中央差分プライバシーの解釈：整理と提案

紀伊 真昇[1,a)]

概要：差分プライバシー (differential privacy, DP) は、データに含まれる個人のプライバシーを保護する安全性指標として注目されている。しかし、その直観的な理解は容易ではなく、なぜ「プライバシー保護」の数理的定義として妥当なのか理解し説明することは難しい。そのため DP は長年盛んに研究が行われているにもかかわらず、社会での活用はほとんど進んでいない状況にある。本稿ではこれまでに提案されてきた (Central) DP の解釈を整理し、さらに少しの提案を行う。最初に DP の定義と基本事項を紹介した後、所属推定攻撃による解釈、因果関係による解釈、ベイズ仮説検定による解釈、頻度論的仮説検定による解釈を紹介する。因果関係による解釈では、より DP を理解しやすくする提案も行う。本稿は DP のその社会的意義に悩んでいる者、非専門家への DP の説明に困っている者など、DP の初学者にも経験者にも役立つものになると期待している。

Interpretations of Central Differential Privacy: Survey and Notes

Masanobu Kii[1,a)]

Abstract: Differential privacy (DP) has gained attention as a security metric for protecting the privacy of individuals contained in data. However, it is not easy to understand intuitively, and it is difficult to understand and explain why it is a valid mathematical definition of "privacy protection." As a result, despite being actively researched for many years, DP has not been widely adopted in society. This paper organizes the interpretations of Central DP that have been proposed to date and makes a few additional proposals. After introducing the definition and basic concepts of DP, we present interpretations based on membership inference attacks, causality, Bayesian hypothesis testing, and frequentist hypothesis testing. In the final section, we propose an interpretation based on causal relationships that aims to make DP more accessible. We hope this paper will be useful for both beginners and experienced researchers in DP, including those who are struggling with the social significance of DP or those who find it challenging to explain DP to non-experts.

1. 導入

差分プライバシー (differential privacy, DP) はデータに含まれる個人のプライバシーを保護する安全性指標として注目されている。しかし、その直観的な理解は容易ではなく、なぜ「プライバシー保護」の数理的定義として妥当なのか理解することは難しい。DP は長年盛んに研究が行われているにもかかわらず、社会での活用はほとんど進んでいない状況にある。本稿の目標は、DP は「プライバシー保護」の数理的定義として隅々まで合理的で、DP の定義が社会的に意義があることを示すことである。

DP は国内外で長年盛んに研究が行われているにもかかわらず、初学者や非専門家が DP の社会的意義に納得することは難しい状況になっている。国内で見受けられる差分プライバシーの説明は、筆者が目にした限り、表層的な短い説明と定義だけ説明するものの他には所属推定攻撃に頼って説明するものしかない。しかし所属推定攻撃による説明で全員が納得出来るわけではない（筆者は出来なかった）し、所属推定攻撃による説明は少々不正確であり、かつ暗黙の前提がある。これでは「プライバシー保護」の数理的定義としての DP が妥当であることについて研究者自身が納得し、社会へ向けて説明していくことは難しくなってしまう。研究者が DP の限界まで正確に理解して適切に社会実装することは更に難しい。

[1] NTT 社会情報研究所
NTT Social Infomatics Laboratories
[a)] masanobu.kii@ntt.com

本稿はこのような課題を解消するために、DP が「プライバシー保護」の定義としてどのように合理的で妥当であるのかを整理・解説する。これにより研究者や技術者が自身で正確に DP を理解し、自信を持って説明するための基礎を提供する。この最終目標に向けた中間的目標として、DP に関する以下の疑問・批判に答えていくことにする。

(i) なぜ確率の差ではなく、比を考えるのか？
(ii) DP は攻撃者の何についての推測を防ぐのか？
- 所属推定攻撃を防ぐと説明されるが、所属判明が問題にならない場面もある。

(iii) どんな攻撃者なら防げるのか？
(iv) なぜ出力される値が近いことではなく、分布が近いことを要求するのか？
(v) なぜ一レコード違う二つのデータベースを入力したときの出力の差異を考えるのか？

本稿の貢献は以下の二点である。一点目は散在していた DP の解釈を整理したことである。もう一点は新しい観点を加えたことである。因果関係による解釈（5 節）はこれまでぼんやりと共有されてきたアイデアをまとめたものであり、本稿の提案とも言える。特に操作・介入主義の観点から述べたのは新しい試みである。また、DP への攻撃者の目標（4 節）を無効値を使って表現しているのも新しい。

一方で、本稿では以下の重要で広範な課題には触れられない。本稿が以下の課題を解決する足がかりになることを期待する。

- DP の意義を短く、分かりやすく、正確に説明する。
- プライバシーパラメータ ε, δ の意味を簡潔に説明する。
- DP を社会の中で正しく運用する方法を示す。

2. DP の基礎事項

最初の準備として DP の定義と基本的な定理をまとめる。

プライバシー保護技術の初学者・非専門家のために二つの大前提を明示しておく。攻撃者に正解を当てられる確率は 0 にはできない。ノーヒントでも当たる可能性は有る。またプライバシー保護技術は私秘的情報を隠すことしかできないため、プライバシー保護技術の外で公開した／していた情報に基づく攻撃を防ぐことはできない。これを大前提として、DP の定義と解釈を説明していく。

2.1 DP の定義

DP には誰を信頼するかによって中央 (central) モデル、手元 (local) モデル、シャッフルモデルがあるが、**本稿では中央モデルだけ扱う**。

2.1.1 記号と用語

定義に必要な用語と記号を導入する。

(i) ランダム写像 $\mathbb{X} \to \mathbb{Y}$ とは，入力 $x \in \mathbb{X}$ で添字付けられた $\mathbb{Y}$ 上の確率変数のこと．[*1]
(ii) レコードとは複数の値の組 (tuple) である。
(iii) データベース (DB) とは同じ項目を持つレコードの集合あるいは多重集合である。
(iv) 隣接する (adjacent, neighboring) DB の組とは，高々一レコードしか変わらない 2 つの DB の組のこと。
(v) 本稿を通じて $\mathcal{M}$ をランダム写像 $\mathcal{M}\colon \mathbb{X} \to \mathbb{Y}$ とする。
(vi) 本稿を通じて $\varepsilon \in [0, \infty], \delta \in [0, 1]$ とする。

2.1.2 「隣接する」「一レコードしか変わらない」についての注意

DP の文脈では「一レコード」を文字通りの意味で扱うべきではない。実際には DP の定義で言う「一レコード」は保護されるべき情報の単位であり、DB 構築の際に定義されたレコードとは異なる可能性がある。例えば DB 中に一個人に関わるデータが複数あればそれらをまとめて「一レコード」とみなすこともあるし、一つのレコードのうちの各属性値を「一レコード」とみなすこともある。

DB の組 (x_1, x_2) が「一レコードしか変わらない」という言葉にも注意が必要である。DB x_1 から一レコードを追加・削除したものを x_2 とすることもあれば、一レコードを書き換えたものを x_2 とすることもある。レコード数が変わるか否かを元に、それぞれ unbounded DP [7, 16], bounded DP [9, 16] と呼ばれている。

2.1.3 DP の定義

定義 2.1 ([10])**.** *DB* を入力とするランダム写像 $\mathcal{M}\colon \mathbb{X} \to \mathbb{Y}$ を考える。部分集合 $S \subseteq \mathbb{Y}$ について $\mu_x(S) = \Pr[\mathcal{M}(x) \in S]$ とおく。

$\mathcal{M}$ が (ε, δ)-*DP* を満たす（与える）とは、任意の隣接 *DB* の組 (x_1, x_2) と、任意の部分集合 $S \subseteq \mathbb{Y}$ について、次が成り立つということ。

$$\mu_{x_1}(S) \leq e^{\varepsilon} \mu_{x_2}(S) + \delta \tag{1}$$

DP を満たすランダム写像のことを *DP* メカニズムと呼ぶ。

$\delta = 0$ のときは *pure* DP, $\delta > 0$ のときは *approximate* DP という．断りがない限り、**本稿では pure DP に絞って**解釈を検討する。

任意の部分集合 $S \subseteq \mathbb{Y}$ について不等式 (1) が成り立つのだから、この不等式は確率変数あるいは確率分布についての不等式と見るべきであり、測度論の意味でほとんど至るところで成り立てば十分である。したがって例えば確率密度が 0 になる点は、そのような点の集合の濃度が 0 であれば無視できる[*2]。

[*1] 一般かつ厳密には Markov kernel で定義される。
[*2] 例えば論文 [13] はこの誤解のために筆者ら自身により撤回されてしまった。

2.2 確率の比を用いることと δ の意味について

DP が確率分布の差異の定義として確率の差ではなく、比あるいは自己情報量 $(-\log \mu_x(S))$ の差を使っている理由はいくつか挙げられる。第一の理由は、DP は一貫して最悪ケースを考えているためである。もしも $\mu_{x_1}(S) = 0$ かつ $\mu_{x_2}(S) > 0$ である事象 S があるなら、どれだけ $\mu_{x_2}(S)$ が小さくても、万が一事象 S が起きてしまえば（所属推定攻撃の意味で）x_2 の情報が漏れてしまう。他にも、確率の差を採用すると確率的に生レコードを出力するランダム写像が「安全」ということになってしまう、と [21] で解説されている。第二の理由は、確率の比を用いることにすると次節の「基本的な定理」が容易に証明できるためである。この二つの定理は DP が重要視される理由の大部分である。

本稿ではほとんど δ を扱わないが、δ の意味を理解するには、hockey-stick divergence を通じた明示的な δ の計算式 [18, 1] を見たり、よくある δ への誤解と正確な DP の定義の違いを書いた Meister の教育的な論文 [21] を読むのが良いだろう。

2.3 基本的な定理

DP が持つ著しい特徴である、合成則と事後処理定理を紹介する。この二つの定理は非常に重要視されており、これらをプライバシー保護の数理的定義が満たすべき公理 (axiom) とする研究者もいる（例えば [5]）。

一つは、DP メカニズムを組み合わせても DP メカニズムになる、という定理である。

定理 2.2 (直列合成定理)**.** *n 個の DP メカニズム $\mathcal{M}_i \colon \mathbb{X}_i \to \mathbb{Y}_i$ がそれぞれ ε_i-DP を満たすならば、これらを組みあわせてつくられるランダム写像*

$$\mathcal{M}^{\mathrm{seq}} = \big(\mathcal{M}_1(x), \dots, \mathcal{M}_k(x)\big) \colon \prod_{i=1,\dots,n} \mathbb{X}_i \to \prod_{i=1,\dots,n} \mathbb{Y}_i$$

も $(\sum_{i=1,\dots,n} \varepsilon_i)$-DP を満たす。

この直列合成定理は最も基本的な形の合成定理であり、DP メカニズムの組み合わせ方に応じて様々な合成定理が研究されている。

定理 2.3 (事後処理定理 [8, 2])**.** *$\mathcal{M} \colon \mathbb{X} \to \mathbb{Y}$ が ε-DP を満たす DP メカニズムならば、別のランダム写像 $F \colon \mathbb{Y} \to \mathbb{Z}$ との合成 $F \circ \mathcal{M} \colon \mathbb{X} \to \mathbb{Y} \to \mathbb{Z}$ も ε-DP である。ただし、$\mathcal{M}$ の入力と確率的従属関係にある変数のうち、F が参照できるのは $\mathcal{M}$ の出力のみとする。*

この定理は情報処理不等式 (data processing inequality) に非常によく似ている。証明は本稿最後に紹介する頻度論的仮説検定を通じたものが最も簡潔だろう [2]。事後処理定理の利用では F が他のデータ加工処理とされることが多いが、ランダム写像なら何でも良いので、例えば攻撃を F とみなしても良いし、計算不可能な写像や詳細が定義されない自然現象を当てはめても良い。

3. 所属推定攻撃による解釈

所属推定攻撃 (membership inference attack) は、攻撃者が処理の出力に対して あるデータがその処理に使われたかどうか（入力データに含まれているか否か）を判別する攻撃である。DP を利用することで所属推定攻撃の成功率・アドバンテージを制御できるため、「DP を使うと所属すらわからなくする事ができる（他の情報は言わずもがな）」と説明されることがある。

当初は処理として機械学習が想定されており、Shokri らによって 2017 年に提案された [23] が、現在はデータ合成への所属推定攻撃も提案されている [24]。ここでは機械学習モデルへの攻撃だけを想定する。

3.1 所属推定攻撃の定義

機械学習に対する所属推定攻撃にはいくつかの設定がありうる。その設定項目は以下の 4 つである。

攻撃対象　攻撃される機械学習モデルは何か。

所属するか推定されるデータの選び方　外部から指定される、ランダムに選ぶ、攻撃者が選ぶなど。

攻撃者が使用する情報　攻撃されるモデルの構造や学習方法、モデルパラメータ、攻撃対象についての背景知識など。

評価指標　正答率などから攻撃者の危険性をどのように評価するか。

標準的な実験手順は以下の通りである。

定義 3.1 (所属推定攻撃の実験 [26, 14, 3])**.** このゲームは、挑戦者 C と敵対者 A の間で次のように進行する。機械学習モデルを f_θ（θ はモデルパラメータ）、訓練データの分布を $\mathcal{D}$、データの個数を n で表す。

(1) 挑戦者は訓練データセット $S \leftarrow \mathcal{D}^n$ をサンプリングし、S を用いてモデル f を学習する。

(2) 挑戦者はビット b をランダムに選ぶ。もし $b = 0$ なら分布 $\mathcal{D}$ から新しい点 $t \leftarrow \mathcal{D}$ をサンプリングする（ただし $t \notin S$）。そうでなければ、点 $t \leftarrow \mathcal{D}$ をランダムに選択する。

(3) 挑戦者は t を敵対者に送る。

(4) 敵対者は分布 $\mathcal{D}$ およびモデル f へのアクセスを得て、ビット $\hat{b} \leftarrow A_{\mathcal{D},f}(x, y)$ を出力する。

(5) 実験は、もし $\hat{b} = b$ なら 1 を出力し、そうでなければ 0 を出力する。

敵対者の評価はアドバンテージ *(advantage)* $|\Pr[\hat{b} = 0 \mid b = 0] - \Pr[\hat{b} = 1 \mid b = 1]|$ などで行われる。

3.2 アドバンテージの理論的・実験的評価と DP の解釈

防御の観点では、所属推定攻撃の成功率・アドバンテージの上限に興味がある。そのため DP で保護されたシステ

ムへの攻撃成功率・アドバンテージの理論的な上限が研究され、導出されている [12, 25]。この成果をもって「DP を使うと所属すらわからなくする事ができる（他の情報は言わずもがな）」と説明されることがある。

一方で実験的に所属推定攻撃を行うと、具体的なシステムが実際にどれほど防御できているのか分かる。こちらについては現在も新しい攻撃手法が盛んに研究されており、例えば [11] が大規模なサーベイを提供している[*3]。逆に攻撃実験の成功率と 4.2 節の定理を用いると、システムのプライバシーパラメータ ε の上限が求められる [6]。

4. 仮説検定による解釈

仮説検定による解釈を通じて次の問いに答える。

- DP は攻撃者の何についての推測を防ぐのか？
- DP はどんな攻撃者なら防げるのか？

ここでは頻度論的仮説検定、ベイズ的仮説検定の両方で解釈を紹介する。特に後者の解釈は攻撃者が持つと想定されるレコードの分布についての事前知識を明示する。

以下では両者での DP の解釈は、概念の上では一致している。すなわち、特定の個人のレコードについての 2 つの仮説「レコード = 特定の値 t」「レコードは削除されている」について、攻撃者がどれだけ強く正しく確信できるのかを、DP はコントロールできる。

最初にこの節で共通して用いる記号を定める。真のデータベース $\mathbf{DB}$（これは確率変数）に含まれうるレコード全体の集合を $\mathcal{T}$ とする。個人が N 人居るとして、個人 i についての真のレコードを $\mathbf{R}_i$ $(i = 1, \dots, N)$ とする。$\mathbf{R}_i$ は常に $\mathbf{DB}$ に属すものの、無効値 (null value) $\perp$ をとる可能性があるとする[*4]。ただし $\perp \notin \mathcal{T}$ とし、ランダム写像はつねに無効値 $\perp$ を無視することとする。

4.1 仮説の設定

頻度論的仮説検定、ベイズ的仮説検定のいずれの解釈でも、真の個人 i のレコード $\mathbf{R}_i$ の値についての二つの仮説

- $\mathbf{R}_i$ は有効で具体的な値 $t \in \mathcal{T}$ である
- $\mathbf{R}_i$ は無効値 $\perp$ である

のどちらが確からしいのかを考える。

この 2 つの仮説は属性推定攻撃で考える「レコードが入っているか否か」すなわち「$\mathbf{R}_i = \perp$ か $\mathbf{R}_i \neq \perp$ か」とは異なる点に注意してほしい。「$\mathbf{R}_i = \perp$ か $\mathbf{R}_i \neq \perp$ か」は非対称性が大きすぎる仮説設定であるため、適切に扱うのは難しい。

なお、「$\mathbf{R}_i = t$ か $\mathbf{R}_i = \perp$ か」という仮説設定は 2.1.2 節で触れた unbounded DP に対応している。「$\mathbf{R}_i = t_1$ か $\mathbf{R}_i = t_2$ か」にすれば bounded DP に対応した定式化になるだろう。後者は半ば属性推定攻撃にも思える。

4.2 頻度論的仮説検定による解釈

頻度論的仮説検定による解釈では、攻撃者（検定する主体）は実際のデータベースのうち、個人 i のレコード以外の部分 x_{-i} を攻撃者の背景知識とし、隣接 DB の組 $(x_1, x_2) = (x_{-i} \cup \{t\}, x_{-i} \cup \{\perp\})$ を考える。ランダム写像 $\mathcal{M} \colon \mathbb{X} \to \mathbb{Y}$ への入力 DB を表す確率変数 $\mathbf{DB}(= x_1 \text{ or } x_2)$ とその出力 y について、次の二つの仮説を考える。もちろん x_1, x_2、H_0, H_1 の添字は入れ替えても良い。

帰無仮説 H_0**:** 出力 y は $\mathbf{DB} = x_1$ から生じた

対立仮説 H_1**:** 出力 y は $\mathbf{DB} = x_2$ から生じた

間違った仮説を選ぶ（頻度論的）確率には、誤って H_0 を採択する確率（偽陽性率 FPR）と誤って H_0 を棄却する確率（偽陰性率 FNR）が有り、次のように定義される。S は検定者により任意に選ばれた部分集合 $S \subset \mathbb{Y}$ である。

$$\mathrm{FPR} = \Pr[\mathcal{M}(x_2) \in S], \quad \mathrm{FNR} = \Pr[\mathcal{M}(x_1) \notin S]$$

次の定理が示す通り、$\mathcal{M}$ が DP メカニズムなら、FPR と FNR を同時に小さくすることはできない。したがって DP は上記の仮説検定を行ったときの正解率を制限することが出来る。

定理 4.1 ([15], [2])**.** *ランダム写像 $\mathcal{M}$ が (ε, δ)-DP を満たすことと、任意の隣接 DB の組 (x_1, x_2) について次が成り立つことは同値。*

$$\mathrm{FPR} + e^{\varepsilon}\,\mathrm{FNR} \geq 1 - \delta \quad \wedge \quad e^{\varepsilon}\,\mathrm{FPR} + \mathrm{FNR} \geq 1 - \delta$$

4.3 ベイズ仮説検定による解釈

この節では pufferfish privacy [17] という安全性の定義の枠組みを参考に、ベイズ仮説検定による解釈を述べる。

定義 4.2. 個人 i と値 $t \in \mathcal{T}$ について、$\sigma_{i,t}$ を命題「レコード $\mathbf{R}_i$ の値は t である」とする。

$\mathbb{S}^{\mathrm{DP}}, \mathbb{S}^{\mathrm{DP}}_{\mathrm{pair}}, \mathbb{D}^{\mathrm{DP}}$ を以下のように定める。

潜在的秘密の集合 $\mathbb{S}^{\mathrm{DP}}$ を、集合 $\{\sigma_{i,t} \mid i = 1, \dots, N, t \in \mathcal{T} \cup \{\perp\}\}$ とする。

潜在的秘密のペアの集合 $\mathbb{S}^{\mathrm{DP}}_{\mathrm{pair}}$ を、集合 $\{(\sigma_{i,t}, \sigma_{i,\perp}) \mid i = 1, \dots, N, t \in \mathcal{T}\}$ とする。$\perp \notin \mathcal{T}$ に注意せよ。

想定されるレコードの分布の集合 $\mathbb{D}^{\mathrm{DP}}$ を、確率 $\Pr[\sigma_{i,t}]$ $(i = 1, \dots, N, t \in \mathcal{T} \cup \{\perp\})$ の可能な選び方全てからなる集合とする。事象 $\sigma_{i,t}$ は i, t ごとに独立であるとし、条件付き確率 $\Pr[\sigma_{i,t}]$ は i, t ごとに異なりうるとする。これは攻撃者（検定者）が持ちうる背景知識の集合とも呼べる。

$\mathbb{D}^{\mathrm{DP}}$ の定義から、真のデータベース $\mathbf{DB}$（これは確率変数）が実現値 db をとる確率は次のように分解できると想

[*3] 著者らは 2025 年 8 月現在も攻撃手法を収集している。https://github.com/HongshengHu/membership-inference-machine-learning-literature

[*4] 4.3 節の原論文 [17] では $\mathbf{R}_i = \perp$ を $\mathbf{R}_i \notin \mathbf{DB}$ と表現していたが、そうすると仮説の対称性が読み取りにくいので、表現を変えた。

定されていることが分かる。各々の確率はレコードの分布（背景知識）$\mathcal{D} \in \mathbb{D}^{\mathrm{DP}}$ で指定される。ここで記号 r_i は db におけるレコード $\mathbf{R}_i$ の実現値を表している。

$$\Pr[\mathbf{DB} = db \mid \mathcal{D}] = \prod_i \Pr[\sigma_{i,r_i}] \tag{2}$$

頻度論的仮説検定での攻撃者の前提知識が実現値 x_{-i} であるのとは違い、ベイズ的仮説検定では攻撃者の前提知識はレコードの分布である。そのため事象 $\sigma_{i,t}$ の独立性も表現できている。

4.3.1 PP を通じた DP の定義とベイズ因子

DP を pufferfish privacy の枠組みを通して定義すると次のようになる。

定理 4.3 ([17] 定理 6.1)**.** ランダム写像 $\mathcal{M}\colon \mathbb{X} \to \mathbb{Y}$ について、$\mathcal{M}$ が *ε-DP* であることと、以下が成り立つことは同値である。

- すべての可能な出力 $\omega \in \mathbb{Y}$ 、
- すべての潜在的秘密のペア $(\sigma_{i,t}, \sigma_{i,\perp}) \in \mathbb{S}^{\mathrm{DP}}_{\mathrm{pair}}$ 、
- $\Pr[\sigma_{i,t}] \neq 0, \Pr[\sigma_{i,\perp}] \neq 0$ を満たすすべての出力レコードの分布（検定者が持つ背景知識） $\mathcal{D} \in \mathbb{D}^{\mathrm{DP}}$

に対して、次が成立する：

$$e^{-\varepsilon} \leq \frac{\Pr[\sigma_{i,t} \mid \mathcal{M}(\mathbf{DB}) = \omega, \mathcal{D}]}{\Pr[\sigma_{i,\perp} \mid \mathcal{M}(\mathbf{DB}) = \omega, \mathcal{D}]} \Big/ \frac{\Pr[\sigma_{i,t} \mid \mathcal{D}]}{\Pr[\sigma_{i,\perp} \mid \mathcal{D}]} \leq e^{\varepsilon} \tag{3}$$

ここで条件 $\Pr[\sigma_{i,t}] \neq 0, \Pr[\sigma_{i,\perp}] \neq 0$ は条件付き確率 $\Pr[\sigma_{i,\_} \mid \mathcal{M}(\mathbf{DB}) = \omega, \mathcal{D}]$ が定義されることを保証する。同時に、この条件は攻撃者は想定しているレコードの分布（背景知識）$\mathcal{D}$ だけではペア $(\sigma_{i,t}, \sigma_{i,\perp})$ について依然として不確実性を持っていることを意味する。

4.3.2 ベイズ因子

式 (3) の中央項はベイズ因子 (Bayes factor) あるいはオッズ比 (odds ratio) と呼ばれるものであり、頻度論的仮説検定で言う p 値のような、ベイズ仮説検定仮説において仮説の確からしさを示す値である。仮説 A と仮説 B を比較するベイズ因子の大きさはそのまま仮説 A/B の確からしさを意味するので、頻度論的仮説検定のように検出力を考える必要はない。仮説 A と仮説 B を比較するベイズ因子が 1 より小さければ小さいほど仮説 A が確からしく、1 より大きければ大きいほど仮説 B が確からしい。

したがって上記定理は、DP は二つの仮説「$\mathbf{R}_i = t$」と「$\mathbf{R}_i = \perp$」両方の確からしさを制御できる、ということを意味している。

4.4 攻撃者の背景知識について

想定されるレコードの分布の集合 $\mathbb{D}^{\mathrm{DP}}$ の定義に明示されている通り、レコード間の相関関係を攻撃者（検定する主体）が知っていてはならない、という点に注意してほしい。もしも攻撃者がレコード間の相関関係を想定すると、DP は攻撃者の目的を定量的に妨げられなくなる（[17] 定理 6.1）。また、この攻撃者における前提は各値 $t \in \mathcal{T}$ について「$\mathbf{R}_i = t$ か $\mathbf{R}_i = \perp$ か」が適切な仮説設定であるためにも必要である。レコード間に依存があると攻撃者が想定していると、他のレコードの値を見るだけで $\mathbf{R}_i$ の値についての情報が得られ、どちらの仮説がより確からしいか分かってしまうからである。この攻撃者における前提については pufferfish privacy 以外の文献 [12, 14] でも指摘されている。

データベースから一部をランダムに取り出して使う (subsampling) ことで、この前提はある程度強制できる。

5. 因果関係による解釈

この解釈は以下の疑問に答えてくれる。

- なぜ一レコード違う二つのデータベースを入力したときの出力の差異を考えるのか？
- なぜ出力される値の近いことではなく、分布が近いことを要求するのか？

また、DP は私秘情報は隠すが集団の傾向は隠さないことが、この解釈を通じて理解できる。

Dwork は DP を提案した論文 [7] で次のように DP の目的を説明している。

> In order to sidestep this issue we change from absolute guarantees about disclosures to relative ones: any given disclosure will be, within a small multiplicative factor, just as likely whether or not the individual participates in the database. (...) Note that a bad disclosure can still occur, but our guarantee assures the individual that it will not be the presence of her data that causes it, nor could the disclosure be avoided through any action or inaction on the part of the user.

本節は、この説明の背後にある直観を明らかにする。

この節では情報提供者の「私が情報を提供することが原因となって、将来なにか悪いことが起きるだろうか？」という問いから出発し、因果論の操作・介入主義を頼って最初の二つの疑問に答える。

5.1 出発点となる直観

個人 i がある情報処理システムに、自身の情報を提供するかどうか選択するときの判断を中心に考えてみよう。その個人は「私が情報を提供することが原因となって、将来なにか悪いことが起きるだろうか？」と考える。この問は情報提供者の合理的な懸念を反映している。もちろん情報提供に関わらず何かしら悪いことは起きるのだから、前半の「……が原因となって」という条件は本質的である。

情報提供が行われて／行われず、システムが情報を処理して何かを出力・公開し、人々の手を渡って、何か「悪い」

ことが起きる。この時系列を通じて、個人の判断と「悪い」ことの間の因果関係について考えてみよう。そのためには因果関係をどのように捉えるかを特定する必要がある。

5.2 因果論の操作・介入主義

この節は参考文献 [19, 27] を参考にしている。因果論とは「因果関係とは何か、どのように定義すべきか」を中心問題とする哲学分野である。そのうち操作・介入主義の基本的なアイデアは、「X が Y の原因であるとは、X を操作することが Y を起こしたり妨げたりする効果的戦略であるということ」というものである [27]。確率を用いて表現すれば次のようになる（なお操作・介入主義は必ずしも確率を必要としない）。周辺状況を変えないまま X を操作した／操作しなかったときに事象 Y が起こる確率の違いが大きいとき、「X が Y の原因である」という。確率の違いが大きいことは因果的影響力が大きいことを示している。X を操作できる行為者がいることが因果関係の前提になっていることが、因果論の他の説と異なる部分である。このアイデアは実験により因果関係を確認する科学の営みから着想している。

なお、ここでの確率は頻度論的に、すなわち何度も試行が行われたときに事象が起こる頻度として考えるのがふさわしい。

5.3 操作・介入主義的な DP の解釈

情報提供者の問い「私が情報を提供することが原因となって、将来なにか悪いことが起きるだろうか？」(5.1 節) を操作・介入主義の観点で見てみよう。

紙面節約のため、重要な事象に表 1 の通り記号をつける。時系列は上から順とする。例えば B は C, P のあとにしか起き得ない。事象 $\neg C$ とは「特定の個人 i がシステムに情報提供しない」という事象である。

1	事象 C	個人 i がシステムに情報提供すること
2	事象 P	システムが特定の値を公開すること
3	事象 B	明確に定義された「悪い」(bad) こと

表 1 事象 C, P, B の定義

5.2 節で述べた操作・介入主義の立場では最初の問いは、選択 $C/\neg C$ と事象 B の因果関係を確かめる実験を通してリファインされる。まず、システムや他人の選択などの周辺状況（記号 γ で表す）を変えずに、事象 C と $\neg C$ のどちらが起きたか（個人 i が選ぶか）だけを変える操作・介入を行う。また、個人 i は気にかけている将来の「悪い」ことを明確・明瞭に選んで、これを事象 B とする。事象 C が選ばれたときに事象 B が起きる条件付き確率と、事象 $\neg C$ が選ばれたときに事象 B が起きる条件付き確率はそれぞれ

$$\Pr[B \mid C, \gamma], \quad \Pr[B \mid \neg C, \gamma]$$

と表記される。すると、この二つの確率の違いが個人 i の選択 $C/\neg C$ の「悪い」こと B への因果的影響力（因果関係の強さ）である。したがって最初の問いは「この二つの確率の違いが大きいか」と言い換えられる。

選択 $C/\neg C$ と周辺状況 γ をまとめて表現すると、より DP に近くなる。周辺状況のうち他人の選択以外は、値が公開される瞬間の $\mathcal{M}$ の暗黙のパラメータと考えることにする。残る他人の選択は、入力データベース $\mathbf{DB}$ に入る／入っている、個人 i 以外のレコードのことである。これは確率変数ではない、固定されたものとして $others$ と書くことにする。また、個人 i が提供するか悩んでいるレコードは r_x と書くことにしよう。そうすると上記の二つの確率はそれぞれ次のようになる。

$$\Pr[B \mid \mathbf{DB} = others \cup \{r_i\}], \quad \Pr[B \mid \mathbf{DB} = others]$$

一方で、DP が述べているのは「悪い」こと B（個人 i が曖昧さなく指定している）が起きる確率ではなく、事象 P が起きる確率の違いである。このギャップは 2 節で紹介した事後処理則を使えば埋めることが出来る。事後処理則を言い換えれば、「選択 $C/\neg C$ に事象 P だけを通じて依存している事象についても、DP はその生起する確率の違いを抑えることが出来る」となる。

以上から、DP はこの二つの確率の違いを小さくすることで、情報提供者の選択 $C/\neg C$ の「選択 $C/\neg C$ に事象 P だけを通じて依存している」事象 B への因果的影響力（因果関係の強さ）をコントロールしている、と説明できる。

事象 P「システムが特定の値を公開すること」について、どの値が公開されることに注意するか、という点はここまで言及していない。この点は自由であり、上記の確率の違いが最大となるように選ぶことにしても良いし、出力される値を確率的だとして確率の違いの期待値を考えても良い。ただし確率の違いが大きいほど選択 $C/\neg C$ と「悪い」こと B の因果関係が強いと考えるのだから、前者を選ぶのが妥当だろう。これは標準的な DP と同じ選択である。なお、後者を選ぶのが妥当だとする研究者もいる（[5], 3.2 節）。

5.4 他人のレコードをどう設定するか

前節では操作・介入主義の立場から、他人のレコード $others$ を周辺状況の一部として固定し、確率変数ではないとした。しかしここで「選択 $C/\neg C$ と事象 B の因果関係を確かめる実験」を行っているのは誰か、という問題が生じる。情報提供者や攻撃者の立場では他人のレコードを指定することはできないので、実際の他人のレコードを把握している存在が、デモンストレーションとして実験を行っている、と想定される。

では情報提供者だけの立場ではシステムの安全性を確認できないのだろうか。これには二つの回答がある。一つ

は、「選択 $C/\neg C$ と事象 B の因果関係」を最も強くしてしまう周辺状況を選ぶ、という選択である。これは DP が選んでいるもので、「最悪ケースを考える」としばしば説明されることである。この場合、情報提供者は最悪の状況を用意して実験を行うことになる。もう一つの回答は、他人のレコードを確率変数 **Others** とする、という選択である。この場合でも DP は $C/\neg C$ の B への因果的影響力（因果関係の強さ）をコントロールできる。

$$\begin{aligned}
&\Pr[B \mid \mathbf{DB} = \mathbf{Others} \cup \{r_i\}] \\
&\quad \le e^{\varepsilon} \Pr[B \mid \mathbf{DB} = \mathbf{Others}] \\
&\sum_o \Pr[B \mid \mathbf{DB} = o \cup \{r_i\}] \Pr[\mathbf{Others} = o] \\
\Longrightarrow \quad &\quad \le e^{\varepsilon} \sum_o \Pr[B \mid \mathbf{DB} = o] \Pr[\mathbf{Others} = o] \\
\Longleftrightarrow \ &\Pr[B \mid r_i \in \mathbf{DB}] \le e^{\varepsilon} \Pr[B \mid r_i \notin \mathbf{DB}]
\end{aligned}$$

ここで **DB** は入力データベースを表す確率変数であることに注意せよ。こちらの場合は因果論のうち操作・介入主義よりも「確率上昇説」に近くなる [19]。

5.5 操作・介入主義的な DP の解釈から分かること

将来の「悪い」ことを考えるのではなく、将来の「良い」ことを中心に考えることも出来る。個人 i の選択 $C/\neg C$ の「良い」ことへの因果的影響力は、「良い」ことが産出されたことへの個人 i の貢献度合い、とも理解できる。すると DP は選択 $C/\neg C$ にかかわらず、全員が平等・公平に「良い」ことへ貢献するようにする、と解釈できる。もしも C を選んだ人全員が平等にコストを払ったり報酬を得ているなら、誰も突出することなく全員が平等に貢献することは望ましいことである。

$C, \neg C$ のどちらを選んでも「悪い」こと B が起こる確率が変わらないことがある、というのは注意するべきである。例えば個人 i が他の場所で情報提供していたり、すでに情報を公開していた場合には、そちらが原因となって「悪い」ことが起きるかもしれない。あるいは他人のレコードなどの周辺状況から推測できることについては、DP は全く隠そうとしない。

この DP の特徴を McSherry [20] は “your secret” と “secret about you” の違いとして説明している。“your secret” は「あなた」しか知らず推測もできない私秘的な、プライベートな情報であり、“secret about you” は「あなた」の周辺から推測できることである。“your secret” は個人の秘密、“secret about you” は集団内の傾向とも表現できる。例えば「あなた」の好きな色は通常は “your secret” であるが、「あなた」が属す文化では特定の色が好まれる傾向があるなら、これは “secret about you” になる。そして McSherry は DP は “your secret” を守るが、“secret about you” については全く守らない、と説明している。

また、“your secret” は隠されるべきだが、公益のために統計的に研究される事実は “secret about you” だと思われる。DP がこの二つの区別を導入したのは、DP が歴史的に偉大である理由の一つである。

6. DP の情報理論による解釈：文献紹介

DP の情報理論による解釈については紙幅が足りずここでは紹介できなかった。この解釈を通じて DP と属性推定攻撃の関係を考えることが出来る。文献としては情報理論を活用する [22, 4] が基本的である。

7. 結論

本稿では、差分プライバシー (DP) の理解を深めるために、既存の多様な解釈を整理し、新たに因果関係の観点からの解釈を提案した。特に、操作・介入主義に基づく因果的な説明は、DP が「私がデータを提供することによって何か悪いことが起きるか？」という実務的かつ社会的に本質的な問いに対して、どのように保証を与えるかを明確にしたものである。また、ベイズ／頻度論的仮説検定による解釈も紹介することで、DP の定義がどのような攻撃者に対して、どの程度の保護を提供しているのかを具体的に把握できるようにした。

今後の課題としては、DP の社会的意義を非専門家にも伝わるように簡潔に説明する方法の確立や、プライバシーパラメータ ε, δ の現実的な運用指針の構築が挙げられる。本稿がそのような取り組みの一助となり、DP の実装と社会への応用がより進むことを期待する。

参考文献

[1] Borja Balle, Gilles Barthe, and Marco Gaboardi. “Privacy Profiles and Amplification by Subsampling”. In: *Journal of Privacy and Confidentiality* 10.1 (1 Jan. 15, 2020). ISSN: 2575-8527. DOI: 10.29012/jpc.726.

[2] Borja Balle et al. “Hypothesis Testing Interpretations and Renyi Differential Privacy”. Oct. 8, 2019. arXiv: 1905.09982 [cs, stat].

[3] Nicholas Carlini et al. *Membership Inference Attacks From First Principles*. Apr. 12, 2022. DOI: 10.48550/arXiv.2112.03570. arXiv: 2112.03570 [cs]. Prepublished.

[4] Paul Cuff and Lanqing Yu. “Differential Privacy as a Mutual Information Constraint”. In: *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. CCS ’16. New York, NY, USA: Association for Computing Machinery, Oct. 24, 2016, pp. 43–54. ISBN: 978-1-4503-4139-4. DOI: 10.1145/2976749.2978308.

[5] Damien Desfontaines and Balázs Pejó. “SoK: Differential Privacies”. July 10, 2020. arXiv: 1906.01337 [cs].

[6] Zeyu Ding et al. “Detecting Violations of Differential Privacy”. In: *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*. CCS ’18: 2018 ACM SIGSAC Conference on Computer and Communications Security. Toronto Canada: Association for Computing Machinery, Oct. 15, 2018, pp. 475–489. ISBN: 978-1-4503-5693-0. DOI: 10.1145/3243734.3243818.

[7] Cynthia Dwork. “Differential Privacy”. In: *Automata, Languages and Programming*. Ed. by Michele Bugliesi et al. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2006, pp. 1–12. ISBN: 978-3-540-35908-1. DOI: 10.1007/11787006_1.

[8] Cynthia Dwork and Aaron Roth. “The Algorithmic Foundations of Differential Privacy”. In: *Foundations and Trends® in Theoretical Computer Science* 9.3–4 (3–4 2014), pp. 211–407. ISSN: 1551-305X. DOI: 10.1561/0400000042.

[9] Cynthia Dwork et al. “Calibrating Noise to Sensitivity in Private Data Analysis”. In: *Theory of Cryptography*. Ed. by Shai Halevi and Tal Rabin. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2006, pp. 265–284. ISBN: 978-3-540-32732-5. DOI: 10.1007/11681878_14.

[10] Cynthia Dwork et al. “Our Data, Ourselves: Privacy Via Distributed Noise Generation”. In: *Advances in Cryptology - EUROCRYPT 2006*. Ed. by Serge Vaudenay. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2006, pp. 486–503. ISBN: 978-3-540-34547-3. DOI: 10.1007/11761679_29.

[11] Hongsheng Hu et al. “Membership Inference Attacks on Machine Learning: A Survey”. In: *ACM Comput. Surv.* 54 (11s Sept. 9, 2022), 235:1–235:37. ISSN: 0360-0300. DOI: 10.1145/3523273.

[12] Thomas Humphries et al. “Investigating Membership Inference Attacks under Data Dependencies”. Dec. 21, 2021. arXiv: 2010.12112 [cs].

[13] Hafiz Imtiaz and Anand D. Sarwate. “Symmetric Matrix Perturbation for Differentially-Private Principal Component Analysis”. In: *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. ICASSP 2016. Shanghai, China: IEEE Press, Mar. 1, 2016, pp. 2339–2343. DOI: 10.1109/ICASSP.2016.7472095.

[14] Bargav Jayaraman et al. “Revisiting Membership Inference Under Realistic Assumptions”. Jan. 13, 2021. DOI: 10.48550/arXiv.2005.10881. arXiv: 2005.10881 [cs, stat].

[15] Peter Kairouz, Sewoong Oh, and Pramod Viswanath. “The Composition Theorem for Differential Privacy”. In: *IEEE Transactions on Information Theory* 63.6 (6 June 2017), pp. 4037–4049. ISSN: 0018-9448, 1557-9654. DOI: 10.1109/TIT.2017.2685505.

[16] Daniel Kifer and Ashwin Machanavajjhala. “No Free Lunch in Data Privacy”. In: *SIGMOD ’11*. 2011. DOI: 10.1145/1989323.1989345.

[17] Daniel Kifer and Ashwin Machanavajjhala. “Pufferfish: A Framework for Mathematical Privacy Definitions”. In: *ACM Transactions on Database Systems* 39.1 (1 Jan. 2014), pp. 1–36. ISSN: 0362-5915, 1557-4644. DOI: 10.1145/2514689.

[18] Masanobu Kii, Atsunori Ichikawa, and Takayuki Miura. “Lightweight Two-Party Secure Sampling Protocol for Differential Privacy”. In: *Proceedings on Privacy Enhancing Technologies* (2025). ISSN: 2299-0984. URL: https://petsymposium.org/popets/2025/popets-2025-0003.php (visited on 12/05/2024).

[19] Douglas Kutach. **現代哲学のキーコンセプト 因果性**. Trans. by 相松 慎也. 岩波書店, Dec. 19, 2019. 230 pp. ISBN: 978-4-00-061380-4.

[20] Frank McSherry. *Lunchtime for Data Privacy*. Aug. 16, 2016. URL: https://github.com/frankmcsherry/blog/blob/master/posts/2016-08-16.md (visited on 06/10/2022).

[21] Sebastian Meiser. *Approximate and Probabilistic Differential Privacy Definitions*. 277. 2018. URL: http://eprint.iacr.org/2018/277 (visited on 03/18/2022).

[22] Darakhshan J. Mir. “Information-Theoretic Foundations of Differential Privacy”. In: *Foundations and Practice of Security*. Ed. by Joaquin Garcia-Alfaro et al. Red. by David Hutchison et al. Vol. 7743. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 374–381. ISBN: 978-3-642-37118-9. DOI: 10.1007/978-3-642-37119-6_25.

[23] Reza Shokri et al. “Membership Inference Attacks against Machine Learning Models”. Mar. 31, 2017. arXiv: 1610.05820 [cs, stat].

[24] Theresa Stadler, Bristena Oprisanu, and C. Troncoso. “Synthetic Data - Anonymisation Groundhog Day”. In: USENIX Security Symposium. Nov. 13, 2020.

[25] Anvith Thudi et al. *Bounding Membership Inference*. Dec. 18, 2022. DOI: 10.48550/arXiv.2202.12232. arXiv: 2202.12232. Pre-published.

[26] Samuel Yeom et al. “Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting”. May 4, 2018. arXiv: 1709.01604 [cs, stat].

[27] 山本 展彰. “介入主義を応用した法的因果関係論の構想”. In: **阪大法学** 72.6 (Mar. 31, 2023), pp. 136–98. DOI: 10.18910/91001.