公募シンポジウム
シンポジウム4
情報の安全な利活用とセキュリティ基盤技術
2021年11月20日(土) 09:10 〜 11:10 C会場 (2号館1階211)

[3-C-1-03] 歯列データによる擬似縦断データ構築の検討
A study of data integration by dentitions

*三本 知明[1,2]、橋本 真幸[1]、横山 浩之[1]、宮地 充子[2] （1. 国際電気通信基礎技術研究所、2. 大阪大学）
*Tomoaki Mimoto[1,2], Masayuki Hashimoto[1], Hiroyuki Yokoyama[1], Atsuko Miyaji[2] （1. Advanced Telecommunications Research Institute International, 2. Osaka University）
キーワード：privacy, dentitions, longitudinal data

あらゆる分野でデータ解析が進められているが、同時にプライバシに関する問題も顕在化しつつある。特に各個体を追跡した時系列データである縦断データはプライバシ性が高く、プライバシを保護しながら利用するには大きくデータを撹乱させる必要があり、有用性を損なう可能性がある。そこで同一個体を追跡するのではなく、類似する個体の別時間のデータを繋ぎ合わせることで有用かつプライバシを保護した疑似的な縦断データの構築が可能と考える。本稿ではその第一ステップとして、個人の状態を表現しうる歯列に着目し、実際の歯列データを用いたデータ統合を検討する。

歯列データによる擬似縦断データ構築の検討

三本知明[*1,3]、橋本真幸[*1]、横山浩之[*1]
野崎一徳[*2] 、玉川裕夫[*2]、宮地充子[*3]
*1 国際電気通信基礎技術研究所、*2 大阪大学歯学部、
*3 大阪大学大学院工学研究科

A Study on Data Integration by Dentitions

Tomoaki Mimoto[*1,3] , Masayuki Hashimoto [*1] , Hiroyuki Yokoyama [*1]
Kazunori Nozaki [*2], Hiroo Tamagawa[*2], Atsuko Miyaji [*3]
*1 Advanced Telecommunication Research Institute International,
*2 Graduate School of Dentistry Osaka University,
*3 Graduate School of Engineering Osaka University

Longitudinal data is high dimensional and contains a lot of information, which can be utilized in various fields, such as medical, finance, and advertising. Personal data including longitudinal data is converted to anonymized datasets, which need to strike a balance between both privacy and utility. We consider pseudo-longitudinal data that is a dataset that tracks similar individuals over the course of long time may achieve the balance. In this paper, we focus on dentitions and discuss the possibility for constructing pseudo-longitudinal data using dentitions.

Keywords: privacy, dentitions, longitudinal data

1. はじめに

近年様々な分野において、個人データの利活用に向けた取り組みが進められている。それと同時にプライバシ保護にも注目が集まっており、国内外で法改正やプライバシ保護に向けた様々な活動が行われている[1]。

個人データは大きく横断データと縦断データの二種類に分類される。前者はある一時点のデータをあらわし、そのデータを分析することで、その時点のデータ間の関係や全体の傾向などを知ることができる。後者は時系列データとも呼ばれ、時間軸に沿ったデータ表現であり、縦断データを用いることで、各個人の行動パターンや時間経過による変化を分析することが可能となる。通常、縦断データは個人名と生年月日などの組み合わせや個人 ID 等の識別子を用いて個人を同定し、時間毎の同一個人のデータを結合することで縦断データとする。縦断データは複数の横断データから構築されるため、横断データと比較して有用性が高いデータであると同時に含まれるプライバシ情報も多くなる。そのため、プライバシ強度を高めるためにデータ加工を実施すると、各タイムスロットにおいて必要以上にデータが加工され、結果として利用価値の低いデータとなる可能性がある。

本稿では縦断データのプライバシと有用性の両立に向けて、疑似縦断データの可能性を検討する。疑似縦断データでは、従来のように識別子ではなく、個人を確率的に識別可能なキー（以下では確率的識別子と呼ぶ。）を用いたデータ突合を行う。この疑似縦断データは類似する確率的識別子を有する個人から構築されるため、適切に確率的識別子を選択することで、同一個人でなくとも同じ特徴を備えた個人のデータから構築されることとなる。したがって、疑似縦断データを利用することで、プライバシリスクと有用性の両立ができる可能性がある。本稿では確率的識別子として歯列データに着目し、どのように擬似縦断データを構築するかを考える。歯列データは身元確認活動での個人識別に用いられるなど、個人を確率的に識別することができる確率的識別子である。また、歯列状態と健康状態に関わりがあるように、歯列データからある程度の個人の特徴を捉えることができる。

2. 既存研究

プライバシ保護と情報利活用の両立に向けて、これまで様々な研究が行われている。k-匿名化[2]や差分プライバシ[3]を用いたデータ撹乱は主要なプライバシ保護技術であり、すでに様々なシーンで利用されている[4,5]。しかしこれらの技術はデータの次元が大きくなると、プライバシと有用性を両立させることが難しくなる。特に縦断データは追跡時間が長くなるとそれだけ次元数が増加し、従来のプライバシ保護技術を用いると有用性が著しく低下するという課題がある。

3. 目的

本稿では同一個人ではなく、類似した個人のデータを紐付けた擬似縦断データを用いることで、プライバシと有用性の両立ができる可能性があると考え、その第一歩として疑似縦断データと確率的識別子との関係を実験的に評価する。具体的には確率的識別子の一つである歯列データに着目し、歯列のみからどの程度正しく個人識別ができるかを評価、疑似縦断データ構築の可能性を検討する。特に歯列は時間経過による変化があり、また歯列パターンによって将来の歯列変化に差があるという特徴があるため、これらの特徴を考慮した評価を実施する。

4. 準備 4.1 実験データ

本稿では、実際の 2016 年から 2020 年のある地方自治体における歯科検診データセットを利用する。なお、当該データセットに含まれる被験者からはデータの公衆衛生学的目的での利用について同意を得た上で検診を実施している。歯列データセットには約 12,400 件のレコードが含まれており、各レコードには患者のイニシャルと性別、年齢、誕生日、検診日、医療機関、歯列、喫煙に関する情報、その他既往歴等が含まれている。同一個人の判定は、患者のイニシャル、性別、年齢、誕生日を元にハッシュ値を生成し、この値が一致するかどうかで決定する。まず、データセットから 5 年間に複数回検査を受けている個人を調べ、更にその中から最も古い検診レコードと最も新しい検診レコードを抽出し、その歯列データのみを実験に利用する。このデータセットをDとする。歯列デ

ータは検診レコードのうち歯の有無のみを抽出し、それぞれ1,0とした。

4.2 表記

本稿で扱う表記法を記す。

・D:歯科検診データセット。データセットには 5 年間で複数回検診を受けた個人の最も古い検査と最も新しい検査の歯列データのみが含まれている。

・$r_i = \{\boldsymbol{r}_i^{old}, \boldsymbol{r}_i^{new}\}$:データセットに含まれる$i$番目の個人。$\boldsymbol{r}_i^{old}, \boldsymbol{r}_i^{new}$はそれぞれ$D$における$r_i$の最も古い歯列と最も新しい歯列をあらわし、$\boldsymbol{r}_i^* \in \{0,1\}^{28} (* \in \{old, new\})$とする。

・r_{ij}: $\boldsymbol{r}_i^{old}, \boldsymbol{r}_j^{new}$間の類似度。$r_{ii}$は同一患者の最も古い歯列と最も新しい歯列の類似度をあらわす。

・$\boldsymbol{r}_i[k]: \{0,1\} \to \{0,1\}: \boldsymbol{r}_i^{old}[k]$から$\boldsymbol{r}_i^{new}[k]$への歯列の変化。例えば$\boldsymbol{r}_i[k]: 1 \to 0$は$\boldsymbol{r}_i^{old}[k] = 1 \to \boldsymbol{r}_i^{new}[k] = 0$をあらわす。

5. 実験

歯科検診データセットDを用いて歯列データが確率的識別子としてどの程度機能するかを実験的に評価する。評価にあたり、各歯列データ間の類似度r_{ij}を用いる。以下ではデータセット$D' \subseteq D$から次の行列式を用いる。

$$A_{D'} = \begin{pmatrix} r_{11} & \cdots & r_{1n} \\ \vdots & \ddots & \vdots \\ r_{n1} & \cdots & r_{nn} \end{pmatrix}$$

ここでr_{ij}はr_i^{old}とr_j^{new}の類似度を表し、$r_{ij} = d\left(r_i^{old}, r_j^{new}\right) = 1 - \frac{H\left(r_i^{old}, r_j^{new}\right)}{28}$とする。ここで$H(\cdot)$はハミング距離を表し、$0 \le r_{ij} \le 1$を満たす。

5.1 歯列パターンを考慮した同定実験

まず、元々の歯科検診データセットDからA_Dを構築した。図1はA_Dの一部を視覚化したものであり、$\forall r_{ij} = \max_{1 \le j \le n} d\left(r_i^{old}, r_j^{new}\right)$に色を付けている。したがって$r_{ii}$、すなわち対角線に位置する要素が色付けされている場合、他の個人の歯列と比較して同一個人の歯列の距離が最も近いことを示している。

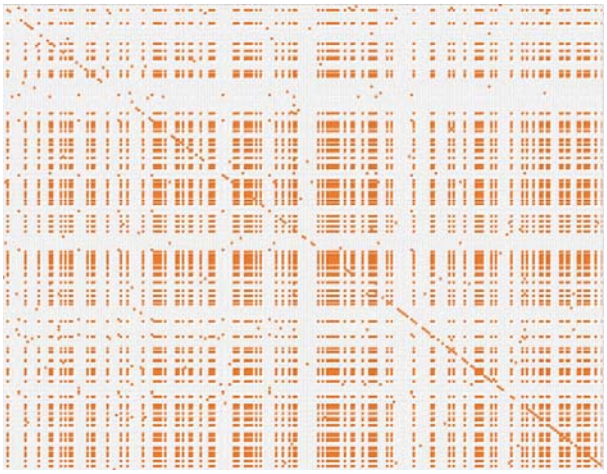

図1:A_Dの一部

図1では同一個人の新旧の歯列の類似度が最も高い傾向にあるが、それと同時に一部の歯列データは別の個人の歯列とも最も高い類似度を示していることがわかる。類似度計算には歯の有無のみを利用するため、パターン数としては約 2 億 7 千万(=2^{28})通り考えられるが、歯列パターンには偏りがあると考えられる。Dについて歯列パターンを確認したところ、$\boldsymbol{r}_i^* = \{1\}^{28}, \{0\}^{28}$、すなわちすべての歯が揃っているレコードとすべての歯を失っているレコードがそれぞれ全体の 30.5%, 14.5%存在しており、この2パターンのみで全体の半分弱を占めていることが確認された。これらの歯列以外のパターンは 828 パターン存在し、各パターン平均 10.3 個のレコードが確認できた。図2は$\boldsymbol{r}_i^* = \{1\}^{28}, \{0\}^{28}$を除いた歯列パターンの分布を表す。なお、この中でもいずれかの臼歯を失っているパターンが特に頻出していた。

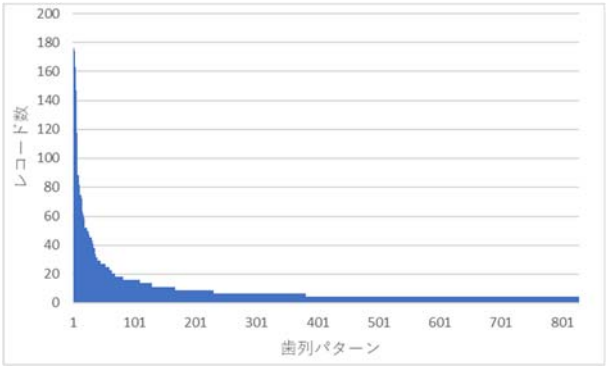

図2:歯列パターン数と各パターンのレコード数

今回は歯列の有無のみを確率的識別子として扱うことを想定しているため、図1の結果からも分かるように、$\boldsymbol{r}_i^* = \{1\}^{28}, \{0\}^{28}$であるような$r_i$を含むデータセットから精度の高い疑似縦断データを構築することは困難である。しかし$\boldsymbol{r}_i^* = \{1\}^{28}, \{0\}^{28}$を除いた歯列パターンは比較的分散しているため、次にDから$\boldsymbol{r}_i^* = \{1\}^{28}, \{0\}^{28}$であるような$r_i$を削除したデータセット$X$から$A_X$を構築した。図3は$A_X$の一部を視覚化したものであり、図1と同様$\forall r_{ij} = \max_{1 \le j \le n} d\left(r_i^{old}, r_j^{new}\right)$に色を付けている。結果から大半の同一個人の新旧の歯列の類似度が最も高い傾向にあることがわかる。また図1と比較して、別の個人の歯列との類似度が最も高い、すなわち$r_{ij}(i \ne j)$が最大となるi, jの組が大幅に削減されている。しかし依然として$r_{ij}(i \ne j)$が最大となるi, jの組が多数存在しており、より精度の高い疑似縦断データが求められる場合、さらなる検討が必要である。

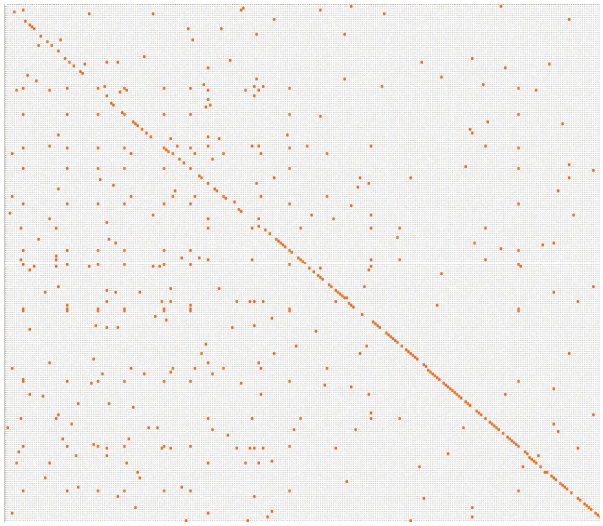

図3:A_Xの一部

以下では確率的識別子の特性を元に、個人同定の確率の変化を確認する。具体的には1. 咬合支持のない歯は将来的に抜けやすい、2. 歯列には不可逆性があり、$\boldsymbol{r}_i[k]:0\rightarrow 1$となる$r_i$は存在しない、という二つの仮定を元に、類似度の計算方法を調整する。

5.2 咬合支持を考慮した同定実験

データセットXを確認したところ、$\boldsymbol{r}_i[k]:1\rightarrow 0$となる要素を持つレコード、すなわち二回の検査の間に任意の歯を喪失した個人は 147 件（全体の 34.7%）存在し、$\left(r_i^{old}[k]=0\right)\wedge\left(\boldsymbol{r}_i[\hat{k}]:1\rightarrow 0\right)$（$r[\hat{k}]$は$r[k]$の噛み合わせ）を満たす個人、すなわち咬合支持がない歯が二回の検査の間で喪失した個人は 109 件（全体の 25.7%）存在することが分かった。これは歯を喪失した個人の実に 74.1%に相当するため、咬合支持の有無による歯の喪失確率の差を考慮することで確率的識別子としての精度の向上が見込める。そこでr_i^{old}とr_j^{new}に含まれる$\left(\boldsymbol{r}_i^{old}[k]=0\right)\wedge\left(\boldsymbol{r}_i^{old}[\hat{k}]=1\right)\wedge\left(\boldsymbol{r}_i^{new}[\hat{k}]=0\right)$となる$\boldsymbol{r}_i^{new}[\hat{k}]=0$の数を$\delta_{ij}$とし、$\Delta_{ij}^1=\delta_{ij}/28$としたとき、$r_i^{old}$と$r_j^{new}$の類似度を$r_{ij}=d\left(r_i^{old},r_j^{new}\right)+\Delta_{ij}^1$として置き換え、$A_X$を計算した。図 4 は$A_X$の一部を視覚化したものであり、これまで同様最も類似度が高いr_{ij}に色付けしてある。

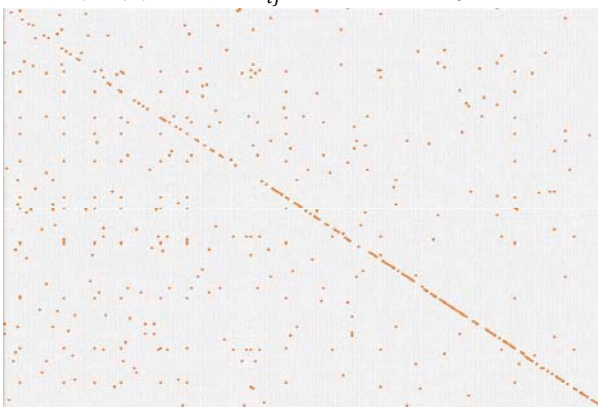

図4：A_Xの一部

図4は図3と比較して、r_{ii}が最大の類似度となる組は同程度であり、一方でr_{ij}が最大となるi,jの組は削減されていることがわかる。したがって、咬合支持を考慮することで、より精度の高い疑似縦断データの構築が可能であると考えられる。

5.3 不可逆性を考慮した同定実験

データセットDは 39.5%ものr_iが$\boldsymbol{r}_i[k]:0\rightarrow 1$と変化する要素を持つことが確認された。このことから補綴治療後の歯は 0 ではなく 1 として判定されていることが予想される。今回はどのように確率的識別子を扱うことで、どの程度正確な疑似縦断データの構築が可能であるかを検討することを目的としているが、人為的に変化させた歯列は確率的識別子の突合確率を低下させる可能性が高い。そこで、$\boldsymbol{r}_i[k]:0\rightarrow 1$と変化する要素を持つレコードを、補綴治療を受けた個人と仮定し、これらのデータを削除したデータセットYを構築する。また、類似度は不可逆性を考慮した重み付け処理を追加する。具体的には、$\left(\boldsymbol{r}_i^{old}[k]=0\right)\wedge\left(\boldsymbol{r}_j^{new}[k]=1\right)$となる$k$が存在するとき、$\Delta_{ij}^2=-\infty$、それ以外は$\Delta_{ij}^2=0$とし、$r_i^{old}$と$r_j^{new}$の類似度を$r_{ij}=\max\left(0,d\left(r_i^{old},r_j^{new}\right)+\Delta_{ij}^2\right)$として置き換え、$A_Y$を計算した。図5は$A_Y$の一部を視覚化したものであり、これまで同様最も類似度が高いr_{ij}に色付けしてある。

図5は同一個人の歯列が高確率で高い類似度を持ち、さらに別の個人の歯列間の類似度が最大となるケースが大幅に減っており、これまでの実験結果と比較して飛躍的に精度が向上していることが分かる。このようにかなり強い条件が与えられた場合、確率的識別子を用いて精度の高い疑似縦断データの構築が可能になると考えられる。

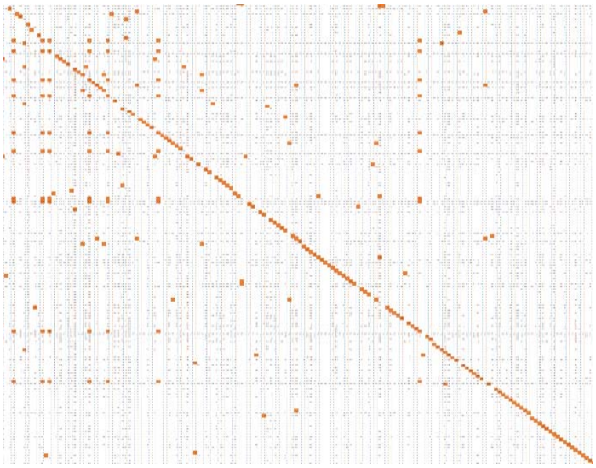

図5：A_Yの一部

5.4 定量評価

最後にこれまでの条件から歯列のみを確率的識別子として用いた場合、どの程度正確な縦断データの構築が可能であるかを$A_{D'}$を用いて定量的に評価する。具体的には、①各パターンのレコードの組数、②$r_{ii}=\max_{1\le j\le n} d\left(r_i^{old},r_j^{new}\right)$となる$r_i$の数、すなわち同一個人の歯列が最も高い類似度となった組数、③$r_{ii}=r_{ij}=\max_{1\le j\le n} d\left(r_i^{old},r_j^{new}\right)(i\neq j)$となる$r_j$の数、すなわち同一個人の他に最も高い類似度となる歯列を持つ別の個人の数、を表1にまとめた。各パターンの詳細は次の通り。

・パターン 1：データセットDに対して類似度を$r_{ij}=d\left(r_i^{old},r_j^{new}\right)$として$A_D$を構築。

・パターン 2：データセットDから$\boldsymbol{r}_i^*=\{0\}^{28},\{1\}^{28}$を含む$r_i$を削除した$X$に対して類似度を$r_{ij}=d\left(r_i^{old},r_j^{new}\right)$として$A_X$を構築。

・パターン 3：データセットXに対して類似度を$r_{ij}=d\left(r_i^{old},r_j^{new}\right)+\boldsymbol{\Delta}^1$として$A_X$を構築。

・パターン 4：データセットXから$\boldsymbol{r}_i[j]:0\rightarrow 1$となる$j$を持つ$r_i$を削除した`$Y$に対して類似度を$r_{ij}=d\left(r_i^{old},r_j^{new}\right)+\boldsymbol{\Delta}^2$として$A_Y$を構築。

・パターン 5：データセットYに対して類似度を$r_{ij}=d\left(r_i^{old},r_j^{new}\right)+\boldsymbol{\Delta}^1+\boldsymbol{\Delta}^2$として$A_Y$を構築。

表1：各パターンの評価結果

パターン	①	②	③
1	424	266	93.65
2	305	158	1.13
3	305	143	0.93
4	176	123	0.63
5	176	121	0.63

まず、パターン1では歯列データをそのまま確率的識別子として扱っており、およそ 62.7%（266/424 組）が最も古い検査と最も新しい検査の歯列の類似度が最大となる一方で、同等の類似度を持つr_jが 93.65 件存在する。これは歯列パターンの偏りが大きく、特に$r_i^* \in \{0\}^{28}, \{1\}^{28}$を含む$r_i$の存在によるところが大きい。次にパターン 2 では、$r_i^* \in \{0\}^{28}, \{1\}^{28}$を含む$r_i$を削除しており、その結果対象データが 71.9%（305/424 組）程度まで減少している。また、同一個人の歯列の類似度が最大となる組は 51.8%(158/305 組)あり、パターン1よりも下がっている。しかし、最大の類似度となる歯列パターンを持つ別の個人は 1.13 組と大幅に削減できており、パターン1よりも高い確率で同一個人のデータを用いた疑似横断データの構築が可能となる。これは歯列データのみを利用して疑似縦断データを構築する際、46.9%(1/2.13)の確率で同一個人のレコードが突合されることを意味しており、高い精度の疑似縦断データの構築が可能である。

パターン3ではさらに咬合支持を考慮することで最大の類似度となる歯列パターンを持つ別の個人は 0.93 組まで削減できている。

パターン 4 は更に強い条件として歯列の不可逆性を考慮している。本稿では補綴治療後の歯は 1 として判定されると想定されるため、$\boldsymbol{r}_i[j]: 0 \rightarrow 1$となる$j$を持つ$r_i$を取り除いて評価を実施している。したがって対象データは更に減少し、176 組である。しかし、同一個人の歯列の類似度が最大となる組は 69.8%(123/176)あり、パターン1よりも高い割合となる。さらに最大の類似度となる歯列パターンを持つ別の個人も 0.63 組に抑えられている。これは歯列データのみを利用して疑似縦断データを構築する際、61.3%(1/1.63)の確率で同一個人のレコードが突合されることを意味しており、高い精度の疑似縦断データの構築が可能である。

パターン 5 は咬合支持の特性と不可逆性を考慮しているが、咬合支持の特性に比べて不可逆性の影響が強く、ほとんどパターン 4 と同等の結果となった。

6. まとめ

本研究では疑似縦断データの構築に向け、確率的識別子としての歯列データの有効性を評価した。実験では歯列データrのパターンは理論上2^{28}と大量にあるが、偏りが大きく、特に$r = \{0\}^{28}, \{1\}^{28}$のパターンが多くを占める。それ以外のパターンに関しては比較的分散しており、単純な距離を用いた類似度評価であっても、歯列データが確率的識別子として機能しうることが分かった。さらに実験では歯列の特性を考慮することで、7 割ほどのレコードが 60%程度の確率で同一個人の識別が可能であることが確認できた。本稿では歯列データを$r = \{0,1\}^{28}$で評価したが、齲蝕歯の有無など他の情報を用いることで更に精度の向上を見込める。

今後の課題としては、実際に確率的識別子に対する類似度評価手法と、生成される疑似縦断データのプライバシ強度の関係を定量的に評価する必要があり、さらに生成した疑似縦断データのプライバシおよび有用性の関係も明らかにする必要がある。

参考文献

1) https://www.jst.go.jp/kisoken/crest/project/45/14532150.html
2) L. Sweeney. ″k-anonymity: a model for protecting privacy″ International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), pp. 557-570, 2002.
3) C. Dwork. Differential privacy. In Proc. of ICALP 2006, Vol. 4052, pp. 1-12, 2006.
4) Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. 2014. Rappor: Randomized aggregatable privacy-preserving ordinal response. In Proceedings of the 2014 ACM SIG.
5) Differential Privacy Team, Apple. Learning with privacy at scale. 2017.