In Search of an Entity Resolution OASIS: Optimal Asymptotic Sequential Importance Sampling

Neil G. Marchant and Benjamin I. P. Rubinstein
School of Computing and Information Systems
University of Melbourne, Australia
{nmarchant, brubinstein}@unimelb.edu.au

ABSTRACT

Entity resolution (ER) presents unique challenges for evaluation methodology. While crowdsourcing platforms acquire ground truth, sound approaches to sampling must drive labelling efforts. In ER, extreme class imbalance between matching and non-matching records can lead to enormous labelling requirements when seeking statistically consistent estimates for rigorous evaluation. This paper addresses this important challenge with the OASIS algorithm: a sampler and F-measure estimator for ER evaluation. OASIS draws samples from a (biased) instrumental distribution, chosen to ensure estimators with optimal asymptotic variance. As new labels are collected OASIS updates this instrumental distribution via a Bayesian latent variable model of the annotator oracle, to quickly focus on unlabelled items providing more information. We prove that resulting estimates of F-measure, precision, recall converge to the true population values. Thorough comparisons of sampling methods on a variety of ER datasets demonstrate significant labelling reductions of up to 83% without loss to estimate accuracy.

1. INTRODUCTION

The very circumstances that give rise to entity resolution (ER) systems—lack of shared keys between data sources, noisy/missing features, heterogeneous distributions—explain the critical role of evaluation in the ER pipeline [9]. Production systems rarely achieve near-perfect precision and recall due to these many inherent ambiguities, and when they do, even minute increases to error rates can lead to poor user experience [22], lost business [26], or erroneous diagnoses and public health planning [16]. It is thus vital that ER systems are evaluated in a statistically sound manner so as to capture the true accuracy of entity resolution. This paper addresses this challenge with the development of an algorithm based on adaptive importance sampling, which we call 'OASIS'.

While crowdsourcing platforms provide inexpensive provisioning of annotations, sampling items for labelling must proceed carefully. A key challenge in ER is the inherent imbalance between matching and non-matching records which can be as high as $1 : n$ when matching two sources of n records (*e.g.,* reaching the millions). Researchers leverage several existing practices to evaluate such an ER system: (i) Label samples drawn from all candidate matches uniformly at random (*e.g.,* record pairs in two-source integration): while yielding unbiased estimates, this can take thousands of samples before finding one match-labelled sample, and many tens of thousands of labels before estimates converge. (ii) Balance inefficient passive sampling with cheap crowdsourcing resources: while crowdsourcing facilitates ER evaluation, large nonstationary datasets require constant refresh and can quickly drive costs back up. (iii) Exploit blocking schemes or search facilities to reduce non-match numbers: such filtering injects hidden bias into estimates.

By contrast, OASIS offers a principled alternative to evaluating F-measure, precision, recall—robust measures under imbalance—given an ER system's set of output similarity scores. OASIS forms an instrumental distribution from which it samples record pairs non-uniformly, minimising the estimator's asymptotic variance. This instrumental distribution is based on estimates of latent truth due to a simple Bayesian model, and is updated iteratively. By stratifying the pool of record pairs by similarity score, OASIS transfers performance estimates and samples fewer points. By ensuring our sampler may (with non-zero probability) sample any stratum, we manage the explore-exploit trade-off, admitting guarantees of statistical consistency: our estimates of F-measure, precision, recall converge to the true population parameters with high probability.

The unique characteristics of OASIS together yield a rigorous approach to ER evaluation that can use orders-of-magnitude fewer labels. This is borne out in thorough comparisons of baselines across six datasets of varying sizes and class imbalance (up to over 1:3000).

Contributions. 1) The novel OASIS algorithm for efficient evaluation of ER based on adaptive importance sampling. This algorithm has been released as an open-source Python package at https://git.io/OASIS ;

2) Theoretical guarantee that OASIS yields statistically consistent estimates, made challenging by the non-independence of the samples and the non-linearity of the F-measure; and

3) A comprehensive experimental comparison of OASIS with existing state-of-the-art algorithms demonstrating superior performance *e.g.,* 83% reduction in labelling requirements under a class imbalance of 1:3000.

Proceedings of the VLDB Endowment, Vol. 10, No. 11
Copyright 2017 VLDB Endowment 2150-8097/17/07.

2. BACKGROUND

Motivated by the challenges of accurate but efficient evaluation of ER, we begin by reviewing the key features of ER.

2.1 Entity resolution

Definition 1 (ER problem). *Let $\mathcal{D}_1$ and $\mathcal{D}_2$ denote two databases, each containing a finite number of records n_1, n_2 representing underlying entities; and let fixed, unknown relation $\mathcal{R} \subseteq \mathcal{D}_1 \times \mathcal{D}_2$ describe the matching records across the databases,* i.e., *pairs of records representing the same entity. The* entity resolution problem *is to approximate $\mathcal{R}$ with a predicted relation $\hat{\mathcal{R}} \subseteq \mathcal{D}_1 \times \mathcal{D}_2$.*

Remark 1. *For simplicity we focus on two-source ER, however our algorithms and theoretical results apply equally well to multi-source ER on relations over larger product spaces, and deduplicating a single source.*

An abundant literature describes the typical ER pipeline: *preparation* amortising record canonicalisation; *blocking* for reducing pair comparisons through a linear database scan; *scoring*, the most expensive stage, in which pair attributes are compared and summarised in similarity scores; and *matching* where sufficiently high-scoring pairs are used to construct $\hat{\mathcal{R}}$. Further normalisation pre- or post-linkage such as schema matching or record merging, while non-core, are important also. We refer the interested reader to review articles [28, 8, 15] and the references therein.

2.1.1 Similarity scores

ER is often cast as a binary classification problem on the set of record pairs $\mathcal{Z} = \mathcal{D}_1 \times \mathcal{D}_2$. A pair $z \in \mathcal{Z}$ has true Boolean label 1 if a "match", that is $z \in \mathcal{R}$, and label 0 if a "non-match", that is $z \notin \mathcal{R}$. In this work, we leverage the similarity scores produced in typical ER pipelines:

Definition 2. *A* similarity score $s(z) \in \mathbb{R}$ *quantifies the level of similarity that a given pair $z \in \mathcal{Z}$ exhibits,* i.e., *the predicted confidence of a match.*

Similarity scores originate from a variety of sources. The scoring phase of typical ER pipelines combine attribute-level dis/similarity measures *e.g.,* edit distance, Jaccard distance, absolute deviation, etc., into similarity scores. The combination itself is often produced by hand-coded rules or supervised classification, fit to a training set of known non/matches. Unlike in evaluation, data used for training need not be representative: heuristically-compiled training sets may be used when learning discriminative models. Any *confidence-based classifier*, *e.g.,* the support vector machine, or *probabilistic classifier*, *e.g.,* logistic regression or probability trees, produces legitimate similarity scores. Scores from probabilistic classifiers may or may not be *calibrated*:

Definition 3. *A scoring function $s(\cdot)$ is* calibrated *if, of all the record pairs mapping to $s(z) = \rho \in [0, 1]$, approximately $100 \times \rho$ percent are truly matching. For example, 60% of pairs with a score of 0.6 should be matches.*

2.2 Evaluation measures for ER

All ER evaluation methods produce statistics that summarise the types of errors made in approximating $\mathcal{R}$ with $\hat{\mathcal{R}}$. Arguably the most popular among these statistics is the pairwise F-measure which we focus on in this work. The F-measure is particularly well suited to ER, unlike accuracy for example, as its invariance to true negatives makes it more robust to class imbalance. The F-measure is a weighted harmonic mean of precision and recall; and in terms of Type I and Type II errors, the statistic on T labels is

$$F_{\alpha,T} = \frac{\mathrm{TP}}{\alpha(\mathrm{TP}+\mathrm{FP}) + (1-\alpha)(\mathrm{TP}+\mathrm{FN})} , \quad (1)$$

where $\alpha \in [0, 1]$ is a weight parameter; TP, FP, FN are true positive, false positive, false negative counts respectively.

$$\mathrm{TP} = \sum_{t=1}^{T} \ell_t \hat{\ell}_t \ , \ \mathrm{FP} = \sum_{t=1}^{T} (1-\ell_t)\hat{\ell}_t \ , \ \mathrm{FN} = \sum_{t=1}^{T} \ell_t (1-\hat{\ell}_t) \ ,$$

where $z_1, \ldots, z_T \sim p$ are query pairs sampled i.i.d from some underlying distribution p of interest on $\mathcal{Z}$ such as the uniform distribution; the ℓ_t denote ground truth labels recording (possibly noisy) membership of z_t within $\mathcal{R}$; and $\hat{\ell}_t$ indicates $z_t \in \hat{\mathcal{R}}$. When $\alpha = 1$, $F_{\alpha,T}$ reduces to precision, $\alpha = 0$ produces recall, and $\alpha = 1/2$ yields the balanced F-measure, with equal importance on precision and recall.[1]

Our goal will be to estimate the asymptotic limit of $F_{\alpha,T}$ as label budget $T \to \infty$. For finite pools $\mathcal{Z}$ this corresponds to labelling of all record pairs with sufficient repetition to account for (any) noise in the ground truth labels ℓ_t.

Remark 2. *The pairwise F-measure is termed "pairwise" to highlight the application of the measure to record pairs. Pairwise measures work well when there are only a few records across the databases which correspond to a particular entity. In such cases one should not use accuracy, due to significant class imbalance (*cf. *Section 3). For cases where most entities have many matching records, one may leverage transitivity constraints while looking to cluster-based measures for evaluation [20]. See [2] for a summary on evaluation.*

3. PROBLEM FORMULATION

Suppose we are faced with the task of evaluating an ER system as described in the previous section. Given that we do not know $\mathcal{R}$, how can we efficiently leverage labelling resources to estimate the pairwise F-measure?

Definition 4 (Efficient evaluation problem). *Consider evaluating a predicted ER $\hat{\mathcal{R}} \subseteq \mathcal{Z} = \mathcal{D}_1 \times \mathcal{D}_2$, equivalently represented by predicted labels $\hat{\ell}(z) = \mathbf{1}[z \in \hat{\mathcal{R}}]$ for $z \in \mathcal{Z}$. We are given access to:*

- *a pool[2] $P \subseteq \mathcal{Z}$ of record pairs,* e.g., *$P = \mathcal{Z}$;*
- *a similarity scoring function $s : \mathcal{Z} \to \mathbb{R}$; and*
- *a randomised labelling $\mathtt{Oracle} : \mathcal{Z} \to \{0, 1\}$, which returns labels $\ell(z) \sim \mathtt{Oracle}(z)$ indicating membership in $\mathcal{R}$. The oracle's response distribution is parametrised by* oracle probabilities $p(1|z) = \Pr[\mathtt{Oracle}(z) = 1]$.

With this setup, the efficient evaluation problem *is to devise an estimation procedure for F_α, which samples record pairs $z_1, \ldots, z_T \in P$ and makes use of the corresponding labels provided by the oracle. We adopt integer index notation on $\hat{\ell}, \ell$ and s to denote their values at the t-th query;* e.g., *$\hat{\ell}_t = \hat{\ell}(z_t)$ for query z_t.*

[1] The relationship to the β-parametrisation is $\alpha = 1/(1+\beta^2)$.

[2] We introduce the pool P for flexibility. It can be taken to be the entire $\mathcal{Z}$, or a proper subset for efficiency.

Solutions should produce estimates $\hat{F}_{\alpha,T}$ exhibiting:

(i) ***consistency:*** *convergence in probability to the true value F_α on pool P with respect to underlying distribution p*

$$F_\alpha = \lim_{T\to\infty} F_{\alpha,T} \ \ ; \ \ \textit{and} \tag{2}$$

(ii) ***minimal variance:*** *vary minimally about F_α.*

In other words, solutions should produce precise estimates whilst minimising queries to the oracle, since it is assumed that queries come at a high cost. Computational efficiency of the estimation procedure is not a direct concern, so long as the response time of the oracle dominates (typically of order seconds in a crowdsourced setting).

ER poses unique challenges for efficient evaluation.

Challenge: Extreme class imbalance. The inherent class imbalance in ER presents a challenge for estimation of F-measure. For deduped databases $\mathcal{D}_1, \mathcal{D}_2$, the minimum possible class imbalance occurs when both DBs contain n records and there is a matching record in $\mathcal{D}_1$ for every record in $\mathcal{D}_2$. In this case, the *class imbalance ratio* (ratio of non-matches to matches) is $n-1$. This is problematic for passive (uniform i.i.d.) sampling even for modest-sized databases, since $\mathcal{O}(n)$ expected pairs would be sampled for every match found. As F_α depends only on matches (both predicted and true), many queries to the oracle would be wasted on labels that don't contribute. The problem becomes one of searching for an oasis within a desert when $n \sim 10^6$ or more.

Approach: Biased sampling. One response to class imbalance is *biased sampling*, that is, sampling from a population (or space more generally) in a way that systematically differs from the underlying distribution [24, Chapter 5]. Biased sampling methods have found broad application in areas as diverse as survey methodology, Monte Carlo simulations, and active learning, to name a few. They work by leveraging known information about the system—here the similarity scores and the pool of record pairs—to obtain more precise estimates using fewer samples. One of the most effective biased sampling methods is *importance sampling (IS)*, which we illustrate below:

EXAMPLE. *Consider a random variable X with probability density $p(x)$ and consider the estimation of parameter $\theta = \mathrm{E}[f(X)]$. The standard (passive) approach draws an i.i.d. sample from p and uses the Monte Carlo estimator $\hat{\theta} = \frac{1}{T}\sum_{i=1}^{T} f(x_i)$. Importance sampling, by contrast, draws from an* instrumental distribution *denoted by q. Even though the sample from q is biased (i.e. not drawn from p), an unbiased estimate of θ can be obtained by using the bias-corrected estimator $\hat{\theta}^{\mathrm{IS}} = \frac{1}{T}\sum_{i=1}^{T} \frac{p(x_i)}{q(x_i)} f(x_i)$.*

An important consideration when conducting IS is the choice of instrumental distribution, q. If q is poorly selected, the resulting estimator may perform worse than passive sampling. If on the other hand, q is selected judiciously, so that it concentrates on the "important" values of X, significant efficiency dividends will follow.

4. A NEW ALGORITHM: OASIS

This section develops our new algorithm for evaluating ER—*Optimal Asymptotic Sequential Importance Sampling (OASIS)*. In designing an adaptive/sequential importance sampler (AIS), we proceed in two stages: (i) choosing an appropriate instrumental distribution to optimise asymptotic variance of the estimator, see Section 4.1; and (ii) deriving an appropriate update rule and initialisation process for the instrumental distribution, now restricted to score strata, see Sections 4.2 and 4.3. Section 4.4 brings all of the components of OASIS together, presenting the algorithm in its entirety. Section 5 presents a thorough theoretical analysis of OASIS.

4.1 Selecting the instrumental distribution

We begin by defining an estimator for the F-measure which corrects for the bias of AIS. It is based on the standard estimator of Eqn. (1), with the addition of importance weights.

DEFINITION 5. *Let $\{x_t = (z_t, \ell_t)\}_{t=1}^{T}$ be a sequence of record pairs and labels, where the t-th record pair in the sequence is drawn from pool P according to an instrumental distribution q_t, which may depend on the previously sampled items $\boldsymbol{x}_{1:t-1} = \{x_1, \ldots, x_{t-1}\}$ and labels $\ell_t \sim \mathtt{Oracle}(z_t)$. Then the* AIS estimator for the F-measure *is given by*

$$\hat{F}_\alpha^{\mathrm{AIS}} = \frac{\sum_{t=1}^{T} w_t \ell_t \hat{\ell}_t}{\alpha \sum_{t=1}^{T} w_t \hat{\ell}_t + (1-\alpha)\sum_{t=1}^{T} w_t \ell_t} \ , \tag{3}$$

where $w_t = p(z_t)/q_t(z_t)$ is the importance weight associated with the t-th item, and p denotes any underlying distribution on the record pairs from which the target F_α is defined.

This definition assumes that the record pairs are drawn from an, as yet, unspecified sequence of instrumental distributions $\{q_t\}_{t=1}^{T}$. It is important that these instrumental distributions are selected carefully, so as to maximise the sampling efficiency. Later, we justify the choice of $\hat{F}_\alpha^{\mathrm{AIS}}$ by proving that it is consistent for F_α (*cf.* Theorem 2).

REMARK 3. *In ER we take: $P \subseteq \mathcal{Z}$ typically a DB product space $\mathcal{D}_1 \times \mathcal{D}_2$ which is finite (but possibly massive); and the p through which $F_{\alpha,T}$ is most naturally defined is the uniform distribution on P* i.e., *placing uniform mass $1/N$ where $N = |P|$. However OASIS and its analysis actually hold more generally: pools P of* instances *that could be uncountably infinite in size; and arbitrary marginal distributions p on P.*

4.1.1 Variance minimisation

A common approach for instrumental distribution design is based on the principle of variance minimisation [24]. In the ideal case, a single instrumental distribution (for all t) is selected that minimises the variance of the estimator:

$$q^\star \in \arg\min_{q} \mathrm{Var}(\hat{F}_\alpha^{\mathrm{AIS}}[q]) \ . \tag{4}$$

This optimisation problem is difficult to solve analytically, in part due to the intractability of the variance term. However, by replacing variance with the *asymptotic* variance (taking $T \to \infty$), a solution is obtained as

$$\begin{aligned} q^\star(z) \propto p(z) \Big[&(1-\alpha)(1-\hat{\ell}(z)) F_\alpha \sqrt{p(1|z)} \\ &+ \hat{\ell}(z)\sqrt{\alpha^2 F_\alpha^2 (1-p(1|z)) + (1-F_\alpha)^2 p(1|z)} \Big] , \end{aligned} \tag{5}$$

where $p(z)$ is the underlying distribution on P (see Remark 3) and $p(1|z)$ is the oracle probability (see Definition 4). The proof of this result is given in [25]. We call $q^\star(z)$ the *asymptotically optimal instrumental distribution*, owing to its relationship with asymptotic minimal variance.

4.1.2 *Motivation for adaptive sampling*

Close examination of (5) reveals that the asymptotically optimal instrumental distribution depends on the true F-measure F_α and true oracle probabilities $p(1|z)$, both of which are unknown a priori. This implies that an adaptive procedure is well-suited to this problem: we estimate $q^\star$ at iteration t using *estimates* of F_α and $p(1|z)$, which themselves are based on the previously sampled record pairs and labels $\boldsymbol{x}_{1:t-1}$. As the sampling progresses and labels are collected, the estimates of F_α and $p(1|z)$ should approach their true values, and $q_t^\star$ should in turn approach $q^\star$.

In order to implement this adaptive procedure, we must devise a way of iteratively estimating F_α and $p(1|z)$. There is a natural approach for F_α: we simply use $\hat{F}_\alpha^{\mathrm{AIS}}$ at the current iteration. However, the oracle probabilities present more of a difficulty. We outline one approach in Section 4.2.

4.1.3 *Exploration vs. exploitation*

In the subsequent analysis of OASIS (*cf.* Section 5), we show that the asymptotically optimal instrumental distribution given in Eqn. (5) does not guarantee consistency (convergence in probability). This is because it permits zero weight to be placed on some items, meaning that parts of the pool may never be explored. Consequently, we propose to replace $q^\star$ by an ε-greedy distribution

$$q(z) = \varepsilon \cdot p(z) + (1 - \varepsilon) \cdot q^\star(z) \ , \qquad (6)$$

where $0 < \varepsilon \leq 1$. For ε close to 0, the sampling approaches optimality (it exploits), whereas for ε close to 1, the sampling approaches passivity (it explores). This bears resemblance to explore-exploit trade-offs commonly encountered in online decision making (*e.g.,* multi-armed bandits) [6].

4.2 Estimating the oracle probabilities

In this section, we propose an iterative method for estimating the oracle probabilities, which are required for the estimation of $q^\star$. Our proposed method brings together two key concepts: stratification and a Bayesian generative model of the label distribution.

4.2.1 *Stratification*

Stratification is a commonly used technique in statistics that involves dividing a population into homogeneous subgroups (called strata) [10]. Often the process of creating the strata is achieved by *binning* according to a variable, or *partitioning* according to a set of rules. Our use of stratification is somewhat atypical, in that we are not using it to estimate a population parameter, but rather as a *parameter reduction* technique. Specifically, we aim to map the set of oracle probabilities $\{p(1|z) : z \in P\}$ (of size $N = |P|$ in ER) to a smaller set of parameters of size K, essentially one per stratum.

Parameter reduction. Consider a partitioning of record pair pool P into K disjoint strata $\{P_1, \ldots, P_K\}$, such that the pairs in a stratum share approximately the same values of $p(1|z)$.[3] If this ideal condition is satisfied, then our work in estimating the set of probabilities $\{p(1|z) : z \in P\}$ is significantly reduced, because information gained about a particular pair $z \in P_k$ is immediately transferable to the other pairs in P_k. As a result, we can effectively replace the set of probabilities $\{p(1|z) : z \in P_k\}$ for the record pairs in P_k, by a single probability $p(1|P_k)$.

Algorithm 1 Cumulative $\sqrt{F}$ (CSF) stratification [12]

```
Input:  $P$          pool of record pairs
        $s$          similarity score function : $P \to \mathbb{R}$
        $\tilde{K}$  desired number of strata
        $M$          number of bins (for estimating score dist.)
Output: strata $P_1, \ldots, P_K$ (not guaranteed $K = \tilde{K}$)
 1: Pool scores: $S \leftarrow \{s(z) | z \in P\}$
 2: Distribution of scores ($F$) using $M$ bins:
    $\texttt{counts}, \texttt{score\_bins} \leftarrow \mathrm{histogram}(S, \mathrm{bins} = M)$
 3: Cum. dist. of $\sqrt{F}$: $\texttt{csf} \leftarrow \left[\sum_{i=1}^{m} \sqrt{\texttt{counts}[i]}\right]_{m=1:M}$
 4: Bin width on cum. $\sqrt{F}$ scale: $w \leftarrow \texttt{csf}[M]/\tilde{K}$
 5: for $k \in \{1, \ldots, \tilde{K} + 1\}$ do
 6:     Bins on cum. $\sqrt{F}$ scale: $\texttt{csf\_bins}[k] \leftarrow (k-1)w$
 7: end for
 8: $K \leftarrow 1$
 9: for $j \in \{1, \ldots, M\}$ do
10:     if $K = \tilde{K}$ or $j = M$ then
11:         Append $\texttt{score\_bins}[\tilde{K}]$ to $\texttt{new\_bins}$
12:         break
13:     end if
14:     if $\texttt{csf}[j] \geq \texttt{csf\_bin}[K]$ then
15:         Append $\texttt{score\_bins}[j]$ to $\texttt{new\_bins}$
16:         $K \leftarrow K + 1$
17:     end if
18: end for
19: Allocate record pairs $P$ to strata $P_1, \ldots, P_K$ based on
    $\texttt{new\_bins}$ (remove any empty strata, updating $K$)
20: return $P_1, \ldots, P_K$
```

Relaxing the homogeneity condition. In reality, we don't know which record pairs in P (if any) have roughly the same values of $p(1|z)$. Fortunately, it turns out that this condition does not need to be satisfied too strictly in order to be useful. Previous work [3, 14] has demonstrated that the homogeneity condition can be satisfied in an approximate sense by using similarity scores as a proxy for true oracle probabilities. In other words, *we regard a stratum to be approximately homogeneous if the pairs it contains have roughly the same similarity scores.* The more this proxy holds true, the more efficient OASIS becomes in practice; however critically, our guarantees hold true regardless.

Stratification method. In order to stratify the record pairs in P according to their similarity scores, we shall use the *cumulative* $\sqrt{F}$ *(CSF) method*, originally proposed in [12] and previously used in the present context in [14]. The CSF method has a strong theoretical grounding, in that it aims to achieve minimal intra-stratum variance in the scores.

For completeness, we have included an implementation of the method in Algorithm 1. It proceeds by constructing an empirical estimate of the cumulative square root of the distribution of scores (lines 2–3). Then the strata are defined as equal-width bins on the CSF scale (lines 4–7). Finally, the bins are mapped from the CSF scale to the score scale (lines 8–18), so that the scores (record pairs) may be binned in the usual way (line 19). We note that any stratification method could be used in place of the CSF method (*cf. e.g.,* the *equal size method* described in [14]).

[3] This is the meaning of "homogeneity" which we adopt.

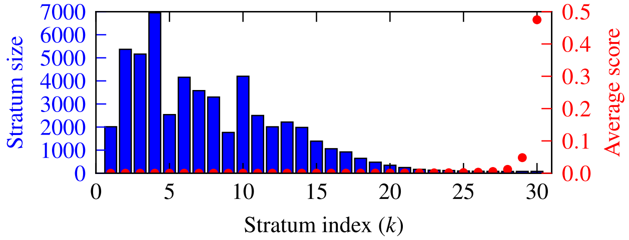

Figure 1: Size and mean score of the CSF strata for the Abt-Buy pool, using calibrated (probabilistic) scores.

Selecting the number of strata. The number of strata K represents a trade-off: For large K, estimates of the oracle probabilities enjoy finer granularity and can better approach their true values; however large K leads to more parameters and hence more labels required for convergence of estimates.

In practice for ER evaluation, we find that the there is often a "natural" range of K for the CSF method. The example in Figure 1 shows that we typically construct very large strata with low similarity scores, and very small strata with high similarity scores: a form of heavy-tailed distribution due to the extreme class imbalance. If K is set too large, then we immediately discover the strata corresponding to the higher similarity scores become too small (they may contain only 1 or 2 record pairs). We find a range of K from roughly 30–60 to work well for most datasets considered in Section 6.

4.2.2 A Bayesian generative model

Having partitioned the record pairs in P into K strata $\{P_1, \ldots, P_K\}$, our goal is to estimate $p(1|P_k)$ (for all k) using the collected `Oracle` labels. For notational convenience, we denote the true value of $p(1|P_k)$ by π_k and a corresponding estimate by $\hat{\pi}_k$. We shall adopt a generative model for observed labels which regards π_k as a latent variable.

Model of a stratum. Consider a label ℓ received from the oracle for a record pair in P_k. We assume that the label is generated from a Bernoulli distribution with probability π_k of being a match (binary label '1'), *i.e.*,

$$\ell \sim \text{Bernoulli}(\pi_k) \ . \tag{7}$$

Since the Bernoulli distribution is conjugate to the beta distribution, we adopt a beta prior for π_k:

$$\pi_k \sim \text{Beta}(\gamma_{0,k}^{(0)}, \gamma_{1,k}^{(0)}) \ , \tag{8}$$

where $\gamma_{0,k}^{(0)}$ and $\gamma_{1,k}^{(0)}$ are the prior hyperparameters. We describe how to choose the prior hyperparameters in Sections 4.3 and 4.4.

Joint model of strata. To model each stratum independently but not identically—we do not transfer information across strata but grant each a prior—we factor the joint prior distribution as a product of the marginal K priors. We collect the π_k's into a vector $\boldsymbol{\pi} = [\pi_1, \pi_2, \ldots, \pi_K]$ and the prior hyperparameters into a $2 \times K$ matrix:

$$\boldsymbol{\Gamma}^{(0)} = \begin{bmatrix} \gamma_{0,1}^{(0)} & \gamma_{0,2}^{(0)} & \cdots & \gamma_{0,K}^{(0)} \\ \gamma_{1,1}^{(0)} & \gamma_{1,2}^{(0)} & \cdots & \gamma_{1,K}^{(0)} \end{bmatrix} \ . \tag{9}$$

The posterior distribution of $\boldsymbol{\pi}$, given the labels received from the oracle up to iteration t, is a product of the K corresponding independent beta posterior distributions. Continuing with the previous notation, we store the posterior hyperparameters at iteration t in a matrix $\boldsymbol{\Gamma}^{(t)}$.

Iterative posterior updates. To obtain a new estimate of $\boldsymbol{\pi}$ per iteration, we iteratively update the posterior hyperparameters upon arrival of `Oracle` label observations ℓ_t. Suppose label ℓ_t is observed as a result of querying with a record pair from stratum $P_{k^\star}$. Then the update involves:

$$\begin{aligned} \text{copy old values}: \quad & \boldsymbol{\Gamma}^{(t)} \leftarrow \boldsymbol{\Gamma}^{(t-1)} \\ \text{if } \ell_t = 1: \quad & \gamma_{0,k^\star}^{(t)} \mathrel{+}= 1 \\ \text{if } \ell_t = 0: \quad & \gamma_{1,k^\star}^{(t)} \mathrel{+}= 1 \end{aligned} \tag{10}$$

A point estimate of $\boldsymbol{\pi}$ can be obtained at iteration t via the posterior mean

$$\hat{\boldsymbol{\pi}}^{(t)} = \text{E}[\boldsymbol{\pi}|\ell_1, \ldots, \ell_t] = \frac{\boldsymbol{\Gamma}_{0,:}^{(t)}}{\boldsymbol{\Gamma}_{0,:}^{(t)} + \boldsymbol{\Gamma}_{1,:}^{(t)}} \ . \tag{11}$$

Here the notation $\boldsymbol{\Gamma}_{i,:}^{(t)}$ represents the i-th row of matrix $\boldsymbol{\Gamma}^{(t)}$, and the division is carried out element-wise.

Remark 4. *As a practical modification to speed up convergence of $\hat{\boldsymbol{\pi}}$, we can decrease our reliance on the prior as labels are received. For each column $\boldsymbol{\Gamma}_{:,k}^{(0)}$ we can retroactively multiply by a factor $1/n_k$ where n_k is the number of labels sampled from P_k thus far. Anecdotally we also observe that this improves robustness to misspecified priors.*

4.2.3 Stratified instrumental distribution

Since the estimation method for the oracle probabilities produces estimates over the strata, rather than for individual pairs in the pool, it is appropriate to estimate the instrumental distribution in the same way. Akin to the mapping from $p(1|z)$ to $\boldsymbol{\pi}$, we therefore propose to map $q(z)$ to a vector $\boldsymbol{v} = [v_1, \ldots, v_K]$ based on our Bayesian stratified model estimates instead of (unknowable) population parameters. Adapting Eqn. (5), the stratified asymptotically optimal instrumental distribution $\boldsymbol{v}^\star$ is defined at iteration t as

$$\begin{aligned} v_k^{\star(t)} \propto \omega_k \Bigg[& (1-\alpha)(1-\lambda_k)\hat{F}_\alpha^{(t-1)} \sqrt{\hat{\pi}_k^{(t-1)}} \\ & + \lambda_k \sqrt{(\alpha \hat{F}_\alpha^{(t-1)})^2 (1-\hat{\pi}_k^{(t-1)}) + (1-\hat{F}_\alpha^{(t-1)})^2 \hat{\pi}_k^{(t-1)}} \Bigg] , \end{aligned}$$

where $\omega_k = |P_k|/N$ is the *weight* associated with P_k and $\lambda_k = \frac{1}{|P_k|}\sum_{i \in P_k} \hat{\ell}_i$ is the mean prediction in P_k. It follows that the ε-greedy distribution at iteration t is given by

$$v_k^{(t)} = \varepsilon \cdot \omega_k + (1-\varepsilon) \cdot v_k^{\star(t)} \ . \tag{12}$$

Having adopted a stratified representation for the instrumental distribution, sampling a record pair is now a two-step process. First a stratum index is drawn from $\{1, \ldots, K\}$ according to $\boldsymbol{v}$. Then a record pair is drawn uniformly at random from the resulting stratum.

4.3 Initialisation

OASIS requires a set of prior hyperparameters $\boldsymbol{\Gamma}^{(0)}$ and a guess for the F-measure $\hat{F}_\alpha^{(0)}$ for initialisation purposes. We elect to set these quantities based on the information contained within the similarity scores. Our approach depends

Algorithm 2 Initialisation of Bayesian model

```
Input:  $0 \leq \alpha \leq 1$          F-measure weight
        $P$                            pool of record pairs
        $\hat{\mathcal{R}}$            predicted ER
        $s$                            similarity score function : $P \to \mathbb{R}$
        $\tau$                         $\mathbb{R}$-valued score threshold (optional)
        $\{P_k\}_{k=1}^K$              stratum allocations
Output: $\hat{F}_\alpha^{(0)}$         initial F-measure
        $\hat{\boldsymbol{\pi}}^{(0)}$ prior hyperparameters

1: for $k \in \{1, \ldots, K\}$ do
2:     Mean score per stratum: $\hat{\pi}_k^{(0)} \leftarrow \frac{1}{|P_k|} \sum_{z \in P_k} s(z)$
3:     if scores are not probabilities in $[0, 1]$ then
4:         Transform: $\hat{\pi}_k^{(0)} \leftarrow \text{logit}(\hat{\pi}_k^{(0)} - \tau)$
5:     end if
6:     Mean pred. per stratum: $\lambda_k \leftarrow \frac{1}{|P_k|} \sum_{z \in P_k} \hat{\ell}_z$
7: end for
8: $\hat{F}_\alpha^{(0)} \leftarrow \frac{\sum_{k=1}^K |P_k| \pi_{0,k} \lambda_k}{\alpha \sum_{k=1}^K |P_k| \lambda_k + (1-\alpha) \sum_{k=1}^K |P_k| \pi_{0,k}}$
9: return $\hat{F}_\alpha^{(0)}$, $\hat{\boldsymbol{\pi}}^{(0)}$
```

centrally on a guess for the oracle probabilities $\hat{\boldsymbol{\pi}}^{(0)}$, in that once $\hat{\boldsymbol{\pi}}^{(0)}$ is available, the values of $\boldsymbol{\Gamma}^{(0)}$ and $\hat{F}_\alpha^{(0)}$ immediately follow. The details of the initialisation are contained in Algorithm 2, with further explanation given below.

Oracle probabilities (lines 2–5). A reasonable guess for $\boldsymbol{\pi}$ can be obtained by taking the mean of the similarity scores in each stratum. If the scores are not probabilities, they should be mapped to the $[0, 1]$ interval. This can be achieved by applying the logistic function.

F-measure (lines 6 & 8). The calculation of $\hat{F}_\alpha^{(0)}$ depends on the guess for $\boldsymbol{\pi}$ described above and the mean prediction per stratum $\boldsymbol{\lambda}$. Breaking down the calculation term-by-term, one begins by estimating the probability of finding a true positive in P_k as $\hat{\pi}_k^{(0)} \lambda_k$, so that the total number of true positives may be approximated by $\sum_{k=1}^K |P_k| \hat{\pi}_k^{(0)} \lambda_k$. Similarly, the total number of actual positives (TP + FN) may be approximated by $\sum_{k=1}^K |P_k| \hat{\pi}_k^{(0)}$. The total number of predicted positives (TP + FP) is known exactly and can be written in terms of $\boldsymbol{\lambda}$ as $\sum_{k=1}^K |P_k| \lambda_k$. Using these estimates in Eqn. (2) yields the guess for $\hat{F}_\alpha^{(0)}$ in line 8.

Prior hyperparameters. We also set $\boldsymbol{\Gamma}^{(0)}$ based on $\hat{\boldsymbol{\pi}}^{(0)}$

$$\boldsymbol{\Gamma}^{(0)} = \eta \begin{bmatrix} \hat{\boldsymbol{\pi}}^{(0)} \\ 1 - \hat{\boldsymbol{\pi}}^{(0)} \end{bmatrix} \; .$$

Here $\eta > 0$ is an adjustable parameter that controls the strength of the prior. For ease of presentation, this step is included in Algorithm 3 (line 1) rather than Algorithm 2.

4.4 Bringing everything together

Having introduced all of the components of OASIS, we are now ready to explain how they fit together. Recall that the evaluation process begins with three main inputs: the pool of record pairs P, similarity scores $s(\cdot)$, and predicted ER $\hat{\mathcal{R}}$. A summary of the main steps involved is as follows:

(i) Generate a set of strata $P_1, \ldots, P_K$ partitioning P using the CSF method (Algorithm 1).

(ii) Generate initial estimates using the strata, $\hat{\mathcal{R}}$ and $s(\cdot)$ (Algorithm 2).

(iii) Conduct AIS to estimate F_α (Algorithm 3).

Algorithm 3 OASIS for estimation of the F-measure

```
Input:  $T > 0$                        number of iterations
        $0 \leq \alpha \leq 1$          F-measure weight
        $0 < \varepsilon \leq 1$        greediness parameter
        $\eta > 0$                     prior strength parameter
        $\hat{F}_\alpha^{(0)}$         initial guess for F-measure
        $\hat{\boldsymbol{\pi}}^{(0)}$ initial guess for pos. probabilities
        $\hat{\mathcal{R}}$            predicted ER
        $\{P_k\}_{k=1}^K$              stratum allocations
        Oracle                         randomised (noisy) true labels
Output: $\hat{F}_\alpha^{(T)}$         F-measure estimate

1:  $\boldsymbol{\Gamma} \leftarrow \eta \begin{bmatrix} \hat{\boldsymbol{\pi}}^{(0)} \\ 1 - \hat{\boldsymbol{\pi}}^{(0)} \end{bmatrix}$        ▷ initialise Bayesian model
2:  for $t \in \{1, \ldots, T\}$ do
3:      Calculate $\boldsymbol{v}^{(t)}$ using Eqn. (12)
4:      Draw $k^\star$ from $\{1, \ldots, K\}$ according to $\boldsymbol{v}^{(t)}$
5:      Draw $z^\star$ from $P_{k^\star}$ uniformly
6:      $w_t \leftarrow \omega_k / v_k^{(t)}$                        ▷ importance weight
7:      $\ell_t \leftarrow$ Oracle($z^\star$)                        ▷ query label from oracle
8:      $\hat{\ell}_t \leftarrow \hat{\ell}(z^\star)$                ▷ record prediction
9:      $\boldsymbol{\Gamma}_{:,k^\star} \leftarrow \boldsymbol{\Gamma}_{:,k^\star} + \begin{bmatrix} \ell_t \\ 1 - \ell_t \end{bmatrix}$        ▷ update posterior
10:     $\hat{\boldsymbol{\pi}}^{(t)} \leftarrow \boldsymbol{\Gamma}_{0,:} \;./\; (\boldsymbol{\Gamma}_{0,:} + \boldsymbol{\Gamma}_{1,:})$        ▷ update $\boldsymbol{\pi}$ estimate
11:     $\hat{F}_\alpha^{(t)} \leftarrow \frac{\sum_{\tau=0}^t w_\tau \ell_\tau \hat{\ell}_\tau}{\alpha \sum_{\tau=0}^t w_\tau \hat{\ell}_\tau + (1-\alpha) \sum_{\tau=0}^t w_\tau \ell_\tau}$
12: end for
13: return $\hat{F}_\alpha^{(T)}$
```

Summary of Algorithm 3. At each iteration t: sample a stratum according to $\boldsymbol{v}^{(t)}$, then a record pair within that stratum uniformly at random. Query Oracle for a label of the record pair. Use the observed label (and the predicted label) to update the oracle probabilities (using Eqn. 10) and the F-measure estimate (using Eqn. 3). Stop after T iterations and return the final estimate $\hat{F}_\alpha^{(T)}$.

5. CONSISTENCY OF OASIS

A fundamental requirement of any well-behaved estimation procedure is *consistency*, that is, given enough samples we want the estimate to be close to the true value with high probability. Nominated as one of our objectives in designing the OASIS algorithm in Section 3, we now prove that OASIS is statistically consistent.

Before we begin, we acknowledge previous theoretical work on the consistency of other AIS algorithms, notably Population Monte Carlo (PMC) [5, 13, 4] and Adaptive Multiple Importance Sampling (AMIS) [11, 19]. Unfortunately, we cannot directly apply these results here owing to the following differences in our setup:

(i) we do not discard and re-draw the entire sample at each iteration since it would waste our label budget;

(ii) we permit the instrumental distribution to be updated based on samples from *all* previous iterations (unlike [13, 4] which are restricted to the previous iterate);

(iii) we examine consistency as $T \to \infty$ (others assume that the sample size increases at each iteration and examine consistency in this limit).

Due to the dependent nature of the sample and the non-linear form of the F-measure, the proof is relatively involved and requires some build-up. In Section 5.1, we first consider simple AIS estimators based on sample averages, and show that strong consistency follows so long as some reasonable conditions are met. Then in Section 5.2 we extend these results to the non-linear F-measure estimator. Until this point, we assume a general instrumental distribution and updating mechanism, before finally specialising to the OASIS method in Section 5.3.

5.1 Simple AIS estimators

Consider a random variable X with probability density $p(x)$ and consider the estimation of parameter $\theta = \mathrm{E}[f(X)]$ using AIS. This involves constructing sample $\{x_1, x_2, \ldots, x_T\}$ by drawing each item sequentially from a separate instrumental distribution. Specifically, we assume that the t-th sample x_t is drawn from an instrumental distribution with density $q_t(x_t|\boldsymbol{x}_{1:t-1})$ which depends on the $t-1$ previously sampled items $\boldsymbol{x}_{1:t-1} = \{x_1, \ldots, x_{t-1}\}$.[4] The AIS estimator of θ is then defined as:

$$\hat{\theta}^{\mathrm{AIS}} = \frac{1}{T} \sum_{t=1}^{T} w_t f(x_t), \tag{13}$$

which may be interpreted as an importance-weighted sample average. Here the importance weights are given by $w_t = p(x_t)/q_t(x_t)$ (we omit the conditioning on $\boldsymbol{x}_{1:t-1}$ for notational simplicity).

In order to prove that $\hat{\theta}^{\mathrm{AIS}}$ is consistent for θ, we rely on the following lemma, which generalises the law of large numbers (LLN) to history-dependent random sequences.

Lemma 1. *Let $\{U_t\}_{t=1}^{\infty}$ be a sequence of random variables and let $\boldsymbol{U}_{1:T} = \{U_1, U_2, \ldots, U_T\}$ denote the sequence up to index $t = T$. Suppose that the following conditions hold:*

(i) $\mathrm{E}[U_1] = \theta$*;*

(ii) $\mathrm{E}[U_t|\boldsymbol{U}_{1:t-1}] = \theta$ *for all $t > 1$; and*

(iii) $\mathrm{E}[U_t^2] \leq C < \infty$ *for all $t \geq 1$.*

Then $\frac{1}{T}\sum_{t=1}^{T} U_t \to \theta$ almost surely.

The proof of this lemma is given in the full report [18], and relies on a more general theorem due to Petrov [23].

By observing that the summands in Eqn. (13) obey conditions (i) and (ii) of Lemma 1, we can establish the following theorem on the strong consistency of $\hat{\theta}^{\mathrm{AIS}}$.

Theorem 1. *The estimator in Eqn.* (13) *is* strongly consistent*, that is, $\hat{\theta}^{\mathrm{AIS}} \to \theta$ almost surely, provided the following conditions are met for all $t \geq 1$:*

(i) $q_t(x) > 0$ *whenever* $f(x)p(x) \neq 0$*, and*

(ii) $\underset{\substack{X_t \sim p \\ \boldsymbol{X}_{1:t-1} \sim g}}{\mathrm{E}} \left[\frac{p(X_t)}{q_t(X_t)} f(X_t)^2\right] \leq C < \infty.$

Proof. Let $U_t = \frac{p(X_t)}{q_t(X_t|\boldsymbol{X}_{1:t-1})} f(X_t)$ and $\theta = \mathrm{E}[f(X)]$. The almost sure convergence follows by checking the conditions of Lemma 1. For condition (ii) of the lemma, we find

$$\begin{aligned}
\mathrm{E}[U_t|\boldsymbol{U}_{1:t-1}] &= \mathrm{E}\left[\frac{p(X_t)}{q_t(X_t|\boldsymbol{X}_{1:t-1})} f(X_t) \middle| \boldsymbol{X}_{1:t-1}\right] \\
&= \int_{\mathcal{X}} \frac{p(x_t)}{q_t(x_t|\boldsymbol{x}_{1:t-1})} f(x_t) q_t(x_t|\boldsymbol{x}_{1:t-1})\, dx_t \\
&= \int_{\mathcal{X}} f(x)p(x)\, dx \qquad\qquad \text{(by condition (i))} \\
&= \mathrm{E}[f(X)] = \theta.
\end{aligned}$$

Condition (i) of the lemma follows by a similar argument.

Finally we check condition (iii): that the second moment is bounded. Denoting the joint density of $\boldsymbol{X}_{1:t-1}$ by g and considering $t > 1$, we have

$$\begin{aligned}
&\mathrm{E}\big[U_t^2\big] \\
&= \mathrm{E}\big[\mathrm{E}[U_t^2|\boldsymbol{U}_{1:t-1}]\big] \\
&= \mathrm{E}\left[\mathrm{E}\left[\left(\frac{p(X_t)}{q_t(X_t|\boldsymbol{X}_{1:t-1})} f(X_t)\right)^2 \middle| \boldsymbol{X}_{1:t-1}\right]\right] \\
&= \iint_{\mathcal{X}} \left(\frac{p(x_t)f(x_t)}{q_t(x_t|\boldsymbol{x}_{1:t-1})}\right)^2 q_t(x_t|\boldsymbol{x}_{1:t-1}) dx_t g(\boldsymbol{x}_{1:t-1}) d\boldsymbol{x}_{1:t-1} \\
&= \iint_{\mathcal{X}} \frac{p(x_t)f(x_t)^2}{q_t(x_t|\boldsymbol{x}_{1:t-1})} p(x_t)\, dx_t\ g(\boldsymbol{x}_{1:t-1}) d\boldsymbol{x}_{1:t-1} \qquad \text{(by (i))} \\
&= \underset{\substack{X_t \sim p \\ \boldsymbol{X}_{1:t-1} \sim g}}{\mathrm{E}} \left[\frac{p(X_t)}{q_t(X_t|\boldsymbol{X}_{1:t-1})} f(X_t)^2\right]
\end{aligned}$$

which is bounded above by assumption. This also holds for $t = 1$ (by the above argument without the sampling history). Thus all of the conditions of Lemma 1 are satisfied, and the proof is complete. □

5.2 The AIS F-measure estimator

The AIS estimator for the F-measure, $\hat{F}_{\alpha}^{\mathrm{AIS}}$, is less straightforward to analyse because it cannot be expressed as a sample average like the estimators studied in Section 5.1. Instead, we regard $\hat{F}_{\alpha}^{\mathrm{AIS}}$ as a *ratio* of sample averages:

$$\hat{F}_{\alpha}^{\mathrm{AIS}} = \frac{\frac{1}{T}\sum_{t=1}^{T} w_t f_{\mathrm{num}}(x_t)}{\frac{1}{T}\sum_{t=1}^{T} w_t f_{\mathrm{den}}(x_t)},$$

(*cf.* Eqn. 3) where $x_t = (z_t, \ell_t)$ denotes a record pair and its observed label, and the functions are

$$\begin{aligned}
f_{\mathrm{num}}(x_t) &= \ell_t \hat{\ell}_t \ ; \quad \text{and} \\
f_{\mathrm{den}}(x_t) &= \alpha\hat{\ell}_t + (1-\alpha)\ell_t \ .
\end{aligned} \tag{14}$$

We leverage Theorem 1 to show that the numerator and denominator both converge to their respective true values, which is sufficient to establish convergence of $\hat{F}_{\alpha}^{\mathrm{AIS}}$.

Theorem 2. *Let $X = (Z, L)$ denote a random record pair Z and its corresponding label L, and let the density of X be $p(x) = p(\ell|z)p(z)$. Suppose AIS is carried out to estimate the F-measure and assume that the conditions of Theorem 1 are satisfied by $p(x)$ and $q_t(x)$ for both functions defined in Eqn.* (14)*. Assume furthermore that the instrumental density can be factorised as $q_t(x_t|\boldsymbol{x}_{1:t-1}) = p(\ell_t|\boldsymbol{z}_t)q_t(z_t|\boldsymbol{x}_{1:t-1})$ for all $t \geq 1$. Then $\hat{F}_{\alpha}^{\mathrm{AIS}}$ is* weakly consistent *for F_{α}.*

[4] Beginning with an initial sampling distribution $q_1(\boldsymbol{x}_1)$.

Table 1: Datasets in decreasing order of class imbalance. The size of the dataset is the number of record pairs it contains and the imbalance ratio is the ratio of non-matches to matches. The $\star$ indicates that the dataset is not from the ER domain.

Dataset Name	Size	Imb. Ratio	No. Matches
`Amazon-GoogleProducts`	4,397,038	3381	1300
`restaurant`	745,632	3328	224
`DBLP-ACM`	5,998,880	2697	2224
`Abt-Buy`	1,180,452	1075	1097
`cora`	1,675,730	47.76	34,368
$\star$ `tweets100k`	100,000	1	50,000

Proof. Observe that for the numerator of $\hat{F}_\alpha^{\mathrm{AIS}}$,

$$\frac{1}{T}\sum_{t=1}^{T}\frac{p(Z_t)}{q_t(Z_t)}f_{\mathrm{num}}(X_t) = \frac{1}{T}\sum_{t=1}^{T}\frac{p(X_t)}{q_t(X_t)}f_{\mathrm{num}}(X_t)$$

using the factorised form of $q_t(x)$. This converges in probability to $\mathrm{E}[f_{\mathrm{num}}(X)]$ by Theorem 1. The same is true for the denominator (replace f_{num} by f_{den}). Invoking Slutsky's theorem, we have

$$\hat{F}_\alpha^{\mathrm{AIS}} = \frac{\frac{1}{T}\sum_{t=1}^{T}\frac{p(Z_t)}{q_t(Z_t)}f_{\mathrm{num}}(X_t)}{\frac{1}{T}\sum_{t=1}^{T}\frac{p(Z_t)}{q_t(Z_t)}f_{\mathrm{den}}(X_t)} \xrightarrow{P} \frac{\mathrm{E}[f_{\mathrm{num}}(X)]}{\mathrm{E}[f_{\mathrm{den}}(X)]}$$

It is straightforward to show that the expression on the right-hand side reduces to F_α by evaluating the expectations with respect to p for finite pool P. For the more general case, it can be shown that the F-measure statistics $F_{\alpha,T}$ converge to the right-hand side population-based F-measure [25]. □

5.3 Application to OASIS

Theorem 2 tells us about the convergence of $\hat{F}_\alpha^{\mathrm{AIS}}$ for *any choice of instrumental distribution and update mechanism meeting the conditions.* Our final remaining task is to show that these conditions are met by Algorithm 3.

Theorem 3. *Algorithm 3 (OASIS) produces a consistent estimate of F_α, that is $\hat{F}_\alpha^{(T)} \xrightarrow{P} F_\alpha$.*

The proof is straightforward, while lengthy, and so is relegated to the full report [18]. It proceeds by checking that the conditions of Theorem 2 are satisfied by the OASIS instrumental distribution.

Remark 5. *It is now apparent why we adopt the ε-greedy instrumental distribution: while $q_t^\star(z)$ can go to zero when $p(z) \neq 0$, violating condition (i) of Theorem 1, ε-greedy cannot. For example, if $\hat{\pi}_k = 0$ and $\lambda_k = 0$ then $q_t^\star(z) = 0$ for all $z \in P_k$, whilst $p(z) = 1/N \neq 0$. The ε-greedy instrumental distribution does not vanish since $q_t(z) = \varepsilon/N > 0$.*

6. EXPERIMENTS

In this section, we examine whether OASIS addresses our main objective of reducing labelling requirements for evaluating ER. We run comprehensive experiments comparing OASIS with established methods, which conclusively establish that OASIS is generally superior, requiring significantly fewer labels to achieve a given precision of estimate.

6.1 Experimental setup

6.1.1 Datasets

We use five publicly available ER datasets as listed in Table 1. All datasets come with true resolution $\mathcal{R}$. `Abt-Buy` [17] and `Amazon-GoogleProducts` [17] are from the e-commerce domain; `cora` [1] and `DBLP-ACM` [17] relate to computer science citations; and `restaurant` contains listings from two restaurant guidebooks [1]. We note that `cora` is unique among these datasets, in that it does not arise from two separate DBs. Technically, it is an example of *de-duplication*, which may be cast as ER on the DB matched with itself.

In addition to these five datasets, we have also included `tweets100k` [21] from outside the ER domain. It is included to test whether the sampling methods are competitive in the *absence* of class imbalance.

Pooling. Although evaluation is ideally conducted with respect to the entire pool, $P = \mathcal{Z}$, a key baseline sampling method (IS, introduced in Section 6.2) is prohibitively slow for such large pools (*cf.* Section 6.3.5) since its instrumental distribution is defined on each record pair. OASIS does not suffer from this drawback and runs efficiently on entire pools. However to complete a fair comparison, we opt to conduct the evaluation with respect to smaller pools drawn randomly from $\mathcal{Z}$, which are listed in Table 2. This does not affect the validity of the theory/algorithm; indeed relative to (significant) randomised pools, F_α is with high probability exceedingly close to that defined relative to $\mathcal{Z}$.

Oracle. We implement an oracle based on the ground truth resolution $\mathcal{R}$ provided per dataset. Since only one label is provided per record pair, we are in the regime of a deterministic `Oracle` *i.e.,* with probabilities $p(1|z) \in \{0,1\}$.

6.1.2 ER pipeline

We build a simple ER pipeline with the following features:

Pre-processing. Strings are normalised by removing symbols, accents & capitalisation. Numeric fields are converted to floats and missing values are imputed using the mean.

Similarity features. For each pair of fields (*e.g.,* the 'Name' fields of $\mathcal{D}_1$ and $\mathcal{D}_2$) we calculate a scalar feature based on some measure of their similarity. For short textual fields we the Jaccard distance based on trigrams and for long textual fields we use cosine similarity with a tf-idf vector representation. For numeric fields we use the normalised absolute difference.

Record pair classifier. At the core of the ER pipeline is a binary classifier, which operates on the space of similarity features. We generally use a linear SVM (L-SVM), trained on a random subset of the entire dataset (including ground truth labels). Since we would like to test the evaluation in a range of circumstances, we don't always aim for the best classifier—we instead aim for a range of classifiers with excellent performance through to poor.

6.2 Baseline methods

We compare OASIS with three baseline methods.

Passive. This simple method samples record pairs uniformly at random from the pool with replacement. At each iteration, the F-measure is estimated using Eqn. (1), based only on the record pairs/labels sampled so far.

Stratified. This method has been used previously in [14] for estimating balanced F-measures. It involves partitioning the pool of record pairs into strata (we set $K = 30$) using Algorithm 1. Record pairs are then sampled by drawing a stratum according to the stratum weights ($\omega_k = |P_k|/N$), then sampling within the stratum uniformly. The F-measure is estimated using a stratified version of Eqn. (2) (see [14]).

IS. Non-adaptive importance sampling has been used for evaluating F-measures in [25]: record pairs are sampled according to a *static* instrumental distribution which aims to approximate Eqn. (5). IS may be far from optimal depending on score reliability, since the approximation replaces $p(1|z)$ with the similarity scores (mapped to the unit interval). The estimate of the F-measure is obtained at each iteration using a static version of Eqn. (3).

6.3 Results

Since each estimation method is randomised, we study their behaviour statistically. For each pool in Table 2, we run each estimation method 1000 times, recording the history of estimates for each run in a vector: $[\hat{F}_\alpha^{(t)}]_{t=1:T}$. In all of the experiments, we set $\alpha = 1/2$, $\eta = 2K$ and $\varepsilon = 10^{-3}$.

6.3.1 Label budget savings

To compare the labelling requirements of the different estimation methods, we plot the expected absolute error $\mathrm{E}[|\hat{F}_\alpha - F_\alpha|]$ (abbreviated as abs. err.) as a function of the label budget.[5] To compute abs. err. we average over 1000 repeats for fixed P. The true F-measure, F_α, is calculated on P using Eqn. (1), assuming all labels are known immediately. The results are presented in Figure 2 for each pool in Table 2. Below the abs. err. plot, we have also plotted the standard deviation of the estimate, which is useful for checking whether the variance reduction methods (IS and OASIS) are operating as designed.

Winning method. OASIS beats the other methods, significantly improving on the state-of-the-art, both in terms of the abs. err. and the variance, on all of the ER datasets except `cora` where it is competitive. The reason for the anomalous behaviour on `cora` is likely due to the fact that the class imbalance is far less pronounced.

Inadequacy of passive sampling. The experiments confirm our claim that passive sampling is a poor choice for evaluating ER. Compared to IS and OASIS, passive sampling demonstrates significantly slower convergence, and is less reliable due to the high variance. In fact, passive sampling often cannot produce any estimate at all until a significant label budget has been consumed (*cf. e.g.,* `DBLP-ACM`). This is because the F-measure remains undefined until a match (or predicted match) is sampled for the first time. We only begin plotting the curve when the estimate has a probability exceeding 95% of being well-defined.

Stratified method. This method does not fare much better than passive sampling, casting doubt on its effectiveness for efficient evaluation as proposed in [14]. We expect that the reason for the poor performance is due to the fact that the sampling is not biased (merely proportional to ω_k).

Balanced classes. For the case of more balanced classes, as in `tweets100k`, and to a lesser extent `cora`, there is effectively no difference between the methods. This implies that the advantage of IS and OASIS over the other methods diminishes as the imbalance ratio decreases. It is important to note however, that the balanced regime is of little relevance to ER—we merely include it for completeness.

6.3.2 Calibrated vs. uncalibrated scores

In the experiments thus far (in Figure 2), we have been evaluating ER pipelines based on linear SVMs. The similarity scores from such systems are distances from the decision hyperplane, which are not intended to approximate the oracle probabilities $p(1|z)$ accurately (they are “uncalibrated” *cf.* Definition 3). As such, we expect the performance of IS to be less favourable, because the instrumental distribution will be further from optimality if $s_i \approx p(1|z_i)$ is not satisfied. Much less degradation is expected under OASIS.

In order to assess whether this has an appreciable effect, we compared running IS and OASIS with calibrated versus uncalibrated similarity scores. The calibrated (probabilistic) scores are obtained using a built-in costly feature of LIBSVM, which runs five-fold cross-validation at training time [7]. The uncalibrated scores are distances from the decision hyperplane used previously. The results in Figure 3 show that the calibrated scores yield significantly better performance, particularly for IS. However, the difference is less pronounced for OASIS, which does a good job of learning the true oracle probabilities from the incoming labels.

6.3.3 Convergence of the model parameters

We have observed excellent convergence properties for OASIS in terms of the F-measure estimate. An interesting supplementary question is whether the estimates of the oracle probabilities (and in turn the instrumental distribution) also converge rapidly to their true (optimal) values. Although we have not studied this question theoretically, we have observed convergence in a limited number of experiments with `Abt-Buy`. An example is depicted in Figure 4. Heatmap plot (b) demonstrates that the estimates of the oracle probabilities for this run converge quite rapidly: after $\sim$ 4000 labels are consumed. However, the instrumental distribution takes longer to converge, because it is very sensitive to slight errors in the estimates. It does not reach optimality until after $\sim$ 8500 labels are consumed. This is easiest to see in the KL divergence plot (d), where a value of zero indicates convergence.

6.3.4 Effectiveness for different classifiers

Although we have focussed on evaluating ER based on linear SVMs so far, there is essentially no limitation on the types of classifiers that can be evaluated, so long as they produce some kind of similarity scores. To this end, we re-run our experiments on the `Abt-Buy` pool using four additional types of classifiers: a neural network (multi-layer perceptron) with one hidden layer (NN), a boosted decision tree AdaBoost (AB), logistic regression (LR), and SVM with a RBF kernel (R-SVM). We implement the classifiers using scikit-learn with the default parameter options.

The expected estimation error for each method (Passive, Stratified, IS and OASIS) is evaluated after 5000 labels are

[5] Note that the label budget is not equivalent to the number of iterations. Since we are sampling with replacement, the same record pair may be drawn at multiple iterations, however it only counts towards the label budget the first (and only) time its label is queried from the oracle.

Table 2: Pools sampled from the datasets in Table 1, along with the true performance measures.

Associated Dataset	Size	Imb. ratio	No. matches	Classifier	Precision	Recall	$F_{1/2}$
`Amazon-GoogleProducts`	676,267	3381	200	L-SVM	0.597	0.185	0.282
`restaurant`	149,747	3328	45	L-SVM	0.909	0.888	0.899
`DBLP-ACM`	53,946	2697	20	L-SVM	1.0	0.9	0.947
`Abt-Buy`	53,753	1075	50	L-SVM	0.916	0.44	0.595
`cora`	328,291	47.76	6874	L-SVM	0.841	0.837	0.839
⋆ `tweets100k`	20,000	0.9903	10049	L-SVM	0.762	0.778	0.770

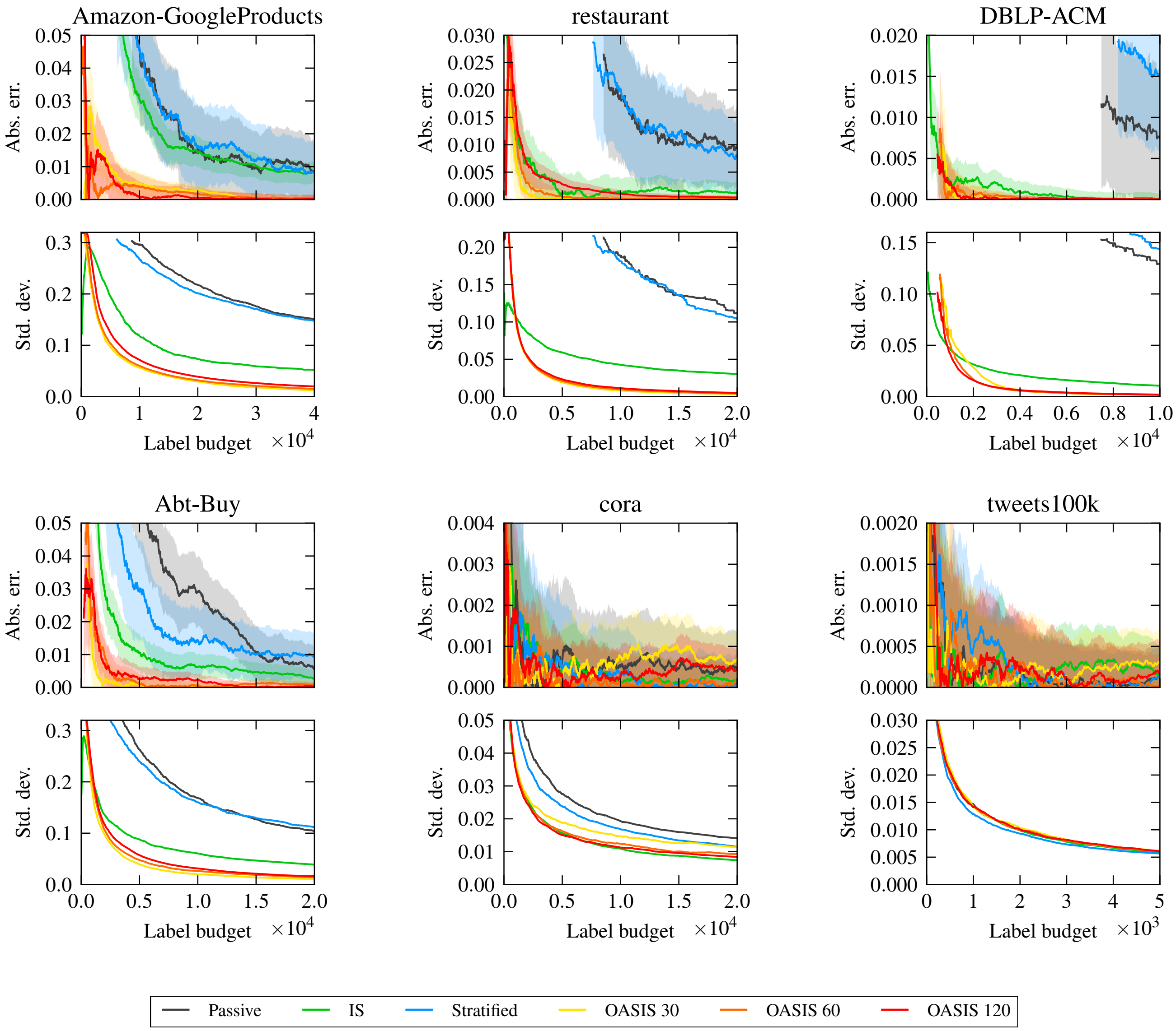

Figure 2: Plots showing the expected absolute error (abs. err.) and the standard deviation (std. dev.) of $\hat{F}_{1/2}$ for the different estimation methods (Passive, Stratified, IS, OASIS) as a function of label budget. The OASIS method is run with $K = 30, 60$ and 120 (except on `tweets100k` where $K = 10, 20$ and 40). This figure is best viewed in colour.

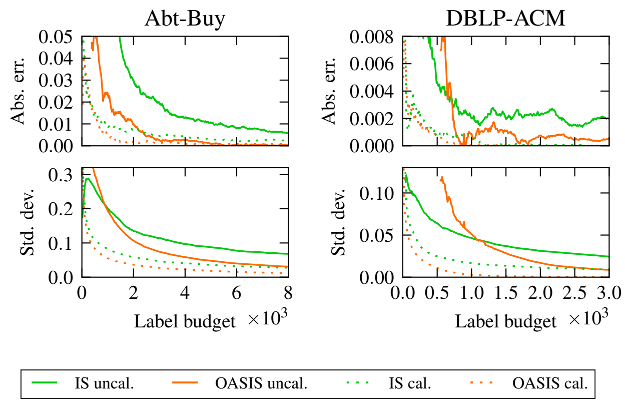

Figure 3: Comparison of calibrated vs. uncalibrated scores for IS & OASIS (run with $K = 60$).

Table 3: CPU times for the cora experiment.

Sampling method	Avg. CPU time per run (s)	Avg. CPU time per iteration (s)
Passive	0.512	2.483×10^{-5}
IS	69.854	3.149×10^{-3}
OASIS 30	3.612	1.228×10^{-4}
OASIS 60	3.281	1.123×10^{-4}
OASIS 120	2.978	1.093×10^{-4}
Stratified	1.967	9.502×10^{-5}

consumed and the results are plotted in Figure 5. We see that OASIS generally outperforms the other methods, yielding an estimate of $F_{1/2}$ which is one order of magnitude more precise than IS.

6.3.5 Runtime

We present evidence that the IS method scales poorly to large pools in Table 3, which lists the average CPU times for experiments on the `cora` dataset (pool size $N \sim 10^5$). The experiments were run on an HP EliteBook 840 G2 with 2.6GHz Core i7 and 16GB RAM. Note that the times listed for the OASIS and Stratified methods exclude precomputation of the strata, which takes less than 0.1 s. Looking at the results, we see that IS is an order of magnitude slower than OASIS—in fact, the timing for IS appears to scale linearly in N based on other timing data (not shown due to space constraints). The reason for this, is that IS samples from a non-uniform distribution over the entire pool (a computation linear in size N), whilst OASIS samples from a smaller non-uniform distribution over the strata (of size K). It appears that the extra operations OASIS requires to update the model are negligible in comparison.

7. RELATED WORK

Efficient evaluation. Previous work has considered efficient evaluation for general classifiers, through approaches such as importance sampling [25], stratified sampling [3, 14] and semi-supervised inference of Bayesian generative models [27]. However, none of this work accounts for the specific features of ER evaluation, namely extreme class imbalance, and the availability of auxiliary information in the form of similarity scores.

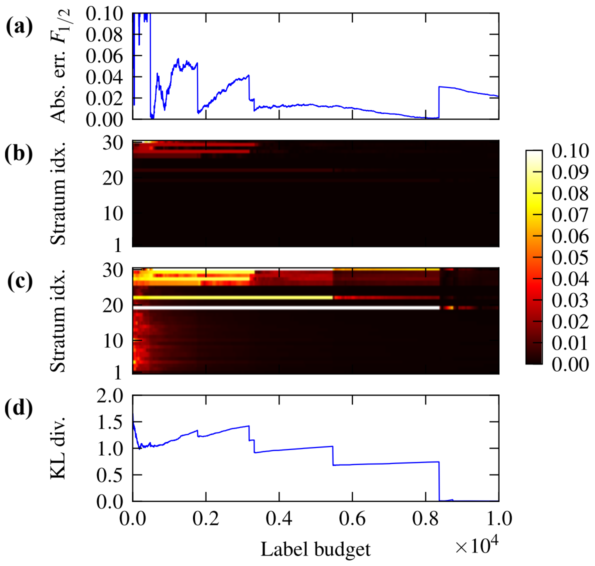

Figure 4: Convergence of the F-measure, oracle probabilities and instrumental distribution for a run of OASIS on the Abt-Buy SVM dataset (with calibrated scores and $K = 30$): (a) absolute error in $F_{1/2}$; (b) absolute error in π; (c) absolute error in $v^\star$; (d) KL divergence from $v^\star$ to the estimate $v^{\star(t)}$.

Bennett & Carvalho [3] outline an adaptive method for estimating precision that stratifies on classifier scores, sampling points with probability proportional to the stratum population and a dynamic estimate of the variance in the labels. However, their method does not incorporate recall and is not proven to be optimal. Druck & McCallum [14] extend the work of [3] to facilitate estimation of vector-valued and non-linear functions (including token-based accuracy and F-measure). Both of these approaches are adaptive and biased, although they rely purely on stratified sampling, which is known to be less effective at variance minimisation than importance sampling [24]. We also note an exception in [14]: the method proposed specifically for estimating F-measure is based on proportional stratified sampling, which is neither adaptive nor biased.

Welinder *et al.* [27] propose an estimation procedure for precision-recall curves, based on a Bayesian generative model. Their method is semi-supervised and makes use of the classifier scores, but it doesn't incorporate biased sampling or adaptivity, making it unsuited to problems with class imbalance. It also imposes a restrictive assumption on the joint distribution of scores and labels, requiring the user to guess an appropriate parametric distribution. Another non-adaptive approach is the IS method of Sawade *et al.* [25]. It facilitates the estimation of F-measures, relying on the asymptotically optimal distribution of equation (5). The authors address the instrumental distribution's dependence on unknown quantities by estimating them using classifier scores. However if the scores are inaccurate or merely uncalibrated, the method will be sub-optimal as it does not actively adapt using incoming labels.

Adaptive importance sampling (AIS). A broad literature covers AIS, however to our knowledge, no prior work

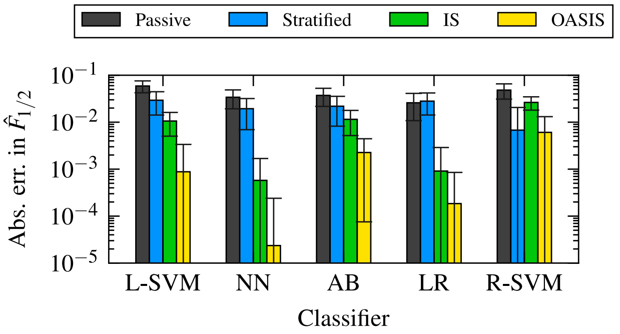

Figure 5: Expected absolute error in $\hat{F}_{1/2}$ for five classifiers trained on the Abt-Buy dataset. The error is measured after 5000 labels are consumed by each method (Passive, Stratified, IS, OASIS). The error bars are approx. 95% confidence intervals.

specialises these techniques to evaluation. A significant drawback of previous AIS algorithms, is that they discard and resample at each iteration, which is prohibitively wasteful when performing efficient evaluation. One of the earliest AIS algorithms is Population Monte Carlo (PMC), which maintains an entire *population* of instrumental distributions, updating them using propagation and resampling steps [5]. Standard formulations of PMC use only the previous sample when updating distributions, reducing statistical efficiency. Previous proofs of consistency also assume that the population grows to an infinite size [13, 4]. A more recent AIS algorithm is Adaptive Multiple Importance Sampling (AMIS) which is "aimed at an optimal recycling of past simulations in an iterated importance sampling (IS) scheme" [11]. Unlike PMC, AMIS makes use of the entire history of samples and instrumental distributions, to update the importance weights and instrumental distribution. However, it is not applicable in the efficient evaluation context because it requires an increasing sample to be drawn at each iteration, which would consume realistic label budgets too quickly.

8. CONCLUSIONS

We have proposed a novel adaptive importance sampler OASIS for estimating the F-measure of ER pipelines. We leverage ER similarity scores through a stratified Bayesian generative model, to update an instrumental sampling distribution that optimises asymptotic variance. Statistical consistency establishes correctness of OASIS, while extensive experimentation demonstrates significant reduction to label budget relative to existing approaches.

9. ACKNOWLEDGEMENTS

N. Marchant acknowledges the support of an Australian Government Research Training Program Scholarship. B. Rubinstein acknowledges the support of the Australian Research Council (DP150103710).

10. REFERENCES

[1] Duplicate Detection, Record Linkage, and Identity Uncertainty: Datasets. `http://www.cs.utexas.edu/users/ml/riddle/data.html`. Accessed: Dec 2016.

[2] M. Barnes. A Practioner's Guide to Evaluating Entity Resolution Results. *arXiv:1509.04238 [cs, stat]*, 2015.

[3] P. N. Bennett and V. R. Carvalho. Online Stratified Sampling: Evaluating Classifiers at Web-scale. In *CIKM*, pages 1581–1584, 2010.

[4] O. Cappé, R. Douc, A. Guillin, J.-M. Marin, and C. P. Robert. Adaptive importance sampling in general mixture classes. *Stat. Comput.*, 18(4):447–459, 2008.

[5] O. Cappé, A. Guillin, J. M. Marin, and C. P. Robert. Population Monte Carlo. *J. Comp. Graph. Stat.*, 13(4):907–929, 2004.

[6] N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

[7] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM TIST*, 2(3):27, 2011.

[8] P. Christen. *Data Matching*. Data-Centric Systems and Applications. Springer Berlin Heidelberg, 2012.

[9] P. Christen and K. Goiser. Quality and complexity measures for data linkage and deduplication. In *Quality Measures in Data Mining*, pages 127–151. Springer, 2007.

[10] W. G. Cochran. *Sampling Techniques*. Wiley, 1977. 3rd ed.

[11] J.-M. Cornuet, J.-M. Marin, A. Mira, and C. P. Robert. Adaptive Multiple Importance Sampling. *Scand. J. Stat.*, 39(4):798–812, 2012.

[12] T. Dalenius and J. L. Hodges. Minimum Variance Stratification. *JASA*, 54(285):88–101, 1959.

[13] R. Douc, A. Guillin, J.-M. Marin, and C. P. Robert. Convergence of Adaptive Mixtures of Importance Sampling Schemes. *Ann. Stat.*, 35(1):420–448, 2007.

[14] G. Druck and A. McCallum. Toward Interactive Training and Evaluation. In *CIKM*, pages 947–956, 2011.

[15] L. Getoor and A. Machanavajjhala. Entity resolution for big data. In *KDD*, pages 1527–1527, 2013.

[16] K. Harron, A. Wade, R. Gilbert, B. Muller-Pebody, and H. Goldstein. Evaluating bias due to data linkage error in electronic healthcare records. *BMC Med. Res. Methodol.*, 14(1):36, 2014.

[17] H. Köpcke, A. Thor, and E. Rahm. Evaluation of Entity Resolution Approaches on Real-world Match Problems. *PVLDB*, 3(1):484–493, 2010.

[18] N. G. Marchant and B. I. P. Rubinstein. In Search of an Entity Resolution OASIS: Optimal Asymptotic Sequential Importance Sampling. *arXiv:1703.00617 [cs, stat]*, 2017.

[19] J.-M. Marin, P. Pudlo, and M. Sedki. Consistency of the Adaptive Multiple Importance Sampling. *arXiv:1211.2548 [math, stat]*, 2012.

[20] D. Menestrina, S. E. Whang, and H. Garcia-Molina. Evaluating Entity Resolution Results. *PVLDB*, 3(1):208–219, 2010.

[21] B. Mozafari, P. Sarkar, M. Franklin, M. Jordan, and S. Madden. Scaling Up Crowd-sourcing to Very Large Datasets: A Case for Active Learning. *PVLDB*, 8(2):125–136, 2014.

[22] S. N. Negahban, B. I. P. Rubinstein, and J. Gemmell. Scaling multiple-source entity resolution using statistically efficient transfer learning. In *CIKM*, pages 2224–2228, 2012.

[23] V. V. Petrov. On the Strong Law of Large Numbers for a Sequence of Dependent Random Variables. *J. Math. Sci.*, 199(2):225–227, 2014.

[24] R. Y. Rubinstein and D. P. Kroese. *Simulation and the Monte Carlo Method*. Wiley, 2007.

[25] C. Sawade, N. Landwehr, and T. Scheffer. Active Estimation of F-measures. In *NIPS*, pages 2083–2091. 2010.

[26] V. S. Verykios, M. G. Elfeky, A. K. Elmagarmid, M. Cochinwala, and S. Dalal. On the accuracy and completeness of the record matching process. In *Proc. of the 2000 Conf. on Information Quality*, 2000.

[27] P. Welinder, M. Welling, and P. Perona. A Lazy Man's Approach to Benchmarking: Semisupervised Classifier Evaluation and Recalibration. In *CVPR*, pages 3262–3269, 2013.

[28] W. E. Winkler. Overview of record linkage and current research directions. Techreport, Bureau of the Census, 2006.