Using Provenance to Evaluate Risk and Benefit of Data Sharing

Taeho Jung
University of Notre Dame

Seokki Lee
University of Cincinnati

Wenyi Tang
University of Notre Dame

Abstract

Data sharing is becoming increasingly important, but how much risk and benefit is incurred from the sharing is not well understood yet. Certain existing models can be leveraged to partially determine the risk and benefit. However, such naïve ways of quantification are inaccurate because they fail to capture the context and the history of the datasets in data sharing. This paper suggests utilizing the data provenance to accurately and quantitatively model the risk and benefit of data sharing between two parties, and describes preliminary approaches as well as further issues to consider. **We limited the paper to 4 pages to allow a future full-length publication.**

1 Introduction

Data sharing is a valuable part in data intensive and collaborative environment due to the synergies created by multimodal datasets generated from different sources. We focus on the case where one party (*owner*) sends his/her structured datasets to another party (*recipient*). For notational simplicity, we assume the party that possesses datasets own them and ignore the distinction between possession and true ownership since the distinction is orthogonal to the problem discussed in this paper. Such sharing between two entities (e.g., companies and organizations) incurs security issues due to the sensitive information about individuals (e.g., health or financial information) or organizations (e.g., business secrets) contained in the datasets. For example, privacy researchers have demonstrated that diverse types of data, e.g., electricity meter readings [6] and hospital visits [21], can be linked back to individuals even when they are shared without names or other overt identifiers such as addresses or dates of birth. Due to such reasons, it is required that the owners sanitize the datasets before releasing them to ensure the disclosure risk is minimized without significantly damaging the utility of dataset. Ideally, the participants in data sharing want to maximize the benefit of data sharing while minimizing the risk of sharing. For that, they may want to quantitatively learn how much benefit they gain from the received datasets and how much risk exists in the sharing. Several aspects need to be considered to accurately gauage them.

• **Existing privacy evaluation can be improper.** Existing models for evaluating the privacy risk associated with the data (e.g., k-anonymity [19], l-divsertiy [16], t-closeness [14]) are modeled for the publication of a single table, and certain pre-processing (e.g., generalization and suppression) is performed to cluster individual records into several *equivalence classes* to ensure the table satisfies the privacy definition(s). In data sharing, multiple tables that are correlated to each other (e.g., derived from the same table) may be shared. However, such correlation is not considered in the data pre-processing. Furthermore, ensuring the privacy definitions (i.e., k, l, t values in the aforementioned models) for individual tables may not result in the desired privacy level. For example, as shown in $\mathbf{T_1}$ and $\mathbf{T_2}$ in Figure 1, one's record may be generalized/suppressed to different equivalence classes due to different distributions of the tables. Such inconsistent pre-processing may lead to either reduced utility (when the same record is represented by different equivalent classes, e.g., $\mathbf{T}^{\cup}$ in Figure 1) or reduced privacy (when different equivalent classes are linked and yield more specific equivalent classes, e.g., $\mathbf{T}^{\bowtie}$ in Figure 1).

• **The risk may change after the recipient integrates the shared datasets with the datasets s/he possesses.** Therefore, the publisher needs to take into account the recipient's context, i.e., the datasets s/he possesses, but the existing models do not capture it.

• **To the best of our knowledge, there are no such models that measure the benefits of data sharing.** Intuitively, the benefits of sharing are based on new information obtained from the integration of datasets between participants. The degree of new information can be estimated by measuring the (dis)similarity between datasets (e.g., [1]). However, it may result in incorrect

$\mathbf{T_1}$

Country	Gen.	Age	Income	Occup.
North America	F	< 30	>100k/yr	Faculty

$\mathbf{T_2}$

Country	Gen.	Age	Income	Occup.
Canada	XX	29	<140k/yr	Higher Ed.

$\mathbf{T}^{\cup}$ (when $\mathbf{T_1}, \mathbf{T_2}$ are merged without linkage)

Country	Gen.	Age	Income	Occup.
North America	F	< 30	>100k/yr	Faculty
Canada	XX	29	<140k/yr	Higher Ed.

* The same person appears in two records (utility loss).

$\mathbf{T}^{\bowtie}$ (when $\mathbf{T_1}, \mathbf{T_2}$ are integrated with linkage)

Country	Gen.	Age	Income	Occup.
Canada	F	29	100k/yr-140k/yr	Faculty

* The equivalent class becomes more specific (privacy risk).

Figure 1: Example of improper pre-processing.

estimation without considering where the information comes from (i.e., data provenance). For example, assume that, in Figure 1, $\mathbf{T_1}$ is a shared data and $\mathbf{T_2}$ is an owned data. Measuring the new information gained from the integrated data $\mathbf{T}^{\bowtie}$ over $\mathbf{T_2}$, e.g., using Jaccard similarity [10], would determine the entire records in $\mathbf{T}^{\bowtie}$ as new information (the fused values are not detected properly). Thus, we use provenance to capture such information, e.g., F in `Gen.` of $\mathbf{T}^{\bowtie}$ is derived from $\mathbf{T_1}$ but 29 in `Age` is actually from $\mathbf{T_2}$, to accurately measure the degree of new information.

This paper investigates how provenance can be used to (1) accurately sanitize correlated data for better utility and privacy, and (2) accurately evaluate the risk and benefit of data sharing. The core ideas can be applied to any sanitization approaches and any modeling of risk and benefit that are data-centric (i.e., defined for individual datasets). An in-depth understanding of risk and benefit in data sharing can inform the data owners to make better decisions and contribute to the development of safe and secure data sharing ecosystem.

2 Related Work

Modeling of Disclosure Risk k-Anonymity [19], l-divsertiy [16], and t-closeness [14] are well-known models which quantify the re-identification risk of a given sanitized dataset. These models capture different types of privacy guarantees. k-Anonymity requires each equivalence class contain at least k records with identical quasi-identifiers and aims at preventing identity disclosure. l-Diversity further considers the sensitive attributes by requiring that in any equivalence class, each sensitive value can occur with a frequency of at most $\frac{1}{l}$. t-Closeness addresses the l-diversity's limitations of overly-strong/weak confidence bound by requiring the sensitive attribute distribution in each equivalence class be close to that in the overall data.

There are also models quantifying the re-identification risk by *record linkability* (e.g., based on values [17], rules [22], probabilistic analysis [8], and data mining and machine learning [5]), which measures the ability of matching between original and shared records and, thus, quantifies the privacy level of shared records.

For the tables that are released with ϵ-differential privacy or ϵ, δ-differential privacy [15, 20, 23], it is also possible to quantify the disclosure risk with the ϵ and/or δ. Though providing rigorous theoretical guarantees, such protection towards data sharing has a specific query function or utility function towards which randomization/perturbation mechanisms are tailored. Therefore, it is not suitable for general-purpose data sharing.

Modeling of Information Loss In general, there are mainly two approaches to measure utility of data. (1) One measures the amount of utility that is remained in sanitized data. This includes measures such as the average size of the equivalence classes [16] and the discernibility metric [2], and the approach of evaluating data utility in terms of data mining workloads. (2) The other measures the loss of utility due to data anonymization. This is measured by comparing the anonymized data with the original data. It includes measures such as the number of generalization steps and the KL-divergence between the reconstructed distribution and the true distribution for all possible quasi-identifier values [11].

3 Problem Definition

Suppose there exist two parties, data owner $\mathfrak{o}$ and third party $\mathfrak{p}$. The owner $\mathfrak{o}$ shares a set of tables $\mathcal{T}_{\mathfrak{o}} = \{T_1, T_2, \cdots\}$ with $\mathfrak{p}$ who possesses another set of tables $\mathcal{T}_{\mathfrak{p}} = \{T'_1, T'_2, \cdots\}$. We assume a finite set of algorithms of data integration is available to both $\mathfrak{o}$ and $\mathfrak{p}$ which is denoted as $\mathcal{A} = \{A_1, \cdots, A_n\}$. Finally, we use $\mathcal{T}_{\mathfrak{p} \bowtie_{\mathcal{A}} \mathfrak{o}} = \{T_1^{\bowtie}, T_2^{\bowtie}, \cdots\}$ to denote the resulting datasets achieved by integrating $\mathcal{T}_{\mathfrak{o}}$ and $\mathcal{T}_{\mathfrak{p}}$ over $\mathcal{A}$.

We informally define the problem as follows: (1) sanitizing $\mathcal{T}_{\mathfrak{o}}$ in a holistic way such that the aforementioned inconsistent sanitization is prevented (i.e., the same record is represented by the same equivalence class); (2) defining a quantitative model that measures the *additional risk* (AR) caused at $\mathfrak{p}$'s end over $\mathcal{T}_{\mathfrak{p} \bowtie_{\mathcal{A}} \mathfrak{o}}$; and (3) defining a quantitative model that measures the *information gain* (IG) over $\mathcal{T}_{\mathfrak{p} \bowtie_{\mathcal{A}} \mathfrak{o}}$.

4 Sanitization with Provenance

As mentioned in Section 1, if data sanitization (e.g., generalization and suppression) is performed on each individual table independently, it may result in inconsistent records. This occurs when two tables T_i and T_j have correlated data (e.g., both are derived from the

same table or one is derived from the other). We use provenance as follows. For each table T_i in $\mathcal{T}_{\mathfrak{o}}$, the owner $\mathfrak{o}$ analyzes the provenance of T_i and identifies all the tables that have dependencies. The tables with dependencies are clustered into several groups, and the tables in each group are integrated into one single table that represents the group. For example, if T_1 and T_2 are related and T_2 and T_3 are related, but T_1 and T_3 may (not) be directly related, we cluster all T_1,T_2, and T_3 into one group. By doing so, one record will be represented by one equivalence class only, and we will not have inconsistency issues in the pre-processing. We use $\mathcal{T}_{\mathfrak{o}\bowtie_{\mathcal{A}}} = \{T_1^{\circ}, T_2^{\circ}, \cdots\}$ to denote the set of representative tables derived from such integration performed by $\mathfrak{o}$. Then, $\mathfrak{o}$ sanitizes each table in $\mathcal{T}_{\mathfrak{o}\bowtie_{\mathcal{A}}}$, e.g., such that they have desired k-anonymity, l-diversity, and/or t-closeness based on existing approaches and shares $\mathcal{T}_{\mathfrak{o}\bowtie_{\mathcal{A}}}$ with $\mathfrak{p}$. Note that any other data-centric approaches (e.g., [3,18]) can be applied with/instead of the one described above.

5 Measuring AR

Any data-centric risk metric that is defined for the given dataset can be used, and we will leverage the aforementioned disclosure risk of sanitized data in this paper for illustration purposes only. We measure AR by the worst-case differences of the privacy metrics (e.g., k,l,t values in the k-anonymity, l-diversity, and t-closeness) between the original sanitized representative datasets published by $\mathfrak{o}$, i.e., $\mathcal{T}_{\mathfrak{o}\bowtie_{\mathcal{A}}}$ and the resulting datasets $\mathcal{T}_{\mathfrak{p}\bowtie_{\mathcal{A}}\mathfrak{o}}$. We stress that, before $\mathfrak{p}$ integrates $\mathcal{T}_{\mathfrak{p}}$ with $\mathcal{T}_{\mathfrak{o}\bowtie_{\mathcal{A}}}$, the recipient $\mathfrak{p}$ needs to analyze the provenance records of $\mathcal{T}_{\mathfrak{p}}$ and generate the representative tables to avoid the aforementioned inconsistency. We use $\mathcal{T}_{\mathfrak{p}\bowtie_{\mathcal{A}}}$ to denote the set of representative tables generated by $\mathfrak{p}$.

In our example, one can define AR with the difference between the minimum k,l,t values in $\mathcal{T}_{\mathfrak{o}\bowtie_{\mathcal{A}}}$ and the minimum k,l,t values in $\mathcal{T}_{\mathfrak{p}\bowtie_{\mathcal{A}}\mathfrak{o}}$ (which is generated by integrating $\mathcal{T}_{\mathfrak{o}\bowtie_{\mathcal{A}}}$ and $\mathcal{T}_{\mathfrak{p}\bowtie_{\mathcal{A}}}$). It is thus defined as a tuple of three values (Equation (1)). Note that $k(T)$,$l(T)$, and $t(T)$ denote the k,l,t values of T in k-anonymity, l-diversity, and t-closeness respectively. This metric can be used to quantify the additional risk caused by the sharing. The higher the AR is in each dimension, the greater the additional risk is. Note that there can be negative values when the values increase after the sharing, which indicates the risk is actually reduced.

$$AR = \Big(\min_{T_i^{\circ} \in \mathcal{T}_{\mathfrak{o}\bowtie_{\mathcal{A}}}} \big(k(T_i^{\circ})\big) - \min_{T_i^{\bowtie} \in \mathcal{T}_{\mathfrak{p}\bowtie_{\mathcal{A}}\mathfrak{o}}} \big(k(T_i^{\bowtie})\big), \min_{T_i^{\circ} \in \mathcal{T}_{\mathfrak{o}\bowtie_{\mathcal{A}}}} \big(l(T_i^{\circ})\big) - \min_{T_i^{\bowtie} \in \mathcal{T}_{\mathfrak{p}\bowtie_{\mathcal{A}}\mathfrak{o}}} \big(l(T_i^{\bowtie})\big), \min_{T_i^{\circ} \in \mathcal{T}_{\mathfrak{o}\bowtie_{\mathcal{A}}}} \big(t(T_i^{\circ})\big) - \min_{T_i^{\bowtie} \in \mathcal{T}_{\mathfrak{p}\bowtie_{\mathcal{A}}\mathfrak{o}}} \big(t(T_i^{\bowtie})\big) \Big) \quad (1)$$

6 Measuring IG using Provenance

The purpose of data sharing lies partly in the record linkages resulted from the sharing of multimodal datasets, because $\mathfrak{p}$ gains extra information from $\mathcal{T}_{\mathfrak{p}\bowtie_{\mathcal{A}}\mathfrak{o}}$. We define information gain (IG) based on the new values appearing in $\mathcal{T}_{\mathfrak{p}\bowtie_{A_i}\mathfrak{o}}$ over $\mathcal{T}_{\mathfrak{o}}$ and $A_i \in \mathcal{A}$. To measure IG of sharing, we capture different types of provenance, e.g., where and why provenance and column dependencies over $\mathcal{T}_{\mathfrak{p}}$ and $\mathcal{T}_{\mathfrak{o}}$ that contribute to those in $\mathcal{T}_{\mathfrak{p}\bowtie_{A_i}\mathfrak{o}}$ for each A_i. Assuming $\mathbf{T_1}$ in Figure 1 is the shared data, IG of the record in $\mathbf{T}^{\bowtie}$ should be calculated based on the values in `Gen.`, `Income`, and `Occup.` attributes that are derived from the record in $\mathbf{T_1}$.

Adapting the concept of extensional (changes in rows) and intensional (changes in columns) completeness in [4] and using provenance, we measure newly introduced values from common columns that exist in $T_l' \in \mathcal{T}_{\mathfrak{p}}$, $T_m \in \mathcal{T}_{\mathfrak{o}}$, and $\mathcal{T}_{\mathfrak{p}\bowtie_{A_i}\mathfrak{o}}$ and through newly added columns from shared $\mathcal{T}_{\mathfrak{o}}$. We extend the informativeness metric in [13] for computing the IG of each record in $\mathcal{T}_{\mathfrak{p}\bowtie_{A_i}\mathfrak{o}}$ and use it as a baseline method for measuring column changes of each record. We, then, generalize the IG of $\mathcal{T}_{\mathfrak{p}\bowtie_{A_i}\mathfrak{o}}$ as the harmonic mean of measures of changes in rows and columns. This method can be applied to any data-centric metric for measuring IG that may exist. Note that the generalization method should be independent of $\mathcal{A}$ such that the computed IG of each integrated data is directly comparable. For example, we may be able to compute IG of $\mathcal{T}_{\mathfrak{p}\bowtie_{A_i}\mathfrak{o}}$ based on the IG of each record in it. However, generalized IG over $\forall A_i \in \mathcal{A}$ may not be directly comparable because IG of $\mathcal{T}_{\mathfrak{p}\bowtie_{A_i}\mathfrak{o}}$ and $\mathcal{T}_{\mathfrak{p}\bowtie_{A_j}\mathfrak{o}}$ may differ although they contain the same amount of new information.

7 Issues to Consider Further

Overhead of Model Evaluation To use the aforementioned AR and IG modeling, the owner $\mathfrak{o}$ and the third party $\mathfrak{p}$ engaged in data sharing need to (1) perform integration on their own datasets before the sharing to generate representative tables and (2) compute the metrics over all integrated datasets $\mathcal{T}_{\mathfrak{p}\bowtie_{\mathcal{A}}\mathfrak{o}}$ after the sharing. The proposed models would not be practical if their computation complexities are prohibitively high and the datasets are too large. To address them, we apply the following solutions.

Firstly, the dataset similarity measurement [10], which is based on the MinHash algorithms [12], will be leveraged. These approaches generate semantic-preserving hash values from a table, based on which the similarity between two tables can be measured extremely efficiently. Due to the law of large numbers, the similarity between

tables can be approximated with bounded errors and complexity that are independent from the size of the tables. Since the hash values from the MinHash algorithms can be used to approximately estimate how many common data elements (e.g., tuples and subtuples in each table) the tables have, we will use such hash values to estimate the difference of the aforementioned k,l,t values between the shared dataset $\mathcal{T}_{\mathfrak{o}}$ and the integrated datasets $\mathcal{T}_{\mathfrak{p}\bowtie_{\mathcal{A}}\mathfrak{o}}$. This is possible since the k,l,t values are based on the tuples and the subtuples of the tables, whose similarity has been successfully estimated via the MinHash algorithms. We will use the similarity between the tuples and subtuples of $\mathcal{T}_{\mathfrak{o}}$ and $\mathcal{T}_{\mathfrak{p}\bowtie_{\mathcal{A}}\mathfrak{o}}$ to estimate the k,l,t values efficiently at the errors and the complexity independent from the size of the datasets.

Secondly, approximate summaries of provenance records [13] will be leveraged to efficiently evaluate IG of $\mathcal{T}_{\mathfrak{p}\bowtie_{\mathcal{A}}\mathfrak{o}}$. We can naively measure the IG using the aforementioned model such that evaluating all the changes in rows (quantity of information) and columns (quality of information) from each integrated data $\mathcal{T}_{\mathfrak{p}\bowtie_{A_i}\mathfrak{o}}$ comparing against $\mathcal{T}_{\mathfrak{p}}$ and $\mathcal{T}_{\mathfrak{o}}$ (then, $\mathcal{T}_{\mathfrak{p}\bowtie_{\mathcal{A}}\mathfrak{o}}$ is the set of IGs for all $\mathcal{T}_{\mathfrak{p}\bowtie_{A_i}\mathfrak{o}}$ where $A_i \in \mathcal{A}$). However, it is computationally not feasible. By extending summarization technique in [13], we trade-off quality of IG for computational performance. We compute a provenance summary for $\mathcal{T}_{\mathfrak{p}\bowtie_{A_i}\mathfrak{o}}$ using a sample of provenance and patterns which represent sets of provenance. We, then, evaluate approximate IG of $\mathcal{T}_{\mathfrak{p}\bowtie_{A_i}\mathfrak{o}}$ over the summaries while assuring that the approximate IG is as close as possible to the actual IG. The provenance summaries are also used to reduce the computational space. Assume that we have a set of patterns p_1 computed over $T'_j \bowtie_{A_i} T_l$ and p_2 generated over $T'_j \bowtie_{A_i} T_m$ where $T'_j \in \mathcal{T}_{\mathfrak{p}}$, $\{T_l, T_m\} \in \mathcal{T}_{\mathfrak{o}}$, and $A_i \in \mathcal{A}$. If $p_2 \subset p_1$, we do not have to measure the IG of $T'_j \bowtie_{A_i} T_m$. There are several challenges: (1) While computing patterns over sample(s) of provenance of $\mathcal{T}_{\mathfrak{p}\bowtie_{A_i}\mathfrak{o}}$, we may lose some new information that is obtained over A_i. Thus, we should keep new information as much as possible in order to achieve the approximate IG close enough to the actual IG; (2) Obtaining the minimal set of summaries over $T'_j \bowtie_{A_i} T_l$ for all $T'_j \in \mathcal{T}_{\mathfrak{p}}$ and $A_i \in \mathcal{A}$ is challenging because there may exist overlaps among the integrated datasets. We can find a superset of the optimal set by estimating the bounds of IGs and develop an efficient method to speed up the evaluation by reusing the provenance (summaries).

Retroactive Evaluation While such evaluation can help mitigate the high risk caused by data sharing and also allow third party $\mathfrak{p}$ to understand the benefit, it is a retroactive evaluation with certain limitations.

Firstly, the data sharing that incurs high AR (i.e., the k,l,t values become significantly lower in the datasets integrated by $\mathfrak{p}$ after the sharing) may result in individual disclosure. If $\mathfrak{p}$ is adversarial (e.g., due to insider threats or compromise attacks), the attackers inside $\mathfrak{p}$ may have access to the integrated datasets before the rest of $\mathfrak{p}$ take further steps to perform extra sanitization after receiving the dataset. The attackers thereby gain extra benefits in deanonymizing the sanitized data.

Secondly, if the data sharing is part of trading transactions (e.g., $\mathfrak{p}$ pays $\mathfrak{o}$ fees/royalties to receive the datasets), there can be fairness issues when the extra benefit is too low or too high. When the extra benefit is shown to be too low, malicious $\mathfrak{o}$ may refuse to refund the transaction amounts or share additional datasets to ensure $\mathfrak{p}$ receives enough benefit from the sharing. On the other hand, when the extra benefit is too high, malicious $\mathfrak{p}$ may refuse to pay extra fees/royalties.

All of issues arise because the evaluation can be done in a retroactive manner only. One potential solution towards this limitation is to perform the evaluation before sharing the actual datasets. This is possible if one leverages trusted execution environment such as Intel SGX [7] to allow the two parties to learn the additional risk and information gain without learning each other's datasets. Namely, they can both securely their datasets to each others' SGX enclaves such that the AR/IG evaluation can be done securely within the enclaves that are inaccessible by $\mathfrak{o}$ or $\mathfrak{p}$. The calculated AR/IG can be returned to $\mathfrak{o},\mathfrak{p}$ with the signatures generated inside the enclaves, which would show the integrity. Though being technically feasible, such an approach incurs prohibitively large memory I/O overhead due to the small capacity of the SGX enclaves, therefore further investigation is needed to make such approaches more practical.

Honesty Requirement This paper assumes the owner $\mathfrak{o}$ and the third party $\mathfrak{p}$ honestly perform the calculation to measure the AR and IG correctly. If either party is malicious, the AR and IG values would be meaningless. For example, in the scenario of dataset trading, $\mathfrak{p}$ who pays for the dataset access may be motivated to claim the extra benefit is too low to get monetary benefit. Certain authentication mechanisms need to be in place to vet the parties and ensure that they will not engage in such misbehavior.

Alternatively, it is technically feasible to rely on zero-knowledge proofs such as GROTH16 [9] which is one of the most efficient implementations of Zero-Knowledge Succinct Non-Interactive Argument of Knowledge (zk-SNARK). Namely, both $\mathfrak{o}$ and $\mathfrak{p}$ can prove to each other without disclosing their datasets that the AR and IG have been calculated correctly. However, similar to any other cryptographic approaches, such an approach would incur prohibitively larger overhead.

References

[1] Nikolaus Augsten and Michael H Böhlen. Similarity joins in relational database systems. *Synthesis Lectures on Data Management*, 5(5):1–124, 2013.

[2] Roberto J Bayardo and Rakesh Agrawal. Data privacy through optimal k-anonymization. In *ICDE*, pages 217–228. IEEE, 2005.

[3] Khalid Belhajjame. Lineage-preserving anonymization of the provenance of collection-based workflows. In *EDBT*, pages 229–240, 2020.

[4] Jens Bleiholder and Felix Naumann. Data fusion. *ACM computing surveys (CSUR)*, 41(1):1–41, 2009.

[5] Katherine E Boronow, Laura J Perovich, Latanya Sweeney, Ji Su Yoo, Ruthann A Rudel, Phil Brown, and Julia Green Brody. Privacy risks of sharing data from environmental health studies. *Environmental health perspectives*, 128(1):017008, 2020.

[6] Erik Buchmann, Klemens Böhm, Thorben Burghardt, and Stephan Kessler. Re-identification of smart meter data. *Personal and ubiquitous computing*, 17(4):653–662, 2013.

[7] Victor Costan and Srinivas Devadas. Intel sgx explained. *IACR Cryptol. ePrint Arch.*, 2016(86):1–118, 2016.

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

[9] Jens Groth. On the size of pairing-based non-interactive arguments. In *Annual international conference on the theory and applications of cryptographic techniques*, pages 305–326. Springer, 2016.

[10] Taeho Jung, Xiang-Yang Li, Wenchao Huang, Zhongying Qiao, Jianwei Qian, Linlin Chen, Junze Han, and Jiahui Hou. Accounttrade: Accountability against dishonest big data buyers and sellers. *IEEE Transactions on Information Forensics and Security*, 14(1):223–234, 2019.

[11] Daniel Kifer and Johannes Gehrke. Injecting utility into anonymized datasets. In *SIGMOD*, pages 217–228, 2006.

[12] David Koslicki and Hooman Zabeti. Improving minhash via the containment index with applications to metagenomic analysis. *Applied Mathematics and Computation*, 354:206–215, 2019.

[13] Seokki Lee, Bertram Ludäscher, and Boris Glavic. Approximate summaries for why and why-not provenance. *VLDB*, 13(6).

[14] Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *ICDE*, pages 106–115. IEEE, 2007.

[15] Qi Li, Yuqiang Li, Guicai Zeng, and Aihua Liu. Differential privacy data publishing method based on cell merging. In *2017 IEEE 14th International Conference on Networking, Sensing and Control (ICNSC)*, pages 778–782. IEEE, 2017.

[16] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. *TKDD*, 1(1):3–es, 2007.

[17] Gonzalo Navarro. A guided tour to approximate string matching. *ACM computing surveys (CSUR)*, 33(1):31–88, 2001.

[18] Mehmet Ercan Nergiz, Christopher Clifton, and Ahmet Erhan Nergiz. Multirelational k-anonymity. *IEEE Transactions on Knowledge and Data Engineering*, 21(8):1104–1117, 2008.

[19] Pierangela Samarati and Latanya Sweeney. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. 1998.

[20] Jordi Soria-Comas and Josep Domingo-Ferrert. Differential privacy via t-closeness in data publishing. In *2013 Eleventh Annual Conference on Privacy, Security and Trust*, pages 27–35. IEEE, 2013.

[21] Latanya Sweeney. Matching known patients to health records in washington state data. *Available at SSRN 2289850*, 2013.

[22] William E Winkler. Matching and record linkage. *Business survey methods*, 1:355–384, 1995.

[23] Tianqing Zhu, Gang Li, Wanlei Zhou, and S Yu Philip. Differentially private data publishing and analysis: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 29(8):1619–1638, 2017.