応専

データベース問合せにおけるプライバシー保護モデル

荒井ひろみ（理化学研究所情報基盤センター）
佐久間淳（筑波大学大学院システム情報工学研究科）

データベース問合せにおけるプライバシーの問題

個人情報や機密情報を含むデータベースを利活用するための一方式として，データベース問合せにおけるプライバシー保護についての理論とモデルが整備されつつある．

データベース問合せにおいて，たとえ応答が個人情報を明示的に含まなくとも，情報利用者の応答を用いた推測攻撃による情報漏洩の危険性は常に存在する．たとえば，統計クエリの応答について，情報利用者が複数のクエリ応答やデータベースに関する事前知識を利用する場合がこれにあたる．本稿ではデータベース問合せにおけるプライバシー保護として，データベースの複数回の応答がデータベースの機密情報を含むかどうかを検査する「クエリ監査」の問題およびデータベース問合せに応答して開示される統計情報に対するプライバシーの定義である「差分プライバシー」の問題について対応技術と関連研究を紹介する．

従来研究では統計クエリに対する問題が多く扱われてきたが，近年移動履歴データにおけるこのようなプライバシー保護技術の応用研究も行われており，本稿ではこれらについても紹介する．

データ開示における監査

データベース問合せにおけるプライバシー保護には，出力をモニタしデータベースに含まれるプライベートな情報を漏洩するか否かを検出する方式がある．本稿ではこのような方式をクエリ監査，監査を行う方式を監査役と呼ぶことにする．クエリ監査はその処理結果に応じて漏洩を防ぐように応答を抑制することを含む場合もある（**図 -1 上**）．

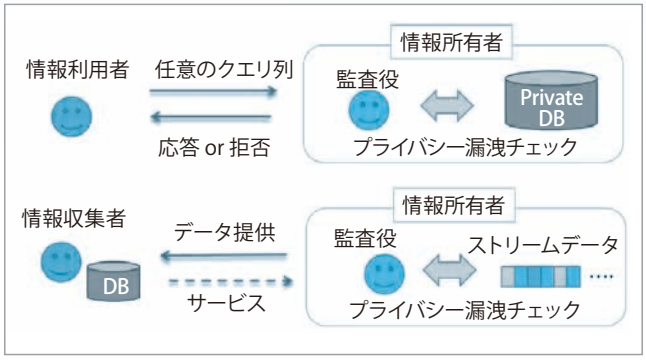

図 -1　出版監査におけるクエリ監査（上）とストリームデータ出版（下）

クエリ監査では，問合せを行う情報利用者がデータベースに含まれる秘匿性の高い情報を推定する攻撃者と想定し，情報利用者に対するプライバシー保護を行う．ここでプライバシー保護とは，データベースのいかなるプライベートな情報も，あるプライバシー基準（一意推定や高い確率で推定されないことなど）を満たすことである．

データベースのクエリ応答によるプライベート情報の漏洩の例として，患者のデータベースに対する年齢や性別，疾患などの条件を指定した患者数の問合せを考える．これは条件によって指定されるデータベースレコードの部分集合に対するカウントクエリの応答として得ることができる．ここで個人の疾患情報は秘匿性が高いプライベート情報であるとする．このカウントクエリに対する応答は統計値であるため一見個人の疾患情報を漏洩していないように見える．しかし，たとえば 20 代の患者の数というクエリに対する応答が 1，20 代の疾患 A の患者数というクエリに対する応答も 1 である場合にはその差分から「20 代の患者は疾患 A の患者である」と一意に

特定されてしまう．そのようなプライベート情報の漏洩が起きないよう，これまでのデータベースのクエリ応答のログを検証し，また漏洩を防ぐように応答を抑制するための技術がクエリ監査である．

また，クエリ監査に関連する最近の話題として，移動履歴などのストリームデータにおけるプライバシー保護データ出版を紹介する（**図-1 下**）．近年個人の移動履歴を情報収集者に集約，分析することで，さまざまな位置情報サービスの品質が向上している．しかし，個人の詳細な移動履歴を情報収集者に提供すると，個人の秘匿したい情報が直接あるいは間接的に漏洩する可能性がある．たとえば　バス停Aから病院を経て薬局近くの喫茶店Bに行った場合に情報提供者が病院に行った事実のみを隠しても，同様の経路をとる人が頻繁にいると知られている場合には行動パターンから病院に行った事実が推測されてしまう．推論攻撃を防ぐデータ提供のために，提供データを監査，抑制する方式が提案されている．

◆ データベース問合せの応答における監査

クエリ監査の理論は主にビット値または実数値などで表現されるデータベースに対するsum, maxなどの集約クエリを対象として研究が進められてきた．これらは比較的単純なデータ形式ではあるが，上記の患者データベースの例も含め，多くの集約クエリを扱う問題がこのモデルで表現される．

さらにクエリ監査は，クエリの受付方式や応答の抑制方式の差異によって，オフライン監査とシミュレータブル監査に区別される．以下にそれぞれを解説する．

オフライン監査

オフライン監査では監査役は情報利用者から受け付ける複数のクエリに対する応答からプライベート情報が漏洩するかどうかを判断する．これは，企業が保有するデータベースがクエリに応答したときに，これまでのクエリ応答がプライバシーに関するコンプライアンスが遵守されているか応答後に確認するためなどに用いることができる．

これまでのオフライン監査研究では，攻撃者の事前知識がない場合を想定し，データベースが持つプライベートな情報の値が攻撃者によって一意に特定される（完全開示）か否かを判定する方式が1種類または2種類の集約クエリの組合せについて提案されている[1)]．オフライン監査問題は計算量が大きいという難点があるが，実数値データベースに対するsumクエリまたはmaxクエリのみの監査には多項式時間で終了するアルゴリズムが存在する．たとえばsumクエリに対するオフライン監査はクエリに対応する線形方程式として記述されるため，その解の一意性を検証するアルゴリズムを用いて監査が可能である．以上のようにオフライン監査ではプライバシー保護基準やクエリの種類が限定した場合は効率的に実現できる．今後は監査手法の改良により，適用できる問題の種類や規模が拡張する方向で研究が進むと期待される．

シミュレータブル監査

逐次的に発行されるクエリに対する逐次的な監査としてオンライン監査問題が提唱されている．監査役はこれまでのクエリに対する応答の列，すなわち真のクエリ応答または拒否の列を保持している．新たにクエリが発行された場合に，監査役はそれに対する応答がプライバシー基準を守るかを検証する．この場合の検証は一見オフライン監査の場合と同様の方法で行える．しかし，クエリ応答を拒否したという事実そのものがプライベートな情報を漏洩する可能性がある．

たとえば上述の患者データベースの場合において，データベースが20代の患者数に応答した後に20代の疾患Aの患者数のクエリを拒否したとしよう，このとき，監査役による「クエリ応答を拒否する」という判断は，監査役が「クエリに応答する」ことによってプライバシー情報が漏洩する，と判断したことを示唆してしまう．上の例では，監査役が「クエリ応答を拒否する」と判断した場合，情報利用者に「20代の患者は全員疾患Aの患者である」と推測させ，意図したプライバシー保護は結局達成できない．このような推測の可能性を考慮した監査方式

としてシミュレータブル監査[4]が提案されている．

シミュレータブル監査ではプライベート情報を漏洩しないクエリもランダムに拒否することで，クエリ拒否が意図せず引き起こす情報漏洩を防ぐ．これは暗号理論におけるシミュレーションパラダイムから着想を得たアプローチである．直観的には，シミュレータブル監査に基づいてクエリを拒否するデータベースは情報利用者にとってほぼ区別ができないようなクエリを拒否するシミュレータが構成できる．

シミュレータブル監査の実行には，クエリを受け付けるごとにこれまでの応答を満たし得るデータベースの解空間および，現在受け付けたクエリへの応答によって変化した解空間のサイズを評価する必要がある．いくつかの問題設定においてシミュレータブル監査の方式が提案されている．解の一意特定を防ぐ監査が sum クエリ，max クエリについて可能である．また，情報利用者のプライベート情報に対する推定確率を数値的に制限するプライバシー基準も提案されている．この基準については攻撃者が事前知識を持つとき，データベースの出力を得て推定するプライベートな情報の値の事後確率と事前確率の差を制限する方式が，ビット値のデータベースに対する sum クエリについて提案されている．

シミュレータブル監査は逐次的なクエリ監査が可能であるが，解空間サイズが大きい場合，計算のコストは高く，現状では対応可能なクエリは限られており，今後の発展が期待される．

◆ストリームデータ出版におけるプライバシー保護

クエリ監査と関連深い問題として，情報提供者が移動履歴を情報収集者に渡す際にプライベートにしたい移動先を秘匿する方式を紹介する．文献 3）では，情報提供者として携帯電話のユーザを，情報被提供者として地図情報のアプリケーション業者を，監査役として携帯電話上で動くアプリケーションを想定し，情報提供者の現在地に関する情報の提供におけるプライベート情報の保護について議論している．情報提供者の移動履歴がマルコフモデルでモデル化されるという仮定の下で，情報被提供者が提供された情報および事前知識を用いて行う推測攻撃を防ぐことが目的であり，ランダムまたはシミュレータブル，もしくはそのハイブリッドモデルに基づいて確率的に位置情報の提供の抑制を行う方式が提案されている．

この方式はユーザのマルコフモデルをもとに抑制モデルを計算する部分とそのモデルをもとに抑制を判定する部分からなり，後者のみを携帯電話で行うため実用的な計算コストで実行できる．実移動履歴データを用いた実験においてハイブリッド方式はプライバシー保護と多くのデータを提供できることが示された．また監査の計算コストも小さくスマートフォンにおいて数百 ms 程度であったと報告されている．

以上のように，クエリ監査などの情報開示の抑制によるプライバシー保護方式はこれまでにいくつか提案されてきている．クエリ監査は出版する情報にノイズをのせることなく正確なデータ出版を可能にするため，ランダム化が好まれないような応用が期待される．

出力プライバシー

データベース問合せに対する応答が，データベース中の秘密情報の推測の手がかりを与えることがある．このような，クエリ応答にかかわるプライバシー侵害の問題を，出力プライバシーと呼ぶ．これに対応するためには，応答値をあらかじめ修正したうえで開示する必要がある．本稿ではこのような方式の中でも近年特に注目されている差分プライバシー[2]にスコープを絞り，解説する．差分プライバシーの研究例として特に統計値の開示と移動履歴の開示を取り上げる．

差分プライバシーの目標は，データベースがクエリに応答したことよって可能になった秘密情報の推測があるバウンドで抑制されることを理論的に保証することである．差分プライバシーの問題はデータベースへのクエリとその応答として定式化されるが，

クエリ監査のようにデータベースとクエリ発行者のインタラクティブな応答は通常想定しない．クエリ応答のための技術というよりはむしろデータベースが持つデータの統計情報の公開に伴うプライバシー情報の漏洩を対象にしていると考えれば分かりやすいだろう．

◆差分プライバシー

差分プライバシーは本来数学的な定義に基づいて議論されるべきであるが，ここでは厳密さを若干犠牲にしつつ，最小限の数式のみでなるべく直観的に説明することを目指す．

情報利用者は，データベースから取得した統計情報と自身の背景知識を組み合わせてデータベース内の情報を推測しようとする．情報利用者がどのような背景知識を保持しているかをあらかじめ定めることは通常困難である．差分プライバシーでは，任意の背景知識を有する情報利用者にとってさえも，データベースが保持する情報の推測がある一定の基準の下で制限されることを保証することをもってプライバシー保護が達成されるとの立場を取る．

データベースの統計情報の公開によるプライバシー情報漏洩の例として，職業・性別・地域別の個人の所得データベースに対する，職業・性別・地域を指定した平均所得を表す統計表の公開を考える．公開する統計表の各セルは，上記の属性値を条件としたデータベースレコードの部分集合に関する属性の集約クエリの応答に相当する．このとき，以下の2つの例を考えよう．

1. 東京都渋谷区の女性30代のコンピュータ科学研究者は105人で，その平均所得は612万円である
2. X県Y市の女性30代のコンピュータ科学研究者は2人で，その平均所得は551万円である

個人の所得情報は秘密にするべき情報に属するが，例1のように，開示情報が十分に多い人数についての平均値であれば，さほどプライバシー保護に気を使う必要はないと多くの人は考えるであろう．一方で例2のように，対象となる個人の数が少ない場合には，平均値から所得が特定されるリスクは相対的に大きく，たとえ特定されないにしても，統計値の開示にはプライバシー上の懸念があると考えることは自然であろう．

では，上の2つのシチュエーションにおいて，何が達成されれば，ある一定のプライバシー保護（この場合は出力された所得の平均値からの個別の個人所得の推測を妨げること）に成功したといえるのだろうか．これを考えるために，n人の個人の情報が含まれるデータベースDと，これに類似したデータベースD'を考える．このD'は，Dから任意の1人の情報を取り除いたデータベースである．Dに対するクエリfの応答をr=f(D)と書く．上記の例でいえば，fは平均値を求める集約関数である．またプライバシー保護のために応答値に施される変更をプライバシーメカニズムMで表す．たとえば，応答f(D)は実際にはM(f(D))として外部に公開されるとしよう．

このとき，直観的には「AさんのデータがデータベースDに含まれていようが，いなかろうが」いいかえれば「問合せ対象がDであろうがD'であろうが」出力される統計値が大して変化しないのであれば，統計値を開示すること自体はAさんのプライバシーを侵害しない，と考えよう，というのが差分プライバシーのアイディアである．

数学的には，プライバシーメカニズムMが確率的アルゴリズムであるとき，Dに対するクエリfの応答がrである確率をPr(r=M(f(D))), D'に対するクエリfの応答がrである確率をPr(r=M(f(D')))とする．このとき，任意の応答値rおよび任意の隣接するデータベースの組(D, D')について

$$\Pr(r=M(f(D)))/\Pr(r=M(f(D'))) < e^{\epsilon}$$

であれば，クエリ f は差分プライバシーを満たす，と考える．

定義では任意の隣接するデータベースについて上記のバウンドの達成を要求するため，具体的な A さん，ではなく，それがだれであっても，その人がいようがいなかろうが統計値に影響しないことが必要である．そしてこの「大して」という概念は，パラメータ ε を用いて，定量化されているのである．

ではこの差分プライバシーを充足するには，メカニズム M は応答値にどのような修正を加えればよいだろうか？　数学的な意味づけは省略するが，応答値にラプラス分布から得たランダムノイズを加え，応答値をランダム化することによって，差分プライバシーが達成されることが証明されている（**図 -2**）[2]．平均値を応答するクエリの場合には，ランダムノイズの分散は，評価対象となる人数（上記の例 1 では 105 人，例 2 では 2 人）の逆数に比例させればよいことも同時に示されている．このことは，対象となる人数が多い場合には，分散の小さいノイズを，対象となる人数が少ない場合には，分散の大きいノイズを用いる必要があることを示しており，直観ともよく一致する．

◆ 差分プライバシーに基づく移動履歴情報の公開

差分プライバシーはデータの統計値等，何らかの集約値を出力することを念頭に設計されており，データそのものを公開することは考慮されていないが，近年のオープンデータ政策の流れを受け，データファイル公開におけるプライバシー保護も注目を集めつつある．そこで差分プライバシーの定義に基づくデータ公開の一例として，Machanavajjhala らによる OnTheMap データにおけるプライバシー保護データ公開[5] について紹介する．

US センサスビューローの Longitudinal Employer-Household Dynamics（LEHD）プログラムでは，(id, origin_block, destination_block）をスキーマとする，各レコードが個人の通勤経路を表すデータを提供している．id はテーブルのキーを，origin_block は個人の居住地を，destination_block は個人の勤務地を表している．いずれの地点も調査上の区画を表しており，その数は 800 万に及ぶ．

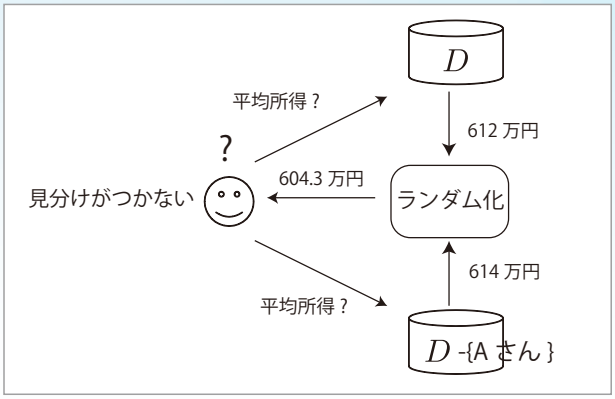

図 -2　データベース D とこれに隣接するデータベース D' に対するクエリ応答のランダム化

勤務地におけるカウントはすでに公開されており，個人の居住地を表す origin_block に関する情報をプライバシーに配慮しつつ公開することが目的である．Machanavajjhala らの実験は，データはミネソタ州に勤務地がある 1,495,415 の異なる出発地・到着地ブロックペア（O/D ペア）を対象とした．

事前の別データを用いた調査で，通勤距離が 200 マイル以上に達するレコードは全体の 1% 以下であることから，これらを除き，233,726 の出発地ブロックを対象とした．

Machanavajjhala らは，加工なしの origin_block に関する情報の代わりに，差分プライバシーを満足するような人工的なデータ合成（synthetic data generation）に基づくデータ公開を試みた．具体的には，データからディリクレ事前分布のもとで推定した多項分布をデータ生成の統計モデルとし，そのモデルのパラメータが差分プライバシーを満足するように調整した上で，その生成モデルから改めてデータ点をサンプルし，これを公開データに用いるアプローチを取っている．差分プライバシーを満足する統計モデルからサンプリングされたデータの公開は，差分プライバシーを満足したデータ出版である，というアイディアである．

プライバシー保護の定義には，前述の ε 差分プライバシーの緩和である (ε, δ)- 差分プライバシーを

用いている．

Machanavajjhala らの結果によれば，合成データにおける各ブロックを勤務地としたときのブロックごとの平均通勤距離の分布は，オリジナルデータから生成した同様の分布と比較において，特にブロックあたりの通勤者数が多い場合には，オリジナルデータの性質をよく保存し，利用に耐える品質を持つと結論づけている．

差分プライバシーは，情報利用者は任意の背景知識を利用可能であるという非常に攻撃者にとって有利な仮定の下でのプライバシー定義であり，このような大規模データについて差分プライバシーを保証しつつ，かつ，分析に耐える品質を維持しながら公開することは容易ではない．Machanavajjhala らは生成頻度の低い O/D ペアを除外するなどいくつかの工夫を加え，これを達成している．

公開データは元のデータとは独立に合成されているため，データ解析者の立場からすれば有用性についての懸念は残るが，このようなデータ合成アプローチは，個々の出力データと，入力データに含まれる各個人に結びつけることが原理的にできないというプライバシー保護データ公開における好ましい性質を持っており，今後有望なアプローチであるといえよう．

参考文献

1) Chin, F. and Ozsoyoglu, G. : Auditing for Secure Statistical Databases : ACM'81, pp.53-59.
2) Dwork, C., McSherry, F., Nissim, K. and Smith, A. : Calibrating Noise to Sensitivity in Private Data Analysis, TCC 2006, pp.265-284.
3) Götz, M., Nath, S. and Gehrke, J. : MaskIt : Privately Releasing User Context Streams for Personalized Mobile Applications, SIGMOD 2012, pp.289-300.
4) Krishnaram, K., Mishra, N. and Nissim, K. : Simulatable Auditing, SIGMOD 2005, pp.118-127.
5) Machanavajjhala, A., Kifer, D., Abowd, J. M., Gehrke, J. and Vilhuber, L. : Privacy : Theory Meets Practice on the Map, ICDE 2008, pp.277-286.

（2013 年 8 月 19 日受付）

荒井ひろみ | hiromi.arai@riken.jp

理化学研究所情報基盤センター基礎科学特別研究員．プライバシー保護，バイオデータマイニングの研究に従事．博士（理学）．

佐久間淳（正会員） | jun@cs.tsukuba.ac.jp

筑波大学大学院システム情報工学研究科准教授．データプライバシーと機械学習の研究に従事．博士（工学）．