シャッフルモデルにおける局所差分プライバシに基づく垂直分割データのプライバシ保護データ合成

紀伊真昇[1,a)]　市川敦謙[1]　三浦尭之[1]　山本充子[1]　千田浩司[2]

概要：各組織が保有するパーソナルデータを横断的に利活用するためには，データを提供する個人のプライバシ保護が欠かせない．そこで筆者らは SCIS2022 において，各組織に対しても匿名性を保ちつつデータ連結できる手法を提案したが，従来の匿名化に基づく場合，属性数が多くなるにつれ有用性が著しく悪化する問題があった．本稿ではその対策として，シャッフルモデルにおける局所差分プライバシに基づく匿名化手法を提案する．提案方式のプライバシ予算の消費量の評価および公開データを用いた実験評価により，提案方式の安全性および有用性を定量的に示す．またプライバシ保護データ合成への応用可能性についても検討する．

キーワード：垂直分割データの協調匿名化，局所差分プライバシ，シャッフルモデル，プライバシ保護データ合成

Privacy-Preserving Data Synthesization from Vertical Partitioned Data with Local Differential Privacy in the Shuffled Model

Masanobu Kii[1,a)]　Atsunori Ichikawa[1]　Takayuki Miura[1]　Juko Yamamoto[1]　Koji Chida[2]

Abstract: Privacy is of utmost concern when utilizing across the personal data sets held by each organization. At SCIS2022, the authors presented a protocol to join the databases while fulfilling anonymity even for the organizations; however, there remains a major obstacle that the conventional anonymization methods applicable in the proposed scheme drastically decrease the utility as the number of attributes increases. As a countermeasure, we propose an anonymization method satisfying local differential privacy in the shuffled model. We quantitatively demonstrate the privacy and utility of the proposed method by evaluating the amount of consumption of privacy budget and an experimental result of our method obtained from public data. We also discuss a privacy-preserving data synthesization algorithm using our method.

Keywords: collaborative anonymization for vertical partitioned data, local differential privacy, shuffled model, privacy-preserving data synthesization

1. はじめに

データ収集デバイスの多様化や AI の進化等に伴い，社会課題の解決やサービスの質向上に資する手段としてパーソナルデータの利活用が注目を集めている．一方，データ提供者となる個人のプライバシ意識も高まっており，パーソナルデータの利活用は，倫理，法規制，運用管理，技術と様々な観点をふまえた適切な対応が強く求められる．特に組織間で同一個人のデータを連結して組織横断的に利活用する場合は，新たな価値創造が期待できる反面，データ連結と匿名性のトレードオフ問題が生じる．すなわち，ある組織がパーソナルデータを匿名化して別の組織に提供する場合，一般に同一個人のデータを連結するためのキー[*1](以

[1] NTT 社会情報研究所
NTT Social Informatics Laboratories
[2] 群馬大学 情報学部
Faculty of Informatics, Gunma University
[a)] masanobu.kii.gw@hco.ntt.co.jp

[*1] 携帯電話番号，会員番号，氏名と住所の組等，特定の個人を一意

降，連結キーと呼ぶ) が匿名性を損ねてしまう．この問題は，**垂直分割データの協調匿名化問題**として知られている [1]．データを提供した個人の同意を得る方法もあるが，データ連結する相手組織が多くなるほど，同意を得ることはより難しくなると考えられる．

筆者らは SCIS2022 において，垂直分割データの協調匿名化問題の対策として，各組織が保有する連結キーの仮名化，その他の属性データの非識別化に加え，セキュアマルチパーティ計算または Pk-匿名化 [2] を適用したプロトコルを提案した [3](詳細は次節)．しかし属性数が多くなると，非識別化や Pk-匿名化による有用性の悪化が大きな問題となる．

本稿では，**シャッフルモデルにおける局所差分プライバシ** [4] に基づくランダム化手法を提案する．シャッフルモデルは，複数人のデータをシャッフル (ランダム置換) することで匿名性を高め，局所差分プライバシ (LDP: Local Differential Privacy) のプライバシ予算を増幅 (言い換えればプライバシ予算の消費量を抑制) できるモデルとして近年注目されている．本稿では，属性間の積演算の推定に適したランダム化のメカニズムを提案し，理論評価および公開データを用いた実験評価によりその適用効果を示す．特に提案方式の安全性および有用性を定量的に評価するため，プライバシ予算の消費量について考察する．また提案方式の応用として，ニューラルネットワークやベイジアンネットワーク等を用いて多属性データの特徴抽出に優れた手法として注目されているデータ合成への適用について検討する．

2. 先行研究

前節で述べた垂直分割データの協調匿名化問題に対して筆者らは，分割データを保有する各組織に対しても匿名性を保ちつつ，同一個人データを連結した匿名化データを作成するプロトコルを提案した [3]．以下にその概要を紹介する．

先ず，組織 A,B はそれぞれ表形式データ (テーブル) からなるパーソナルデータ T_A, T_B を入力し，下記の基本プロトコル (Protocol 1) を実行して，仮名付きの非識別化データ T'_A, T'_B を得る．ここで T_A, T_B の各行 (レコード) $\mathbf{x}_{A,i}$, $\mathbf{x}_{B,j}$ $(i = 1, ..., n_A\ ;\ j = 1, ..., n_B)$ は特定個人のデータとし，T_A, T_B の各列 $\mathbf{c}_{A,u}$, $\mathbf{c}_{B,v}$ $(u = 0, 1, ..., m_A\ ;\ v = 0, 1, ..., m_B)$ は所与の属性のデータ，ただし $\mathbf{c}_{A,0}$, $\mathbf{c}_{B,0}$ は連結キーとする．また非識別化とは，元のデータとの照合が困難になるような処理を指し，例として k-匿名化が挙げられる．仮名は変換表やハッシュ関数に入力するソルト等の追加情報を用いて，プロトコルの実行毎に独立かつランダムに生成される．追加情報はプロトコル内で削除されるため，プロトコル実行後は仮名を生成できず，連結キーと仮名との対応関係も判別できない．すなわち仮名化とは，追加情報 Info をパラメータとする連結キーから仮名への写像 $\mathcal{R}_{\mathrm{Info}}$ であり，Info が与えられない限り任意の連結キー c および仮名 c' について $c' = \mathcal{R}_{\mathrm{Info}}(c)$ が成り立つかどうか判別困難という性質を持つ．

Protocol 1 基本プロトコル [3]

[入力] A: T_A, B: T_B (パーソナルデータ)

[出力] A: T'_A, B: T'_B (仮名付き非識別化データ)

(1) 仮名化に必要な追加情報 Info を共有する．

(2) A,B はそれぞれ以下を行う (‘ * ’ は A,B の何れか)．

- (a) 連結キー $\mathbf{c}_{*,0}$ を仮名化し，得られた仮名 $\mathbf{c}'_{*,0}$ を連結キーと置き換える．
- (b) 仮名以外のテーブル $T_*/\mathbf{c}'_{*,0}$ を非識別化したテーブル $\tilde{T}'_*$ を作成する．
- (c) $\tilde{T}'_*$ の各レコード $\mathbf{x}_{*,i}$ をランダムに置換し (置換関数を π_* とする)，テーブル $T'_* = \pi_*(\tilde{T}'_*)$ を出力する．
- (d) Info, π_* を削除する[*2](これにより T'_* は，組織 A,B に対しても T_* との照合が困難なテーブルとなる)．

ステップ (2b) で得られる仮名は，連結キーと 1 対 1，または別々の連結キーの仮名が等しくなる確率は無視できるほど小さいとする．すると組織 A,B の出力テーブル T'_A, T'_B を共通の仮名 $\mathbf{c}'_{A,0} \cap \mathbf{c}'_{B,0}$ で連結することで，同一個人データの連結が可能となる．また 3 節で詳しく触れるが，ステップ (2c) の置換が，シャッフルモデルにおける局所差分プライバシを満たすための有効な処理となる．

一方，Protocol 1 の出力 T'_A, T'_B を共通の仮名で連結したデータ $\tilde{T}'_{AB}$ は，組織 A,B それぞれの入力データ T_A, T_B とは照合困難であっても，$\tilde{T}'_{AB}$ 自体の匿名性が保証されるものではない (例えば T'_A, T'_B がそれぞれ k-匿名化されていても，仮名で連結すると一般に k-匿名性を満たさないことは明らか)．そこで [3] では，以下の 2 通りの追加処理が提案されている．

(1) セキュアマルチパーティ計算を用いて，互いの出力テーブルを秘匿しつつデータ連結および匿名化を行う (すなわち T'_A, T'_B を暗号化により秘匿しつつ $\tilde{T}'_{AB}$ の匿名化データ T'_{AB} を求める)．

(2) 非識別化処理として，データ連結しても k-匿名性の保証が可能な Pk-匿名化を用いる．

前者のデータ連結の具体例として，連結キーを暗号化により秘匿しつつ共通集合を求める秘匿共通集合プロトコル

に識別する情報．本稿では各組織が保有していると仮定する．

[*2] [3] には π_* の削除について言及されていないが，匿名性の観点で明らかに必要な処理であるため，本稿では追記している．

の応用が挙げられている．なお前記の基本プロトコルで仮名化や非識別化を行わず追加処理 (1) を実行することも可能だが，現状では国内外の関連法規制の多くが個人情報は暗号化の有無に依らないとしていることから，その効果は限定的と言える．

後者の Pk-匿名化は，データのランダム化 (データの確率的な置き換えやノイズ付加等) の強度をパラメータとして k-匿名性を満たす手法である．組織 A,B のテーブル T_A, T_B の属性数 m_A, m_B について，前記の基本プロトコルの非識別化として Pk-匿名化を用いる場合は，組織 A,B はそれぞれ m_A, m_B に依存した強度のランダム化を行う．これに対し，連結後も k-匿名性を満たすためには，組織 A,B がともに $m_A + m_B$ に依存した強度のランダム化を行う．ただし一般に属性数が多いほど強いランダム化が必要となり，有用性が著しく悪化してしまう．

次に Nguyễn らが提案した LDP に基づくメカニズム [5] および Wang らによるその変形版 [6], [7] を紹介する．単一の数値属性 $x_i \in [-1, 1]$ をランダム化する場合における，LDP のパラメータ (プライバシ予算) ε を入力とした Nguyễn らのメカニズムおよび Wang らのメカニズム (PM: Piecewise Mechanism) をそれぞれ Algorithm 1, 2 に記す．

Algorithm 1 Nguyễn らのメカニズム (単一数値属性)[5]

Input: $x_i \in [-1, 1]$, ε
Output: $x'_i \in \{-\frac{e^\varepsilon+1}{e^\varepsilon-1}, \frac{e^\varepsilon+1}{e^\varepsilon-1}\}$

Sample $w \in \{0, 1\}$ with $\Pr[w = 1] = \frac{e^\varepsilon-1}{2e^\varepsilon+2} \cdot x_i + \frac{1}{2}$
if $w = 1$ **then**
  $x'_i \leftarrow \frac{e^\varepsilon+1}{e^\varepsilon-1}$
else
  $x'_i \leftarrow -\frac{e^\varepsilon+1}{e^\varepsilon-1}$
end if
return x'_i

Algorithm 2 Wang らの PM(単一数値属性)[6]

Input: $x_i \in [-1, 1]$, ε
Output: $x'_i \in [-C, C]$ where $C = \frac{e^{\varepsilon/2}+1}{e^{\varepsilon/2}-1}$

Sample $w \in_U [0, 1]$
if $w < \frac{e^{\varepsilon/2}}{e^{\varepsilon/2}+1}$ **then**
  Sample $x'_i \in_U [\ell(x_i), r(x_i)]$
  where $\ell(x_i) = \frac{C+1}{2} \cdot x_i - \frac{C-1}{2}$, $r(x_i) = \ell(x_i) + C - 1$
else
  Sample $x'_i \in_U [-C, \ell(x_i)) \cup (r(x_i), C]$
end if
return x'_i

Algorithm 1, 2 はともに ε-LDP を満たし，出力 x'_i の期待値は x_i となることが示されている．そして Algorithm 1, 2 の出力の分散 $V(x'_i)$ はそれぞれ $\left(\frac{e^\varepsilon+1}{e^\varepsilon-1}\right)^2 - x_i^2$, $\frac{x_i^2}{e^{\varepsilon/2}-1} + \frac{e^{\varepsilon/2}+3}{3(e^{\varepsilon/2}-1)^2}$ となる．特に最悪ケース (Algorithm 1 は $x_i = 0$，Algorithm 2 は $x_i = \pm 1$) において，$\varepsilon \geq 1.29$ であれば Algorithm 2 が優位となる．[6] では，Algorithm 1 は ε がおよそ 2 以下であればラプラスメカニズム [8] を用いた LDP(以降，ラプラス LDP メカニズムと呼ぶ) よりも優れるが，Algorithm 2 は ε の値に関わらずラプラス LDP メカニズムよりも優位であることが示されている．

Wang らは Algorithm 1, 2 を効果的に組み合わせて分散を抑えるアルゴリズム (HM: Hybrid Mechanism) を提案している．また，$\{x_i\}_{i=1}^n$ の平均値 μ を $\mu' = \frac{\sum_{i=1}^n x'_i}{n}$ と推定したとき，Algorithm 1, 2 における平均値の誤差 $|\mu - \mu'|$ は，$\beta \in [0, 1]$ をパラメータとして $1 - \beta$ 以上の確率で $O\left(\frac{\sqrt{\ln(1/\beta)}}{\varepsilon\sqrt{n}}\right)$ を満たすことが示されている．

3. 提案方式

本節では，垂直分割データの協調匿名化問題の対策に向けた [3] の改良方式を提案する．LDP に基づく数値属性のランダム化は，筆者らが知る限り，2 節で紹介した Algorithm 1, 2 およびそれらを組み合わせたメカニズムが平均値の有用性に優れる．しかし，分散や属性間の共分散等に必要な積演算については明らかでない．そこで LDP に基づく数値属性のランダム化について，Algorithm 1 の自然な拡張となる新たなメカニズム (Algorithm 3) を提案し，Protocol 1 の非識別化処理として適用するとともに，積演算の有用性に優れることを示す．

Algorithm 3 提案メカニズム (単一数値属性のランダム化)

Input: $x_i \in [-1, 1]$, $(a, b) \in \mathbb{R}^2$ where $0 < a \leq b$
Output: $x'_i \in \{-\frac{b}{a}, \frac{b}{a}\}$

Sample $w \in \{0, 1\}$ with $\Pr[w = 1] = \frac{ax_i+b}{2b}$
if $w = 1$ **then**
  $x'_i \leftarrow \frac{b}{a}$
else
  $x'_i \leftarrow -\frac{b}{a}$
end if
return x'_i

Algorithm 3 は $a = e^\varepsilon - 1$, $b = e^\varepsilon + 1$ とすれば Algorithm 1 と等価になる．安全性については以下が成り立つ．

Theorem 1.
Algorithm 3 が ε-LDP を満たす $\iff \frac{a+b}{b-a} \leq e^\varepsilon$.

Proof. $\mathcal{M}$ を Algorithm 3 のメカニズムとし，$P_{x'_i}(x) := \Pr[\mathcal{M}(x) = x'_i]$ とする．Algorithm 3 が ε-LDP を満たすことは

$$\max_{x,\bar{x}\in[-1,1]} \frac{P_{x'_i}(x)}{P_{x'_i}(\bar{x})} \leq e^\varepsilon \tag{1}$$

と等価のため，

$$\max_{x,\bar{x}\in[-1,1]} \frac{P_{x'_i}(x)}{P_{x'_i}(\bar{x})} = \frac{a+b}{b-a} \tag{2}$$

を示せばよい．$x'_i = 1$ のとき，$-1 \le x \le 1$ について $P_1(x) = \Pr[\mathcal{M}(x) = 1] = \frac{ax+b}{2b}$ は非負単調増加関数であり，式 (1) の左辺は $P_1(1)/P_1(-1) = (a+b)/(b-a)$ より式 (2) を満たす．同様に $x'_i = -1$ のとき，$-1 \le x \le 1$ について $P_{-1}(x) = \Pr[\mathcal{M}(x) = -1] = 1 - \frac{ax+b}{2b}$ は非負単調減少関数であり，式 (1) の左辺は $P_{-1}(-1)/P_{-1}(1) = (a+b)/(b-a)$ より式 (2) を満たす．□

Theorem 1 より，例えば Algorithm 1 のように $a = e^{\varepsilon} - 1$, $b = e^{\varepsilon} + 1$ とすれば，Algorithm 3 は ε-LDP を満たすことが分かる．

Algorithm 3 の有用性については以下が言える．

Lemma 1. $\mathcal{M}$, $x_i \in [0,1]$ をそれぞれ Algorithm 3 のメカニズムおよび入力とし，$x'_i = \mathcal{M}(x_i)$ とする．このとき $\mathbb{E}[x'_i] = x_i$ が成り立つ．

Proof. [5] の Lemma 2 から明らか．□

Theorem 2. $(\mathcal{M}, x_i), (\mathcal{M}, y_i)$ をそれぞれ Algorithm 3 のメカニズムおよび入力の組とし，$x'_i = \mathcal{M}(x_i), y'_i = \mathcal{M}(y_i)$ とする．このとき $\mathbb{E}[x'_i y'_i] = x_i y_i$ が成り立つ．

Proof. $r = \frac{b}{a}$ とする．$\mathcal{M}$ の出力は $+r, -r$ のいずれかであるから，$\mathbb{E}[x'_i y'_i]$ は以下のように計算できる．

$$
\begin{aligned}
&\mathbb{E}[x'_i y'_i] \\
&= \sum_{\pm_x, \pm_y \in \{+1,-1\}} (\pm_x \pm_y r^2) \Pr[(x'_i, y'_i) = (\pm_x r, \pm_y r)] \\
&= \sum_{\pm_x, \pm_y \in \{+1,-1\}} \left(\pm_x \pm_y \frac{(\pm_x a x_i + b)(\pm_y a y_i + b)}{4b^2} \right) r^2 \\
&= \sum_{\pm_x, \pm_y \in \{+1,-1\}} \frac{a^2 x_i y_i \pm_x b x_i \pm_y b y_i \pm_x \pm_y b^2}{4b^2} \cdot \left(\frac{b}{a}\right)^2 \\
&= x_i y_i
\end{aligned}
$$

□

プライバシ予算の消費量

Protocol 1 に 提案メカニズム (Algorithm 3) を適用した場合のプライバシ予算の消費量 (以降、単にプライバシ消費量と呼ぶ) について考察する．なお以降では話を簡単にするため属性は全て数値属性とする．

シャッフルモデルを用いてプライバシ予算を増幅させるための指標として以下が知られている [9], [10](Theorem 3).

Theorem 3 (シャッフルモデルにおける LDP ([9], Theorem 3.8 の簡素化))**.** $\mathcal{R}^{(i)}$ をドメイン $\mathcal{D}$ から出力空間 $\mathcal{S}^{(i)}$ への $(\varepsilon_0, \delta_0)$-LDP を満たすメカニズム，$\pi$ を $\{1, 2, \ldots, n\}$ の一様ランダム置換，$\mathcal{A}(\pi, \mathbf{x}_1, \ldots, \mathbf{x}_n) := \{\mathcal{M}^{(\pi(1))}(\mathbf{x}_{\pi(1)}), \ldots, \mathcal{M}^{(\pi(n))}(\mathbf{x}_{\pi(n)})\}$ $(\mathbf{x}_i \in \mathcal{D})$ としたとき，メカニズム $\mathcal{A}$ は (ε, δ)-DP を満たす．ただし

$$
\begin{aligned}
\varepsilon &:= \ln\left(1 + \tfrac{e^{\varepsilon_0}-1}{e^{\varepsilon_0}+1}\left(\tfrac{8\sqrt{e^{\varepsilon_0}\ln(4/\delta)}}{\sqrt{n}} + \tfrac{8e^{\varepsilon_0}}{n}\right)\right), \\
\delta &:= \quad \bar{\delta} + (e^{\varepsilon}+1)(1+e^{-\varepsilon_0}/2)n\delta_0,
\end{aligned}
$$

$\bar{\delta} \in [0,1]$ は $\varepsilon_0 \le \ln\left(\frac{n}{16\ln(2/\delta)}\right)$ を満たすものとする．

Theorem 3 に関して，δ はエラーパラメータであり，確率 δ で ε-LDP が成立しない可能性があるため，十分小さくとる必要がある．また，$\varepsilon_0 > 1$ であれば $\varepsilon = O\left(\frac{\sqrt{e^{\varepsilon_0}\ln(1/\delta)}}{\sqrt{n}}\right)$, $\varepsilon_0 \le 1$ であれば $\varepsilon = O\left(\varepsilon_0 \frac{\sqrt{\ln(1/\delta)}}{\sqrt{n}}\right)$ となる．

前記の仮定および Theorem 3 から，Protocol 1 に Algorithm 3 を適用したプロトコルのプライバシ消費量に関して以下が言える．

Theorem 4. Protocol 1 の入力は全て数値属性とする．このとき Protocol 1 に必要なプライバシ消費量は高々以下となる．

$$
\varepsilon^* = \ln\left(1 + \frac{e^{\varepsilon'}-1}{e^{\varepsilon'}+1}\left(\frac{8\sqrt{e^{\varepsilon'}\ln(4/\hat{\delta})}}{\sqrt{n}} + \frac{8e^{\varepsilon'}}{n}\right)\right),
$$

$\delta^* = \hat{\delta}$．ただし $n := \min(n_A, n_B)$，$\varepsilon' := (m_A + m_B)\varepsilon_0$，$m_A, m_B$ は T_A, T_B の属性数，ε_0 は Algorithm 3 のプライバシ予算とし，$\hat{\delta} \in [0,1]$ は $\varepsilon' \le \ln\left(\frac{n}{16\ln(2/\hat{\delta})}\right)$ を満たすものとする．

Proof (sketch). Protocol 1 において Algorithm 3 が実行される回数は $m_A + m_B$ のため，直列合成定理および Theorem 3 より，Protocol 1 の非識別化が Theorem 3 のメカニズム $\mathcal{A}$ と等価であることを示せばよい．Protocol 1 について，一般性を失うことなく $\min(n_A, n_B) = n_A$，および

$$
\mathbf{x}_i = \begin{cases} \mathbf{x}_{A,i} \cup \mathbf{x}_{B,j} \text{ if } \exists j \text{ s.t. } \mathbf{x}_{A,i}[\mathbf{c}_{A,0}] = \mathbf{x}_{B,j}[\mathbf{c}_{B,0}] \\ \mathbf{x}_{A,i} \cup \varnothing \quad \text{otherwise} \end{cases} \tag{3}
$$

とする．式 (3) より $\mathbf{x}_i$ の各要素を Algorithm 3 のメカニズム $\mathcal{M}$ でランダム化した場合，プライバシ消費量は高々 $(m_A + m_B)\varepsilon_0$ となり，$\mathbf{x}_i$ に含まれない $\mathbf{x}_{B,j'}$ は，並列合成定理よりプライバシ消費量には影響無いため無視できる．したがって，$\mathcal{M}(\mathbf{x}_i)$ を $\mathbf{x}_i$ の各要素を $\mathcal{M}$ でランダム化したデータとすれば， Protocol 1 の非識別化データは $\{\mathcal{M}(\mathbf{x}_{\pi_A(1)}), \ldots, \mathcal{M}(\mathbf{x}_{\pi_A(n)})\}$ となり，$\mathcal{A}$ の出力と等価である．□

補足: 組織 A,B が正しく Protocol 1 を実行し，非識別化，ランダム置換，(Info, π_*) の削除を行う場合，すなわち semi-honest モデルにおいて，組織 A,B に対しても Protocol 1 に提案メカニズム (Algorithm 3) を適用した出力 T'_* は Theorem 1 により (ε, δ)-LDP を満たし，(ε, δ) の値は Theorem 4 により評価可能となる．

具体例

組織 A,B がそれぞれ m_A, m_b 属性のテーブルを保有し，データを連結する場合のプライバシ予算について例示する．パラメータはテーブルのレコード数 ($n = n_A = n_B$ とする) およびエラーパラメータ δ^* となる．

先ず Theorem 4 の条件 $\varepsilon' \leq \ln\left(\frac{n}{16\ln(2/\hat{\delta})}\right)$ を満たす Algorithm 3 のプライバシ予算の上限 $\varepsilon_0^* = \frac{\varepsilon'}{m_A+m_B}$ を確認する．ここでパラメータが $n, \delta^* = \hat{\delta}$ と二種類あるため，それぞれ現実的な設定と考えられる $n = 10^3, 10^4, 10^5, 10^6$(図 1)，$\delta^* = 10^{-10}, 10^{-9}, 10^{-8}, 10^{-7}$(図 2) に固定する．図 1,2 より，現実的なレコード数やエラーパラメータの下では，ε' として設定できる値は高々 8 程度であることが分かる．すなわち Algorithm 3 に必要なプライバシ予算 ε_0^* は $\frac{8}{m_A+m_B}$ 程度とできる．

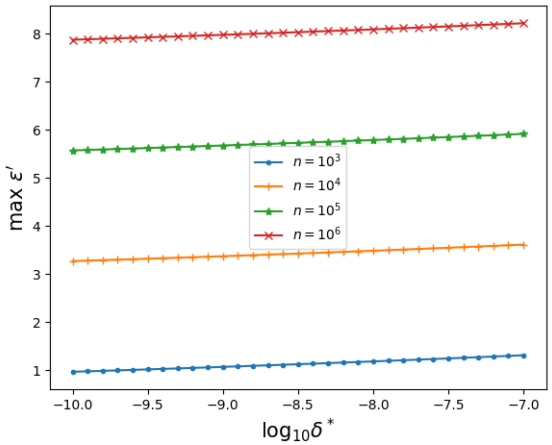

図 1 Theorem 4 の条件を満たすプライバシ予算 ε' の上限 (n を固定)

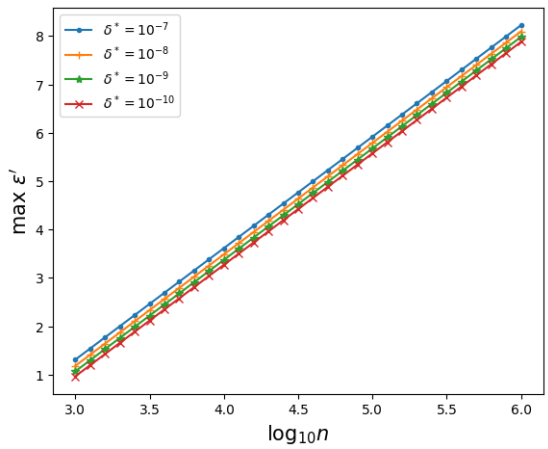

図 2 Theorem 4 の条件を満たすプライバシ予算 ε' の上限 (δ^* を固定)

次に Theorem 4 に基づき Protocol 1 に必要なプライバシ予算 ε^* を確認する．図 1,2 より，$0 < \varepsilon' \leq 8$ について $\delta^* = 10^{-10}$，$n = 10^3, 10^4, 10^5, 10^6$ としたときの ε^* を計算した (図 3)．例えば $n = 10^6$, ε^* =0.1, 0.2, 0.5, 1 のとき，ε' はそれぞれ約 2.4, 3.7, 5.7, 7.6 となっており，シャッフルモデルを用いることでプライバシ消費量をおよそ $\frac{1}{7.6} \sim \frac{1}{24}$ 程度に抑えられることが分かる．

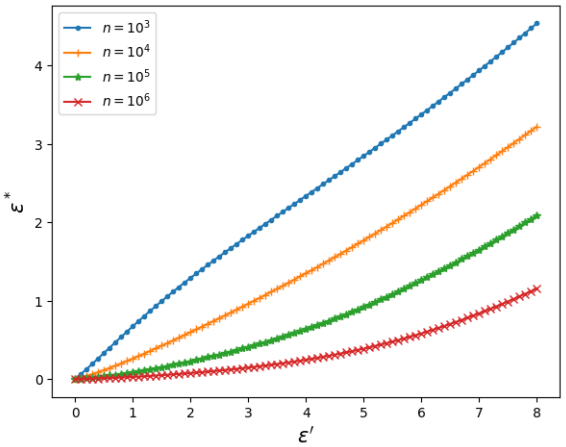

図 3 Protocol 1 に必要なプライバシ消費量 ε^*

4. 実験評価

提案メカニズム (Algorithm 3) の有用性について，公開データセット [11] を用いて実験評価を行った．なお公開データセットは，数値属性の age, education-num, fnlwgt, capital-gain, capital-loss, hours-per-weeks の 6 属性のみを各属性 $[-1, 1]$ の範囲に収まるよう正規化した．このデータセットを D_{orig} とする．

先ず Algorithm 3 が ε-LDP を満たすパラメータ条件

$$\frac{a+b}{b-a} \leq e^{\varepsilon}\ ,\ 0 < a < b \tag{4}$$

を満たす最適な値 a, b を実験評価により確認する．なお式 (4) より $b \geq \frac{e^{\varepsilon}+1}{e^{\varepsilon}-1}a$ となる．

D_{orig} を入力として，Algorithm 3 のプライバシ予算を $\varepsilon/6$ としたときのデータセットを $D_{prop}^{\varepsilon,a,b}$ とする．D_{orig}, $D_{prop}^{\varepsilon,a,b}$ の分散共分散行列 $\Sigma_{orig} := \{x_{ij}\}_{1\leq i,j\leq n}$, $\Sigma_{prop}^{\varepsilon,a,b} := \{x'_{ij}\}_{1\leq i,j\leq n}$ の平均絶対誤差

$$\mathrm{MAE}(\Sigma_{orig}, \Sigma_{prop}^{\varepsilon,a,b}) := \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}|x_{ij} - x'_{ij}| \tag{5}$$

を計算する．ただし Algorithm 3 は確率的アルゴリズムであり $D_{prop}^{\varepsilon,a,b}$ は実行毎に異なるため，MAE は 10 回の試行の平均とし，その標準偏差を求めた．

評価結果を図 4 に示す．図 4 は $\varepsilon = 1$ の場合の条件式 (4) を満たす a, b について式 (5) を計算したときの平均絶

対誤差 (左側) および標準偏差 (右側) を表す．標準偏差は平均絶対誤差と比べ相対的にかなり小さい値となっていることが分かる．また a を固定したとき，b が小さいほど平均絶対誤差も小さくなっている．この傾向は $\varepsilon = 0.1, 10$ でも同様だった．平均絶対誤差や標準偏差が最適となる a の値の傾向は図 4 からは明らかでないが，a を大きくしても平均絶対誤差や標準偏差に大きな変化は見られず，a が小さいときは平均絶対誤差や標準偏差が不安定な結果となった．考察すると，いま $b = \frac{e^\varepsilon+1}{e^\varepsilon-1}a$ としていることから，任意の定数 c について $a' = ac, b' = bc$ としても成立し，かつ Algorithm 3 におけるランダム化の確率 $\frac{ax_i+b}{2b}$ および出力 $x'_i \in \{-\frac{b}{a}, \frac{b}{a}\}$ も変わらないことが分かる．したがって浮動小数点数の誤差の影響が原因と考えられる．特に先行研究 [5] の Algorithm 1 で用いる $a = e^\varepsilon - 1$ は平均絶対誤差や標準偏差が不安定となっていることから，Algorithm 3 の提案メカニズムを用いてより大きな a を選ぶことで誤差の軽減が期待できる．

次に [6] 同様，ラプラス LDP メカニズムと提案メカニズムの有用性を比較する．前記の実験同様，D_{orig} の分散共分散行列との平均絶対誤差および標準偏差を比較する．提案メカニズムのパラメータ a, b は，$\varepsilon \leq 10$ において凡そ安定した結果が得られている $a = 1,000, b = \frac{e^\varepsilon+1}{e^\varepsilon-1}a$ とした．

図 5, 6 より，ε が小さいときほど提案メカニズムの方が平均絶対誤差および標準偏差が相対的に良いことが分かる．一方 $\varepsilon = 3.7$ 付近でラプラス LDP メカニズムの方が優位となった．この傾向は 2 節で触れた Algorithm 1 とラプラス LDP メカニズムの比較結果と類似しており，Algorithm 3 は Algorithm 1 の拡張であるため，その結果は妥当と言える．もし大きな ε を用いるときは，対策として [6] で提案されている Hybrid Mechanism の適用が挙げられる．しかし ε の値に応じて Algorithm 3 以外のメカニズムを使う場合は，積演算の有用性を明らかにする必要があり，今後の課題である．

5. データ合成への応用

Protocol 1 に 提案メカニズム (Algorithm 3) を適用して得られる匿名化データは，ランダム化されているためユーザビリティの点で課題が残る．そこで本節では，提案方式によってランダム化された匿名化データから有用性の悪化を抑えつつデータ合成を行うことを考える．

提案メカニズムは積演算に有効であるため，分散・共分散や相関係数等の統計量，主成分分析，および [6] で例示されている線形回帰，ロジスティック回帰，SVM の計算等への適用が期待できる．データ合成アルゴリズムは，分散共分散行列を用いた方式 [12]，主成分分析を用いた方式 [13]，線形回帰を用いた方式 [14]，SVM を用いた方式 [15] 等，数値属性の積演算を必要とする手法が多数提案されている．

本節では具体例として，[12] への適用について考察する．[12] では，カテゴリ属性と数値属性が混在するテーブルを入力し，カテゴリ属性は One-Hot Encoding により二値属性に変換し，全ての属性について平均，ヒストグラム，分散・共分散を求め，それらの統計量を持つ合成データを作成するアルゴリズムが提案されている．そこで Protocol 1 に 提案メカニズムを適用して得られる匿名化データから，[12] で提案されたアルゴリズムにより合成データを作成することを考える．

Protocol 1 の入力となる，属性数 m_A, m_B の元のパーソナルデータ T_A, T_B について，カテゴリ属性は二値属性に変換されるものとし，変換後の属性数を m'_A, m'_B，$m' := m'_A + m'_B$ とする．なお二値属性は 0,1 の数値属性と見なせるため Algorithm 3 が適用可能だが，入力 x_i が二値の場合は維持確率を $\frac{ax_i+b}{2b}$ とした Randomized Response に他ならない．Randomized Response を用いてヒストグラムを効果的に推定するアルゴリズムは多数提案されており [16]，既存手法を適用してもよい．

プライバシ消費量

Theorem 4 の式に $\varepsilon' = m'\varepsilon_0$ を代入すればよい．ただし δ の条件式は図 1,2 を参照すると，現実的な設定と考えられる $\delta = 10^{-10} \sim 10^7$，$n = 10^3 \sim 10^6$ において，$\varepsilon' = m'\varepsilon_0 \leq 1 \sim 8$ となることから，$\varepsilon_0 \leq \frac{1}{m'} \sim \frac{8}{m'}$ 程度であれば，Theorem 4 に基づくシャッフルモデルにおける LDP が適用可能となる．また $m' \geq 2$ であり，本稿では多属性データを前提としているとともに One-Hot Encoding により属性数も増加することから m' の値はさらに大きくなると考えられる．したがって前記の条件式を満たすような ε_0 は，ラプラス LDP メカニズムよりも提案メカニズムの方が優位となることが期待できる．

有用性

図 7 は，公開データセット ($n = 30,163$) を用いて ε_0 または ε' を横軸として，提案メカニズムにより得られる相関行列の平均絶対誤差[*3](青線，横軸は ε_0) と Theorem 4 に基づくプライバシ予算 ε^* (赤線，横軸は ε') を表している．例えば青線の左端は $\varepsilon_0 = 10^{-5}$ であり，このときの相関行列の平均絶対誤差が 0.054 以下となっていることが分かる．また赤線について横軸の ε' がおよそ 4 のとき ε^* がおよそ 1 になっていることが分かる．これにより例えばプライバシ予算を $\varepsilon^* = 1$，属性数を $m' = 1,000$ としたいとき，赤線から ε' がおよそ 4 であることが分かり，$\varepsilon_0 = \frac{\varepsilon'}{m'} \approx \frac{4}{1000}$ より，青線の横軸が 0.004 の箇所を見れば，$\varepsilon^* = 1$，$m' = 1,000$ としたときに提案メカニズムから得られる相関行列の平均絶対誤差を見積もることができる．$\varepsilon_0 = 0.004$ のときの相関行列の平均絶対誤差は 0.054

[*3] 分散共分散行列の平均絶対誤差は有用性が直感的に分かり辛いため，$[0, 2]$ を値に取る相関行列の平均絶対誤差とした．

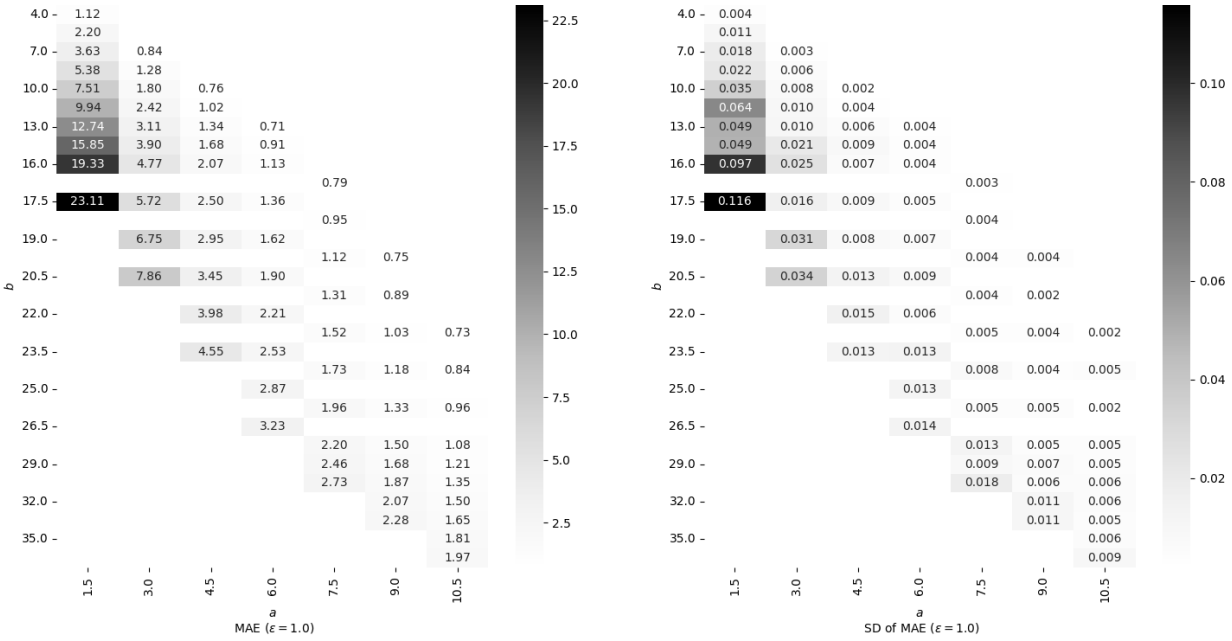

図 4　Alg.3 の a, b と分散共分散行列の平均絶対誤差 (MAE) および標準偏差 (SD) ($\varepsilon = 1.0$)

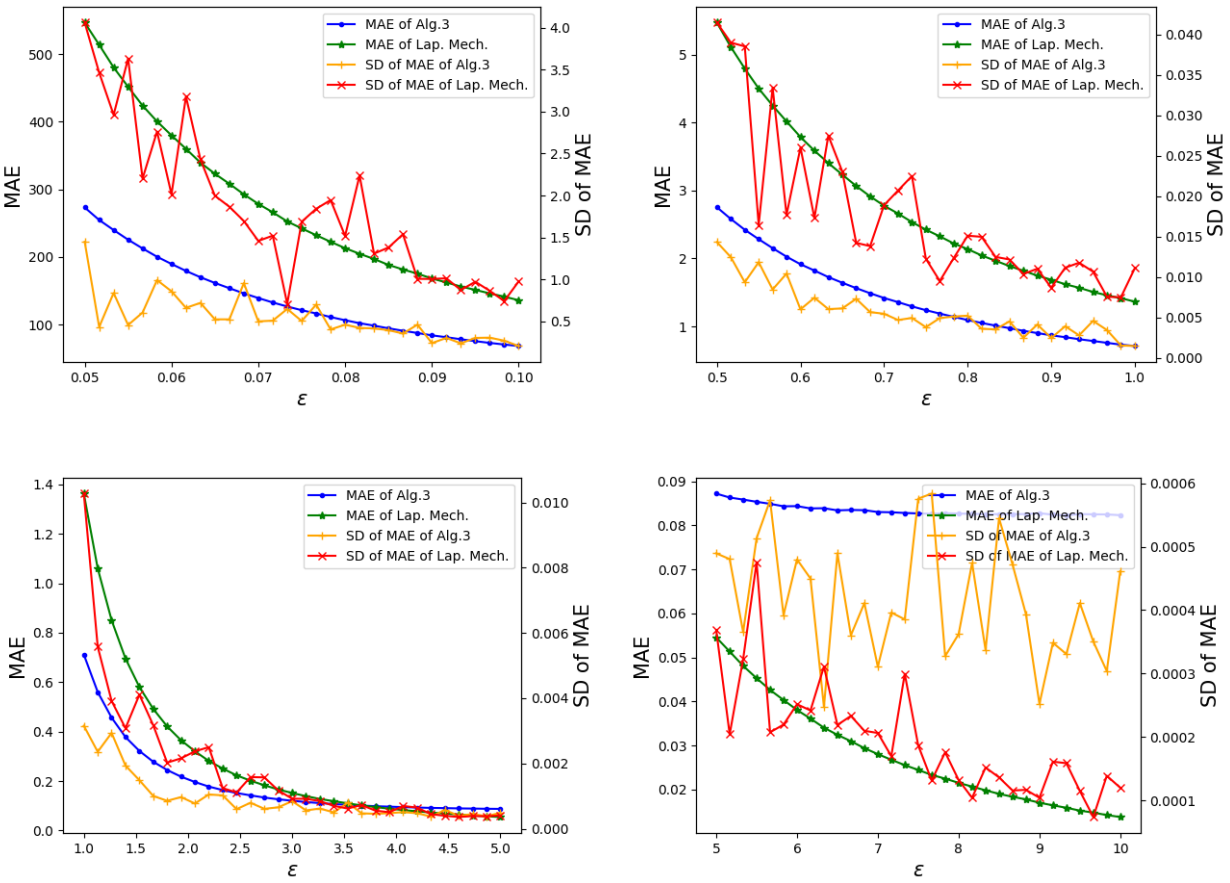

図 5　提案メカニズムとラプラス LDP メカニズムの比較 (ε の範囲を変えて計測)

(誤差 2.7%) 程度となる．この誤差を許容できれば，連結データの属性数が 1,000 であっても有用性と匿名性のバランスのとれた合成データの提供が可能と言える．

一方，平均値や分散・共分散行列の一部の要素は元のデータを保有する組織 A または B が単独で計算可能であるため，通常の計算後に差分プライバシを適用して提供す

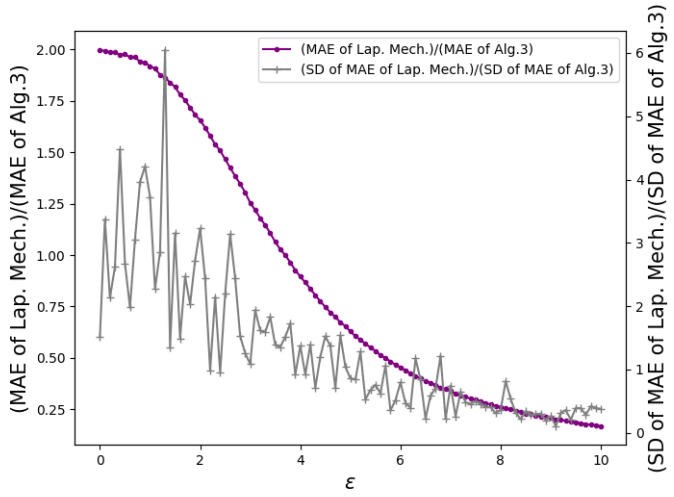

図 6 提案メカニズムとラプラス LDP メカニズムの比較 (平均絶対誤差および標準偏差の比率)

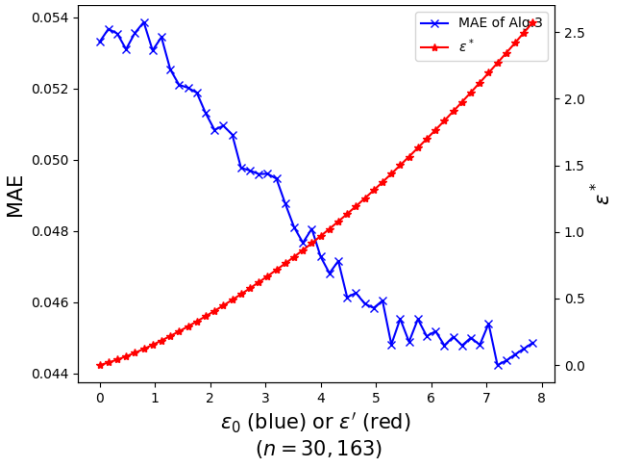

図 7 提案メカニズムによる相関行列の平均絶対誤差とプライバシ予算の関係

る方法も考えられる．しかしその場合，追加でプライバシ予算を消費し，組織 A,B に跨った属性間の計算は直接行えないため，その効果は限定的と言える．

6. まとめ

本稿では，多属性のパーソナルデータについて匿名性を確保しつつ組織横断的に利活用するためのプライバシ保護技術を提案した．従来は多属性データの匿名性確保のために有用性が大きく損なわれる課題があったが，分散共分散行列等の属性間の積演算に適した，シャッフルモデルにおける局所差分プライバシに基づくランダム化手法を提案し，その適用効果を理論評価および実験評価により定量的に示した．またランダム化データを利用するためのユーザビリティ向上手段として，提案方式のデータ合成への応用について考察した．

今後は実際に提案メカニズムを用いたデータ合成アルゴリズムを実装し有用性評価を行うとともに，ランダム化データからの分散共分散行列をより正確かつ精度良く推定するためのアルゴリズムの検討を行う予定である．また直列合成定理よりもタイトなプライバシ予算評価手法も研究が進んでおり，その適用も今後の課題としたい．

参考文献

[1] Fung, B.C.M., Wang, K., Fu, A.W.-C., and Yu, P.S.: *Introduction to Privacy-Preserving Data Publishing – Concepts and Techniques*, CRC Press (2010).

[2] Ikarashi, D., Kikuchi, R., Chida, K., and Takahashi, T.: *k-anonymous Microdata Release via Post Randomisation Method*, IWSEC2015, LNCS9241, Springer-Verlag, pp.225–241 (2015).

[3] 千田浩司, 紀伊真昇, 市川敦謙, 野澤一真, 長谷川慶太, 堂面拓也, 中川智尋, 青野博, 寺田雅之: パーソナルデータの等結合に適した匿名化技術の考案, SCIS2022, 1F3-2 (2022).

[4] Bittau, A., Erlingsson, Ú., Maniatis, P., Mironov, I., Raghunathan, A., Lie D., Rudominer, M., Kode, U., Tinnes, J., and Seefeld, B.: *Prochlo: Strong Privacy for Analytics in the Crowd*, SOSP2017, pp.441–459 (2017).

[5] Nguyen, T.T., Xiao, X., Yang, Y., Hui, S.C., Shin, H., and Shin, J.: *Collecting and Analyzing Data from Smart Device Users with Local Differential Privacy*, CoRR abs/1606.05053 (2016).

[6] Wang, N., Xiao, X., Yang, Y., Zhao, J., Hui, S.C., Shin, H., Shin, J., and Yu, G.: *Collecting and Analyzing Multidimensional Data with Local Differential Privacy*, CoRR abs/1907.00782 (2019)

[7] Wang, N., Xiao, X., Yang, Y., Zhao, J., Hui, S.C., Shin, H., Shin, J., and Yu, G.: *Collecting and Analyzing Multidimensional Data with Local Differential Privacy*, IEEE ICDE 2019, pp.638–649 (2019)

[8] Dwork, C., McSherry, F., Nissim, K., and Smith, A.: *Calibrating Noise to Sensitivity in Private Data Analysis*, The third Theory of Cryptography Conference (TCC 2006), LNCS 3876, Springer-Verlag, pp.265–284 (2006).

[9] Feldman, V., McMillan, A., and Talwar, K.: *Hiding Among the Clones: A Simple and Nearly Optimal Analysis of Privacy Amplification by Shuffling*, CoRR abs/2012.12803 (2021).

[10] Feldman, V., McMillan, A., and Talwar, K.: *Hiding Among the Clones: A Simple and Nearly Optimal Analysis of Privacy Amplification by Shuffling*, FOCS2022, pp.954–964 (2022).

[11] Dua, D., and Graff, C.: *UCI Machine Learning Repository*,(2017).

[12] 岡田莉奈, 正木彰伍, 長谷川聡, 田中哲士: 統計値を用いたプライバシ保護擬似データ生成手法, CSS2017, 3F3-4 (2017).

[13] Sano, N.: *Synthetic Data by Principal Component Analysis*, 20th IEEE International Conference on Data Mining Workshops (ICDMW 2020), pp.101-105 (2020).

[14] Nowok, B., Raab, G. M., and Dibben, C.: *Synthpop: Bespoke Creation of Synthetic Data in R*, Journal of Statistical Software, 74(11) (2016).

[15] Drechsler, J.: *Using Support Vector Machines for Generating Synthetic Datasets*, Privacy in Statistical Databases (PSD) 2010, LNCS 6344, Springer-Verlag, pp.148–161 (2010).

[16] Wang, T., Zhang, X., Feng, J., and Yang, X.: *A Comprehensive Survey on Local Differential Privacy Toward Data Statistics and Analysis*, CoRR abs/2010.05253 (2021).