician, the data is blocked by state of residence and gender. Such a strategy results in 52 merges (26 states times two gender categories). The block size ranges from 309 thousand pairs (Roraima/Female) to 3 billion pairs (São Paulo/Male) with the median value of 52 million pairs (Mato Grosso/Male). To make comparisons possible, I use three levels of agreement (`different, similar, identical (or nearly so)`) for the following variables: first name, last name, and age. For the string-valued variables we used Jaro-Winkler as our measure of distance with 0.88 and 0.94 as the thresholds (see Winkler 1990). In the case of age, the $L1$ norm was used with thresholds: 1 and 2.5 (see Jackman and Spahn 2018 and Enamorado and Imai 2018 for similar choices). The remaining variables: municipality name and marital status, were compared based on whether or not they had an identical value.

[9]The data used in the application can be directly from the TSE here: `http://www.tse.jus.br/eleicoes/estatisticas/repositorio-de-dados-eleitorais-1/repositorio-de-dados-eleitorais`

Note that while the amount of missing information per merging field is not too large (less than 1 percent), if an exact match based on these fields were to be conducted, only 61% of the more than 168 thousand true matches would be recovered due to other sources of noise in the data (e.g., typographical errors).

	Fellegi-Sunter	Active learning	Actual
Match rate	39.98	35.10	34.81
Rate of party switchers	57.75	52.70	52.59
Number of labeled cases	0	2,080	

Table 2: Match Rate and Rate of Party Switchers for Local-level Politicians in Brazil (2012 - 2014). Merging is based on the Fellegi-Sunter model ("Fellegi-Sunter"), the proposed methodology ("Active learning"), and the unique identifier ("Actual"). Overall, the proposed method produce rates that are almost identical to their true counterparts.

Table 2 summarizes the results of the merge. In terms of match rates, estimates from the Fellegi-Sunter model would lead to a match rate that is 5 percentage points larger than the truth. Using the proposed method with $\Omega = 0.01$, the match rate we obtain is only 0.29 percentage points apart from the true. For each of the 52 blocks, I label 20 record pairs, and after only two iterations of the proposed model, convergence was achieved. The latter results in manual labeling the matching status of 2,080 pairs of records, which represents 4% of the clerical review region suggested by the Fellegi-Sunter framework (53,502 cases).

The second row of table 2 shows that if one wanted to study the rate of party-switching between 2012 and 2016, the estimated rate using the Fellegi-Sunter model would be 5 percentage points larger that the truth. In contrast, the party-switching rate obtained from the proposed methodology is virtually the same as the one we would obtained by merging the datasets using the unique identifier. As shown in appendix A.3.1 these results are robust to the selection of Ω conditional on its value being positive but small. Moreover, in the same appendix I explore the effects of combining the proposed procedure with a noisy labeler. I find that if more than 30% manual mis-classifications (completely at random) are made, the rate of matches and party-switchers are as biased as the ones obtained from fitting the Fellegi-Sunter to unlabeled data.

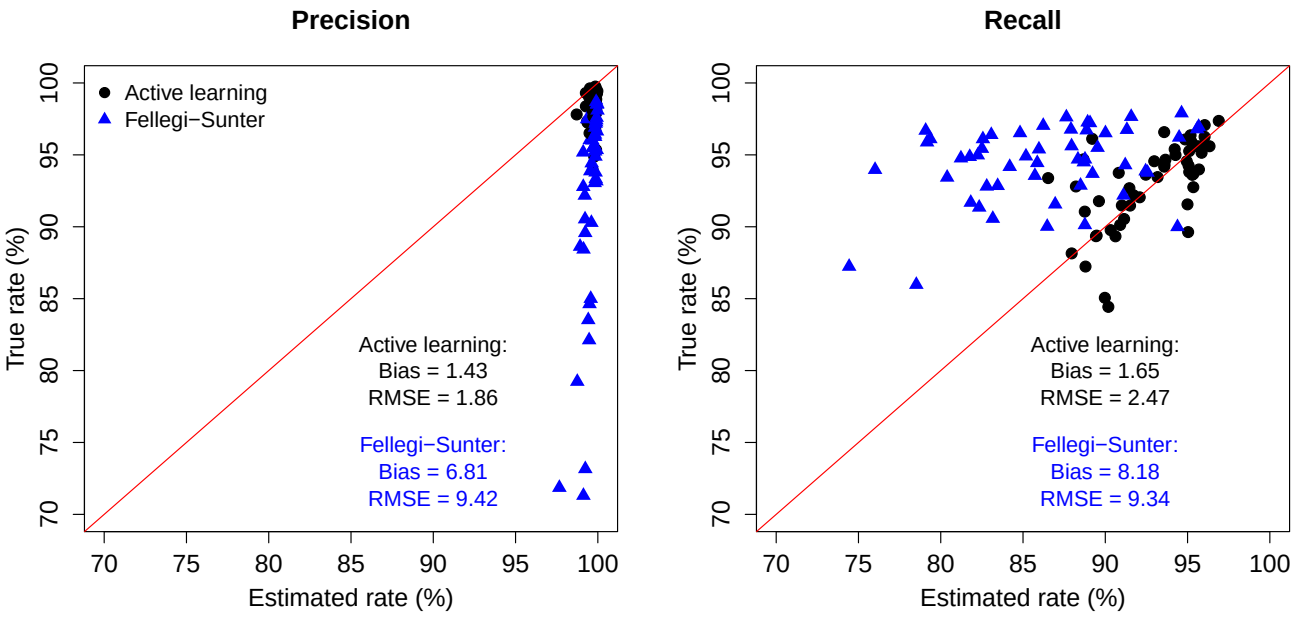


Figure 5: Precision and Recall Rates Across the 52 Blocks Involved in the Merging Process. The recall rate refers to the share of true matches found, while the precision rate represents the share of true matches among declared matches. Overall, the estimates obtained from Fellegi-Sunter (blue triangles) are less accurate when compared to the true if compared to those obtained from the active learning (black dots).

Figure 5 presents precision and recall rates, two common measures used to assess the accuracy of a classifier. The precision rate represents the share of true matches out of all the records classified as matches. The recall rate, on the other hand, denotes the share of true matches that we were able to recover.[10] Figure 5 shows that both in terms of precision (left panel) and recall (right panel) rates, active learning (black solid circles) produces estimates that are much closer to the truth if compared to the ones obtained from the Fellegi-Sunter model alone (blue solid triangles). In terms of bias and root mean squared error, we can see that active learning significantly outperforms the Fellegi-Sunter model by producing estimates that much closer to the truth. All these improvements are important, as accurate parameters estimates in probabilistic record linkage are key to properly account for the uncertainty in the merging process (Enamorado, Fifield, and Imai 2018).

Finally, to illustrate why the proposed sampling scheme is an efficient use of *human judgment* to improve the accuracy of probabilistic record linkage, figure 6 presents precision and

[10]These rates are obtained as a weighted averages using the match probability as weights. Under some mild assumptions, that strategy is shown to be appropriate by Enamorado, Fifield, and Imai (2018). In particular, one of the most important conditions is that the rankings obtained from the Fellegi-Sunter and Enamorado, Fifield, and Imai (2018) are equivalent. As shown in Appendix A.2 this is indeed the case.

recall rates obtained from: 1. the proposed methodology ("Uncertainty sampling"); and 2. the proposed methodology but replacing uncertainty with random sampling ("Random sampling"). As it can be noted in the figure, by labelling the matching status of 20 pairs of records (per block and iteration), the proposed method achieves precision and recall rates of 99% and 93%, respectively. If instead of uncertainty sampling, we use random sampling in `step 3`, we would need to label more than 200 pairs of records (per block and iteration) to achieve similar levels of accuracy. In other words, instead of labeling 2,080 pairs of records, we would require to manually adjudicate the matching status of 10 times that quantity. These results corroborate what the previous literature on active learning has found i.e., random sampling is not an optimal approach to select informative cases (see Settles 2010).

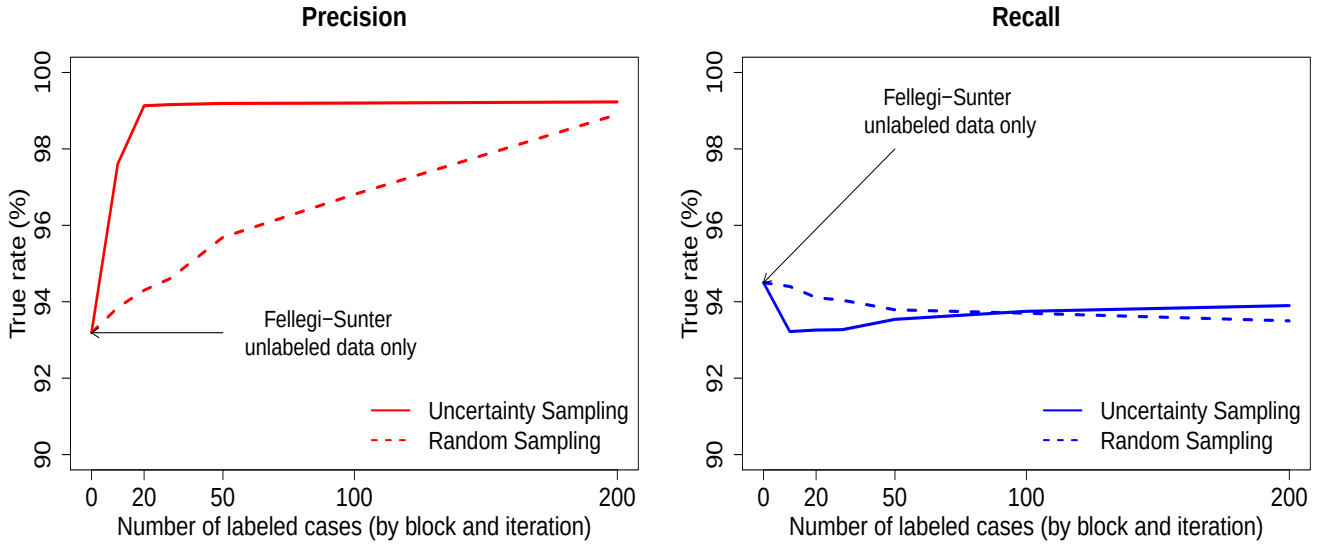


Figure 6: The Power of Uncertainty vs Random Sampling for Manual Labeling. Precision (red) and recall (blue) rates across different amounts of labeled data per iteration and block group used for the merge. The recall rate refers to the share of true matches found, while the precision rate represents the share of true matches among declared matches. The estimates obtained from the proposed method (solid lines) dominate those from the replacing uncertainty with simple random sampling (dashed lines).

4.2 Validating Turnout for the 2016 ANES

The American National Election Studies (ANES) is one of the most prestigious and methodologically rigorous public opinion study in the United States. The target population of the ANES are U.S. citizens eligible to vote. Since 2008, the ANES is carried out via two interview modes: the traditional face-to-face interviews; and the nowadays more common Internet interviews. The face-to-face component of the ANES is a multi-stage stratified cluster sample of

residential addresses, which due to financial constraints does not include Alaska and Hawaii. The Internet component of the ANES is a random sample of residential addresses in all the 50 U.S. states and the District of Columbia – from these addresses, individuals are selected to be part of the study. In 2016, out of 4,271 ANES respondents, 1,181 were face-to-face and 3,090 were Internet respondents; and the corresponding attrition rate from the pre-election to post-election surveys was 10% for the face-to-face and 16% for the Internet component.[11]

Motivated by the large gap between self-reported and validated turnout rates for previous ANES studies, Enamorado and Imai (2018) through a research collaboration agreement with the ANES, merged the 2016 ANES with a nationwide voter file using the Fellegi-Sunter model as implemented in fastLink, in order to validate the self-reported turnout of each ANES respondent. The merge was conducted based on the following common variables: first and last name, age, house number, street name, and postal code.

Merging the ANES with the voter file of more than 180 million observations would result in a total of over 756 billion comparisons. To reduce the number of comparisons, Enamorado and Imai (2018) blocked the data according to gender and state of residence, which resulted in 102 blocks. The block size ranges from 48,315 pairs (Hawaii/Female: ANES = 3, Voter file = 16,105) to 705 million pairs (California/Female: ANES = 225, Voter file = 3,137,276) with the median value of 11 million pairs (Idaho/Male: ANES = 28, Voter file = 426,636). Finally, to make comparisons possible, we selected three levels of agreement (`different, similar, identical (or nearly so)`) for the following variables: first name, last name, street name, and age. For the string-valued variables we used Jaro-Winkler as our measure of distance with the following thresholds: 0.85 and 0.94. In the case of age, the $L1$ norm was used with thresholds: 1 and 2.5. The remaining variables, house number and zip code, were compared based on whether they had an identical value or not.

Enamorado and Imai (2018), conducted an extensive and time-consuming clerical review for each ANES respondent. Unlike the previous empirical application (section 4.1) where the ground truth is known ex-ante, the goal of the clerical review was to approximate the truth (pseudo-truth) by product of strenuous manual coding. The clerical review discarded 280 false-matches for the ANES for which it was found that a survey respondent is matched with a different individual.

The proposed method revisits the results in Enamorado and Imai (2018). In particular, for

[11] For more details on the sampling design see American National Election Studies (2017) and DeBell et al. (2016).

`Step 2` (sampling informative cases) the number of observations selected for manual labeling was set to the minimum between 10 and the total amount of ANES respondents per block-group. In practice, the latter is done because there are states like Hawaii and South Dakota for which the ANES only interviewed a limited number of individuals – in those instances, labeling more pairs than observations available would be inefficient as it would be enough to check the best match available for those observations. In all, by implementing the active learning approach here, only 644 record pairs needed to be labeled, which represents 15% of the original clerical review conducted by Enamorado and Imai (2018) and 10% of the 6215 record pairs the decision rules of Fellegi-Sunter would suggest for labeling. The proposed method cycled through `Steps 1 - 5` two times for the largest states (California, New York, Texas, Florida, Pennsylvania) while only one iteration was necessary for the remaining block-groups. Note that as in the previous empirical application, I present results for the balancing parameter Ω when set equal to 0.01 – however, in appendix A.3.2 other values of Ω are explored as well. In particular, the larger the value of Ω the closer to the results obtained using unlabeled data.

Table 3 summarizes the results for the match rate among the ANES respondents. The match rate is given separately for the face-to-face and Internet samples as well as for the combined sample ("Overall"). The results are based on the Fellegi-Sunter model with unlabeled data ("Fellegi-Sunter"), the original clerical review in Enamorado and Imai (2018) ("clerical review") and the proposed method ("Active learning"). As a reference, two estimates of the registration rate are presented here. The first ("Voter file active") is the total number of registered "active" voters in the voter file divided by the number of eligible voters. Given that the exact definition of active voters varies by state and some states do not distinguish active and inactive voters, these estimates may still overestimate the actual registration rate in the population. The second one is the estimated registration rate based on self-reports from the Voter Supplement of the Current Population Survey (CPS) from the U.S. Census Bureau, which is an additional questionnaire of the CPS focusing on voting and registration that typically produces estimates close to the actual rates observed in the population.

Table 3 shows that the match rates based on the Fellegi-Sunter model using unlabeled data alone ("Fellegi-Sunter") are similar to the registration rates based on active voters – which as noted above might be an overestimate of the actual rate. In contrast, the estimates obtained from the proposed methodology cannot be statistically distinguished from the most comprehensive clerical review estimates. The turnout rates produced from the clerical review of En-

	Post-election Match Rates			Registration rate	
	Fellegi-Sunter	Clerical review	Active learning	Voter file active	CPS
Overall	77.15 (0.67)	69.85 (0.76)	70.88 (0.75)	76.57	70.34 (1.40)
Face-to-face	75.64 (1.27)	69.12 (1.42)	69.03 (1.41)	76.43	70.40 (1.39)
Internet	77.77 (0.79)	70.15 (0.90)	71.64 (0.88)	76.57	70.34 (1.40)
Number of labeled cases	0	4,271	644		

Table 3: Match Rates from the Results of Merging the 2016 ANES with the Nationwide Voter File. The match rates are computed separately for the face-to-face and Internet samples as well as together for the overall sample. Merging is based on the Fellegi-Sunter model ("Fellegi-Sunter'), the original clerical review in Enamorado and Imai 2018 ("Clerical review"), and the proposed methodology ("Active learning"). Standard errors are given within parentheses. As a benchmark the estimated registration rates from the voter files (Voter file active) as well as the self-reported registration rate from the Current Population Survey (CPS) are reported.

amorado and Imai (2018) and the one from the proposed approach are much closer to the self-reported registration rates from the CPS. Overall, there is little difference in the results across interview modes.

To further explore the gains of labeling just a small number of observations, the validated turnout rates are presented in table 4. These validated turnout rates are calculated as a weighted average of the binary turnout variable from the voter file where the match probability is used as a weight. Again, the results obtained from the Fellegi-Sunter model ("Fellegi-Sunter"), the clerical review of Enamorado and Imai 2018 ("clerical review"), and the proposed methodology ("Active learning") are compared with the actual turnout rates based on the voter file ("Voter file") and the United States election project ("Election project"). The standard errors that account for the sampling design of the ANES are given in parentheses.

In table 4, we see a similar pattern as in table 3. Due to not properly accounting for false positives, the validated turnout rate obtained from the Fellegi-Sunter model alone is 5 percentage points larger than the rates from the comprehensive clerical review and the proposed method. While both measures get closer to the actual turnout rates observed in the population, active learning requires only labeling 644 cases, while the comprehensive clerical review of Enamorado and Imai 2018 involved more than 6 times as many. Finally, it is worth noting

that the closeness between validated and actual turnout rates can be considered evidence in favor that through a rigorous sampling procedure the ANES is able to account problems like unit non-response and attrition, and still be representative of its target population.

	Post-election			Actual turnout	
	Fellegi-Sunter	clerical review	Active learning	Voter file	Election project
Overall	64.96 (0.96)	59.77 (1.00)	60.74 (1.01)	57.55	58.83
Face-to-face	67.59 (1.69)	63.07 (1.83)	63.55 (1.91)	57.58	58.86
Internet	63.99 (1.15)	58.55 (1.18)	59.72 (1.91)	57.55	58.83
Number of labeled cases	0	4,271	644		

Table 4: Validated Turnout Rates among the Survey Respondents from the 2016 ANES. The turnout rates are computed separately for the face-to-face and Internet samples as well as together for the overall sample. The validated turnout rates obtained from the probabilistic model alone ("Fellegi-Sunter"), the model plus clerical review ("clerical review"), and the proposed methodology ("Active learning") are compared to the actual turnout rate for the corresponding target population based on the voter file and the data from the United States election project. The standard errors are given in parentheses.

5 Concluding Remarks and Future Work

Incorporating information from multiple sources is at the core of social science research. In scenarios where a unique identifier is missing, the Fellegi-Sunter model of probabilistic record linkage has been proven to be a useful tool, especially in situations where the overlap between the datasets to be merged is large and the amount of noise in the data is moderate. However, in many common situations, the data at hand is far from perfect, which leads to poor estimates for the false match and false non-match rates when using the the Fellegi-Sunter model. To detect such problems, the researcher has to rely on lengthy clerical reviews or ad-hoc methods such as random spot checking.

In this paper, through active learning, I propose a method that efficiently incorporates *human judgment* into parameter estimation of probabilistic record linkage models. The method not only approximates error rates quite well, but offers guidance in how to conduct a clerical review that requires manual labeling a small sample of the most informative cases i.e., those

instances where the probabilistic model has problem adjudicating the matching status of a pair of records. Of course, there is a price: manual labeling the matching status of a few cases. However, as shown in the empirical applications, even when merging large datasets, the accuracy gains outweighs the costs.

While the proposed methodology offers a robust approach to probabilistic record linkage, incorporating such a technique with generative models that aim to model comparisons in a more flexible way i.e., beyond agreement patterns, could be an interesting avenue of future research. In addition, to further improve the proposed methodology, it is critical to find ways about how to find an optimal number of pairs to be labeled at each iteration; an open question in the active learning literature. Finally, even when the Fellegi-Sunter offers an excellent first approximation to many record linkage tasks, the combination of ensemble methods and semi-supervised appears to be a promising way to improve probabilistic record linkage.

References

Acharya, Avidit, Matthew Blackwell, and Maya Sen. 2016. "The Political Legacy of American Slavery." *Journal of Politics.* 78 (3): 621–641.

American National Election Studies. 2017. User's Guide and Codebook for the ANES 2016 Time Series Study. Technical report University of Michigan and Stanford University Ann Arbor, MI and Palo Alto, CA: .
URL: *https://www.electionstudies.org/wp-content/uploads/2018/03/anes_timeseries_2016_userguidecodebook.pdf*

Ansolabehere, Stephen, and Eitan Hersh. 2012. "Validation: What Big Data Reveal About Survey Misreporting and the Real Electorate." *Political Analysis.* 20 (4): 437–459.

Arceneaux, Kevin, Martin Johnson, René Lindstädt, and Ryan J. Vander Wielen. 2016. "The Influence of News Media on Political Elites: Investigating Strategic Responsiveness in Congress." *American Journal of Political Science.* 60 (1): 5–29.

Barbera, Pablo. 2015. "Birds of the Same Feather Tweet Together. Bayesian Ideal Point Estimation Using Twitter Data." *Political Analysis.* 23 (1): 76–91.

Belin, Thomas R., and Donald B. Rubin. 1995. "A Method for Calibrating False-Match Rates in Record Linkage." *Journal of the American Statistical Association.* 90 (June): 694–707.

Bellare, Kedar, Suresh Iyengar, Aditya Parameswaran, and Vibhor Rastogi. 2012. "Active Sampling for Entity Matching." In *Knowledge, Discovery, and Data Mining*.

Berent, Matthew K., Jon A. Krosnick, and Arthur Lupia. 2016. "Measuring Voter Registration and Turnout in Surveys." *Public Opinion Quarterly.* 80 (Fall): 597–621.

Bertrand, Marianne, Matilde Bombardini, and Francesco Trebbi. 2014. "Is It Whom You Know or What You Know? An Empirical Assessment of the Lobbying Process." *American Economic Review.* 104 (12): 3885–3920.

Bilenko, Mikhail. 2006. "Learnable Similarity Functions and Their Application to Record Linkage and Clustering." Ph.D. diss. University of Texas at Austin.

Bombardini, Matilde, and Francesco Trebbi. 2012. "Competition and Political Organization: Together or Alone in Lobbying for Trade Policy?" *Journal of International Economics.* 87 (1): 18–26.

Bonica, Adam. 2018. "Are Donation-Based Measures of Ideology Valid Predictors of Individual-Level Policy Preferences?" *Journal of Politics (Forthcoming)*.

Christen, Peter. 2012. *Data Matching. Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection.* Springer.

Cohen, W. W., P. Ravikumar, and S. Fienberg. 2003. "A Comparison of String Distance Metrics for Name-Matching Tasks." In International Joint Conference on Artificial Intelligence (IJCAI) 18.

De La O, Ana. 2013. "Do Conditional Cash Transfers Affect Electoral Behavior? Evidence from a Randomized Experiment in Mexico." *American Journal of Political Science.* 57 (1): 1–14.

DeBell, Matthew, Michelle Amsbary, Vanessa Meldener, Shelly Brock, and Natalya Maisel. 2016. Methodology Report for the ANES 2016 Time Series Study. Technical report Stanford University and the University of Michigan. Ann Arbor, MI and Palo Alto, CA: .
URL: *https://www.electionstudies.org/wp-content/uploads/2016/02/anes_timeseries_2016_methodology_report.pdf*

DellaVigna, Stefano, and Ethan Kaplan. 2007. "The Fox News Effect: Media Bias and Voting." *Quarterly Journal of Economics.* 122 (3): 1187–1234.

Dempster, Arthur P., Nan M. Laird, and Donald B. Rubin. 1977. "Maximum Likelihood from Incomplete Data Via the EM Algorithm (with Discussion)." *Journal of the Royal Statistical Society, Series B, Methodological.* 39 (1): 1–37.

Enamorado, Ted, Benjamin Fifield, and Kosuke Imai. 2017. *fastLink*. R package version 0.4.
URL: *https://CRAN.R-project.org/package=fastLink*

Enamorado, Ted, Benjamin Fifield, and Kosuke Imai. 2018. "Using a Probabilistic Model to Assist Merging of Large-scale Administrative Records." Social Science Research Network (SSRN).
URL: *https://ssrn.com/abstract=3214172*

Enamorado, Ted, and Kosuke Imai. 2018. "Validating Self-Reported Turnout by Linking Public Opinion Surveys with Administrative Records." Social Science Research Network (SSRN). **URL:** *https://ssrn.com/abstract=3217884*

Fellegi, I. P., and A. B. Sunter. 1969. "A Theory for Record Linkage." *Journal of the American Statistical Association.* 64 (328): 1183–1210.

Hall, Andrew, Connor Huff, and Shiro Kuriwaki. 2018. Wealth, Slave Ownership, and Fighting for the Confederacy: An Empirical Study of the American Civil War. Technical report Stanford University.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning*. Springer Series in Statistics. 2 ed. New York, NY, USA: Springer New York Inc.

Herzog, Thomas H., Fritz Scheuren, and William E. Winkler. 2010. "Record Linkage." *Wiley Interdisciplinary Reviews: Computational Statistics.* 2 (5): 535–543.

Hill, Seth J., and Gregory A. Huber. 2017. "Representativeness and Motivations of the Contemporary Donorate: Results from Merged Survey and Administrative Records." *Political Behavior.* 39 (1): 3–29.

Hopkins, Daniel J., and Jonathan M. Ladd. 2014. "The Consequences of Broader Media Choice: Evidence from the Expansion of Fox News." *Quarterly Journal of Political Science.* 9 (1): 115–135.

Jackman, Simon, and Bradley Spahn. 2018. "Why Does the American National Election Study Overestimate Voter Turnout?" *Political Analysis (Conditionally Accepted)*.

Jaro, Matthew. 1989. "Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida." *Journal of the American Statistical Association.* 84 (406): 414–420.

Kim, In Song. 2017. "Political Cleavages within Industry: Firm-level Lobbying for Trade Liberalization." *American Political Science Review.* 111 (1): 1–20.

Larsen, Michael D., and Donald B. Rubin. 2001. "Iterative Automated Record Linkage Using Mixture Models." *Journal of the American Statistical Association.* 96 (March): 32–41.

Lewis, D. D., and J. Catlett. 1994. "Heterogeneous uncertainty sampling for supervised learning." In Proceedings of the Eleventh International Conference on Machine Learning (ICML-94).

Martin, Gregory J., and Ali Yurukoglu. 2017. "Bias in Cable News: Persuasion and Polarization." *American Economic Review.* 107 (9): 2565–2599.

McVeigh, Brendan S., and Jared S. Murray. 2017. Practical Bayesian Inference for Record Linkage. Technical report Carnegie Mellon University.

Meredith, Marc, and Michael Morse. 2015. "The Politics of the Restoration of Ex-Felon Voting Rights: The Case of Iowa." *Journal of Biomedical Informatics.* 10 (1): 41–100.

Murray, Jared S. 2016. "Probabilistic Record Linkage and Deduplication after Indexing, Blocking, and Filtering." *Journal of Privacy and Confidentiality.* 7 (1): 3–24.

Newcombe, H. B., and J. M. Kennedy. 1962. "Record Linkage: Making Maximum Use of the Discriminating Power of Identifying Information." *Communications of Association for Computing Machinery.* 5 (11): 563–567.

Newcombe, H. B., J. M. Kennedy, S. J. Axford, and A. P. James. 1959. "Automatic Linkage of Vital Records." *Science.* 130: 954–959.

Nigam, K, A. K. McCalum, S. Thrun, and T Mitchell. 2000. "Text classification from labeled and unlabeled documents using EM." *Machine Learning.* 39 (2/3): 103–134.

Rueda, Miguel. 2016. "Small Aggregates, Big Manipulation: Vote Buying Enforcement and Collective Monitoring." *American Journal of Political Science.* 61 (1): 163–177.

Sadinle, Mauricio. 2017. "Bayesian Estimation of Bipartite Matchings for Record Linkage." *Journal of the American Statistical Association.* 112 (518): 600–612.

Sarawagi, S, and A Bhamidipaty. 2002. "Interactive Deduplication Using Active Learning." In *Knowledge, Discovery, and Data Mining*.

Sariyar, M, A Borg, and K Pommerening. 2012a. "Active learning strategies for the deduplication of electronic patient data using classification trees." *Journal of Biomedical Informatics.* 45 (5): 893–900.

Sariyar, Murat, Andreas Borg, and K. Pommerening. 2012b. "Missing Values in Deduplication of Electronic Patient Data." *Journal of the American Medical Informatics Association.* 19: e76–e82.

Settles, Burr. 2010. "Active Learning Literature Survey." Technical Report 2010-09-14, University Wisconsin–Madison.

Spahn, Bradley. 2017. Before The American Voter. Technical report Stanford University.

Steorts, Rebecca C., Samuel L. Ventura, Mauricio Sadinle, and Stephen E. Fienberg. 2014. "A Comparison of Blocking Methods for Record Linkage." In *Lecture Notes in Computer Science*. Vol. 8744 Privacy in Statistical Databases Privacy in Statistical Databases.

Thibaudeau, Yves. 1993. "The Discrimination Power of Dependency Structures in Record Linkage." *Survey Methodology.* 19: 31–38.

Winkler, William E. 1988. "Using the EM Algorithm for Weight Computation in the Fellegi–Sunter Model of Record Linkage." In *Proceedings of the Section on Survey Research Methods, American Statistical Association*. In *Proceedings of the Section on Survey Research Methods, American Statistical Association*.

Winkler, William E. 1990. "String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage." Proceedings of the Section on Survey Research Methods. American Statistical Association.
URL: *https://www.iser.essex.ac.uk/research/publications/501361*

Winkler, William E. 1993. "Improved Decision Rules in the Fellegi-Sunter Model of Record Linkage." In Proceedings of Survey Research Methods Section, American Statistical Association.
URL: *http://ww2.amstat.org/sections/srms/Proceedings/papers/1993_042.pdf*

Winkler, William E. 1995. *Business Survey Methods*. New York: J. Wiley.

Winkler, William E. 2002. Methods for Record Linkage and Bayesian Networks. Research Report Series (Statistics) 2002-05 Statistical Research Division U.S. Census Bureau.

Winkler, William E. 2006. "Automatic Estimation of Record Linkage False Match Rates." In *Proceedings of the Section on Survey Research Methods*. American Statistical Association.

Winkler, William E., and Willian Yancey. 2006. "Record Linkage Error-Rate Estimation without Training Data." In *Proceedings of the Section on Survey Research Methods*. American Statistical Association.

Yancey, Willian. 2005. "Evaluating String Comparator Performance for Record Linkage." Research Report Series. Statistical Research Division U.S. Census Bureau.

Zucco, Cesar. 2013. "When Payouts Pay Off: Conditional Cash Transfers and Voting Behavior in Brazil 2002–10." *American Journal of Political Science.* 57 (4): 810–822.

Zucco, Cesar. 2015. "The Impacts of Conditional Cash Transfers in Four Presidential Elections (2002–2014)." *Brazilian Journal of Political Science.* 9 (1): 135–149.

A Supplementary Appendix

A.1 Expectation Maximization (EM) Algorithm

A.1.1 EM for the Fellegi-Sunter Model

From the E-Step, we can compute the match probability for each pair using the Bayes rule as follows:

$$\begin{aligned} \xi_{ij} &= \Pr(M(i,j) = 1 \mid \boldsymbol{\delta}(i,j), \boldsymbol{\gamma}(i,j)) \\ &= \frac{\lambda \prod_{k=1}^{K} \left(\prod_{l=0}^{L_k - 1} \pi_{k1l}^{\mathbf{1}\{\gamma_k(i,j)=l\}} \right)^{1-\delta_k(i,j)}}{\sum_{m=0}^{1} \lambda^m (1-\lambda)^{1-m} \prod_{k=1}^{K} \left(\prod_{l=0}^{L_k - 1} \pi_{kml}^{\mathbf{1}\{\gamma_k(i,j)=l\}} \right)^{1-\delta_k(i,j)}} \end{aligned} \tag{9}$$

and from the M-Step we obtain:

$$\lambda = \frac{1}{N_{\mathcal{A}} N_{\mathcal{B}}} \sum_{i=1}^{N_{\mathcal{A}}} \sum_{j=1}^{N_{\mathcal{B}}} \xi_{ij} \tag{10}$$

$$\pi_{kml} = \frac{\sum_{i=1}^{N_{\mathcal{A}}} \sum_{j=1}^{N_{\mathcal{B}}} \mathbf{1}\{\gamma_k(i,j) = l)\}(1-\delta_k(i,j))\xi_{ij}^m (1-\xi_{ij})^{1-m}}{\sum_{i=1}^{N_{\mathcal{A}}} \sum_{j=1}^{N_{\mathcal{B}}} (1-\delta_k(i,j))\xi_{ij}^m (1-\xi_{ij})^{1-m}} \tag{11}$$

with a suitable set of starting values we cycle through the E and M-Steps until some convergence criteria is met e.g., the log-likelihood of the model increases less than a certain threshold (see Enamorado, Fifield, and Imai 2018).

A.1.2 EM-Ω

In the case of the proposed method, for unlabeled observations ($t = 0$), the E-step takes the following form:

$$\begin{aligned} \xi_{ij}^0 &= \Pr(M(i,j) = 1 \mid \boldsymbol{\delta}(i,j), \boldsymbol{\gamma}(i,j), T(i,j) = 0) \\ &= \frac{\lambda \prod_{k=1}^{K} \left(\prod_{l=0}^{L_k - 1} \pi_{k1l}^{\mathbf{1}\{\gamma_k(i,j)=l\}} \right)^{1-\delta_k(i,j)}}{\sum_{m=0}^{1} \lambda^m (1-\lambda)^{1-m} \prod_{k=1}^{K} \left(\prod_{l=0}^{L_k - 1} \pi_{kml}^{\mathbf{1}\{\gamma_k(i,j)=l\}} \right)^{1-\delta_k(i,j)}} \end{aligned} \tag{12}$$

For labeled data ($t = 1$), the matching status is fixed at the qualitative assessment made by the researcher:

$$\xi_{ij}^1 = \begin{cases} 1 & \text{if } (i,j) \text{ is labeled as Match} \\ 0 & \text{if } (i,j) \text{ is labeled as Non-match} \end{cases} \tag{13}$$

While the M-step, the parameters of the model are determined by labeled and unlabeled data as follows:

$$\lambda = \frac{\sum_{i=1}^{N_{\mathcal{A}}} \sum_{j=1}^{N_{\mathcal{B}}} \sum_{t=0}^{1} \Omega^{1-t} \mathbf{1}\{T(i,j) = t\} \xi_{ij}^t}{N_L + \Omega(N_{\mathcal{A}} N_{\mathcal{B}} - N_L)} \tag{14}$$

$$\pi_{kml} = \frac{\sum_{i=1}^{N_{\mathcal{A}}} \sum_{j=1}^{N_{\mathcal{B}}} \sum_{t=0}^{1} \Omega^{1-t} \mathbf{1}\{T(i,j) = t\} \mathbf{1}\{\gamma_k(i,j) = l\} (1 - \delta_k(i,j)) (\xi_{ij}^t)^m (1 - \xi_{ij}^t)^{1-m}}{\sum_{i=1}^{N_{\mathcal{A}}} \sum_{j=1}^{N_{\mathcal{B}}} \sum_{t=0}^{1} \Omega^{1-t} \mathbf{1}\{T(i,j) = t\} (1 - \delta_k(i,j)) (\xi_{ij}^t)^m (1 - \xi_{ij}^t)^{1-m}} \tag{15}$$

where N_L represents the number of labeled cases and, consequently, $N_{\mathcal{A}} N_{\mathcal{B}} - N_L$ denotes all the unlabeled pairs. Again, with a suitable set of starting values, we repeat the E-step and M-step until convergence.

A.2 Equivalence between Rankings

Proposition 1: the ranking obtained by sorting agreement patterns according to their Fellegi-Sunter weights (W_{ij}) is the same as the one would obtain by using the probability of being a match (ξ_{ij}) instead.

Proof: Note that for every observed agreement pattern $h \in \{1, \ldots, H\}$, where H is the total number of observed agreement patterns, if we divide the numerator and denominator of equation 12 by $\lambda \Pr(\boldsymbol{\gamma}_h \mid M_h = 0)$ we get:

$$\xi_h = \frac{R_h}{R_h + \frac{1-\lambda}{\lambda}}$$

where $R_h = \frac{\Pr(\gamma_h|M_h = 1)}{\Pr(\gamma_h|M_h = 0)}$. After some rearranging, the equation above is equivalent to:

$$R_h = \kappa \times \frac{\xi_h}{1 - \xi_h}$$

the ratio $\kappa = \frac{1-\lambda}{\lambda}$ is always positive and and in record linkage problems strictly larger than 1. Thus, if ξ_h increases, so does R_h as $\frac{\partial R_h}{\partial \xi_h} = \frac{\kappa}{(1-\xi_h)^2} > 0$.

Therefore, since the logarithm is a monotonically increasing function, that means that for h and h', with $h \neq h'$ and $\xi_h \geq \xi_{h'}$, we have that $W_h = \log R_h \geq \log R_{h'} = W_{h'}$. The latter is true for any pair of agreement patterns, so the ranking induced by ξ_h is preserved by W_h.

Following a similar rationale, one can argue that $W_h = \log R_h \geq \log R_{h'} = W_{h'}$ if and only if $R_h \geq R_{h'}$. The latter directly implies that $\xi_h \geq \xi_{h'}$ for every pairwise comparison between agreement patterns. Therefore the rankings from both approaches are equivalent.QED.

Proposition 1 discards any possible difference in the results obtained from the Fellegi-Sunter model if one were to use either the Fellegi-Sunter decision rules or the one proposed in Enamorado, Fifield, and Imai (2018).

A.3 Additional Empirical Results

A.3.1 A Validation Study: Merging Data on Local-level Politicians in Brazil

	Fellegi-	Active learning			
	Sunter	$\Omega = 0.01$	$\Omega = 0.05$	$\Omega = 0.10$	Actual
Match rate	39.98	35.10	36.36	36.76	34.81
Rate of party switchers	57.75	52.70	53.05	53.46	52.59

Table 5: Match Rate and Rate of Party Switchers for Local-level Politicians in Brazil (2012 - 2014). Merging is based on the Fellegi-Sunter model ("Fellegi-Sunter"), the proposed methodology ("Active learning") for three values of Ω, and a unique identifier ("Actual"). Overall, the proposed method produce rates that are almost identical to their true counterparts.

	Active learning with a mis-classification equal to:			
	0%	15%	30%	Actual
Match rate	35.10	37.00	39.05	34.81
Rate of party switchers	52.70	53.63	56.91	52.59

Table 6: The Effect of a Noisy Manual Labeling Process. Match Rate and Rate of Party Switchers for Local-level Politicians in Brazil (2012 - 2014). Merging is based on the proposed methodology ("Active learning") for three values of manual mis-classification, and a unique identifier ("Actual"). Mis-classifications are added completely at random to the labeled data. For active learning, Ω is set to 0.01.

A.3.2 Validating Turnout for the 2016 ANES

	Post-election Match Rates				
	Fellegi-Sunter	Clerical review	Active learning $\Omega = 0.01$	Active learning $\Omega = 0.05$	Active learning $\Omega = 0.10$
Overall	77.15 (0.67)	69.85 (0.76)	70.88 (0.75)	73.19 (0.72)	73.89 (0.71)
Face-to-face	75.64 (1.27)	69.12 (1.42)	69.03 (1.41)	71.19 (1.37)	71.82 (1.36)
Internet	77.77 (0.79)	70.15 (0.90)	71.64 (0.88)	74.01 (0.85)	74.74 (0.84)

Table 7: Match Rates from the Results of Merging the 2016 ANES with the Nationwide Voter File. The match rates are computed separately for the face-to-face and Internet samples as well as together for the overall sample. Merging is based on the Fellegi-Sunter model ("Fellegi-Sunter"), the original clerical review in Enamorado and Imai 2018 ("Clerical review"), and the proposed methodology ("Active learning") for three values of Ω. Standard errors are given within parentheses. As a benchmark the estimated registration rates from the voter files (Voter file active) as well as the self-reported registration rate from the Current Population Survey (CPS) are reported.

	Post-election Match Rates				
	Fellegi-Sunter	Clerical review	Active learning		
			$\Omega = 0.01$	$\Omega = 0.05$	$\Omega = 0.10$
Overall	64.96	59.77	60.74	62.15	62.59
	(0.96)	(1.00)	(1.01)	(0.95)	(0.95)
Face-to-face	67.59	63.07	63.55	64.72	65.06
	(1.69)	(1.83)	(1.91)	(1.87)	(1.83)
Internet	63.99	58.55	59.72	61.21	61.68
	(1.15)	(1.18)	(1.91)	(1.10)	(1.11)

Table 8: Validated Turnout Rates among the Survey Respondents from the 2016 ANES. The turnout rates are computed separately for the face-to-face and Internet samples as well as together for the overall sample. The validated turnout rates obtained from the probabilistic model alone ("Fellegi-Sunter'), the model plus clerical review ("clerical review"), and the proposed methodology ("Active learning") for three values of Ω. The standard errors are given in parentheses.