3I2-OS-20-7

データを逐次公開する際のプライバシー保護

Some Methods for Privacy-Preserving Re-publication

菊池 亮 Ryo Kikuchi　五十嵐 大 Dai Ikarashi　濱田 浩気 Koki Hamada　千田 浩司 Koji Chida

日本電信電話株式会社
NTT Corporation

1. はじめに

近年，購買履歴や行動履歴等のデータを蓄積・分析し，商品のリコメンドやより良い都市開発等に活かしたいといった要望がある．しかし，このような個人に紐づくデータの利活用はプライバシーの問題があり第三者への提供や分析のアウトソーシング等が難しく，またデータを収集した者にとっても繊細な扱いが要求されるといった問題がある．そのような中，データに対し適切なプライバシー保護措置を行うことでプライバシーの保護と統計分析を両立する匿名化技術が注目されている．

匿名化技術では，データがどの程度プライバシーを保護できているかどうかを定量的に表すため，幾つかのプライバシー指標が提案されている．その中でも k-匿名性 [8]，その派生である ℓ-多様性 [7] 及びそれらを満たすアルゴリズムは，近年最も研究が盛んな匿名化技術の一つである．しかし，k-匿名性及びその派生は確率的手法に適用できないという問題があったため，五十嵐らは k-匿名性を確率空間へと拡張させ，Pk-匿名性 [4] と呼ばれるプライバシー保護指標を提案している．

1.1 データの逐次公開

データを公開する際には様々な公開方法が考えられるが，本論文では，データを逐次公開する状況を想定する．例として，ある地区での人の動きを時系列順に公開することが挙げられる．逐次公開では，各時点での公開データの変化や依存関係が存在する．例えば前述の例では，人の移動による場所移動や，地区への出入りによる人の増減が発生する．

このような逐次公開が行われている場合，一般に k-匿名性及び Pk-匿名性だけでは匿名化処理として十分でない．なぜなら，k-匿名性及び Pk-匿名性は静的なデータを 1 度だけ公開するモデルであり，逐次公開のような動的なデータの変化，例えば時系列ごとの値の変化およびレコードの増減は考慮していないからである．

そのため，近年では逐次公開に対する匿名化技術が研究されており，m-不変性 [9] や ℓ-希少性 [3] といった指標が提案されている．しかし，これらの指標は k-匿名性と同様に確率的手法に適用できない指標となっており，例えば撹乱再構築法 [1] のような確率的手法には適用できない．

またこれらとは別に，五十嵐らは 1 回の公開に一部の属性のみを公開するというモデルにおいて，逐次公開する際に満たすべき指標として多重 Pk-匿名性を提案している [6]．

1.2 本論文の成果

本論文ではまず，既存の逐次公開に対するプライバシー保護指標を概説する．さらに，確率的方式にも適用可能な Pk-匿名性を拡張し，逐次公開するデータが満たすべきプライバシー指標を提案する．

提案するプライバシー指標を用いることで，確率的な匿名化技術，例えば維持置換撹乱 [2] やノイズ付加 [5] を適用したデータが逐次公開して良いかどうか判断することができるようになる．

連絡先: NTT セキュアプラットフォーム研究所, 〒 180-8585 東京都武蔵野市緑町 3-9-11, {kikuchi.ryo, ikarashi.dai, hamada.koki, chida.koji}@lab.ntt.co.jp

2. 準備

本章では，プライバシー保護処理を行う対象となるテーブルや，基本的な保護処理の考え方について説明する．

2.1 匿名化を行うデータ

本論文で扱う匿名化技術は，レコードと属性からなるテーブルをプライバシー保護処理の対象としている．

表 1: テーブルの例

ID	国籍	趣味
1	日本	サッカー
2	中国	野球
3	日本	野球
4	中国	野球

直感的な理解としては，表 1 にあるように，テーブルには国籍，趣味などの属性があり，各行のレコードには各属性に対し属性値が記述されているものである．

また，k-匿名性の文脈においては，属性を準識別子とセンシティブ属性という 2 つに大別している．準識別子とは年齢や住所のように，単独では個人が特定できないものの組み合わせることによって個人を特定できうる情報のことである．センシティブ属性とは，センシティブであるがゆえに誰にも知られておらず，個人特定に全く役に立たない属性を指す．

2.2 k-匿名性

k-匿名性とは，匿名化後テーブルを見ても個人が特定できないことを目指す指標であり，あるテーブルにおいて，“同じ準識別子を持つレコードが，本人以外に少なくとも $k-1$ 人存在する” という状態を表す．例えば前述のテーブル例であれば，ID1 と ID3 のレコードの趣味を “球技” と変更すれば，変更後のテーブルは 2-匿名性を満たす．

2.3 Pk-匿名性

Pk-匿名性とは，あるテーブルにおいて，どのような背景知識を持つ攻撃者であっても “任意の個人を $\frac{1}{k}$ 以上の確率で当てることができない” という性質を指す．紙面の都合上，詳細な定義は [4] を参照されたい．

3. k-匿名化されたデータの逐次公開

本章では既存の k-匿名性の拡張である逐次公開のための指標と対策について述べる．なお，k-匿名法では，あらかじめ属性は準識別子とセンシティブ属性の 2 つに分けられているとする．

3.1 m-不変性

m-不変性では，センシティブ属性に対する知識はないが，各公表時の参加者及び準識別子を知識として持つ攻撃者を想定し

ている．このような場合，各テーブルに通常の k-匿名法を施しただけでは，複数のテーブルのセンシティブ属性をキーにして個人特定が可能になってしまう場合がある．例えば，ある時点のターゲットのセンシティブ属性候補は $\{a, b, c\}$，次の時点でのセンシティブ属性候補は $\{b, d, e\}$ であった場合，ターゲットのセンシティブ属性は b であることがわかってしまう．

これを避けるため，任意の人物について，最初に公開されたデータでセンシティブ属性候補が $\{a, b, c\}$ ならば，それ以降でもかならずセンシティブ属性候補が $\{a, b, c\}$ となるように追加人物，もしくは “おとり”を配置するアルゴリズムを提案している．

しかし，この m-不変性はレコードのセンシティブ属性は変わらないという制約がある．この制約は，多くの応用において強い制約事項となってしまう．

3.2 ℓ-希少性

m-不変性ではセンシティブ属性は不変であったが，ℓ-希少性はセンシティブ属性のうち変化するものと不変なものが混在する場合のプライバシー保護指標である．例えば，センシティブ属性が病歴であった場合，例えば HIV は不変なセンシティブ属性，Flu は可変なセンシティブ属性と考えることができる．このとき，不変なセンシティブ属性を持っていないレコードを “不変なセンシティブ属性を持っている”と仮定し複数レコードを観察すると，矛盾が生じることがある．この矛盾を使い，個人特定が可能となる場合がある．

これを避けるため，テーブルに含まれるレコード群を “Cohort ベース分類の原則” と “Role ベース分類の原則” と呼ばれる 2 原則によって分類・匿名化することで，ℓ-希少性を満たすことを示した．

しかし，この手法はセンシティブ属性がカテゴリ属性である必要がある．

4. Pk-匿名化されたデータの逐次公開

本章では，k-匿名性の確率空間版である Pk-匿名性を，ある特定の逐次公開に耐えるよう拡張した多重 Pk-匿名性を紹介する．なお，以下の Pk-匿名化ではセンシティブ属性は存在せず，どの属性も匿名化処理が行われるという前提とする．

4.1 多重 Pk-匿名性

多重 Pk-匿名性とは直観的に “複数の匿名テーブルの一部を見ても，どのテーブルについても Pk-匿名性を破れない”ことを表している．

この定義は本来，あるテーブルのうち一部の属性を抜き出して匿名化し公開するといった操作を複数回行う際のプライバシー保護を考えるための指標であるが，逐次公開においても適用可能な指標である（詳しくは文献 [6] を参照されたい）．しかし，多重 Pk-匿名性では，攻撃者は各公表時の参加者を知らないという前提を置いている．これは，m-不変性で想定している攻撃者よりも弱いものとなっている．

5. 提案指標：属性多重 Pk-匿名性

m-不変性や ℓ-希少性では確率的手法に適用できず，多重 Pk-匿名性は想定する攻撃者の知識が m-不変性に比べ弱いという問題がある．そこで，本章では確率的手法に適用可能であり，且つ攻撃者が m-不変性と同様の知識を持っていたとしても Pk-匿名性を満たすことができる，というプライバシー保護指標を与える．

まず，テーブル保護を形式的に議論するための定義を与える．データ保護処理前のテーブル τ を $\tau : \mathcal{R} \to \mathcal{V}$ であるような関数とする．ただし R はレコードの集合であり，$\mathcal{V}$ は，$\mathcal{A}$ はテーブルの属性の集合，$\mathcal{V}_a$ を各属性の属性値としたとき $\mathcal{V} = \prod_{a \in \mathcal{A}} \mathcal{V}_a$ である．また，シャッフル関数 $\pi : \mathcal{R} \to \mathcal{R}$，データ保護処理 $\delta : (\mathcal{R} \to \mathcal{V}) \to (\mathcal{R} \to \mathcal{V}')$ が存在し $\delta(\tau) = \tau' \circ \pi$ を満たすものとし，それぞれの関数 τ, τ', δ, π に対応する確率変数を T, T', Δ, Π とおく．さらに攻撃者を確率密度関数 f で表すものとする．

定義 5..1. $\mathcal{R}_0, \ldots, \mathcal{R}_N$ をレコード集合の組，k を 1 以上の任意の実数とし，$T'_0, \ldots, T'_{n-1}$ を，それぞれレコード集合 $\mathcal{R}_0, \ldots, \mathcal{R}_N$ を持つテーブル $T_0, \ldots, T_{n-1}$ に対し，確率的関数 Δ に従う確率過程 $\Delta_0, \ldots, \Delta_{n-1}$ により匿名化し，さらに $\Pi_0, \ldots, \Pi_{n-1}$ にてシャッフルした n 個の匿名化データベースの組であるとする．

このとき，$T'_0, \ldots, T'_{n-1}$ のインスタンスの列 $\{\tau'_0, \ldots, \tau'_{n-1}\}$ が属性多重 Pk-匿名性を持つとは，任意の r，任意の $T_0, \ldots, T_{n-1}$ の同時確率密度関数 $f_{(T_0, \ldots, T_{n-1})}$，任意の $i < n$ と任意の $r' \in \mathcal{R}_i$ について，

$$\Pr\left[\Pi_i(r) = r' \mid T'_0 = \tau'_0 \ldots, T'_{n-1} = \tau'_{n-1}\right] \leq \frac{1}{k}$$

が満たされることである．

この定義は確率空間で定義されており，撹乱再構築法等の確率的手法に適用可能である．また，属性多重 Pk-匿名性と比べ，確率密度関数は同時確率密度関数である．これは，攻撃者は m-不変性で考えられているような，レコードの追加及び削除の情報を持つということである．

6. まとめ

本論文では，データの逐次公開に着目し，既存の結果の整理を行った．さらに，Pk-匿名性を拡張し，確率的手法にも適用可能な逐次公開用プライバシー指標を定義した．この定義を満たすことができれば，撹乱再構築法等を用い，データを逐次公開することができる．

今後の課題としては，提案した指標を満足する方式の立案，例えば指標を満たす維持置換撹乱のパラメータ ρ の導出が挙げられる．

参考文献

[1] Agrawal, R., Srikant, R.: “Privacy-preserving data mining”. 2000 ACM SIGMOD international conference on Management of data, 2000.

[2] Agrawal, R., Srikant, R., Thomas, D.: Privacy Preserving OLAP. 2005 ACM SIGMOD international conference on Management of data,pp. 251-262, 2005.

[3] Bu, Y., Fu, A.W., Wong, R.C.W., Chen, L., Li, J.: “Privacy Preserving Serial Data Publishing by Role Composition”. 2008 Proceedings of the VLDB Endowment, pp.845-856, 2008.

[4] 五十嵐 大, 千田 浩司, 高橋 克巳.: “k-匿名性の確率的指標への拡張とその適用例”. コンピュータセキュリティシンポジウム 2008(CSS2009), 2009.

[5] 五十嵐 大, 千田 浩司, 高橋 克巳.: “数値属性における, k-匿名性を満たすランダム化手法”. コンピュータセキュリティシンポジウム 2011(CSS2011), 2011.

[6] 五十嵐 大, 千田 浩司, 濱田 浩気.: “秘匿計算とランダム化によるハイブリッド匿名化システム”. 暗号と情報セキュリティシンポジウム 2012(SCIS2012), 2012.

[7] Machanavajjhala, A., Kifer, D., Gehrke, J., Venkitasubramaniam, M.: “ℓ-Diversity: Privacy beyond k-anonymity”. ACM Transaction on Knowledge Discovery from Data, Vol. 1, No. 1, Article 3, Publication data, pp.1-52, 2007.

[8] Sweeney, L.: “k-anonymity: a model for protecting privacy”. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, Vol. 10, No. 5, pp.557-570, 2002.

[9] X. Xiao and Y. Tao.: “m-invariance: Towards Privacy-Preserving Re-publication of Dynamic Datasets”. 2007 ACM SIGMOD international conference on Management of data, pp. 689-700, 2007.