Data Privacy for Record Linkage and Beyond

Shurong Lin Eric D. Kolaczyk

Abstract

In a data-driven world, two prominent research problems are record linkage and data privacy, among others. Record linkage is essential for improving decision-making by integrating information of the same entities from different sources. On the other hand, data privacy research seeks to balance the need to extract accurate insights from data with the imperative to protect the privacy of the entities involved. Inevitably, data privacy issues arise in the context of record linkage. This article identifies two complementary aspects at the intersection of these two fields: (1) how to ensure privacy during record linkage and (2) how to mitigate privacy risks when releasing the analysis results after record linkage. We specifically discuss privacy-preserving record linkage, differentially private regression, and related topics.

Keywords: record linkage, differential privacy, privacy-preserving data mining, data integration, secure multi-party computation, federated learning

1 Introduction

Research in data privacy seeks to balance the need to extract accurate insights from data for decision-making with the imperative to protect the privacy of the entities involved. In today's digital age, the unprecedented volume of data has heightened privacy concerns regarding sensitive information such as health records, genomic data, and Census surveys. The recent implementation of the EU General Data Protection Regulation (GDPR) underscores the urgent need to enforce robust privacy protection.

One traditional approach to addressing data privacy issues is through statistical disclosure control (SDC) or statistical disclosure limitation (SDL), which originated in official statistics. In computer science, various other privacy frameworks and techniques have emerged, with differential privacy now recognized as the gold standard. Both SDC and differential privacy have strong statistical foundations and are considered among the most dominant statistical data privacy frameworks (Slavković and Seeman, 2023). Notably, the U.S. Census Bureau has recently adopted differential privacy for its new disclosure avoidance system (U.S. Census Bureau, 2021).

Record linkage, a task with a long history in surveys and censuses (U.S. Census Bureau, 2022), has also been extensively studied in both statistics and computer science (Christen, 2012; Binette and Steorts, 2022). Often, data of the same group of entities are distributed across multiple sources, with unique identifiers unavailable for precise linkage due to non-existence, measurement errors, or privacy restrictions. Record linkage, also known as entity resolution or data matching, aims to find records that refer to the same entity across different data sources. This statistical task has become increasingly essential for better decision-making in a data-driven world.

Naturally, various privacy concerns arise when record linkage is involved. In this article, we explore research questions at the intersection of data privacy and record linkage. There are at least two key facets to consider when connecting these two fields: (1) how to perform record linkage privately for data sets contributed by multiple parties, and (2) how to conduct statistical analysis on linked data in a privacy-preserving fashion. The first facet involves completing the linkage task without disclosing excessive sensitive information among the different parties. The second facet ensures that the downstream analyses on the linked data are conducted privately, regardless of whether the record linkage itself is performed privately. We refer to the first facet as the primary perspective on record linkage and the second as the secondary perspective.

In Section 2, we review record linkage in a non-private setting. In Section 3, we provide an overview of privacy-preserving record linkage. In Section 4, we discuss the key challenges in private analysis following record linkage, featuring the recent advancements in Lin et al. (2024) using differential privacy for this purpose. Lastly, in Section 5, we briefly cover related topics in privacy-preserving data integration.

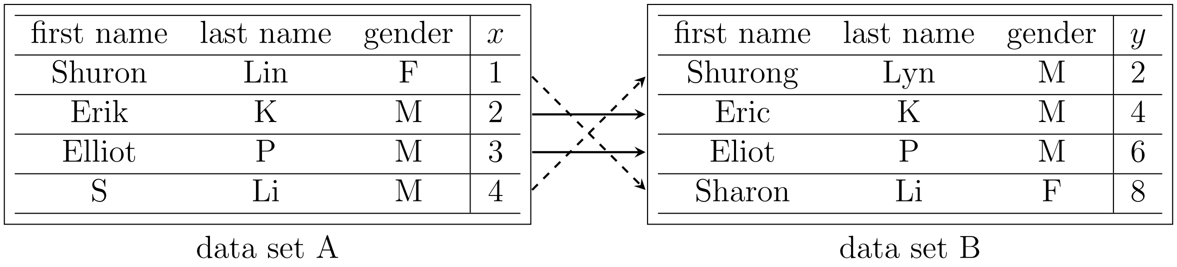
2 Record Linkage Overview

Record linkage refers to the task of linking records that refer to the same entity across different data sources, with lack of unique identifers. The earliest work to formalize record linkage as a statistical and computational problem is Newcombe et al. (1959). A seminal contribution by Fellegi and Sunter (1969) laid the probabilistic foundations for record linkage. With the rise of big data and modern computing, record linkage has become increasingly indispensable for big data analytics. In the following, we define the record linkage problem for two data sets, noting that this concept can be extended to any number of data sets.

2.1 Problem and Strategy

Given are two data sets, A and B, possibly of different sizes, containing information about the same group of entities. Instead of unique identifiers, quasi-identifiers (e.g., name, gender, date of birth) are used to identify the potential matches between the two data sets. These quasi-identifiers are referred to as *linking variables*. Figure 1 provides a toy example where first and last names, along with gender, are available for linkage. Due to possible measurement errors and the non-uniqueness of these linking variables, the linkage problem becomes probabilistic.

A traditional strategy for record linkage is given by: (1) compare linking variables to measure the similarity between records in the two data sets; (2) calculate the probabilities that two records are a match; (3) follow a decision rule to designate pairs as links and nonlinks; (4) defer decisions for ambiguous pairs to a further clerical review. In step (1), the similarity or agreement level between records is



first name	last name	gender	x
Shuron	Lin	F	1
Erik	K	M	2
Elliot	P	M	3
S	Li	M	4

data set A

first name	last name	gender	y
Shurong	Lyn	M	2
Eric	K	M	4
Eliot	P	M	6
Sharon	Li	F	8

data set B

Figure 1: A toy example of record linkage with mismatches (dashed links). (Adapted from Figure 2 in Lin et al. (2024).)

measured using certain metrics. Classic string metrics include Jaro similarity (Jaro, 1989) and Jaro–Winkler similarity (Winkler, 1990). In step (2), the probability that two given records refer to the same entity is calculated based on the similarity score. A seminal work by Fellegi and Sunter (1969) proposed the Fellegi-Sunter model, where the matching probability is derived from the m- and u-probabilities. The m-probability is the probability of an observation given the records are a match, whereas the u-probability is the counterpart for non-matches. Formally:

$$\begin{cases} m = \Pr(\text{Observation} \mid \text{Records match}); \\ u = \Pr(\text{Observation} \mid \text{Records do not match}). \end{cases} \tag{1}$$

Then, in step (3), a decision rule is applied to categorize pairs as links or non-links by choosing cutoffs for the matching probabilities. For pairs that fall between the cutoffs for links and non-links, the records can be forwarded for manual clerical review.

In practice, multiple issues arise when performing record linkage. For instance, it is computationally expensive to compare every possible pair of records between two data sets. A common technique to address this is blocking, which significantly reduces the number of comparisons by only comparing records within the same blocks. Other challenges include determining the appropriate m- and u-probabilities and selecting the optimal cutoffs for classifying links. In addition, there are alternatives to the Fellegi-Sunter model. For example, Steorts et al. (2016) proposed a Bayesian approach, and machine learning models can also be employed for these tasks (Winkler, 2011). We refer interested readers to a comprehensive survey on record linkage by Binette and Steorts (2022).

2.2 Implications on Downstream Analysis

In most cases, record linkage serves as a preprocessing step, as the ultimate goal is to combine information from two data sets to enable better analysis. Due to the probabilistic nature of record linkage, uncertainties are inevitable in downstream analysis. A naive approach might treat the linked set as accurate and proceed with standard analysis. However, studies have shown that ignoring linkage errors can result in substantial bias, even when linkage accuracy is high (Neter et al., 1965; Scheuren and Winkler, 1993).

In the toy example in Figure 1, simple linear regression is performed after record linkage. We aim to regress the variable y in data set B on the variable x in data set A. Prior to regression, linkage variables are used to match records between the two data sets. The true data set is $D_{\text{true}} = \{(1,2),(2,4),(3,6),(4,8)\}$, yielding a slope estimate $\hat{\beta}_1 = 2$, while the linked set is given by $D_{\text{linked}} = \{(1,8),(2,4),(3,6),(4,2)\}$, yielding $\hat{\beta}_1 = -1.6$. This discrepancy shows that mismatches in the linked data can even change the sign of the slope estimate. Therefore, it is crucial to accurately propagate linkage uncertainties to downstream tasks to ensure reliable estimates and informed decision-making. Statisticians have addressed uncertainty propagation in various statistical tasks with linked data, such as regression and small area estimation (Chambers, 2009; Han and Lahiri, 2019; Chambers et al., 2021).

3 Privacy-Preserving Record Linkage

In this section, we discuss the primary perspective on record linkage where privacy constraints are a major concern during the process. When two sensitive data sets held by different parties need to be linked, the field of privacy-preserving record linkage (PPRL) comes into play. PPRL, which sits at the intersection of record linkage and privacy-preserving data mining (Hall and Fienberg, 2010), aims to mitigate the risk of inadvertently disclosing private information. During the linkage process, it is crucial that non-linked records, which may contain sensitive data, are not revealed to the other party.

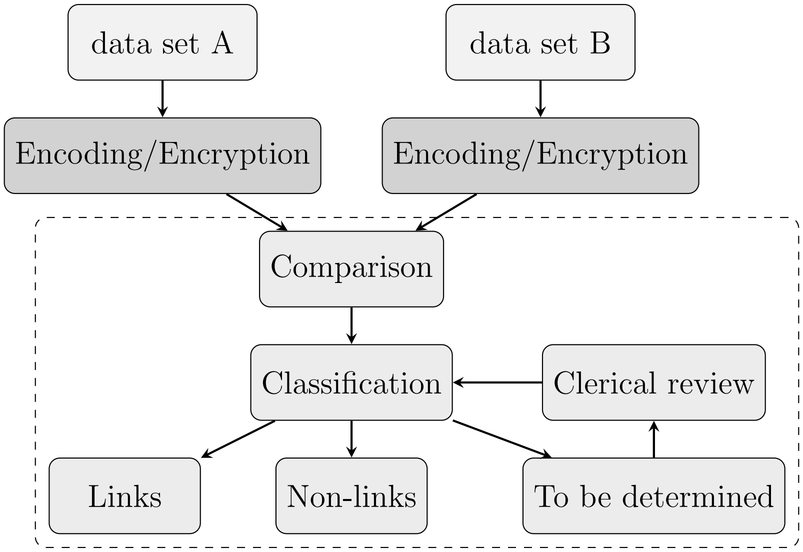


Figure 2: A simplified process of privacy-preserving linking two databases. The dashed box presents a typical process of record linkage. (Adapted from Figure 3.6 in Christen et al. (2020).)

Figure 2 presents a simplified version of the general process of PPRL with two data sets. The diagram can include additional steps such as data preprocessing before linkage, blocking for more efficient comparison, and linkage evaluations. As indicated, the primary distinction between PPRL and non-privacy record linkage is the inclusion of an encoding or encryption step before the comparison. Suitable encoding methods, based on various approaches such as hashing, Bloom filters, secure multi-party computation, and differential privacy, offer provable privacy (Christen et al., 2020). These privacy and security techniques, thus PPRL, are extensively studied in computer science.

4 Private Analysis of Linked Data

From the secondary point view on record linkage, we focus on incorporating privacy protection into the downstream analysis *after* the linkage has been completed as a pre-processing step. The linkage process itself may or may not be privacy-preserving. The primary goal of this facet is to mitigate privacy risks when the results of the analysis based on the linked data are released to certain audiences for decision-making.

As mentioned in the introduction, differential privacy is now regarded as the gold standard for statistical releases. Therefore, in the following, we primarily focus on differential privacy for ensuring privacy protection in the analysis of linked data.

4.1 Differential Privacy

First proposed by Dwork et al. (2006), differential privacy (DP) is a formal mathematical framework designed to ensure privacy when releasing statistical analyses on sensitive data. Central to DP is the concept of neighboring data sets or neighbors, which differ by only a single record. The goal of DP is to ensure that outputs, f and f', from any two such neighboring data sets, D and D', are similarly distributed, making it difficult to determine which output corresponds to which data set. This intuition is illustrated in Figure 3, where the random algorithm A satisfies DP by generating outputs with similar probabilities for any pair of neighboring data sets. Because the neighboring data sets are arbitrary, a DP algorithm provides privacy protection for any individual record in the data set.

The privacy protection level of a DP algorithm is quantified by measuring the distance between the probability distributions of the outputs from D and D'. A smaller distance indicates higher distinguishability, thus implying a lower privacy loss, i.e., stronger privacy protection. Mathematical definitions and properties of DP are detailed in sources such as Dwork and Roth (2014).

To construct differentially private algorithms, independent randomness is introduced in a calibrated manner to mitigate privacy risks. A typical approach involves injecting random noise into certain phases of data analysis. The amount of noise is

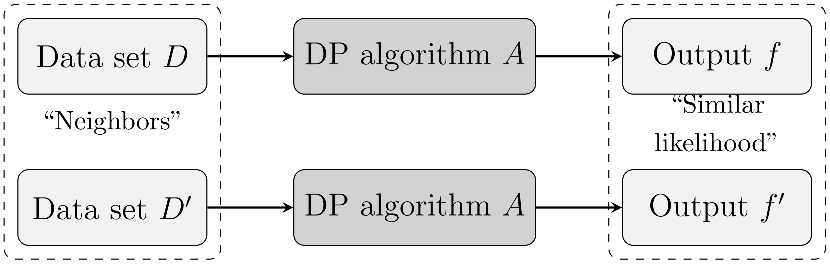


Figure 3: The intuition of differential privacy (DP).

determined by a designated limit for privacy loss (referred to as the *privacy budget* in DP) and the specifics of the analysis (referred to as the *sensitivity* in DP). For instance, Gaussian noise is a popular choice.

4.2 Differentially Private Regression on Linked Data

Given two data sets that provide the independent variable x and the dependent variable y, respectively, they share common linking variables Φ_x and Φ_y that are used to perform linkage. Ideally, if no linkage is needed or perfect linkage is feasible, one would have the data set $D_{\text{true}} = (x, y)$, and standard DP regression could be performed (see, e.g., Sheffet (2017); Wang (2018); Cai et al. (2021)). In the case of linked data, instead of having D_{true}, we have $D_{\text{linked}} = (x, z)$, where z is a permutation of y that depends on the comparison of Φ_x and Φ_y. Consequently, implementing DP for regression on linked data requires a more complex notion of neighboring data sets than simply swapping a row in D_{true} or D_{linked}. In fact, a pair of neighboring data sets should be defined as $D = (x, \Phi_x, y, \Phi_y)$ and $D' = (x', \Phi_{x'}, y', \Phi_{y'})$, where one entity's quasi-identifier differs, in addition to the values in variables x and y.

After articulating the neighboring relation for linked data, DP algorithms can possibly be designed accordingly. To the best of our knowledge, Lin et al. (2024) is the first work to study DP regression on linked data that accounts for upstream linkage uncertainties. Specifically, they propose two algorithms to perform linear regression while providing DP guarantees. The first algorithm, a noisy gradient descent method, introduces noise into the gradient descent process. The second approach, sufficient statistics perturbation, adds noise to the sufficient statistics used for estimation. These methods propagate the linkage uncertainties under a classical probabilistic linkage model, which has been employed in non-private settings by Lahiri and Larsen (2005) and Chambers et al. (2021).

One key challenge in designing DP algorithms for linked data is determining the appropriate level of injected noise, which depends on the complexities involved in analyzing linked data. At first glance, the uncertainties from upstream linkage might appear to offer some privacy protection. Indeed, data swapping, a statistical disclosure control technique, can be differentially private, as discussed in James Bailie (2024). However, in the case of record linkage, the randomness introduced by linkage errors is not independent of the data. Instead, it is the quasi-identifiers, which are part of the data itself, that give rise to these linkage uncertainties. Therefore, the randomness due to linkage does not directly provide privacy guarantees.

For linear regression, Lin et al. (2024) determined the scale of noise to be added for both post-record linkage algorithms and analyzed the finite-sample error bounds for the private estimators. Their theoretical results show that more injected noise is needed for regression on linked data compared to non-linked data, due to the complexity caused by linkage in the neighboring relations. A larger sample size and smaller intrinsic regression error both help reduce the amount of noise needed for privacy protection. The finite-sample error can be decomposed into two parts: linkage-regression errors, which are independent of DP, and DP-specific errors. As the sample size becomes sufficiently large, the linkage-regression error term dominates. Higher linkage accuracy decreases estimation error. The numerical results also confirm that treating the linked data (x, z) as the ground truth, ignoring linkage errors, leads to noticeable bias, even when the linkage accuracy is higher than 90%.

Even though Lin et al. (2024) focus specifically on linear regression, their methodology of propagating linkage uncertainties and implementing DP could potentially be extended to general supervised learning problems. In particular, the noisy gradient descent method is well-suited for broader applications.

5 Related Topics

We briefly cover two relevant approaches designed for private computation and analysis over multiple sensitive databases. Additionally, we provide an overview of data integration, of which record linkage is a key component.

5.1 Secure Multi-party Computation

Secure multi-party computation (SMC) is a cryptographic technique that allows multiple parties to jointly perform computations on sensitive data while keeping their individual data private from one another. SMC involves complex multi-party protocols that provide strong privacy guarantees. In the context of PPRL, SMC-based techniques have been devised, as described in Christen et al. (2020). In addition, SMC can be used to perform linear regression on vertically partitioned data, where multiple parties share the same set of records but have different sets of features. In contrast to regression with linked data, a unique identifier is available, and thus no linkage errors are present.

5.2 Federated Learning

Federated learning (FL) (McMahan et al., 2017) is a distributed machine learning framework that collaboratively trains a shared model while ensuring that data from multiple sources remain decentralized. By performing computations locally and aggregating model updates, federated learning reduces privacy risks. It can offer enhanced privacy protection when combined with SMC (Mugunthan et al., 2019) and DP (Ouadrhiri and Abdelhadi, 2022). The most relevant type of FL to record linkage is vertical federated learning (VFL) where FL is applied to vertically partitioned data. Most works that combine VFL and record linkage train models using one-to-one deterministic linkage as opposed to probabilistic linkage, while Wu et al. (2022) integrates one-to-many linkage into the VFL training process. Nonetheless, the statistical implications and procedures in this area remain unexplored.

5.3 Privacy for Data Integration

Data integration refers to the process of combining data from multiple sources into a single, unified format. While it encompasses a broad range of activities, record linkage is a specific statistical task that plays a critical role within data integration. The main challenge in data integration is resolving heterogeneity at various levels, such as differences in data sources, schemas, data types, and semantics.

Privacy concerns in data integration can also be broadly divided into two categories: (1) privacy and security issues that arise during the integration process and (2) privacy risks associated with running statistical analyses on the integrated view. Existing works on privacy-preserving data integration (PPDI) have established a wide range of techniques to manage multi-layered heterogeneity (Shelake and Shekokar, 2017). PPDI primarily addresses the first set of challenges, while the design of private algorithms on the integrated view that quantify privacy loss and account for upstream integration uncertainties has not yet been explored.

References

Olivier Binette and Rebecca C. Steorts. (almost) all of entity resolution. *Science Advances*, 8(12):eabi8021, 2022. doi: 10.1126/sciadv.abi8021.

T. Tony Cai, Yichen Wang, and Linjun Zhang. The cost of privacy: Optimal rates of convergence for parameter estimation with differential privacy. *The Annals of Statistics*, 49(5):2825 – 2850, 2021.

Raymod Chambers, Enrico Fabrizi, and Nicola Salvati. Small area estimation with linked data. *Journal of the Royal Statistical Society Series B, Royal Statistical Society*, 83(1):78–107, 2021.

Raymond Chambers. Regression analysis of probability-linked data. *Technical report, Official Statistics Research, Statistics New Zealand*, 2009.

Peter Christen. *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection.* Springer Publishing Company, Incorporated, 2012. ISBN 3642311636.

Peter Christen, Thilina Ranbaduge, and Rainer Schnell. *Linking Sensitive Data: Methods and Techniques for Practical Privacy-Preserving Information Sharing.* Springer Cham, 01 2020. ISBN 978-3-030-59705-4. doi: 10.1007/978-3-030-59706-1.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2014.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In Shai Halevi and Tal Rabin, editors, *Theory of Cryptography*, pages 265–284. Springer Berlin Heidelberg, 2006.

Ivan P. Fellegi and Alan B. Sunter. A theory for record linkage. *Journal of the American Statistical Association*, 64(328):1183–1210, 1969. doi: 10.1080/01621459.1969.10501049.

Rob Hall and Stephen E. Fienberg. Privacy-preserving record linkage. In Josep Domingo-Ferrer and Emmanouil Magkos, editors, *Privacy in Statistical Databases*, pages 269–283. Springer Berlin Heidelberg, 2010. ISBN 978-3-642-15838-4.

Ying Han and Partha Lahiri. Statistical Analysis with Linked Data. *International Statistical Review*, 87(S1):139–157, 2019.

Xiao-Li Meng James Bailie, Ruobin Gong. Can swapping be differentially private? a refreshment stirred, not shaken (working paper). *NBER*, 2024.

Matthew A. Jaro. Advances in record-linkage methodology as applied to matching the 1985 census of tampa, florida. *Journal of the American Statistical Association*, 84:414–420, 1989.

Partha Lahiri and Michael D Larsen. Regression analysis with linked data. *Journal of the American Statistical Association*, 100(469):222–230, 2005. doi: 10.1198/016214504000001277.

Shurong Lin, Elliot Paquette, and Eric D. Kolaczyk. Differentially Private Linear Regression With Linked Data. *Harvard Data Science Review*, 6(3), jul 31 2024. https://hdsr.mitpress.mit.edu/pub/4if53bjq.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. In Aarti Singh and Jerry Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 1273–1282. PMLR, 20–22 Apr 2017. URL https://proceedings.mlr.press/v54/mcmahan17a.html.

Vaikkunth Mugunthan, Antigoni Polychroniadou, David Byrd, and Tucker Hybinette Balch. Smpai: Secure multi-party computation for federated learning. In *Proceedings of the NeurIPS 2019 Workshop on Robust AI in Financial Services*, volume 21. MIT Press Cambridge, MA, USA, 2019.

John Neter, E. Scott Maynes, and Ramu Ramanathan. The effect of mismatching on the measurement of response errors. *Journal of the American Statistical Association*, 60:1005–1027, 1965.

H. B. Newcombe, J. M. Kennedy, S. J. Axford, and A. P. James. Automatic linkage of vital records. *Science*, 130(3381):954–959, 1959. doi: 10.1126/science.130.3381.954.

Ahmed El Ouadrhiri and Ahmed Abdelhadi. Differential privacy for deep and federated learning: A survey. *IEEE Access*, 10:22359–22380, 2022. doi: 10.1109/ACCESS.2022.3151670.

Fritz Scheuren and William Winkler. Regression analysis of data files that are computer matched. *Survey Methodology*, 19:39–58, 1993.

Or Sheffet. Differentially private ordinary least squares. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning, ICML 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 3105–3114. PMLR, 2017.

Vijay Maruti Shelake and Narendra Shekokar. A survey of privacy preserving data integration. In *2017 International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques (ICEECCOT)*, pages 59–70, 2017. doi: 10.1109/ICEECCOT.2017.8284559.

Aleksandra Slavković and Jeremy Seeman. Statistical data privacy: A song of privacy and utility. *Annual Review of Statistics and Its Application*, 10 (Volume 10, 2023):189–218, 2023. ISSN 2326-831X. doi: https://doi.org/10.1146/annurev-statistics-033121-112921. URL `https://www.annualreviews.org/content/journals/10.1146/annurev-statistics-033121-112921`.

Rebecca C. Steorts, Rob Hall, and Stephen E. Fienberg. A bayesian approach to graphical record linkage and deduplication. *Journal of the American Statistical Association*, 111(516):1660–1672, 2016.

U.S. Census Bureau. Disclosure Avoidance for the 2020 Census: An Introduction, 2021.

U.S. Census Bureau. Annual Report of the Center for Statistical Research and Methodology, 2022.

Yu-Xiang Wang. Revisiting differentially private linear regression: optimal and adaptive prediction & estimation in unbounded domain. *Conference on Uncertainty in Artificial Intelligence (UAI)*, 49, 2018.

William Winkler. Machine learning and record linkage. In *58th World Statistics Congress ISI*, 2011.

William E. Winkler. String comparator metrics and enhanced decision rules in the fellegi-sunter model of record linkage. 1990. URL `https://api.semanticscholar.org/CorpusID:54580585`.

Zhaomin Wu, Qinbin Li, and Bingsheng He. A coupled design of exploiting record similarity for practical vertical federated learning. *Advances in Neural Information Processing Systems*, 35:21087–21100, 2022.