DEIM2021 G33-2

多次元データへのカウントクエリに適した差分プライバシ

加藤 郁之† 高橋 翼†† 曹 洋† 吉川 正俊†

† 京都大学 〒 606–8501 京都府京都市左京区吉田本町
†† LINE 株式会社 〒 160–0004 東京都新宿区四谷 1-6-1 四谷タワー 23 階
E-mail: †fumiyuki@db.soc.i.kyoto-u.ac.jp, ††tsubasa.takahashi@linecorp.com,
†††{yang,yoshikawa}@i.kyoto-u.ac.jp

あらまし 差分プライバシを満たしたデータ探索はプライバシ性の高いデータに対するマイニングの方針を決定する上で重要である．一方で，これまでに提案されている有効な手法は，(1) ワークロードに非依存，(2) 多次元に適用可能，のいずれかを満たさず，多次元データに対して差分プライバシを保護したデータ探索を提供することはできない．そこで本論文では，差分プライバシを満たしてかつ高い有用性を保持するマテリアライズドビューを作成するアルゴリズム，DP-Mondrian を提案する．DP-Mondrian は多次元データの再帰的な分割によって摂動後でも有用性を維持するブロック分割を効率よく探索する．実データを用いた実験によって，DP-Mondrian によって得られたマテリアライズドビューが様々なレンジカウントクエリに対して高い有用性をもつことを示す．さらに，得られたビューからサンプリングされたデータを用いてクラス分類タスクを評価することで，提案手法が元のデータの特徴を効果的に保存することができることを示す．

キーワード 差分プライバシ，データベース

1 はじめに

データ探索 (Data Exploration) プロセスは，データの特徴や統計的な性質を理解し，効果的なデータマイニングのアルゴリズムの構築などを行うために，特にデータ分析の初期の段階において重要である．データ分析者は様々なカウントクエリを発行することで，それらの特徴を把握する．一方で，プライバシ性の高いデータに対するこのような探索プロセスはプライバシ保護のために一般に大きく制限される．

プライバシ性の高い多次元データが与えられた場合，どのようにデータ探索を提供することが可能であるだろうか？我々はこのようなシナリオでデータ探索を行うためにプライバシ保護型マテリアライズドビュー (p-view と呼ぶ) を構築する方法について研究する (図 1)．特に本研究では，ランダムノイズによる摂動を用いて数学的に厳密なプライバシ保証を与える差分プライバシ (DP) [2] を保護した p-veiw を構築することを目指す．p-view は以下のような性質を満たす必要がある．

- **クエリに独立**: データ分析者は p-view に対して様々なクエリを発行したいと考えているため，回答できるクエリは事前に固定されていない．
- **レンジカウントクエリのノイズの蓄積を低減**: DP を保証するレンジカウントクエリは，一般に保護対象データのドメインサイズの増大に伴いノイズの蓄積も増大し応答誤差が大きくなるため，ノイズの増大を抑えるメカニズムが必要である．
- **データの特徴を維持**: 属性間の相関関係などのデータの特徴は，データ分析の手法や統計モデルの設計に重要であり，p-view においても元のデータの特徴を維持している必要がある．

一方で，レンジカウントクエリへの応答を差分プライベートに公開する state-of-the-arts は，事前に与えれたクエリ集合 (ワークロード) に対して摂動を最適化するか [6,7]，もしくは低次元データを想定している [4,6,9]．データ分割によって摂動誤差を縮小する手法 [6,9] はマテリアライズドビューを提供する一方で，分割を決定する最適化が低次元 (1,2 次元) を想定しており，多次元データに対して有効な手法は提案されていない．特に，複雑な最適化は多次元データに対してスケールせず，多くのプライバシバジェットを消費する可能性がある．したがって，多次元データに対する p-veiw の構築が課題となっている．

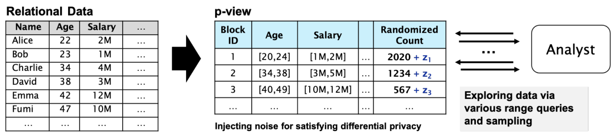

図 1: プライバシ保護型マテリアライズドビュー (p-view) を用いた多次元データに対するデータ探索．データ分析者はデータ分析の方針を設計するために p-view に対して元のデータのプライバシを厳密に保護しつつ任意のカウントクエリを発行可能．

提案． この課題を解決するために，多次元データに対して有用性の高い多次元ブロック分割を発見する手法，DP-Mondrian を提案する．DP-Mondrian はデータを小さなブロックに分割し，ブロックごとにカウント値を集約して公開することで摂動による誤差を縮小することができる．この原則は [6,9] と共通である一方で，提案手法はより簡潔で効率的な探索手法を提案する．DP-Mondrian は DP を満たした，分割点の決定と分割の収束判定を提供することで，プライバシバジェットの制限下で再帰的にブロックを分割して最適なブロック分割を探索することができる．我々の手法は簡潔で効率的であり，多次元

	DP-Mondrian	Identity [1]	Identity_est	DAWA [1]	HDMM	Privbayes
Average Relative RMSE	**1.00**	1.39	16724395.39	3.95	30.76	5.00

表 1: DP-Mondrian は様々なレンジカウントクエリに対して平均的に高い精度で回答可能.

	Identity [2]	DAWA [6]	HDMM [7]	PrivBayes [8]	DP-GAN [3]	**DP-Mondrian(提案手法)**
クエリに独立	✓	✓		✓	✓	✓
カウントクエリのノイズ低減		1(,2) 次元のみ	✓			✓
データの特徴を維持				✓		✓

表 2: 提案手法のみが多次元データに対するデータ探索 (p-view) における全ての要件を満たす.

データに対しても効果的かつ高速に動作する.

貢献. 前述の p-view の要件を満足するに加えて，DP-Mondrian は以下のような性質を提供する.

- **有用性**: DP-Mondrian は様々な多次元レンジカウントクエリに対して既存手法を上回る精度で回答を公開することができる．さらに，p-view からサンプリングされるデータはクラス分類タスクに対しても高い精度を示す.
- **スケーラビリティ**: DP-Mondrian の実行時間はデータのドメインサイズに対して劣線形である.
- **空間効率**: DP-Mondrian は元のカウントテンソルに対してコンパクトなサイズの p-view を生成することができる．Adult データでは約 10^{12} 分の 1 のサイズとなる.

評価. 表 1 は 8 種類のワークロード (レンジカウントクエリの集合) と 8 つの実データセットに対する Average Relative RMSE (平均相対二乗平均平方根誤差) を示す [1]．DP-Mondrian は全てのアルゴリズムの中で最も小さなスコアを達成している．他のアルゴリズムは特定のデータセットとワークロードに対して DP-Mondrian のスコアを上回ることもあるが，平均的には DP-Mondrian が最も良い結果となる．これは提案手法がデータ探索タスクに適していることを示す．さらに，5 つの実データセットを用いたクラス分類タスクの評価実験によって，DP-Mondrian が生成する p-view から有用性の高いデータをサンプリング可能であることを示す.

提案手法はデータ探索をプライバシを保護下で行えることを可能にする．このとき，プライバシ保護型データ分析を行う実践的かつ合理的な戦略として以下が考えられる．最初に，提案手法を用いてデータ探索を行い，ワークロードや DP アルゴリズムの選択・設計を行う．その後，そのタスクに最適化された DP アルゴリズムを用いて差分プライバシを満たした統計データや統計モデルなどを得る．このようにすることでデータ分析のワークフロー全体に対して厳密にプライバシ保護を行うことが可能となる．本研究はデータ探索の段階に焦点を当てており，全てのワークフローを考慮することは今後の課題である.

本論文の構成は以下の通りである．2 節では準備，3 節では問題設定，4 節では提案手法，5 節では評価について述べる．最後に 6 節では本論文の結論を述べる.

1 : DAWA と Identity は低次元のデータセットに対する結果のみを報告する.

1.1 関連研究

ここでは，既存の state-of-the-arts と DP-Mondrian との比較について述べる．要約を表 2 に示す.

データ分割. 差分プライベートにレンジカウントクエリを公開する最もナイーブなアプローチはカウントベクトルに対してラプラスノイズを直接載せることである．この方法は簡単に p-view を作成できる一方で，各カウントに対するノイズが集約されることでデータの次元の増加に伴い指数的なノイズの増加を引き起こしてしまう．DAWA [6] と AHP [9] の提供するパーティショニング手法はデータの分布に基づいてデータを分割し，パーティションごとにデータを集約することで摂動ノイズを縮小させる．一方で，データの集約によって集約誤差が発生する．よって，これらのトレードオフの中で最も小さな合計の誤差を与えるパーティショニングを発見する．しかし，これらの手法は 1 次元もしくは 2 次元の低次元データに対してのみしか機能しない.

ワークロード最適化. 与えれたワークロードに対する期待誤差を最小化することでレンジカウントクエリの精度を高める手法もこれまで盛んに研究されている．Li [1] らが提案した Matrix Mechanism は，行列形式で表現されるワークロードに対して誤差を最小化するクエリ戦略を近似計算することを可能にしている．HDMM [7] はクロネッカー積を利用することで MM を多次元データに対してもロバストな手法へと拡張する．これらの手法は多次元データに対する 1 つの有望なアプローチを提供する一方で，固定のワークロードに対してしか機能しないため，p-view の生成に用いることはできない.

プライバシ保護型データ生成. 別の方法として，プライバシ保護型の生成モデルを学習してデータを生成することで，そのデータを用いてデータ探索を行う方法がある．Privbayes [8] は差分プライバシを保証しつつ，ヒューリスティックにベイジアンネットワークを学習する手法である．近年差分プライバシを満たした深層生成モデルも注目を集めているが，多くの研究が画像データに焦点を当てており，テーブルデータで十分な性能を提供するものは提案されていない．Fan ら [3] は generative adversarial network (GAN) をテーブルデータに適用し，広範な実験によって評価した．彼らの実験結果は，差分プライベートな GAN はテーブルデータに対して Privbayes よりも低い性能を示すと報告している.

2 準　　備

2.1 表 記 法

最初に本論文全体で使用する表記法について記載する．X を n 個のレコードを持ち属性集合 $\mathbf{A}$ から構成される入力データベースとする．$A \subseteq \mathbf{A}$ は属性の部分集合であり d 個の属性 $A = (a_1, \ldots, a_d)$ をもつ．属性 a のドメイン $dom(a)$ は有限個の離散データであり，その数は $|dom(a)|$ で表される．a が連続値だった場合は，ビン化によって離散値をもつドメインに変換する．よって，A の全体のドメインサイズは $dom(A) = \Pi_{i\in[d]}|dom(a_i)|$, ただし $[d] = (1, \ldots, d)$，とかける．

ここで，データベース X を d モードのカウントテンソル $\mathcal{X}_A$ に変換する．$\mathcal{X}_A[i_1, \ldots, i_d]$ は $(a_1 = i_1, \ldots, a_d = i_d) \in X$ を満たすレコードの個数を表す．以降，$\mathcal{X}_A$ を単に $\mathcal{X}$ と表す．また，$x(\in \mathcal{X})$ を用いて $\mathcal{X}$ のカウント値を表す．さらに，$\mathcal{X}$ の部分テンソルを $\mathcal{B}$ $(\subseteq \mathcal{X})$ と表し，これをブロックとよぶ．ブロック $\mathcal{B}$ は $\mathcal{X}$ と同様に d モードのカウントテンソルであるが，$\mathcal{B}$ のそれぞれのドメインは，$\mathcal{X}$ のドメインに対して等しいかもしくは小さい．ブロック $\mathcal{B}$ のドメインサイズを $|\mathcal{B}|$ とかく．

q をカウントクエリとし，$\mathbf{W}$ をワークロードとする．$\mathbf{W}$ は $|\mathbf{W}|$ 個のレンジカウントクエリの集合で $\mathbf{W} = \{q_1, ..., q_{|\mathbf{W}|}\}$ と表される．$q(\mathcal{X})$ は $\mathcal{X}$ 上のカウントクエリ q に対する結果を返す．

2.2 差分プライバシ

差分プライバシは厳密なプライバシ保証を提供する数学的なプライバシの定義である．データベースからのアウトプットをランダム化することで確率的にデータのプライバシを保証する．

定義 1 (ϵ-差分プライバシ)**.** $d_H(D, D') = 1$ を満たす任意のデータベースの組 $D, D' \in \mathcal{D}$，および任意の出力の部分集合 $Z \subseteq \mathcal{Z}$ に対し，ランダム化メカニズム $\mathcal{M} : \mathcal{D} \to \mathcal{Z}$ が以下の式を満たしているとき，$\mathcal{M}$ は ϵ-DP を満たす．

$$\Pr[\mathcal{M}(D) \in Z] \leqq \exp(\epsilon) \Pr[\mathcal{M}(D') \in Z].$$

ただし $d_H(D, D')$ は D と D' のハミング距離を表す．

ランダム化メカニズム $\mathcal{M}$ はある関数 f に対して DP を保証するために使用される．メカニズム $\mathcal{M}$ は f の敏感度に応じて f の出力に対する摂動を行う．f の敏感度は，$d_H(D, D') = 1$ を満たす任意の D と D' に対して f の出力がとりうる最大の差によって定義される．

定義 2 (敏感度)**.** f の敏感度は，$d_H(D, D') = 1$ を満たす任意のデータベースの組 $D, D' \in \mathcal{D}$ に対して以下のように定義される:

$$\Delta_f = \sup_{D,D'\in\mathcal{D}} \|f(D) - f(D')\|.$$

ただし $||\cdot||$ は f' の出力のドメイン上に定義されるノルムを表す．

このように，f の敏感度に基づいて差分プライバシを保証するノイズを決定することができる．

ラプラスメカニズムと指数メカニズムは標準的なメカニズムである．ラプラスメカニズムは数値データを公開する際の摂動に用いられる．

定義 3 (ラプラスメカニズム)**.** ラプラスメカニズムは，以下のようにラプラス分布からサンプリングされたノイズを $f(D)$ に加算する:

$$f(D) + \mathrm{Lap}(\Delta_f/\epsilon). \tag{1}$$

指数メカニズムは離散データに対するランダムな選択を摂動するために用いられる．各項目の選択確率は，それぞれの項目に対するスコア関数の値によって決定される

定義 4 (指数メカニズム)**.** q を，データベース D に対して，項目 $y \in Y$ が選択される時のスコアを出力する関数とする．指数メカニズムは Y の中から y を決定するために，以下のように定義される重みつきのランダムサンプリングを行う:

$$\Pr[y] \sim \exp(\frac{\epsilon q(D, y)}{2\Delta_q}). \tag{2}$$

複数の出力の公開に対する差分プライバシは，以下の直列合成定理と並列合成定理によって保証される．

定理 1 (直列合成定理 [2])**.** メカニズム $\mathcal{M}_1, \ldots, \mathcal{M}_k$ がそれぞれ ϵ_1-, $\ldots, \epsilon_k$-DP を満たすとする．このとき，$\mathcal{M}_1, \ldots, \mathcal{M}_k$ に対する応答値を出力するメカニズムは $(\sum_{i\in[k]} \epsilon_i)$-DP を満たす．

定理 2 (並列合成定理)**.** メカニズム $\mathcal{M}_1, \ldots, \mathcal{M}_k$ がそれぞれ ϵ_1-, $\ldots, \epsilon_k$-DP を満たすとする．このとき，互いに素なデータベース $D_1, \ldots, D_k$ に対して平行に適用されるメカニズム $\mathcal{M}_1, \ldots, \mathcal{M}_k$ は $(\max_{i\in[k]} \epsilon_i)$-DP を満たす．

3 問 題 設 定

本節では，カウントテンソル $\mathcal{X}$ に対して多次元ブロック分割を行い，差分プライベートなマテリアライズドビュー $\tilde{\mathcal{X}}$ を得るための基礎となる概念について説明する．さらに，多次元ブロック分割を，誤差を最小化する最適化問題として定式化する．

背景． カウントテンソル $\mathcal{X}$ が与えられたとき，$\mathcal{X}$ を m 個の ブロック $\pi = \{\mathcal{B}_1, ..., \mathcal{B}_m\}$ に分割すると考える．ブロックは $\mathcal{B}_i \cap \mathcal{B}_j = \emptyset$ を満たす．ただし $i, j \in [m]$ and $j \neq i$, and $\mathcal{B}_1 \cup \cdots \cup \mathcal{B}_m = \mathcal{X}$．また，$\mathcal{B}_i$ に対するカウント値の合計を $S_i = \sum_{x'\in\mathcal{B}_i} x'$ とし，摂動後の出力を $\tilde{S}_i = S_i + z_i$ とする．$z_i \sim Lap(1/\epsilon)$ はラプラスメカニズムによってサンプリングされ，出力は差分プライバシを満たす．

ブロック $\mathcal{B}_i$ 内の任意のカウント値 x に対しては，摂動誤差 (PE) と集約誤差 (AE) の 2 つの誤差が発生する．いま，任意の $x \in \mathcal{B}_i$ に対して，x を $\bar{x}_i = (S_i + z_i)/|\mathcal{B}_i|$ と置き換えるとすると，x と $\bar{x}_i$ の誤差は以下のように計算できる．

$$|x - \bar{x}_i| = \left|\left(x - \frac{S_i}{|\mathcal{B}_i|}\right) - \frac{z_i}{|\mathcal{B}_i|}\right| \leqq \left|x - \frac{S_i}{|\mathcal{B}_i|}\right| + \left|\frac{z_i}{|\mathcal{B}_i|}\right|. \tag{3}$$

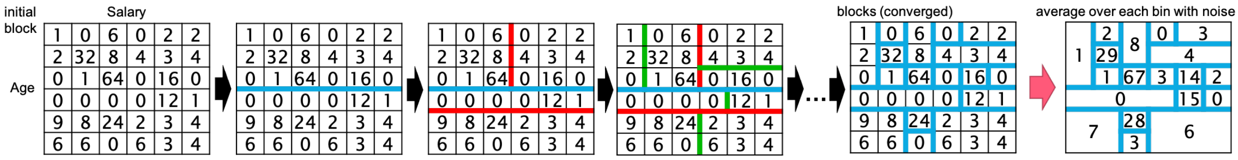

図 2: DP-Mondrian は集約誤差の少ないブロック分割を効率よく発見する (黒矢印)．最後にノイズを加えて平均化する (赤矢印)．p-view はランダム化された値をブロックごとに保持する (最も右の図)．

したがって，ブロック $\mathcal{B}_i$ における誤差の合計値，分割誤差 (SE) は以下のように与えられる．

$$\mathrm{SE}(\mathcal{B}_i) = \sum_{x \in \mathcal{B}_i} |x - \tilde{x}_i| \leqq \mathrm{AE}(\mathcal{B}_i) + \mathrm{PE}(\mathcal{B}_i) \tag{4}$$

ただし，

$$\mathrm{AE}(\mathcal{B}_i) = \sum_{x \in \mathcal{B}_i} \left| x - \frac{S_i}{|\mathcal{B}_i|} \right|, \tag{5}$$

$$\mathrm{PE}(\mathcal{B}_i) = |z_i|. \tag{6}$$

式 (5) と (6) はそれぞれ集約誤差と摂動誤差を表す．

問題． ブロック分割は，元のカウントテンソルに対するラプラスメカニズムの摂動誤差をそれぞれのブロックに対して $\frac{1}{|\mathcal{B}_i|}$ に縮小している．ここで，以下のように SE の期待値を考える．

$$\begin{aligned}
\mathbb{E}\left[\sum_{i \in [m]} \mathrm{SE}(\mathcal{B}_i)\right] &\leqq \mathbb{E}\left[\sum_{i \in [m]} \mathrm{AE}(\mathcal{B}_i)\right] + \mathbb{E}\left[\sum_{i \in [m]} \mathrm{PE}(\mathcal{B}_i)\right] \\
&= \sum_{i \in [m]} \mathrm{AE}(\mathcal{B}_i) + \sum_{i \in [m]} \mathbb{E}\left[\mathrm{PE}(\mathcal{B}_i)\right] \\
&= \sum_{i \in [m]} \mathrm{AE}(\mathcal{B}_i) + m \cdot \frac{1}{\epsilon}.
\end{aligned} \tag{7}$$

したがって，多次元カウントテンソル内で最適なブロック分割を発見する問題は分割誤差 (7) を最小化する最適化問題として以下のように定式化することができる．

$$\begin{aligned}
&\underset{\pi}{\text{minimize}} \sum_{\mathcal{B} \in \pi} \left(AE(\mathcal{B}) + \frac{1}{\epsilon} \right) \\
&\text{subject to } \mathcal{B}_i \cap \mathcal{B}_{j \neq i} = \emptyset \wedge \mathcal{B}_1 \cup \cdots \cup \mathcal{B}_m = \mathcal{X}
\end{aligned} \tag{8}$$

課題． しかし，一般に (8) を満たす最適解をブロック集合の中から発見するのは困難である．解空間は $|dom(A)|^2$ に比例しており，$|dom(A)|$ 自体が次元数に対して指数的な増加をすることを考えると，多次元データに対して実時間で解くことは困難となる．したがって，本論文では，上記の最適化問題に対して，有用性とプライバシの間の優れたトレードオフをもつヒューリスティックな探索手法を考える．

4 提案手法

本節では，提案手法について示す．我々の提案手法は関係データを入力とし p-view を構築する．p-view の有用性の維持と厳密なプライバシ保護のために，誤差を効果的に減少させるブロック分割を差分プライバシを保証しながら探索する．

4.1 概　　要

本手法の課題は，差分プライバシの保証とブロック分割の効率的な探索を同時に行うために，いかにして簡潔で効果的なアルゴリズムを構築するかである．この課題の実現のために我々は DP-Mondrian を提案する．DP-Mondrian は，高速でスケーラブルな k-匿名化アルゴリズムである Mondrian [5] で示された，再帰的な 2 分割の枠組みを用いる．この枠組みが効率の良い手法である一方で，我々の課題は与えられたプライバシバジェットの範囲内で効果的に誤差を減少させるブロック分割を発見することである．

図 2 は提案手法の概要を示す．まず，DP-Mondrian は元のカウントテンソル $\mathcal{X}$ を最初のブロック $B^{(0)}$ とする．次に，ブロック B (最初は $B = B^{(0)}$) を 2 つの互いに素なブロック B_L と B_R ($B_L \cup B_R = B$ かつ $B_L \cap B_R = \emptyset$) に 2 分割する．ただし，その分割前に B 内の集約誤差が十分に小さいかどうか確認する．もし十分に小さい場合は，そこで B に対する再帰的分割を止めて，そうでない場合は B を分割する．B を B_L と B_R に分割する際には，DP-Mondrian はある分割点 $p \in dom(a)$, $a \in A$ を選択する．分割後は，B_L と B_R に対してそれぞれ別々に同様の処理を，再帰的に実行する．全てのブロックが収束したら，DP-Mondrian はそれぞれのブロック B_i に対して，カウント値の合計に摂動を行い $S_i + z_i$ を得る (z_i はラプラスノイズ)．最終的に，全ての $x \in B_i$ に対して，ランダム化されたカウント値 $\tilde{x} = (S_i + z_i)/|B_i|$ を得られる．

上記のアルゴリズムは集約誤差を縮小させるブロック分割をヒューリスティックに発見することができる．さらに，この手法はその簡潔さゆえに効率的である．必要な空間計算量はレコード数に対して線形であり，分割回数はドメインサイズに対して劣線形である．しかしながら，大きな課題は，この簡潔で効率的なアルゴリズムに対し，いかにして差分プライバシを保証するかということである．

この課題を解決するために 2 つのメカニズム *random cut* と *random converge* を導入する．これらを用いることで，DP-Mondrian はリーズナブルな分割点を最小限のプライバシの消費で発見することができ，それゆえに全体のプライバシバジェットの消費を抑え，より深く集約誤差を減少させるブロック分割の探索を行うことができる．

ϵ_r を再帰的分割のためのバジェット，ϵ_p を摂動のためのバジェットとし，$\epsilon_B = \epsilon_r + \epsilon_p$ を DP-Mondrian における全体のプライバシバジェットとする．また，DP-Mondrian は分割の深さが κ に達するまで再帰的な 2 分割を続ける．それぞれの 2

アルゴリズム 1 DP-Mondrian

Input: count tensor $\mathcal{X}$, convergence threshold θ, privacy parameters ϵ_r, ϵ_p, privacy budget ratios β, γ

Output: p-view $\tilde{\mathcal{X}}$

$\tilde{n} \leftarrow$ TotalDomainSizeOf($\mathcal{X}$)

$\kappa \leftarrow \beta \log_2 \tilde{n}$ // maximum depth of recursive bisection

$\pi \leftarrow \{\}$; $k \leftarrow 1$

RecursiveBisection($\mathcal{X}, \pi, \epsilon_r, k, \kappa, \theta, \gamma$)

$\tilde{\mathcal{X}} \leftarrow$ AdoptivePerturbation($\pi, \epsilon_r, \epsilon_p$)

return $\tilde{\mathcal{X}}$

アルゴリズム 2 RecursiveBisection

Input: block $\mathcal{B}$, converged blocks π, privacy parameter ϵ_r, current depth k, maximum depth κ, convergence threshold θ, privacy budget ratio γ

if $k == \kappa$ **then**

  $\pi \leftarrow \pi \bigcup \mathcal{B}$

  return

end if

// Random Converge

$\epsilon_{conv} \leftarrow \gamma\epsilon_r/\kappa$

if $AE(\mathcal{B}) + \mathrm{Lap}(1/\epsilon_{conv}) \leqq \theta$ **then**

  $\pi \leftarrow \pi \bigcup \mathcal{B}$

  return

end if

// Random Cut

$\epsilon_{cut} \leftarrow 1 - \epsilon_{conv}$

for all $i \in [d], j \in [|dom(a_i)|]$ **do**

  $Q[i, j] \leftarrow q(\mathcal{B}, a_{ij})$

end for

$(i^*, j^*) \leftarrow$ WeightedSampling(Q)

$(\mathcal{B}_L, \mathcal{B}_R) \leftarrow$ Split(i^*, j^*)

// Repeat Recursively

RecursiveBisection($\mathcal{B}_L, \pi, \epsilon_r, k+1, \kappa, \theta, \gamma$)

RecursiveBisection($\mathcal{B}_R, \pi, \epsilon_r, k+1, \kappa, \theta, \gamma$)

return

分割において，DP-Mondrian は $\gamma\epsilon_r/\kappa$ を random converge に使い，$(1-\gamma)\epsilon_r/\kappa$ を random cut に使う．ただし $0 \leqq \gamma \leqq 1$ である．分割が収束したのち，DP-Mondrian は ϵ_p のバジェットによるラプラスメカニズムを適用して，カウント値に対する摂動を行う．

全体のアルゴリズム (DP-Mondrian) をアルゴリズム 1，再帰的分割 (RecursiveBisection) をアルゴリズム 2 に示す．

4.2 Random Converge

ここでは，差分プライバシを満たしつつ合理的に再帰的分割を止める方法を考える．DP-Mondrian は集約誤差がある閾値 θ よりも大きい場合は再帰的分割を継続する．この停止の判断を差分プライベートに行うために，この集約誤差の評価をラプラスメカニズムを用いて行う．random converge が用いるランダム化された停止の基準は以下の通りである:

$$\mathrm{AE}(B) + \mathrm{Lap}(\Delta_{AE}/\epsilon) \leqq \theta \tag{9}$$

ただし，Δ_{AE} は集約誤差 AE の L1-敏感度である．

定理 3. 集約誤差 AE の L1-敏感度は $2(1 - 1/|\mathcal{B}|)$ である．

Proof. $\mathcal{B}'$ を $\mathcal{B}$ と 1 つだけカウント値が異なるブロックとする．集約誤差 AE$(\mathcal{B}')$ は以下のように計算できる:

$$\mathrm{AE}(\mathcal{B}') = \sum_{i \neq j \in [|\mathcal{B}|]} \left| x_i - \frac{S+1}{|\mathcal{B}|} \right| + \left| x_j + 1 - \frac{S+1}{|\mathcal{B}|} \right|.$$

したがって，AE の L1-敏感度は以下のように導出される:

$$\Delta_{AE} = (|\mathcal{B}| - 1)\frac{1}{|\mathcal{B}|} + 1 - \frac{1}{|\mathcal{B}|} = 2(1 - 1/n)$$

□

4.3 Random Cut

次に，ブロック B において，全ての属性値からいかにして差分プライベートに尤もらしい分割点を発見するかについて説明する．我々は，良い分割点は分割後の 2 つのブロックにおける集約誤差がより小さくなるはずである，という重要な直感を用いる．$B_L^{(p)}$ と $B_R^{(p)}$ を B に対して分割点 p で分割した後の 2 つのブロックとし，B に対して p をスコアリングする関数 q を以下のように定義する:

$$q(B, p) = -(\mathrm{AE}(B_L^{(p)}) + \mathrm{AE}(B_R^{(p)})). \tag{10}$$

全ての属性値 $p \in dom(a)$, $a \in A$ に対してこのスコアを計算する．そして，これらのスコアと指数メカニズムを用いて分割点の重みつきのランダムサンプリングを行う．分割点 p に対するサンプリング確率は以下の値に比例する:

$$\Pr[p* = p] \sim \exp\left(\frac{\epsilon q(B, p)}{2\Delta_q}\right) \tag{11}$$

ただし，Δ_q はスコア関数 q の L1-敏感度である．このとき，q は 2 つの AE の合計値であるので，$\Delta_q = 2\Delta_{AE}$ とできる．

4.4 再帰的分割におけるプライバシの計算

DP-Mondrian はブロックを分割していくため，ある収束したブロックに対してプライバシ消費を計算するためにはその収束に向けての分割処理のパスを辿ればよい．さらに，DP-Mondrian はブロックを互いに素なブロックに分割するため，全体のプライバシ消費は並列合成定理によって計算できる．

補題 1. $\mathcal{B}$ を DP-Mondrian による深さ k の分割によって得られたブロックとし，あるバジェット ϵ_r を用いて，それぞれの random converge に $\gamma\epsilon_r$，それぞれの random cut に $(1-\gamma)\epsilon_r$ のバジェットを消費したとする．このとき，深さ k の 2 分割は $k\epsilon_r$-DP を満たす．

補題 2. κ を DP-Mondrian における再帰的分割の最大の深さとし，$\mathcal{B}_1, \ldots, \mathcal{B}_m \in \pi$ を最終的に収束したブロックの集合とする．また，$i \in [m]$ に対して k_i を B_i の分割の深さとする．このとき，DP-Mondrian の再帰的分割は $\min(\kappa, \max(k_1, \ldots, k_m))\epsilon_r$-DP を満たす．

データセット	レコード数	カラム数 (カテゴリ属性)	ドメイン数	%Positive
Adult[2]	48842	15 (9)	9×10^{19}	0.239
Small-adult	48842	4 (2)	3×10^{5}	N/A
Numerical-adult	48842	7 (1)	2×10^{11}	0.239
Traffic[3]	48204	8 (2)	1×10^{14}	N/A
Bitcoin[4]	500000	9 (1)	4×10^{12}	0.986
Electricity[5]	45312	8 (1)	1×10^{15}	0.425
Phoneme[6]	5404	6 (1)	2×10^{6}	0.293
Jm1[7]	10885	22 (1)	2×10^{21}	N/A

表 3: データセット.

4.5 Adaptive Perturbation

注目すべきこととして，ブロックの中には，早い段階に深さが浅い分割のみで収束するものもある．これらはプライバシの消費が他のブロックよりも少ない．これらのブロックが深さ k で収束したとすると，プライバシバジェット $\epsilon_r(1-k/\kappa)$ を無駄にすることになる．我々は，これらを摂動のバジェットに回すことで十分にバジェットを使う方法を提案する．すなわち

$$\tilde{S}_i = S_i + Lap(1/(\epsilon_p + \epsilon_r(1-k/\kappa))). \quad (12)$$

この adaptive perturbation によって p-view の有用性をさらに高めることができる．

定理 4. 再帰的分割と adaptive perturbation を用いた DP-Mondrian は ϵ_B-差分プライバシを満たす．

定理 4 の証明は，本節で述べた補題より直接導かれる．

5 評　　価

本節では，DP-Mondrian に対する実験評価の結果について示す．DP-Mondrian の構築する p-view の効果を経験的に示すために，2 つのタスク:レンジカウントクエリ (5.2 節) とクラス分類 (5.3 節) を行う．さらに 5.4 節は DP-Mondrian の実行のスケーラビリティと空間効率に対する実験結果を示す．

5.1 実 験 設 定

実験設定について述べる．以降の実験では，DP-Mondrian と比較対象に対して 10 回の試行を行い，その平均を報告してバイアスを除去している．また，DP-Mondrian のパラメータはデータに独立して $(\theta, \frac{\epsilon_r}{\epsilon_B}, \beta, \gamma) = (0, 0.9, 1.2, 0.9)$ で固定されている．このパラメータ設定は，より多くのバジェットを分割の探索に割いた場合でも，adaptive perturbation によって摂動にも多くのバジェットが使用されるために効果的に機能する．

データセット．テーブル 3 に示すように，文献で広く用いられる 8 つの多次元のテーブルデータセットを用いる．Adult[2]は 6 つの数値属性と 9 つのカテゴリ属性をもつ．さらに，age, workclass, race, capital-gain の 4 属性を抽出して Small-adult，数値属性とラベルだけを抽出して Numerical-adult とする．前者は，低次元のデータセットに対する評価を行うため，後者は DP-Mondrian が順序付きの属性に対して特に効果的に機能することを示すために用いる．Traffic[3]は 8 つのカラムをもつ交通量のデータセットであり，Bitcoin[4]はマルウェアのラベルをもつ Bitcoin の取引情報をもつデータである．Electricity[5]は電気料金の価格上下の推定を行うためのデータセットである．Phoneme[6]は音声情報のデータセットであり我々の実験では比較的低次元である．Jm1[7]はソースコードの静的解析結果のデータセットであり，22 個の属性をもつ．

比較対象．DP-Mondrian と比較するアルゴリズムは Identity [2], HDMM [7], Privbayes [8], DAWA [6], そして DP-GAN [3] の 5 種類である．これらを多次元データに適用させて実験を行うために以下のような設定を用いる．Identity は低次元データに対しては直接摂動を行い計測を行う一方で，より高次元のデータに対しては，[7] で導入されているクロネッカー積を用いた暗黙の行列表現とワークロードに基づいた誤差推定を用いることで計測する．実験では，こちらの推定値を Identity_est として示す．HDMM ではテンプレートとして p-Identity strategies [7] を用いる．DAWA のパーティショニングは多次元データをフラットな 1 次元のカウントベクトル表現にして行う．DAWA の提供する v-optimal ヒストグラムに基づく最適化は多次元データにスケールしないため，比較的低次元の Small-adult と Phoneme に対してのみ実行する．また，DP-Mondrian とのパーティショニングの能力を比較するため，DAWA が提案するワークロードに対する最適化は行わない．DP-GAN では [3] で示されるように 10 回のランダムな試行でモデルの形状やハイパーパラメータを各データセットに対して決定する．このプロセスはプライバシ漏洩を引き起こす可能性があるが DP-GAN の最大の能力と比較するためにここでは考慮しない．DP-Mondrian と他のいくつかの比較手法はビン化を必要とする．カテゴリ属性は単に ordinal encoding を適用した．テーブル 3 のドメイン数はビン化した後のドメイン数を示す．また，公平のため Privbayes と DP-GAN は生データを用いて学習を行う．

ワークロード．実験では，以下のワークロードを用いる．*k-way All Marginal* は k 個の属性の全ての組み合わせに対するマージナルカウントクエリである．*Prefix-kD* は k 個の属性の全ての組み合わせに対するプレフィックスカウントクエリである．*Random kD Range query* は任意の k 個の属性に対するレンジカウントクエリである．以降の実験では，3000 個のクエリをランダムに発生させ，さらにその 10 回の試行の平均の結果を報告する．

5.2 レンジカウントクエリ

DP-Mondrian の生成する p-view のデータ探索における有用性を示すためにレンジカウントクエリによる評価を行う．

指標．カウントクエリに対する評価のメトリックとして二乗平均平方根誤差 ($RMSE$) を用いる．カウントテンソル $\mathcal{X}$ が与えられたとき，p-view$\mathcal{X}'$ とワークロード $\mathbf{W}$ に対して $RMSE$

2 : http://archive.ics.uci.edu/ml/datasets/Adult
3 : http://archive.ics.uci.edu/ml/datasets/Metro+Interstate+Traffic+Volume
4 : https://archive.ics.uci.edu/ml/datasets/BitcoinHeistRansomwareAddressDataset
5 : https://www.openml.org/d/151
6 : https://www.openml.org/d/1489
7 : https://www.openml.org/d/1053

図 3: DP-Mondrian は多次元データ上の多様なレンジカウントクエリに対して小さな誤差を示す．

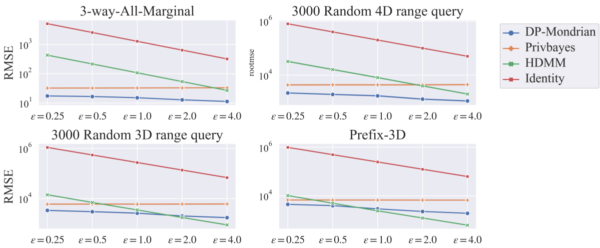

図 4: DP-Mondrian は厳しいバジェット制約下でも小さい誤差を示す (ϵ1)．

は以下のように定義される．

$$RMSE = \sqrt{\frac{1}{|\mathbf{W}|}\sum_{q\in\mathbf{W}}(q(\mathcal{X}) - q(\mathcal{X}'))^2}$$

このメトリックはカウントクエリに対する p-view の有用性を表すとともに，Matrix Mechanism に基づくワークロード最適化 [1,7] の目的関数とも一致しており，それらの提供する推定誤差とも公平に比較することができる．

さらに，各アルゴリズムに対して，DP-Mondrian に対する相対的な RMSE の大きさを計算し，これを全てのデータセットと全てのワークロードに対して平均化した値を示す．このメトリックを *averaged relative RMSE* (ARR) と呼ぶ．

平均的に高い有用性． 1 節のテーブル 1 は ARR による，各アルゴリズムの平均的なパフォーマンスの結果を示す．DP-Mondrian が最も良い精度であり，他のアルゴリズムとは数倍から数桁の差がある．DAWA と Identity は多次元データに対して実行可能ではないので `Small-adult` と `Phoneme` に対する結果のみから算出されおり，比較的低次元なデータに対しても DP-Mondrian が有効であることが示されている．Privbayes の ARR は，顕著に差がある `Small-adult` と `Numerical-adult` の 2 つのデータセットを除いた結果でも **1.31** となる．データ探索においては，多様なワークロードに対応できることが望まれるので平均的な精度は重要である．この結果は，DP-Mondrian がデータ探索に望ましい性質をもっていることを強く示す．

高次元のデータ・クエリに対する有効性． 図 3 は ϵ=1.0 における比較的高次元な述語をもつワークロードとデータセットに対する *RMSE* を示す．DP-Mondrian は多くの箇所で最も小さい誤差を達成しており，Privbayes もいくつかの箇所ではそれに近い誤差を示している．さらに，Privbayes は `Adult` において DP-Mondrian よりもわずかに低い誤差を示す．HDMM は低次元データに対しては極端に小さい誤差を示している．一方で，HDMM はデータとワークロードの次元が大きくなるにつれて誤差が非常に大きくなっている．DP-Mondrian と Privbayes はそのような次元の増加に対してロバストである．ベースラインである Identity_est は高次元データに対して極端に大きい誤差を示す．DP-Mondrian は特に高次元のデータとクエリに対して有効であると言える．また，上に述べたように多くのワークロードとデータセットで平均的に高い精度を示しており，データ探索に有効な手段であるといえる．

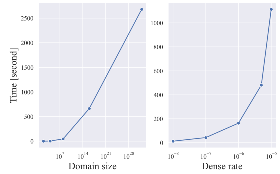

図 5: 実行時間はドメインサイズに対して劣線形である．

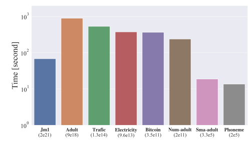

図 6: 実データに対する実行時間．() 内はドメイン数を示す．

バジェットへのロバスト性． 図 4 はプライバシバジェットを $\epsilon = \{0.25, 0.5, 1.0, 2.0, 4.0\}$ と変化させた `Numerical-adult` 上の *RMSE* の結果である．DP-Mondrian はバジェットの増加に対して緩やかに精度を向上させる一方で，Privbayes はほぼ変化しない．HDMM と Identity は厳しいバジェットでは非常に多くの誤差を発生させるが，バジェットの増加に応じて誤差を減少させることができている．この結果からは，DP-Mondrian は厳しいプライバシに対してもロバストであると言える．

5.3 クラス分類

ここでは，p-view からサンプリングされて生成されるデータがどの程度有用性を維持しているのかを評価する．DP-Mondrian の生成する p-view は差分プライベートな n 次元のヒストグラムの近似であり，サンプリングによって差分プライバシを満たしたデータを生成することができる．仮にこれらのデータが元のデータの分布を十分に捉えることができていれば，プライバシの問題なしに任意のデータマイニングタスクに対応可能である．これらの理由により，生成データを用いたクラス分類タスクにおけるパフォーマンスを評価する．

指標． 2 値分類タスクにおいて receiver operating characteristic curve (AUROC) と area under the precision-recall curve (AUPRC) を用いて評価を行う．最初に，データセットを訓練データと評価データに分けて，訓練データを用いてそれぞれのモデル (DP-Mondrian，Privbayes，DP-GAN) を学習させる．次に，サンプリングによって訓練データと同じ数のデータを生成し，それを用いて分類器を学習する．最後に，その分類器を用いて評価データに対してクラス分類を行い評価する．

	AUROC				AUPRC			
	DP-Mondrian	Privbayes	DP-GAN	Original	DP-Mondrian	Privbayes	DP-GAN	Original
`Adult`	0.687983	**0.836833**	0.517572	0.911240	0.422552	**0.605006**	0.244066	0.781548
`Numerical-adult`	**0.750175**	0.694033	0.600376	0.852462	**0.501523**	0.423544	0.313016	0.693934
`Bitcoin`	**0.695888**	0.594983	0.490054	0.807515	**0.992736**	0.989690	0.985909	0.996178
`Electricity`	**0.674922**	0.642064	0.484334	0.878174	**0.590087**	0.566358	0.415145	0.843882
`Phoneme`	**0.787010**	0.777258	0.472719	0.888834	**0.546300**	0.515753	0.265866	0.737750

表 4: DP-Mondrian は特に数値属性をもつデータに対して高い有用性をもつデータを生成する.

データセット	Identity-based	DP-Mondrian
`Adult`	30.99 EB	27.52 MB
`Bitcoin`	1.27 TB	28.11 MB
`Electricity`	1.11 TB	11.24 MB
`Phoneme`	781.34 KB	722.40 KB

表 5: DP-Mondrian の p-view は高い空間効率を示す.

分類器. 4 つ異なる分類器，LogisticRegression，AdaBoostClassifier，GradientBoostingClassifier，XGBoost を用いる．入力データの全ての特徴は one-hot encoding によってエンコードされ，スコアはそれぞれの分類器に対する 10 回の試行の平均を報告する.

結果. 表 4 は ϵ=1.0 におけるクラス分類タスクの結果を示す. DP-Mondrian は `Numerical-adult`, `Bitcoin`, `Electricity` そして `Phoneme` のデータセットで最も高い有用性をもつデータを生成している．Privbayes は `Adult` に対する最も有用性の高いデータを生成する．DP-GAN は全体的にスコアが低い．また，DP-GAN は頻繁に mode collapse を起こしており，GAN をテーブルデータに対して差分プライベートに学習するのは難しいという結果になっている．DP-Mondrian は `Adult` では Privbayes より低いスコアを示す一方で，`Numerical-adult` では高いスコアを示している．よって，DP-Mondrian は順序付きのデータに対して効果的であることが示唆されている．全体として本実験の結果は，いくつかのテーブルデータに対しては DP-Mondrian によって生成された簡潔な多次元の近似ヒストグラムが，グラフィカルモデルや深層生成モデルなどの，より複雑なモデルよりも効果的であるということを示している.

5.4 スケーラビリティと空間効率

図 5 は人工データを用いて，レコード数を 10^5 に固定し，ドメインサイズを 10^2 から 10^{32} に変えた結果と，レコード数を 10^6，ドメインサイズを 10^10 に固定してカウントテンソルに対する充填率を 10^{-8} から 10^{-5} に変えたときの実行時間の推移を示す．結果は，DP-Mondrian がドメインサイズに対して劣線形な実行時間を与えることを示している．一方で，充填率，すなわち非ゼロのカウント値の数に影響を受けてしまっている．図 6 は実データに対する実行時間を示す.

DP-Mondrian によって生成される p-view(図 1) はブロックとカウント値からなり，その空間計算量はブロックの数に比例する．表 5 は ϵ=1.0 の DP-Mondrian によって生成された p-view のサイズと, Identity メカニズムによって摂動された通常のカウントベクトルのサイズを示す．DP-Mondrian は p-view のコンパクトな表現を構築し，そのサイズは `Adult` では 10^{12} 倍小さくなっている.

6 まとめ

本研究は，プライバシ性の高い多次元データが与えられた場合，どのようにデータ探索を提供することが可能であるだろうか？という課題に取り組んだ．我々は，DP-Mondrian を提案して効果的な p-view を構築する方法を提案した．さらに，提案手法が p-view が満たすべき要件とともに，有用性，スケーラビリティそして空間効率に優れていることを示した．また，実データを用いた実験によって提案手法がデータ探索に望ましい性質をもっていることを示した.

文 献

[1] M. H. A. M. V. R. Chao Li, Gerome Miklau. The matrix mechanism: optimizing linear counting queries under differential privacy. In *The VLDB Journal*, page 757–781, 2015.

[2] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006.

[3] J. Fan, J. Chen, T. Liu, Y. Shen, G. Li, and X. Du. Relational data synthesis using generative adversarial networks: A design space exploration. *Proc. VLDB Endow.*, 13(12):1962–1975, July 2020.

[4] I. Kotsogiannis, Y. Tao, X. He, M. Fanaeepour, A. Machanavajjhala, M. Hay, and G. Miklau. Privatesql: a differentially private sql query engine. *Proceedings of the VLDB Endowment*, 12(11):1371–1384, 2019.

[5] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan. Mondrian multidimensional k-anonymity. In *22nd International conference on data engineering (ICDE'06)*.

[6] C. Li, M. Hay, G. Miklau, and Y. Wang. A data-and workload-aware algorithm for range queries under differential privacy. *Proceedings of the VLDB Endowment*, 7(5):341–352, 2014.

[7] R. McKenna, G. Miklau, M. Hay, and A. Machanavajjhala. Optimizing error of high-dimensional statistical queries under differential privacy. *Proc. VLDB Endow.*, 11(10):1206–1219, June 2018.

[8] J. Zhang, G. Cormode, C. M. Procopiuc, D. Srivastava, and X. Xiao. Privbayes: Private data release via bayesian networks. *ACM Transactions on Database Systems (TODS)*, 42(4):25, 2017.

[9] X. Zhang, R. Chen, J. Xu, X. Meng, and Y. Xie. Towards accurate histogram publication under differential privacy. In *Proceedings of the 2014 SIAM international conference on data mining*, pages 587–595. SIAM, 2014.