Potential of Privacy-Preserving Record Linkage for the Statistics of Hidden Populations

Olivier Binette and Andy Demma

Duke University

May 2021

Abstract

Many populations are "hidden" from the point of view of traditional probabilistic surveys. They are populations for which we have no meaningful sampling frame and whose members may be difficult to identify. Examples include victims of human trafficking and civilian casualties in armed conflicts. Understanding these populations is central to policymaking and to prosecuting human rights violations, yet statistical inference remains extremely difficult in practice. One main approach to inference consists of aggregating data from different sources through record linkage and adjusting for undercoverage using multiple systems estimation. In practice, however, organizations may not be willing to disclose private information to conduct such studies. We therefore analyze the potential of privacy-preserving record linkage (PPRL) to assist these studies, comparing the security and privacy guarantees provided by PPRL frameworks to the privacy needs of organizations in practical applications. Our analysis highlights many limitations of current PPRL frameworks and suggests next steps towards practical PPRL for the statistics of hidden populations.

1 Introduction

In the context of statistical research, a population is said to be "hidden" if it may not be reached using traditional survey methodology. Such a population has two main characteristics (Heckathorn, 1997; Spreen, 1992). First, there is no sampling frame — there is no "phone book" or general index which would allow the design of an efficient probabilistic

sampling scheme. The size of the hidden population is unknown and may be relatively small compared to the general population. Second, because of privacy concerns, particular vulnerabilities, or other barriers, members of this population may be difficult to identify or may not be willing to disclose information. Some examples include victims of human trafficking and modern slavery (United Nations, 2001, 2000; Datta and Bales, 2013; Cockayne, 2015; Bales et al., 2015; International Labour Organization, 2017a,b; Landman, 2020), civilian casualties in armed conflicts (Ball and Asher, 2002; Sadinle, 2014; Ball and Price, 2019), and people who inject drugs (Mathers et al., 2008).

There are two main categories of approaches to the statistical study of hidden populations: (1) link-tracing approaches, and (2) approaches based on administrative data sources or convenience samples. Link-tracing, which includes snowball sampling and respondent-driven sampling (Spreen, 1992), relies on an underlying social network in the target population. Members of the population are asked to identify others within their population, and a sample of the population is generated by navigating these relationships. Approaches based on administrative data, on the other hand, rely on existing but incomplete data sources. In the context of human trafficking, many organizations such as the police and non-governmental organizations come into contact with victims. Their records can be aggregated and cleaned through record linkage in order to obtain a more complete picture. There are still many challenges associated with such data, however, such as undercoverage and non-representation (Price and Ball, 2015; Hand, 2018). Multiple systems estimation (MSE) (Fienberg, 1972) and other statistical techniques are therefore used to adjust for undercoverage and bias, to the extent possible. In its simplest form, MSE is used in conjunction with record linkage to estimate the size of a population using data from multiple sources. This has been widely used in epidemiology (Wittes, 1974; Yip et al., 1995; Chao et al., 2001), and human rights statistics (Lum et al., 2013; Manrique-Vallier et al., 2013; Bales et al., 2015; Sadinle, 2018; Bird and King, 2018).

Here we consider the use of record linkage and multiple systems estimation (RL+MSE) to estimate the size of hidden populations. These methods are particularly suited to the problems of estimating casualties in armed conflicts and quantifying human trafficking in countries with strong law enforcement, as link tracing approaches may not be applicable in these cases. A central challenge in RL+MSE studies is to obtain the collaboration of organizations which possess data regarding individuals in the population of interest. In past studies, a trusted third party (or an overarching governmental agency) would perform record linkage and only release the minimal set of statistics necessary for population size estimation

through MSE. However, this poses a disclosure risk if the third party is compromised. Some organizations might also not be willing or legally allowed to share private data with a third party (Bales et al., 2019).

There is therefore a need for secure computation and privacy-preserving technologies to assist RL+MSE studies. This would reduce the risks associated with relying on third parties and could allow the broader implementation of this methodology. The field of privacy-preserving record linkage (Hall and Fienberg, 2010; Vatsalan et al., 2013; Schnell, 2015; Vatsalan et al., 2017) provides solutions to "private" record linkage which are suitable to different kinds of input data, required data releases, and privacy requirements. However, care has to be taken in understanding the security and privacy guarantees provided by PPRL approaches. The field of PPRL mostly focuses on *secure* computation of a certain record linkage output. However, we will see that current PPRL approaches usually do not satisfy strong security guarantees. Futhermore, PPRL approaches do not provide guarantees for the protection of individual privacy. The risks associated with disclosing computation outputs also need to be considered and mitigated (Hall and Fienberg, 2010; Schnell, 2015).

Our paper provides a first step towards the implementation of security and privacy technologies for RL+MSE studies. We review the field of PPRL from the point of view of individual privacy in view of the security and privacy needs of organizations participating in RL+MSE studies. Given the limitations of PPRL, we propose a reframing of the problem through what we call *secure entity resolution* and *collaborative data disclosure*. We define secure entity resolution as the problem of obtaining unique entity identifiers (uIDs) for records spread across organizations, while not revealing any sensitive information, and we propose two approaches to do so securely. Following the obtention of these uIDs, data releases can be orchestrated through secure multiparty data disclosure protocols which respect individual privacy.

The rest of the paper is organized as follows. Section 2 reviews the record linkage and multiple systems estimation problem of interest. We describe the data held by organizations, different record linkage approaches which can be used, the record linkage output needed for MSE, and we formalize the security and privacy need of organizations. In section 3, we provide a broad overview of PPRL. In section 4, we consider PPRL within the context of RL+MSE studies. We define *secure entity resolution* in section 4.1 and we consider possible implementations in section 4.2. Section 4.3 then discusses collaborative data disclosure approaches which can be considered following secure entity resolution. Section 5 concludes with a discussion of the main challenges we encountered and outlines the potential for future

work in the area.

2 Record Linkage and Multiple Systems Estimation

This section reviews the use of record linkage and multiple systems estimation in order to estimate the size of populations. We begin by reviewing MSE and past studies in section 2.1, explaining the logic of the methodology, its requirements, and its limitations. Next in section 2.2, we explain how data from multiple sources can be aggregated through record linkage. We provide an overview of different record linkage frameworks and explain how they relate to how multiple systems estimation is carried out. Finally, in section 2.3, we discuss security and privacy needs of organizations in reasonable practical applications. This provides the basic set of requirements against which PPRL approaches are evaluated in sections 3 and 4.

2.1 Multiple Systems Estimation Studies

Multiple systems estimation refers to a set of statistical tools used to estimate the size of populations using multiple sources of information. The underlying ideas were introduced as early as the 17th century by John Graunt, who estimated London's population at the time by comparing death records to an estimated death rate (Hald, 2005). Laplace (Laplace, 1820; Cochran, 1978) later formalized the approach and provided error bounds through asymptotic theory. These ideas led to capture-recapture methods in population ecology and were generalized by Fienberg (1972), leading to modern MSE. Today, MSE is used to evaluate the accuracy of the United States decennial census, to estimate disease prevalence in epidemiological applications, as well as more broadly in the fields of official statistics and human rights statistics (Bird and King, 2018).

The logic of MSE is best explained with only two sources of information. So, suppose that two organizations have records regarding individuals in a population of interest. Each organization's set of records defines a list of observed individuals. Let n_1 be size of the first organization's list, let n_2 be the size of the second organization's list, and let $n_{1,2}$ be the overlap between the two. A large overlap between the two lists, meaning that $n_{1,2}$ is relatively large, might indicate that the lists have a good coverage of the population of interest. On the other hand, a small overlap can indicate under-coverage. More precisely, the Lincoln-Petersen estimator (Lincoln, 1930; Petersen, 1895) of the population size is

defined as

$$\hat{N} = \frac{n_1 n_2}{n_{1,2}}.$$

If the lists as uncorrelated, meaning that an individual appearing on one list does not affect the chances of that individual appearing on the other list, then the Lincoln-Peterson estimator is approximately unbiased. If the lists are positively correlated, then $\hat{N}$ estimates an *upper bound* of the population size; and if the lists are negatively correlated, then $\hat{N}$ estimates a *lower bound* of the population size (Chao et al., 2008). Therefore, knowledge of the overlap pattern between lists, together with background knowledge of interactions between them, is informative of the population size.

This approach can be extended in two ways. First, we may consider the use of more lists. In this case, all overlap patterns between lists must be considered, and interactions between lists must be carefully modeled. That is, for any set of lists ω, let n_ω be the number of individuals which appeared in ω and in no other list. This provides the set of overlap counts for the data, where $n_\emptyset$ is missing. A model for the set of counts n_ω can then be specified and used to estimate $n_\emptyset$ through maximum likelihood or through other means (Fienberg, 1972). The population size is the estimated as $\hat{N} = n_{\text{obs}} + \hat{n}_\emptyset$, where n_{obs} is the total number of observed individuals. Second, if the target population is composed of heterogeneous groups, then we may account for this through stratification: MSE will be separately applied to different groups. This can help obtain more accurate estimates when heterogeneity induces interactions between lists that are difficult to model directly.

Example 2.1 (Quantifying Modern Slavery in the United Kingdom)**.** Modern slavery is an umbrella term which refers "to situations of exploitation that a person cannot refuse or leave because of threats, violence, coercion, deception, and/or abuse of power" (International Labour Organization, 2017a,b). This includes issues of forced labour, forced sexual exploitation, and forced marriage (International Labour Organization, 2017a), and the term "modern slavery" is commonly considered synonymous to human trafficking.

Silverman (2014); Bales et al. (2015) quantified modern slavery in the United Kingdom in 2013 using multiple systems estimation and data provided by the National Crime Agency Strategic Assessment. The data referred to 2,744 cases of trafficking and modern slavery having been observed by either non-governmental organizations (NG), governmental organizations (GO), the general public (GP), local authorities (LA), the police force (PF) or directly by the National Crime Agency (NCA). These six sources, each corresponding to a list of potential victims of modern slavery, were the basis of the MSE study. The National

Total	Cases observed once					Cases observed twice									3+ times			
2744	54	463	995	695	316	15	19	3	62	19	1	76	11	8	1	1	4	1
LA	×					×	×	×							×	×		×
NG		×				×			×	×	×				×	×	×	×
PFNCA			×				×		×			×	×		×		×	×
GO				×				×		×		×		×		×	×	×
GP					×						×		×	×				

Table 1: Overlap counts for the study of Silverman (2014); Bales et al. (2015) on modern slavery in the United Kingdom. Note that the PF and NCA lists have been merged together as in Silverman (2014). Each column represents the number of cases which appeared only in the lists corresponding to an "×" symbol underneath. For instance, 54 potential victims have been observed only on the LA list, and 15 cases have appeared both on the LA and NG lists.

Crime Agency linked the data in order to obtain the overlap counts n_ω for each non-empty subset of the lists NG, GO, GP, LA, PF and NCA lists. The data is represented in table 1.

Silverman (2014); Bales et al. (2015) modeled the data using a Poisson log-linear model with two-way interactions. That is, each count n_ω is modeled as

$$n_\omega \sim \text{Poisson}\left(\lambda_\omega\right), \quad \log \lambda_\omega = \mu + \sum_{i \in \omega} \alpha_i + \sum_{\{i,j\} \subset \omega} \beta_{i,j}. \tag{1}$$

Here μ is an intercept term, α_i represents main list effects, and the parameters $\beta_{i,j}$, $i \neq j$, are used to model pairwise list interactions. Crucially, this model assumes the absence of higher-order interactions between the lists.

The model parameters were estimated by maximum likelihood with two-way interaction terms selected through stepwise forward p-value thresholding. See Bales et al. (2015) for details and Chan et al. (2020) for methodological improvements.

Given an estimate $\hat{\mu}$ of μ and using the fact that $n_\emptyset \sim \text{Poisson}(\exp\{\mu\})$, we may then estimate $n_\emptyset$ as $\hat{n}_\emptyset = \exp\{\hat{\mu}\}$. In Silverman (2014); Bales et al. (2015), the confidence interval for $\hat{\mu}$ led to an estimate of between 10,000 to 13,000 potential victims of modern slavery in the UK in 2013.

2.1.1 Requirements of MSE Studies

The most basic requirement of MSE studies is a table of overlap counts, as in table 1. This table may be further disaggregated by socio-demographic variables and other information. In its most detailed form, the data consists of one row $x_i = (y_i, z_i)$ for each observed individual i, where y_i is a vector indicating the set of lists on which the individual appears and z_i is a vector of covariates. Obtaining this data requires to perform record linkage of the organizations' databases and to combine their information into a consistent set of covariates z_i. Processes to do so are discussed in section 2.2.

Furthermore, MSE requires sufficient knowledge about the organizations and their data collection process to specify a good model for the aggregated data. This model must be sufficiently informative to link the unknown count of unobserved individuals to estimable model parameters. Typically, MSE studies assume the absence of "highest-order interaction" between lists (Fienberg, 1972; Yip et al., 1995), as is the case with log-linear models (Fienberg, 1972) and decomposable graphical models (Darroch et al., 1980; Madigan and York, 1997). This assumption effectively expresses the expected value of $n_\emptyset$ as a function of other model parameters.

2.1.2 Limitations of MSE Studies

There are many challenges to successfully applying MSE in practice. In many cases, we have limited knowledge of the data collection practices of organizations or of plausible interactions between lists. It may therefore be difficult to justify the choice of a model or the choice of specific assumptions used to infer the number of unobserved individuals. Furthermore, even if a model can be adequately chosen, the data is MSE studies is often limited. In table 1, for instance, we can see how few individuals have appeared on more than one list. For many possible combinations of lists, there is no corresponding data point. This limited data implies that we must resort on prior knowledge and regularization procedures in order to estimate model parameters. This is problematic since the extrapolation formulas used in MSE studies are very sensitive to small changes in estimated parameters. These issues boil down to a problem of underspecification (D'Amour et al., 2020) - many reasonable approaches to MSE can give wildly different results.

These limitations should be carefully accounted for when performing MSE studies securely and privately. MSE cannot be considered a simple black box. While protecting individual privacy, sufficient information need to be disclosed to allow results to be

scrutinized.

2.2 Record Linkage Approaches

The goal of record linkage (also referred to as entity resolution and deduplication) is to identify records which refer to the same entity (Herzog et al., 2007; Christen, 2012; Christophides et al., 2019; Jurek-Loughrey and Deepak, 2019; Binette and Steorts, 2020). That is, the goal is to identify pairs of records which *match* or, equivalently, to cluster together records which refer to the same entity. These records may be spread across one or more databases and the entities of interest can be persons, objects, or events.

The problem is formally defined below under the name of *entity resolution*. We rely on *unique entity identifiers* to formalize the specification of a record linkage or clustering. That is, we formulate the entity resolution problem as one of uniquely assigning records to entities. Given unique entity identifiers, databases may then be joined or cleaned through standard techniques such as SQL "group by" and different "join" operations.

Definition 2.1 (Entity resolution)**.** Entity resolution is the problem of assigning *unique entity identifiers* to a set of records, where a unique identifier is any code which uniquely identifies a given entity.

Example 2.2 (Record linkage of killings and disappearance records)**.** Table 2 shows an illustrative example of records of killings and disappearances (Sadinle, 2014). This was inspired by data from the United Nation Comission Truth Commission for El Salvador (UNTC). Following the 1979 - 1992 civil war (Ball, 2000; Hoover Green, 2011; Green and Ball, 2019; Sadinle, 2014), the United Nations collected records of killings and photocopies were later digitized using optical character recognition. Given that the data was collected many years after the conflict and that photocopies were digitized, there is noise in the data, such as misspellings, optical character recognition errors, and possible inaccuracies. Furthermore, many different records can refer to the same individual. Record linkage approaches are therefore necessary to clean and integrate the data. In the context of an MSE study, where the source of each record would also be identified, record linkage would enable the computation of overlap counts.

A potential clustering of the records of table 2 is shown in table 3, with unique entity identifiers "Person A" and "Person B." Note that there is a lot of uncertainty involved in such a clustering, and this is only of many possible ways in which the records could have been grouped and aggregated. Different record linkage approaches range from providing

Record	Given name	Family name	Year	Month	Day	Municipality
1.	JOSE	FLORES	1981	1	29	A
2.	JOSE	FLORES	1981	2	NA	A
3.	JOSE	FLORES	1981	3	20	A
4.	JULIAN ANDRES	RAMOS ROJAS	1986	8	5	B
5.	JILIAM	RMAOS	1986	8	5	B

Table 2: Example of records of killings and disappearances inspired by the UNTC dataset. The data shows some of the challenges of record linkage. For instance, it is unclear whether or not records 4 and 5 refer to the same individual or if the two individuals could be siblings. Records 1–3 may refer to one or more individuals, depending on the level of noise in the data. This table is reproduced from table 1 of (Sadinle, 2014).

1. / \ 3. —— 2.	4. —— 5.
Person A Jose Flores Municipality A –/–/1981	**Person B** Julian Ramos Municipality B 05/08/1986

Table 3: Possible clustering and merging of the records in table 2. This is reproduced from table 2 of Binette and Steorts (2020).

a simple estimate of the linkage structure (such as in table 3) to quantifying uncertainty through pairwise match probabilities or through probability distributions on the clustering. These different approaches are reviewed next in section 2.2.2, following the introduction of the record linkage pipeline 2.2.1.

2.2.1 The Record Linkage and Data Cleaning Pipeline

Record linkage projects are typically carried out in four main stages. First, records are standardized and aligned so that they can be compared to one another. Second, in the *blocking* phase, records are sorted into blocks according to given rules. The goal is to speed

up record linkage by eliminating obvious non-matches: pairs of records which do not appear in the same block are automatically considered to be non-matches. For example, records may be blocked according to an individual's initials, with disagreeing initials meaning that records are considered to be non-matches. Third, record linkage is performed within blocks as to identify *coreferent* records (records which refer to the same entity). In the fourth stage, *canonicalization* or *data fusion*, coreferent records are merged as to combine their information and resolve inconsistencies.

The quality of a record linkage is usually evaluated using the *precision* and *recall* performance metrics. Precision is the percentage of posited links which are actually a match, while recall is the percentage of matching record pairs which are correctly linked together. Computing precision and recall require *ground truth* data to be available. In practice these values are estimated using unsupervised model-based approaches or through clerical review of record pairs.

2.2.2 Typology of Record Linkage Approaches

We now review three main types of record linkage approaches which are widely used in official statistics and research. These are *deterministic*, *probabilistic*, and *Bayesian* approaches. We also restrict our focus to *unsupervised* record linkage, where no ground truth data or manual review of records is used. Unsupervised record linkage is better suited to the privacy requirements of RL+MSE studies.

Deterministic record linkage: Deterministic approaches provide an estimated linkage, without accounting for possible errors in the results. These approaches often use fixed rules to determine the matching status of record pairs. For example, records may be linked if they are an exact match in all but 1 fields. If typographical errors are expected, string comparison metrics can be used as the basis of more sophisticated linkage rules. The Levenshtein distance between two strings, for instance, computes the minimum number of insertions, deletions or substitutions needed to transform one string into the other. This string similarity function can be thresholded to determine the matching status of certain record pairs. In addition to explicit linkage rules, clustering algorithms can also be used to obtain entity clusters from pairwise record similarities.

Probabilistic record linkage: Probabilistic record linkage approaches focus on estimating the *probability* that record pairs match and provide estimated error rates. The

seminal work of Fellegi and Sunter (1969) introduced the first model for probabilistic record linkage. This model relies on computing a *comparison vector* $\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_k)$ for each record pair, where γ_i is the result of a comparison between the records ith field. For instance, binary comparison can be used as to determine the matching status $\gamma_i \in \{0, 1\}$ of the record pair's ith field. A mixture model is then fitted to the set of comparison vectors (e.g. using the EM algorithm), resulting in estimated match probabilities for each possible value of γ. While there are many variants to this approach, the model of Fellegi and Sunter (1969) remains the most widely-used as the basis of unsupervised probabilistic record linkage.

Bayesian approaches: Bayesian approaches provide a *posterior distribution* on the linkage, allowing full quantification of uncertainty (Sadinle, 2018). Many such approaches are based on the model of Fellegi and Sunter (1969), while others have adopted latent entity modeling which is more closely related to clustering techniques. These require the computation of comparison vectors or the computation of comparisons between vectors and latent entities. These models are fitted using Monte Carlo Markov Chain techniques and the posterior distribution on the linkage is approximated by a set of plausible linkages.

2.3 Security and Privacy Needs

This section details the security and privacy need of organizations which could participate in RL+MSE studies. We define *security* as protection against unintended data disclosures. These disclosures could be the result of a deliberate attack or it could be the result of simple negligence if adequate protections are not taken. *Privacy*, on the other hand, is meant to ensure that even *intended* data disclosures do not inadvertently reveal sensitive personal information. In the context of RL+MSE studies, sensitive personal information includes:

1. which organizations (if any) a given individual has been in contact to, and
2. any more specific information about a given individual collected by these organizations.

None of (1) and (2), excepted for information already available to organizations, should be disclosed in RL+MSE studies. This includes both disclosures due to a security breach and inadvertent disclosures due to deductive re-identification or other inferences.

Given that RL+MSE studies require the collaboration of organizations in order to compute a shared function (an estimate of a total population size), defining security requires

particular care. The notion of secure multiparty computation, reviewed in section 2.3.1, provides adversary models and security definitions suitable to this context.

Furthermore, given the requirements and limitations of MSE studies discussed in sections 2.1.1 and 2.1.2, it is necessary for organizations to share sufficient information for the scrutiny of MSE analysis. At the very least, a table of overlap counts as in table 1 should be released. Protecting against the inadvertent disclosure of sensitive personal information through this table requires the formalizations of privacy discussed in section 2.3.2.

2.3.1 Secure-Multiparty Computation

Secure Multiparty Computation (SMC) (Evans et al., 2017) is the problem of evaluating a function based upon multiple parties' inputs where each input must be kept private from all parties, except for what is inferred from the comparison between a party's own input and the realized output. Each situation in which SMC is used will provide different constraints and definitions of privacy.

Computational security is the security against adversaries implemented by non-uniform, polynomial-time algorithms. A security structure can be defined as computationally secure if an adversary with computationally bounded resources cannot succeed, while an unbounded adversary can break the security of the structure.

Adversary models present in this paper are the semi-honest and malicious adversaries. Semi-honest, also known as honest-but-curious, adversaries are parties which are included in the SMC, which abide by all rules of protocol. However, they will make any attempt to uncover any data about the other parties without straying from the protocol. In contrast with semi-honest adversaries, malicious adversaries have all of the curious tendencies of a semi-honest adversary, but will make attempts to deviate from the protocol. This deviation is achieved through means of not communicating the result of their function, or tampering with the output of the function.

For protection specific to these malicious adversaries, a protocol must abide to descriptors such as guaranteed delivery, and output fairness. Guaranteed delivery ensures that all parties will receive the output of the multiparty function despite malicious adversaries, and output fairness describes that if one party receives the output, all parties receive the output.

The use of Secure-Multiparty Computation in this context of hidden populations is the situation described. There are two separate organizations that are in contact with these hidden populations seeking help. One of these organizations is a governmental organization, and the other is a nonprofit organization. The reason why SMC is useful in this situation

can be where a member of a hidden population only contacted the nonprofit due to the governmental organization is compromised by the human trafficking group. When this member is seeking help with the nonprofit, they are in fear of the retaliation by the human trafficking group if they found their information. This semi-honest governmental organization would not have access to the database of the nonprofit when SMC is used in conjunction with PPRL.

2.3.2 Statistical Disclosure Control and Differential Privacy

Protecting individual privacy in data disclosures is the subject of *statistical disclosure control* (SDC) (Hundepool et al., 2012). Since any useful data release will necessarily reveal some amount of information, the goal of SDC is to ensure that the probability of adverse events is sufficiently small. The concept of *differential privacy* (Dwork, 2008) has emerged in the field to formalize the meaning of protecting individual privacy. The idea is that data releases should be randomized such that the inclusion or exclusion of any one individual from the original data does not substantially affect the odds of any subsequent event.

Formally, let D be a dataset and let M be any algorithm which randomizes a selected function of D. Then M is said to be ϵ-differentially private if, for any two databases D and D' which differ by at most one individual, and for any event E, we have

$$\log \frac{\mathbb{P}(M(D) \in E)}{\mathbb{P}(M(D') \in E)} \leq \varepsilon.$$

3 Overview of Privacy-Preserving Record Linkage

Privacy-preserving record linkage is the problem of securely linking records across organizations without revealing personally identifying information. That is, organizations want to rely on personal information to carry out record linkage, without revealing that information to each other. There are many reasons why this is necessary beyond our case study of RL+MSE studies. Health care providers might be interested in aggregating medical records for research purposes. However, the Health Insurance Portability and Accountability Act of 1996 (HIPAA) restricts the sharing of protected health information. Similar legal restrictions are formulated under the Family Educational Rights and Privacy Act (FERPA) and under the General Data Protection Regulation (GDPR) in the European Union.

Most of the PPRL literature focuses on the problem of releasing, to the organizations or to a third party, a *restricted* set of attributes only for *matching records*. With two databases

for example, only selected information regarding individuals appearing in both databases will be released. The PPRL literature also typically ignores disclosure risks associated with the result of the record linkage. That is, there is a focus on *security* while *privacy* is not directly considered. Section 3.1 reviews the standard conceptualization of PPRL which is used in the literature we discuss.

Implementations of PPRL approaches fall under three main categories which are reviewed in section 3.2. *Hash-based coding approaches* are used to mask identifying attributes while still allowing comparisons between them. These techniques are used in practice because of their simplicity and speed. However they can be susceptible to dictionary, frequency and cryptanalysis attacks. Approaches which satisfy the formal security definition of secure multiparty computation (MPC) fall under *homomorphic encryption* and *other MPC protocols*. Generally, homomorphic encryption is preferred to other MPC protocols as it does not require as much network communication between organizations. Network access to databases is usually restricted when dealing with highly sensitive data.

3.1 Standard Conceptualization of PPRL

As previously stated, most of the literature on PPRL focuses on the problem of releasing a restricted set of attributes for records matching across organizations. Following Vatsalan et al. (2013), the PPRL problem can therefore be defined as follows.

Definition 3.1 (Standard goal of PPRL.)**.** Let $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_m$ be a set of m databases with records $r_{i,j} \in \mathcal{D}_j$, $i = 1, 2, \ldots, N_j$. Let $C : \mathcal{D}_1 \times \mathcal{D}_2 \times \cdots \times \mathcal{D}_1 \rightarrow \{0, 1\}$ be a classification rule to identify tuples of matching records, and let $M = C^{-1}(\{1\})$ bet the set of matching record tuples. Also let $f(r_{i,j})$ be a selected attribute for record $r_{i,j}$. The goal of privacy-preserving record linkage is then to release the set $\{(f(r_1), f(r_2), \ldots, f(r_m)) \mid (r_1, r_2, \ldots, r_m) \in M\}$ of selected attributes for matching records.

The way in which PPRL is carried out may involve a non-colluding third party. Ideally, the security definition of secure multiparty computation discussed in section 2.3.1 should be achieved under reasonable adversary models. Current PPRL approaches typically do not provide such a high level of security. The security of various PPRL approaches is discussed on a case-by-case basis in section 3.2.

3.2 Standard Approaches to PPRL

Before continuing on to review standard approaches to PPRL, let us discuss the underlying framework and its associated terminology in more detail.

Record attributes used to perform record linkage are referred to as *quasi-identifiers*. These can be attributes such as name, date of birth, address, etc. Typically, PPRL involves the organizations (data owners) as well as a third party called a *linkage unit*. The linkage unit's role is to aid organizations in performing the computations necessary for record linkage.

3.2.1 Hash-Based Coding Approaches.

Most of the PPRL literature advocates for so-called masking approach (Christen et al., 2020). Organizations individually mask quasi-identifiers using hash-based codes which are then shared to the linkage unit. The linkage unit then performs record linkage based on these codes and notifies organizations of matching records. Selected attributes for matching records can finally be shared. Below we describe PPRL approaches to deterministic record linkage.

Keyed hash functions for exact matching rules. The simplest deterministic record linkage algorithms rely on the exact matching status of quasi-identifiers. For example, two records may be considered a match if they agree on name and date of birth, or if they agree on name, age and zip code. This only relies on the equality or non-equality of quasi-identifiers, and the algorithms may therefore be implemented on an equality-preserving encryption of the quasi-identifiers. This was first proposed in Dusserre et al. (1995); Quantin et al. (1998).

Formally, the approach is as follows. First, organizations agree on a shared secret seed using any secure multiparty secret key generation protocol. Next, quasi-identifiers are encrypted using a keyed one-way hash function. The encrypted quasi-identifiers are communicated to the linkage unit which performs the record linkage and communicates back the matching status of records.

This approach is susceptible to collusion between the linkage unit and other organizations. Indeed, if the hash function seed is shared with the linkage unit, or if encrypted values held by the linkage unit are shared with other organizations, then a simple dictionary attack can be performed to break the encryption.

A solution to the dictionary attack problem is to secret share the seed and to implement the one-way hash function collectively through commutative one-way hash functions. The protocol is described in Vaidya and Clifton (2005) for the secure set intersection cardinality problem. We are not aware of the use of this technique for PPRL applications.

Bloom filters for string similarity evaluation. Bloom filters are data structures which are used for evaluating the similarity between two strings, and computing the chance that an item is within a set. A bloom filter consists of a bloom vector which contains l bits which are initialized at 0. To fill the bloom filter, a set of k hash functions are utilized to map each of the elements of the set to the bloom vector. Schnell et al. (2009) details the process for string similarity evaluation. At the start of the process, strings that could be similar are broken up into x-grams, which contain x characters each. For example, the name "Alice" in a 3-gram similarity will consist of strings "–A", "-Al", "Ali", "lic", "ice", "ce-", and "e–", and "Alexa" will produce "–A", "-Al", "Ale", "lex", "exa", "xa-", and "a–". Alice's 3-grams will be passed through the hash functions and added to its bloom filter, and Alexa's 3-grams will undergo the same process to its own bloom filter. The string similarity is found when the collisions are compared between the bloom filters. A quick glance at the strings will immediately show they are not the same, but bloom filters are employed to make this calculation while securely encrypted. Described by Vatsalan and Christen (2016), bloom filters can also be used to check if an item is possibly within the set without having access to the set. In this method, each item of the set is added to the bloom vector after being passed through the k hash functions, changing each of the mappings to a 1 bit on the bloom vector. When querying for an element on the bloom filter, the element is run through the same hash functions and compared to the bloom vector. A party can say that the element is definitely not in the set if the queried element's bloom vector has at least one "1" bit that correlates to a "0" bit on the set bloom vector. Through this method of comparison, there is no possibility of false negatives, however, the chance of a false positive increases with each element that is added to the filter.

3.2.2 Homomorphic Encryption

Homomorphic encryption is a scheme in which a party can do calculations on a piece of encrypted data without decrypting the data. Described by Rivest et al. (1978), a bank loan company stores its loan data in a time-sharing server, but has its data encrypted on that server. If the bank loan company would like to calculate trends for their bank loans

while keeping their data on the time-sharing server, it would need to utilize homomorphic encryption to keep the data encrypted while calculating these trends. Homomorphic Encryption is often utilized in conjunction with SMC. For that example, Alice has a database that needs calculations done, and Bob has an input that will complete the calculations, but Bob does not need the outcome of the calculations. Alice can encrypt her data and send it to Bob. Once Bob has the encrypted data, he can complete the required calculations with his input. After completing the calculations, Bob sends the encrypted data back to Alice, and Alice and decrypt her calculated data with her original key. Homomorphic encryption is often used for PPRL when a number of parties would like to find the linkage in their data without utilizing an independent party to handle and complete the linkage.

3.2.3 Multi-Party Computation Protocols

Secure multiparty computation (SMC) protocols which satisfy formal security guarantees have been proposed for the computation of string similarity (Atallah et al., 2003) and for secure collaborative data disclosure (Mohammed et al., 2014) (collaborative data disclosure is further discussed in section 4.3). Another interesting application of SMC is in Laud and Pankova (2018), who proposed a PPRL approach based on the use of a trusted third party, and where the trusted third party is implemented by three independent servers which compute collaboratively through SMC. The decomposition of the trusted third party through three entities is thought to increase security, as all three servers need to be corrupted in order for the third party to be compromised.

Generic SMC protocols, such as Yao's garbled circuits and other protocols (Evans et al., 2017), have not been widely considered for PPRL beyond the previously mentioned cases. This is due to the computational and communication overhead of SMC in practice. In particular, organizations holding sensitive data may not be willing to set up network access to the servers holding sensitive data (Christen et al., 2020).

4 Application to Multiple Systems Estimation Studies

We consider the use of privacy-preserving record linkage in the context of RL+MSE studies. First, we note that the standard framework of PPRL (Definition 3.1) is not well suited to the privacy need of RL+MSE studies discussed in section 2.3. Indeed, in our application we do not want organizations to be able to learn which records they have in common. Furthermore,

in order to construct a table of overlap count as in table 1, we need information regarding non-matching records. Finally, definition 3.1 merges the issue of performing record linkage and of releasing data, thus complicating the analysis of disclosure risks.

Given these challenges, we propose to redefine PPRL in section 4.1, where secure entity resolution (secure ER) now refers to a more specific conceptualization of PPRL. Our new definition clearly separates secure computation from privacy considerations. In section 4.2, we discuss how it may be implemented using the approaches reviewed in section 3.2. Finally, in section 4.3, we discuss how organizations can disclose information while preserving individual privacy.

4.1 Secure Entity Resolution

Given the previously discussed drawbacks of the standard conceptualization of PPRL (Definition 3.1), we define *secure entity resolution* as follows.

Definition 4.1 (Secure ER)**.** *Secure entity resolution* is the problem of securely computing unique entity identifiers (uIDs) for records spread across organizations, where uIDs are defined according to a given record linkage algorithm. The goal is for each organization to obtain uIDs for their own set of records, without anything else being revealed.

Note that secure ER does not involve organizations learning about which of their records match with other organizations' records. There is also no utility in performing secure ER on its own. Data disclosure must be considered as a second step to secure ER.

4.2 Implementing Secure Entity Resolution

The techniques and protocols discussed in section 3.2 and which rely on a linkage unit can be adapted for secure ER. Indeed, the linkage unit can simply return to each organization a set of uIDs, randomly generated under the matching constraint, instead of revealing a set of matching records. However, the problem posed by a colluding linkage unit is exacerbated in the context of secure ER since we do not want to reveal which records are matching.

Here we explore two techniques for secure ER which do not rely on a linkage unit and which are applicable to the simplest forms of deterministic record linkage.

Method 1: Exact matching. The simplest way to perform entity resolution is to directly construct unique entity identifiers from record attributes. For example, name and

date of birth can be concatenated in a standard form in order to obtain an individually identifying code. This corresponds to matching on the conjunction of given attributes and can be a satisfactory solution for some problems.

For secure entity resolution, a keyed hash function can be applied to in order to obtain anonymous unique entity identifiers, where the key is generated through some collaborative protocol.

Method 2: More complex exact matching rules. Method 1 can be extended by considering more complex exact matching rules. For instance, we may match records which agree on name and date or birth, **or** which agree on name, zip code and birth year. For each conjunction of attributes, an identifying code can be generated by concatenation and anonymized through a keyed hash function, as in method 1. Two records will then match if they agree on any of their corresponding identifying code. The remaining problem for entity resolution is to resolve the many codes into single unique entity identifiers. A solution is proposed below.

Formally, we propose an algorithm for secure ER in this context of matching with a disjunction of two code-based rules, and with only two databases. That is, suppose we have two databases D_1 and D_2 with owners O_1 and O_2, and suppose that the matching rule has the following form: for given functions f and g, records x and y match if and only if $f(x) = f(y)$ **or** $g(x) = g(y)$. For simplicity, we assume that there are no duplicates within the databases. As stated in the previous paragraph, the entity resolution problem is to determine, for each record x, which of the codes $f(x)$ or $g(x)$ should be used as the basis of a unique entity identifier (uID). If there exists a record y such that $f(x) = f(y)$ but $g(x) \neq g(y)$, then $f(x)$ should be used instead of $g(y)$ to ensure that the resulting uID respects the matching rule. If x does not match any other records, then any one of the two codes can be used. Each database owner must therefore know, for each of their records, which code they should keep. For secure ER, they must do so without learning or revealing the matching status of their records.

Our solution to this problem relies on additively homomorphic encryption and is outlined in Algorithm 1 (see page 21). The idea is that elements of a set (say the identifying codes for O_2) can be represented as the zeros of a polynomial P (as in (Freedman et al., 2004)). Owner O_1 can then decide which of the codes $f(x)$ or $g(x)$ to keep as a unique entity identifier for record x by computing $P^2(f(x)) - P^2(g(x))$, where P^2 is the square of P. If $f(x) = f(y)$ of $g(x) = g(y)$, then the sign of $P^2(f(x)) - P^2(g(x))$ specifies which is the case.

If both $f(x) \neq f(y)$ and $g(x) \neq g(y)$, then the sign of $P^2(f(x)) - P^2(g(x))$ is arbitrary and any one of $f(x)$ and $g(x)$ can be used as a unique entity identifier. The evaluation of the polynomial is carried out under additively homomorphic encryption (such as Paillier (1999); Freedman et al. (2004)) in order to protect each owner's information.

Note that some information is leaked through this protocol. For instance, suppose a record x in D_1 is such that there exists $y \in D_2$ with $f(x) = f(y)$ and $g(x) = g(y)$. Then $P^2(f(x)) - P^2(g(x)) = 0$ and O_2 learns that one of O_1's record has matched through both identifying codes. Furthermore, the set of evaluations $\{P^2(f(x)) - P^2(g(x)) \mid x \in D_1\}$ may reveal information to D_2 about the records in D_1. Further work (such as based on Freedman et al. (2004) which inspired our protocol) would be required to evaluate and mitigate these risks.

Input : Databases D_1 and D_2 with each N elements and owners O_1 and O_2; numeric functions $f, g : D_1 \cup D_2 \to \mathbb{R}$.

Output : Each owner O_i obtains unique entity identifiers uID_i for its database which satisfy the following matching rule: records x and y match if and only if $f(x) = f(y)$ or $g(x) = g(y)$.

1 **for** $i \in \{1, 2\}$, *owner* O_i **do**
2 Construct polynomial P_i with zeros $\{f(x) \mid x \in D_i\} \cup \{g(x) \mid x \in D_i\}$.
3 Encrypt P_i^2 to $\tilde{P}_i^2$ using an additively homomorphic scheme.
4 Send $\tilde{P}_i^2$ to O_{3-i}.
5 **end**
6 **for** $i \in \{1, 2\}$, *owner* O_i **do**
7 Initialize vector $\tilde{r}_i$ of length N.
8 **for** $k = 1, 2, \ldots, N$ **do**
9 Let x be the kth record of database D_i.
10 $\tilde{r}_i[k] \leftarrow \tilde{P}_{3-i}^2(f(x)) - \tilde{P}_{3-i}^2(g(x))$
11 **end**
12 Send randomly permuted vector $\tilde{r}_i$ to O_{3-i}.
13 **end**
14 **for** $i \in \{1, 2\}$, *owner* O_i **do**
15 Initialize vector r_{3-i} of length N.
16 **for** $k = 1, 2, \ldots, N$ **do**
17 Decrypt $\tilde{r}_{3-i}[k]$ to $r_{3-i}[k] = P_{3-i}^2(f(x)) - P_{3-i}^2(g(x))$.
18 **end**
19 Send $\text{sign}(r_{3-i})$ to O_{3-i}.
20 **end**
21 **for** $i \in \{1, 2\}$, *owner* O_i **do**
22 Initialize vector uID_i of length N.
23 Unpermute r_i given step 12.
24 **for** $k = 1, 2, \ldots, N$ **do**
25 Let x be the kth record of database D_i.
26 **if** $r_i[k] < 0$ **then** $\text{uID}_i[k] = f(x)$
27 **else** $\text{uID}_i[k] = g(x)$
28 **end**
29 **end**

Algorithm 1: Computation of unique entity identifiers using an additively homomorphic encryption scheme.

4.3 Collaborative Data Disclosure

Following secure ER, data can be disclosed through additional collaborative protocols which are secure and which respect individual privacy. In the context of RL+MSE studies, one goal is to obtain a table of overlap counts as in table 1. This can be achieved using protocols for secure set intersection cardinality (Vaidya and Clifton, 2005).

Other studies can require different kinds of information. For instance, a risk analysis study may rely on covariates which needs to be integrated across databases. However, we do not want to reveal which cases come from which databases, and we might want to output an integrated database which is differentially private. Data is then protected from re-identification attacks (even attacks from any one of the participating organizations). This is the subject of the growing field of secure multiparty data release under differential privacy (Mohammed et al., 2014). Secure ER is the necessary first step to such secure collaborative data diclosure protocols.

5 Discussion

There is no agreed-upon definition of PPRL, nor of the kind of security and privacy guarantees which are required. This motivated us to formulate the notion of *secure ER*, which clearly separates secure record linkage from privacy-preserving data disclosure. However, unless we rely on a trusted third party, secure ER is a more difficult problem than PPRL.

Furthermore, most of the PPRL literature proposes solutions which are either impractical or insecure. Hash-based coding approaches are susceptible to different attacks, and the risk of a colluding third party is real in many applications (e.g. when working with law enforcement agencies which could coerce the third party).

Given the state of the field, there are opportunities to clarify the goals and meaning of PPRL, and to possibly propose practical and secure solutions. Much more work is needed in this area. In particular, it would be interesting to explore the practical applicability of the secure ER approaches which we propose, and to extend them to more broadly useful record linkage algorithms.

References

Atallah, M. J., F. Kerschbaum, and W. Du (2003). Secure and private sequence comparisons. In *Proceedings of the 2003 ACM Workshop on Privacy in the Electronic Society*, pp. 39–44.

Bales, K., O. Hesketh, and B. W. Silverman (2015). Modern slavery in the UK: How many victims? *Significance 12*(3), 16–21.

Bales, K., L. T. Murphy, and B. W. Silverman (2019). How many trafficked people are there in Greater New Orleans? Lessons in measurement. *Journal of Human Trafficking*, 1–13.

Ball, P. (2000). The Salvadoran human rights commission: Data processing, data representation, and generating analytical reports. In P. Ball, H. F. Spirer, and L. Spirer (Eds.), *Making the Case: Investigating Large Scale Human Rights Violations Using Information Systems and Data Analysis*, pp. 15–24. American Association for the Advancement of Science.

Ball, P. and J. Asher (2002). Statistics and Slobodan: Using Data Analysis and Statistics in the War Crimes Trial of Former President Milosevic. *Chance 15*(4), 17–24.

Ball, P. and M. Price (2019). Using Statistics to Assess Lethal Violence in Civil and Inter-State War. *Annual Review of Statistics and Its Application 6*(1), 63–84.

Binette, O. and R. C. Steorts (2020). (Almost) All of entity resolution. *arXiv e-prints*, 1–53. arxiv:2008.04443.

Bird, S. M. and R. King (2018). Multiple Systems Estimation (or Capture-Recapture Estimation) to Inform Public Policy. *Annual Review of Statistics and Its Application 5*(1), 95–118.

Chan, L., B. W. Silverman, and K. Vincent (2020). Multiple systems estimation for sparse capture data: Inferential challenges when there are nonoverlapping lists. *Journal of the American Statistical Association*, 1–10.

Chao, A., H. Y. Pan, and S. C. Chiang (2008). The Petersen - Lincoln estimator and its extension to estimate the size of a shared population. *Biometrical Journal 50*(6), 957–970.

Chao, A., P. K. Tsay, S. H. Lin, W. Y. Shau, and D. Y. Chao (2001). The applications of capture-recapture models to epidemiological data. *Statistics in Medicine 20*(20), 3123–3157.

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

Christen, P., T. Ranbaduge, and R. Schnell (2020). *Linking Sensitive Data*. Springer.

Christophides, V., V. Efthymiou, T. Palpanas, G. Papadakis, and K. Stefanidis (2019). End-to-end entity resolution for big data: A survey. *arXiv e-prints*. arxiv:1905.06397.

Cochran, W. G. (1978). Laplace's ratio estimator. *Contributions to Survey Sampling and Applied Statistics*, 3–10.

Cockayne, J. (2015). *Unshackling Development: Why we need a global partnership to end modern slavery*. Number December.

D'Amour, A., K. Heller, D. Moldovan, B. Adlam, B. Alipanahi, A. Beutel, C. Chen, J. Deaton, J. Eisenstein, M. D. Hoffman, et al. (2020). Underspecification presents challenges for credibility in modern machine learning. *arXiv preprint arXiv:2011.03395*.

Darroch, J. N., S. L. Lauritzen, and T. P. Speed (1980). Markov fields and log-linear interaction models for contingency tables. *Annals of Statistics 8*(3), 522–539.

Datta, M. N. and K. Bales (2013). Slavery in europe: Part 1, estimating the dark figure. *Human Rights Quarterly 35*(4), 817–829.

Dusserre, L., C. Quantin, and H. Bouzelat (1995). A one way public key cryptosystem for the linkage of nominal files in epidemiological studies. *Medinfo. MEDINFO 8*, 644–647.

Dwork, C. (2008). Differential privacy: A survey of results. In *International conference on theory and applications of models of computation*, pp. 1–19. Springer.

Evans, D., V. Kolesnikov, and M. Rosulek (2017). A pragmatic introduction to secure multi-party computation. *Foundations and Trends® in Privacy and Security 2*(2-3).

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

Fienberg, S. E. (1972). The multiple recapture census for closed populations and incomplete 2k contingency tables. *Biometrika 59*(3), 591–603.

Freedman, M. J., K. Nissim, and B. Pinkas (2004). Efficient private matching and set intersection. In *International conference on the theory and applications of cryptographic techniques*, pp. 1–19. Springer.

Green, A. H. and P. Ball (2019). Civilian killings and disappearances during civil war in el salvador (1980–1992). *Demographic Research 41*, 781–814.

Hald, A. (2005). *A history of probability and statistics and their applications before 1750*, Volume 574. John Wiley & Sons.

Hall, R. and S. E. Fienberg (2010). Privacy-preserving record linkage. In *International conference on privacy in statistical databases*, pp. 269–283. Springer.

Hand, D. J. (2018). Statistical challenges of administrative and transaction data. *Journal of the Royal Statistical Society. Series A: Statistics in Society 181*(3), 555–605.

Heckathorn, D. D. (1997). Respondent-driven sampling: A new approach to the study of hidden populations. *Social Problems 44*(2), 174–199.

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

Hoover Green, A. (2011). *Repertoires of Violence Against Noncombatants: The Role of Armed Group Institutions and Ideology*. Ph. D. thesis, Yale University, Department of Political Science.

Hundepool, A., J. Domingo-Ferrer, L. Franconi, S. Giessing, E. S. Nordholt, K. Spicer, and P.-P. De Wolf (2012). *Statistical disclosure control*. John Wiley & Sons.

International Labour Organization (2017a). *Global Estimates Of Modern Slavery: Forced Labour And Forced Marriage*. Geneva. https://www.ilo.org/wcmsp5/groups/public/---dgreports/---dcomm/documents/publication/wcms_575479.pdf.

International Labour Organization (2017b). *Methodology of the global estimates of modern slavery: Forced labour and forced marriage*. Geneva. https://www.ilo.org/wcmsp5/groups/public/@ed_norm/@ipec/documents/publication/wcms_586127.pdf.

Jurek-Loughrey, A. and P. Deepak (2019). *Semi-supervised and Unsupervised Approaches to Record Pairs Classification in Multi-Source Data Linkage*, pp. 55–78. Cham: Springer International Publishing.

Landman, T. (2020). Measuring Modern Slavery: Law, Human Rights and New Forms of Data.

Laplace, P.-S. (1820). *Théorie analytique des probabilités* (3 ed.). Paris, France.

Laud, P. and A. Pankova (2018). Privacy-preserving record linkage in large databases using secure multiparty computation. *BMC medical genomics 11*(4), 33–46.

Lincoln, F. C. (1930). Calculating waterfowl abundance on the bais of banding returns. Technical report, United States Department of Agriculture.

Lum, K., M. E. Price, and D. Banks (2013). Applications of multiple systems estimation in human rights research. *American Statistician 67*(4), 191–200.

Madigan, D. and J. C. York (1997). Bayesian methods for estimation of the size of a closed population. *Biometrika 84*(1), 19–31.

Manrique-Vallier, D., M. E. Price, and A. Gohdes (2013). Multiple Systems Estimation Techniques for Estimating Casualties in Armed Conflicts. *Counting Civilian Casualties*, 165–181.

Mathers, B. M., L. Degenhardt, B. Phillips, L. Wiessing, M. Hickman, S. A. Strathdee, A. Wodak, S. Panda, M. Tyndall, A. Toufik, and R. P. Mattick (2008). Global epidemiology of injecting drug use and HIV among people who inject drugs: a systematic review. *The Lancet 372*(9651), 1733–1745.

Mohammed, N., D. Alhadidi, B. C. Fung, and M. Debbabi (2014). Secure two-party differentially private data release for vertically partitioned data. *IEEE Transactions on Dependable and Secure Computing 11*(1), 59–71.

Paillier, P. (1999). Public-key cryptosystems based on composite degree residuosity classes. In *International conference on the theory and applications of cryptographic techniques*, pp. 223–238. Springer.

Petersen, C. G. J. (1895). The yearly immigration of young plaice into the limfjord from the german sea, etc.

Price, M. and P. Ball (2015). Selection bias and the statistical patterns of mortality in conflict. *Statistical Journal of the IAOS 31*(2), 263–272.

Quantin, C., H. Bouzelat, F. Allaert, A.-M. Benhamiche, J. Faivre, and L. Dusserre (1998). How to ensure data security of an epidemiological follow-up: quality assessment of an anonymous record linkage procedure. *International journal of medical informatics 49*(1), 117–122.

Rivest, R. L., L. Adleman, M. L. Dertouzos, et al. (1978). On data banks and privacy homomorphisms. *Foundations of secure computation 4*(11), 169–180.

Sadinle, M. (2014). Detecting duplicates in a homicide registry using a Bayesian partitioning approach. *The Annals of Applied Statistics 8*(4), 2404–2434.

Sadinle, M. (2018). Bayesian propagation of record linkage uncertainty into population size estimation of human rights violations. *The Annals of Applied Statistics 12*(2), 1013–1038.

Schnell, R. (2015). Privacy-preserving record linkage.

Schnell, R., T. Bachteler, and J. Reiher (2009). Privacy-preserving record linkage using bloom filters. *BMC medical informatics and decision making 9*(1), 1–11.

Silverman, B. W. (2014). Modern slavery: an application of multiple systems estimation. *Home Office, London*. Available from https://www.gov.uk/government/publications/modern-slavery-an-application-of-multiple-systems-estimation.

Spreen, M. (1992). Rare populations, hidden populations, and link-tracing designs: What and why? *Bulletin de Méthodologie Sociologique 36*, 34–58.

United Nations (2000). Protocol to prevent, suppress and punish trafficking in persons, especially women and children, supplementing the united nations convention against transnational organized crime. https://www.refworld.org/docid/4720706c0.html.

United Nations (2001). United Nations Convention against Transnational Organized Crime (A/RES/55/25). pp. 51. https://www.unodc.org/pdf/crime/a_res_55/res5525e.pdf.

Vaidya, J. and C. Clifton (2005). Secure set intersection cardinality with application to association rule mining. *Journal of Computer Security 13*(4), 593–622.

Vatsalan, D. and P. Christen (2016). Multi-party privacy-preserving record linkage using bloom filters. *arXiv preprint arXiv:1612.08835*.

Vatsalan, D., P. Christen, and V. S. Verykios (2013). A taxonomy of privacy-preserving record linkage techniques. *Information Systems 38*(6), 946–969.

Vatsalan, D., Z. Sehili, P. Christen, and E. Rahm (2017). Privacy-preserving record linkage for big data: Current approaches and research challenges. In *Handbook of Big Data Technologies*, pp. 851–895. Springer.

Wittes, J. T. (1974). Applications of a multinomial capture-recapture model to epidemiological data. *Journal of the American Statistical Association 69*(345), 93–97.

Yip, S. F., M. Richard, S. E. Fienberg, B. W. Junker, R. E. Laporte, and I. M. Libman (1995). Capture-recapture and multiple-record systems estimation I: History and theoretical development. *American Journal of Epidemiology 142*(10), 1047–1058.