Computer Security Symposium 2012
30 October – 1 November 2012

# ユーザ存在／不在確率の範囲を限定した分散匿名化手法と医療データによる評価

竹之内 隆夫 †‡　　川村 隆浩 ‡　　大須賀 昭彦 ‡

† 日本電気株式会社 情報・ナレッジ研究所
211-8666 神奈川県川崎市中原区下沼部 1753
takenouchi@bu.jp.nec.com

‡ 電気通信大学 大学院情報システム学研究科
182-8585 東京都調布市調布ヶ丘 1-5-1

**あらまし** 複数機関が保持するユーザのパーソナル情報を結合・分析し，新たな知見を得ることが期待されている．特に医療情報のようなパーソナル情報はプライバシに関わるため，結合のための情報開示を必要最小限にすることや個人特定を防ぐことが求められ，そのための技術として分散匿名化が注目されている．しかし既存手法では，双方の機関のユーザ集合が一致しない場合にユーザのパーソナル情報がその機関に保持されているか否かというユーザ存在が漏洩する問題があった．そこで本論文では，ユーザ存在を隠蔽した分散匿名化手法を導入し，実際の診療機関のレセプトデータを用いて疾病の相関ルール抽出や診療回数の相関分析を行う際の有効性評価を行う．

# Distributed Anonymization with Limited Range of User Presence/Absence and Evaluation with Medical Information

Takao TAKENOUCHI†‡　　Takahiro KAWAMURA‡　　Akihiko OHSUGA‡

†Knowledge Discovery Research Laboratories, NEC Corporation
1753 Shimonumabe Nakahara-ku, Kawasaki, Kanagawa 211-8666, JAPAN
takenouchi@bu.jp.nec.com

‡Graduate School of Information Systems, The University of Electro-Communications
1-5-1 Chofugaoka, Chofu, Tokyo 182-8585, JAPAN

**Abstract** Personal information collected by service providers is expected to combine for getting valuable knowledge. Because personal information, such as medical information, is privacy related information, it is required to combine it with minimum disclosure and anonymize it to prevent revealing sensitive information. Thus, several researchers have been investigating distributed anonymization methods. However, when sets of the users in providers are different, the presence of users might be revealed. In this paper, we introduce a distributed anonymization method with limited range of users' presence/absence and show the evaluation results of association rule mining and correlation analysis on real medical information.

## 1　はじめに

近年，複数の機関が保持するユーザのパーソナル情報を結合・分析し，新たな知見を得ることが期待されている．例えば，複数の医療機関が保持する医療データを結合・分析することで，医療の質向上に役立てることが期待せれている [1]．本論文では，一般病院である病院 A と専門病院である病院 B が持つ患者の疾病情報を結合

し，研究機関 C に提供する場合を考える．例えば，病院 A は表 1(a) のテーブル ($T_A$) のように病院 A の患者の疾病コード「疾病 A」を，病院 B は表 1(b) のテーブル ($T_B$) のように病院 B の患者の疾病コード「疾病 B」と，その疾病の進行区分「分類」(例:ガンの進行ステージ等) を保持しているとする．そして，$T_A$ と $T_B$ を結合した結合匿名テーブル $T^*$ を研究機関 C へ送る．すると，研究機関 C は病院 A と病院 B の両方に通院する患者を紐付けた分析が可能となる．

このとき，例えば米国の HIPAA(Health Insurance Portability and Accountability Act) 法における必要最小限の情報開示の要件 (minimum necessary requirements)[2] のように，医療情報のようなパーソナル情報を結合する際の情報開示は最小限にする必要がある．また，パーソナル情報を組み合わせると，その組み合わせからのユーザの特定が可能になり，他人に知られたくない情報が知られてしまう恐れもある．そこで機関が持つ情報を必要最小限の開示に留めながら結合し，新たな情報を生成する，分散匿名化手法が注目されている [3, 4].

しかし既存の分散匿名化の手法では，双方の機関のユーザ集合が一致しない場合に，ユーザのパーソナル情報がその機関に保持されているか否かというユーザ存在が他方の機関に漏洩してしまう問題があった．例えば，性病の専門病院等への通院を他の一般の診療所等には知られたくないと考えられるため，ユーザ存在はユーザのプライバシに関わる情報といえる．

本論文の主な貢献は次の 2 つである．まず，著者らが [5] で提案した *δ-max-site-presence* を拡張し，ユーザ存在だけでなくユーザ不在の確率も指定できる指標として *δ-site-presence* を提案する．次に，この指標を満たしたユーザ存在/不在確率の範囲を限定した分散匿名化手法を導入し，実際の患者のレセプトデータを用いて有用性を評価する．本論文は，以下のような構成になっている．まず，2 章で関連研究を示す．次に，3 章で分散匿名化におけるユーザ存在の隠蔽の課題について説明する．続いて，4 章で *δ-site-presence* を提案し，この指標を満たす分散匿名化手法を導入する．そして 5 章で提案手法の有効性を，6 章で計算量と通信量を評価し，最後に 7 章で本論文の内容をまとめる．

# 2 関連研究

匿名化とは，ユーザを特定できないようにパーソナル情報を加工することであり，匿名化の指標として $k$-匿名性 [6] が知られている．$k$-匿名性とは，複数組み合わせるとユーザを特定できる可能性のある属性の組合せを準識別子 (quasi-identifier,QI)，他人に知られたくない属性をセンシティブ属性 (sensitive attribute,SA) とした時，あるテーブルにおいて QI の属性値によって識別されるレコード (ユーザ) が $k$ 個以上あるような状態のことを言う．$k$-匿名性を満たすために QI の属性値を一般化して，より抽象的な値にする方法が知られている．

複数の機関が保持するテーブルを結合して匿名化する処理を分散匿名化と呼ぶ．[3, 4] では，Top Down アプローチとセキュア計算 (secure computation)[7] を組み合わせた手法で，分散匿名化を実現している．これは，QI の属性値を最も一般化されている状態 (「*」など) から徐々に詳細化する手法である．ここで詳細化とは，QI の属性値で識別されるユーザ集合を，ある境目で分割することである．この分割の境目となる属性値を分割点と呼ぶ．例えば，年齢を「20 才」という分割点で分割すると，「20 才以上」と「20 才未満」に分割することになる．分割後のユーザ集合のユーザ ID は，分割をした機関から相手の機関に送信され，共有される．

# 3 分散匿名化における課題

## 3.1 分散匿名化

本論文では，機関 A,B が以下のようなテーブル ($T_A$,$T_B$) を持ち，結合匿名テーブル ($T^*$) を生成する前提とする．

$$T_A(ID, QI_A), T_B(ID, QI_B, SA), T^*(QI_A, QI_B, SA)$$

ここで，$ID$ は機関 A,B で共通のユーザ ID，$QI_A$，$QI_B$ は機関 A,B が持つ QI である．

分散匿名化では，必要最小限の開示に留めながらテーブルを結合し匿名化を行う．これは，異なる機関で完全な信頼関係を築くのは困難であり，テーブルを全て開示するのは危険であると考えられているためである．但し，本論文では機関 A,B は semi-honest[8] であるとする．

表 1: 結合匿名テーブルによるユーザ存在の漏洩

(a) 病院Aの$T_A$

| ID | 疾病A |
|---|---|
| User 1 | 110 |
| User 2 | 140 |
| User 3 | 155 |
| User 6 | 165 |
| User 7 | 171 |
| User 8 | 190 |

(b) 病院Bの$T_B$

| ID | 疾病B | 分類 |
|---|---|---|
| User 1 | 550 | I |
| User 2 | 540 | II |
| User 4 | 580 | I |
| User 5 | 560 | II |
| User 6 | 500 | I |
| User 7 | 521 | II |
| User 9 | 520 | I |
| User 10 | 510 | II |
| User 11 | 525 | I |
| User 12 | 500 | II |

(c) $T^*$(漏洩する場合)

| 疾病A | 疾病B | 分類 |
|---|---|---|
| 000-149 | 530-599 | I |
| 000-149 | 530-599 | II |
| 150-199 | 500-529 | I |
| 150-199 | 500-529 | II |

(d) $T^*$(漏洩しない場合)

| 疾病A | 疾病B | 分類 |
|---|---|---|
| 000-159 | 530-599 | I |
| 000-159 | 530-599 | II |
| 160-199 | 500-529 | I |
| 160-199 | 500-529 | II |

### 3.2 ユーザ存在の漏洩の課題

既存の分散匿名化では，双方の機関のユーザ集合が一致している前提であった．しかしユーザ集合が一致しない場合，自機関のテーブル ($T_A$,$T_B$) と結合匿名テーブル ($T^*$) の比較によってユーザ存在を推測できてしまう問題が発生する．例えば病院 A の $T_A$ が表 1(a)，病院 B の $T_B$ が表 1(b) であり，$T^*$ が表 1(c) であったとする．この時，$T^*$(表 1(c)) では疾病 A が 000-149 である患者は 2 名，$T_A$(表 1(a)) でも疾病 A が 000-149 に該当する患者は 2 名である．このことから病院 A は，User1,2 は確実に $T^*$(表 1(c)) に含まれていると推測できる．さらに，$T^*$(表 1(c)) に含まれるユーザは病院 A と病院 B の双方に存在する共通ユーザであることから，病院 A は，User1,2 の 2 名が確実に機関 B にも存在すると推測できる．それに対し，$T^*$ が表 1(d) のように疾病コードが「160」で分割されていた場合，病院 A は，User1,2,3 の 3 名のうち 2 名が病院 B に存在することまでしか推測できない．

## 4 提案指標と解決手法

### 4.1 $\delta$-*site-presence* の提案

ユーザ存在を推測できてしまう問題を解決するために，新たな指標を提案する．我々は [5] において，分散匿名化におけるユーザ存在の推測の可能性を示す指標として $\delta$-*max-site-presence* という指標を提案した．しかしこの指標は，ユーザ存在の可能性は評価できたが，ユーザ不在の可能性については評価できなかった．そこで我々は，新たに $\delta$-*site-presence* を提案する．

**定義 1** $T_A$,$T_B$ を機関 $A$,$B$ が持つテーブル，$T^*$ を結合匿名テーブルとする．そして，$T^*$ のうち機関 $n \in \{A, B\}$ が持つ属性の属性値の組合せの集合を $\{v_{n,1}, \cdots, v_{n,m_n}\}$ とし，$v_{n,i} \in \{v_{n,1}, \cdots, v_{n,m_n}\}$ とおく．また，$v_{n,i}$ で識別されるテーブル $T_n$ のレコード数を $|T_n[v_{n,i}]|$，$v_{n,i}$ で識別されるテーブル $T^*$ のレコード数を $|T^*[v_{n,i}]|$ と表現する．この時，以下の式で示されるように，機関 $n \in \{A, B\}$ の各 $v_{n,i}$ によるユーザ存在の推測の可能性が $\delta_{max,n}$ 以下かつ $\delta_{min,n}$ 以上である時，$T^*$ は $\{\delta_{min,A}, \delta_{max,A}, \delta_{min,B}, \delta_{max,B}\}$-*site-presence* を満たすと定義する．

$$\delta_{min,n} \leq \frac{|T^*[v_{n,i}]|}{|T_n[v_{n,i}]|} \leq \delta_{max,n} \quad \forall n \in \{A, B\} \quad (1)$$

例えば表 1(d) のうち機関 A が持つ属性 (疾病 A) の属性の属性値の組合せの集合 $\{v_{A,1}, v_{A,2}\}$ は {000-159, 160-199} である．まず,「000-159」について考える．$T^*$(表 1(d)) のうち疾病 A が「000-159」であるレコードは 2 つであるので，$|T^*[v_{A,1}]|$=2 である．そして，$T_A$(表 1(a)) のうち疾病 A が「000-159」を満たすレコードは 3 つであるので，$|T_A[v_{A,1}]|$=3 である．よって，疾病 A の「000-159」についてはユーザ存在の推測の可能性は 2/3 である．同様に「160-199」についてや，機関 B が持つ属性 (疾病 B, 分類) の属性値の組合せの集合についても計算すると，表 1(d) は $\{\frac{2}{3}, \frac{2}{3}, \frac{1}{3}, \frac{1}{2}\}$-*site-presence* を満たすテーブルであることがわかる．

このようなユーザ存在／不在確率の範囲を限定した分散匿名化を実現するための分散匿名化手法は，以下の要件を満たしつつ，できるだけ詳細な $T^*$ を出力必要がある．

**要件 1** $T^*$ は $k$-匿名性と $\delta$-*site-presence* を満たすこと

**要件 2** 通信内容から $T^*$ よりも詳しい情報が極力漏れないこと

### 4.2 ユーザ存在／不在確率の範囲を限定した分散匿名化手法の導入

ユーザ存在／不在確率の範囲を限定した分散匿名化手法として，著者らが [5] において提案

した*δ-max-site-presence*を満たすダミーユーザ手法を導入し，さらに*δ-site-presence*を満たすように拡張する．なお，ダミーユーザ手法は既存の Mondrian アルゴリズム [9] を拡張した手法である．ダミーユーザ手法では，ダミーユーザという自機関に存在しないがあたかも存在するかのように扱うユーザを導入している．これにより，相手機関に通知されるユーザ ID が，存在するユーザなのか存在しないユーザなのかの区別を困難にでき，ユーザ存在を隠蔽したままの分散匿名化を実現できる．

ダミーユーザ手法は，まず機関 A,B 間で通信し，各機関内で内部匿名テーブル$T_n^*(n \in \{A, B\})$を生成する．その後機関 C が，機関 A,B から$T_n^*$を取得・結合し，$T^*$ を得る．以降で機関 A,B 間での$T_n^*$を生成方法について説明する．

#### 4.2.1 Step1:ダミーユーザの割当と初期化

最初に機関 A，B は自機関のダミーユーザを割り当てる．本手法では，双方の機関のユーザを包含する母集団ユーザ集合$U$を事前に知っているという前提を置く．ここで$U$は，機関 A に存在するユーザ集合を$U_A$，機関 B に存在するユーザ集合を$U_B$，機関 A,B のどちらにも存在しないユーザ集合を$U_O$としたき$U = U_A \cup U_B \cup U_O$ $(U_O \neq \phi, U_A \cap U_B \neq \phi)$となる．そして，機関 A が持つダミーユーザは$U - U_A$，機関 B は$U - U_B$となる．

次に内部匿名テーブル$T_n^*$を初期化し，最も一般化された状態にする．例えば，$U$={user1-15} であったとすると，$T_A^*$={user1-15,0-999}, $T_B^*$={user1-15,0-999} となる．

#### 4.2.2 Step2:分割点の決定と分割処理

続いて，$T_n^*$を分割していく分割処理を行う (図 1)．まず，機関 A,B は自機関のダミーユーザの準識別子$(QI_A, QI_B)$の属性値に適切な値を割り当てる．この値をダミー値と呼ぶ．ダミー値は，分割対象のユーザにおける存在ユーザの準識別子の属性値の分布に沿って割り当てられる．

次に，4.2.4 節で説明する分割点決定関数を用いて分割点を決定する．そして，決定した分割点で分割しても$k$-匿名性と*δ-max-site-presence*を満たせるかを確認し，指標を満たしている場合のみ$T_A^*$，$T_B^*$を分割する．例えば，機関 A が持つ$QI_A$が疾病コードであり，「150」で分割を行う場合は，$T_A^*$={{user1-10,0-149},{user11-15,150-999}} のように 2 レコードに分割される．そして，機関 A は機関 B に分割後のユーザ ID を送信する．機関 B は，受信した内容に従って$T_B^*$を分割する．この例では，$T_B^*$={{user1-10,0-999},{user11-15,0-999}} となる．

**function** split($U_p$:分割対象となるユーザ集合の $IDs$)
1: $U_p$ のダミーユーザのダミー値を更新
2: $d \Leftarrow$ 分割点決定関数を用いて分割点を決定
3: **if** $k$-匿名性と *δ-site-presence* を満たせない **then**
4: $U_p$ についての split 処理終了
5: **endif**
6: **if** $d$ は自機関の $T_n^*$ の分割点 **then**
7: $T_n^*$ を $d$ で分割し，分割後の $IDs$ を相手機関へ送信
8: **else**
9: 相手から分割後の $IDs$ を受信し，$T_n^*$ を分割
10: **endif**
11: $U_h, U_l \Leftarrow$ 分割後の $IDs$. split($U_h$),split($U_l$) を実行

図 1: Step2(分割処理) のアルゴリズム

ここで，$k$-匿名性と*δ-site-presence*を満たしているかを確認するために，*secure set intersection* [10] というセキュア計算 [7] のプロトコルを用いる．このプロトコルは，機関 A,B が持つ集合をお互いに隠蔽しながら，それらの集合の積集合や，積集合の要素数 (*set cardinality*) や，積集合の要素数と指定した値との大小関係 (*cardinality threshold*) を求めることが出来る．本論文では，紙面の都合で詳細は記載しないが，このプロトコルを用いる事で，ユーザ存在や属性値を隠蔽したままの指標確認を実現できる [5].

#### 4.2.3 Step3:ダミーユーザの削除

全ての分割処理が完了したらダミーユーザを削除し，共通ユーザ数を求める．この処理は，*secure set intersection*を用いることで求めることができる [5]．そして，最終的に生成される$T_A^*$と$T_B^*$は，例えば$T_A^*$={{user1-10,0-149},{user11-15,150-999}} と，$T_B^*$={{user1-10,0-599,「s1:2 名，s2:3 名」},{user11-15,600-999,「s1:1 名，s2:1 名」}} のようになる．

以上のような Step1〜3 までの処理によって，機関 A,B はお互いにユーザ存在を隠蔽しながら

内部匿名化テーブル $T_A^*$，$T_B^*$ を分割していく．

#### 4.2.4 ダミーユーザを考慮した分割点決定

ダミーユーザ手法の分割点決定関数は，従来の $k$-匿名性を満たすための Mondrian の分割点決定関数を拡張し，新たに $\delta$-*site-presence* も満たし易いようにしている．

まず，従来の Mondrian と同様に normalized range が最大となる属性を選ぶ．そして，その属性における分割点の候補となる属性値 ($x_i \in X$) を分割点候補 $c_i$ として，以下のように定義したダミーユーザのエントロピー ($E$) を計算する．

$$E(c,n) = -\sum_{U_i \in U_h, U_l} \frac{|du(n,U_i)|}{|U_i|} log(\frac{|du(n,U_i)|}{|U_i|}) \quad (2)$$

ここで $c$ は分割点候補の属性値であり，分割前のユーザ集合 $U_p$ を上位 $U_h$ と下位 $U_l$ へ分割することを意味する．また，$du(n,U_i)$ はユーザ集合 $U_i$ のうち機関 $n$ のダミーユーザの集合である．$E$ は，ダミーユーザが分割後のグループ内に偏りなく入る時に大きくなる．

次に，以下に定義したスコア値 $S$ を計算する．

$$\begin{aligned} S(c_i) &= \alpha(\frac{-L(c_i)}{\max_{x_j \in X}(L(x_j))}) \\ &+(1-\alpha)\frac{1}{2}\sum_{n \in A,B}(\frac{E(c_i,n)}{\max_{x_j \in X}(E(x_j,n))}) \quad (3) \\ L(c_i) &= \sum_{x_j \in X} |x_j - c_i| \end{aligned}$$

ここで $\alpha (0 \leq \alpha \leq 1)$ は，$E$ の影響を調整するための重みである．また，$L$ は $c_i$ の属性の各属性値 $x_i$ と $c_i$ の距離の和を意味する．median とは $L$ が最小となる点と言い換えることができるため，$\alpha$=1 とした時は $c_i$ が median の時に $S$ が最大となり，従来の Mondrian と同様に median が分割点に決定される．このように定義した $S$ を最大化させる分割点で分割を行うことで，分割後のユーザ集合にダミーユーザが偏りがなく入り，結果的に $\delta$-*site-presence* を満たしつつ多くの分割が可能になることが期待される．

また，分割点決定関数は，属性値やユーザ存在を隠蔽したまま計算する必要があるため，セキュア計算を用いる．本論文では紙面の都合で詳細は記載しないが，機関 A,B で分割点候補 $c_i$ の $E$ を計算する際には，*secure set intersection* を用いる．これにより，ユーザ存在や属性値を隠蔽した分割点決定が可能になる [5].

## 5 評価実験

提案手法をプロトタイプ実装し，有効性を評価した．実装は Java 1.6 で行い，仮想的に双方の機関で通信を行う構成で動作させた．

評価データには，JMDC(株式会社日本医療データセンター) が提供している，実際のレセプトデータ（診療報酬明細書）を用いた．このデータは，個人の特定はできないが病院間での個人データの結合はできるように，氏名や地域に関する情報は別コードに置き換えられている．このレセプトデータから，異なる診療科の病院のうち，共通の患者数が一番多い病院を病院 A，病院 B として抽出した．病院 A は約 3500 人の患者 ($U_A$) のデータを持つ内科の病院であり，病院 B は約 300 人患者 ($U_B$) のデータを持つ耳鼻科の病院である．そして，これら病院の共通の患者 ($U_A \cap U_B$) は約 230 人である．評価では，これらの患者とは別に，病院 A,B に通院していない患者 ($U_O$) を約 1430 人抜き出し，母集団の患者 ($U$) を約 5000 人とした．そして，2 つの病院間で患者の疾病履歴を結合して病気の相関を調べるというユースケースを想定し，$T_A$(ID, 病名 A1, 病名 A2)，$T_B$(ID, 病名 B1, 病名 B2, 分類) というデータ形式のテーブルを生成した．ここで ID は病院間で共通な患者の識別子,「病名 A1」と「病名 A2」は病院 A における直近に診療した 2 件の疾病の疾病コードである．病院 B の「病名 B1」と「病名 B2」も同様である．また,「分類」は疾病の進行を想定しており，今回の評価結果に影響が無いため疑似的に「I」と「II」をランダムに生成した．なお，結合テーブルの形式は $T^*$(病名 A1, 病名 A2, 病名 B1, 病名 B2, 分類) であり，{ 病名 A1, 病名 A2, 病名 B1, 病名 B2} が QI，分類が SA である．

評価は，$T^*$ に対してデータマイニングを行った場合に，マイニング結果にどの程度の誤差が発生するのかという観点で行った．評価手法は，既存の匿名化の研究 [11] と同様に，ある条件に合致するユーザ数をカウントするクエリ (“select

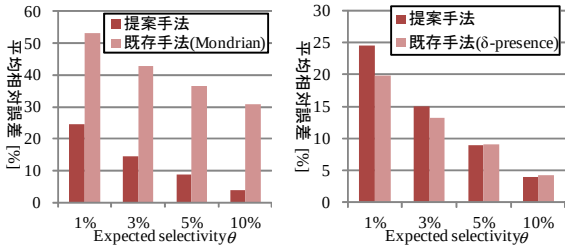


図 2: vs. 既存分散型




図 3: vs. 既存集中型

count(*) from $T^*$ where 条件部") の結果の相対誤差を計測するという手法である．この評価手法では，まずカウントされるユーザ数の割合の期待値 (*expected selectivity*) を $\theta(0\% < \theta < 100\%)$ とおいて，条件部に指定する検索範囲が全体の $\theta$ 倍になるようなクエリをランダムに生成する．つまり，$\theta$=10%と設定したら $T^*$ のユーザの約 10%が検索されるクエリとなる．そして，生成したクエリを用いて，匿名化前の結合テーブル $T_{AB}$($T_A$ と $T_B$ を単純に内部結合したテーブル) に対して得られたユーザ数を $act$，結合匿名テーブル $T^*$ に対して得られたユーザ数を $est$ とし，その相対誤差を $|act - est|/act$ で計算する．なお，条件部に利用する属性は 2 つとし，ランダムに選択した．また各評価値は，ランダムなクエリを 10,000 回生成し相対誤差を計測した値の平均である．

評価は大きく 3 つの観点で行った．まず，既存の分散匿名化手法と提案手法を比較し，提案手法の有用性の評価を行った．続いて，既存の集中型のユーザ存在隠蔽の匿名化手法と提案手法を比較評価した．最後に，QI に相関がある場合のマイニング結果に与える影響を評価した．

## 5.1 既存の分散匿名化との比較

最初に，提案手法となるダミーユーザ手法の有効性を評価するために，既存手法となる Mondrian を単純に分散環境に対応させた分散対応 Mondrian との比較を行う．この分散対応 Mondrian は，提案手法と比較するために $k$-匿名性だけでなく $\delta$-*site-presence* も満たしている際に分割を行い，最終結果では共通ユーザだけを出力する分散匿名化手法である．

図 2 に，提案手法と既存手法のそれぞれについて，$k$=2, $\delta_{max,A}$=$\delta_{max,B}$=0.99, $\delta_{min,A}$ = $\delta_{min,B}$ = 0.01, $\theta$={1%, 3%, 5%, 10%} として平均相対誤差を計測した結果を示す．なお，重み $\alpha$ は 0.5 として $E$ の影響を半分にしている．

| |
|---|
| A:急性気管支炎 ⇒B:急性副鼻腔炎 [s=3.1%,c=71.4%] |
| A:急性気管支炎 ⇒B:アレルギー性鼻炎 [s=3.1%,c=57.1%] |
| A:急性上気道炎 ⇒B:急性咽頭喉頭炎 [s=2.2%,c=60.0%] |

図 4: 病院 A と病院 B の疾病の相関ルール

この結果が示すように，$\theta$=3%の時の既存手法の相対誤差は約 40%と大きいが，提案手法の相対誤差は約 15%程度と小さい．これは，ダミーユーザのエントロピー ($E$) の追加や分割後のダミー値の更新により，ユーザ存在が隠蔽できるような分割点が選ばれたためである．

相対誤差は約 15%というのは，例えば相関ルールマイニングを行った際に得られる相関ルールの支持度 (support) や確信度 (confidence) の相対誤差が約 15%程度である事を意味している．図 4 は，匿名化前の結合テーブル $T_{AB}$ に対して相関ルールマイニングを行い，支持度 (図の s) が 2%以上，確信度 (図の c) が 50%以上となる疾病についての相関ルールを，支持度が高い順に出力した結果である[1]．この結果に示したように，支持度が 3.1%と 2.2%の相関ルールが得られている．もし，匿名結合テーブル ($T^*$) に対して相関ルールマイニングを行った場合は，これらの相関ルールの支持度に 15%の誤差が入るので 3.1%と 2.2%の相関ルールの支持度は約 2.6〜3.6%と約 1.9〜2.5%になる．この程度誤差であれば，得られた相関ルールに大きな差は無く，提案手法は十分有用であると考える．

続いて，機関 A,B の $\delta_{min}$ か $\delta_{max}$ を設定を変化させて相対誤差を計測した．図 5 に $\theta$=3%とした際の提案手法と既存手法の平均相対誤差を示す．なお，例えば $\delta_{max,A}$ を設定している際は他の $\delta$ の設定していない．

この結果が示すように，提案手法も既存手法も機関 A から見たユーザ存在の隠蔽の限界値 ($\delta_{max,A}$ として設定できる値の最小値，$\delta_{min,A}$ として設定できる値の最大値) が 0.75($\simeq$ 230/300) や，機関 B から見た限界値が 0.06($\simeq$ 230/3500) に近づくと急激に誤差が大きくなる．しかし，既

[1]図 4 に示した疾病は，鼻や咽喉頭の炎症が気道や気管支に到達した際に起こる良く知られた合併症である．

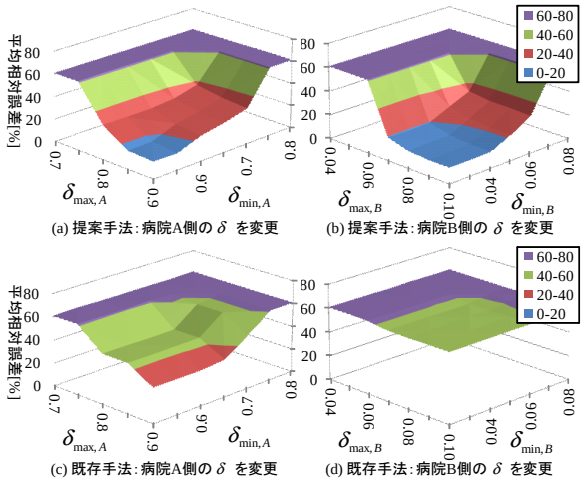


図 5: $\delta$ を変化させた際の相対誤差

存手法 (図 5(b)(d)) は $\delta_{max}$ や $\delta_{min}$ を限界値近くに設定していなくとも，誤差が 30〜50%もある．それに対し，提案手法 (図 5(a)(c)) は限界値付近でなければ相対誤差が小さい．つまり，提案手法は $\delta_{max}$ や $\delta_{min}$ を限界値近くに設定しなければ，相対誤差が小さくなるような有効な匿名化が行えることがわかった．

## 5.2 集中型のユーザ存在隠蔽との比較

次に，集中型 (非分散環境の匿名化) でのユーザ存在の隠蔽手法である $\delta$-*presence* を満たすための MPALM アルゴリズム [12] と比較し，分散型 (分散環境の分散匿名化) に対応した提案手法の有用性がほぼ同等であることを示す．集中型での既存手法は，あるテーブルと匿名テーブルにおけるユーザ存在を隠蔽する手法であり，提案手法のように機関 A と機関 B の双方からみた，ユーザ存在の推測を防ぐというものではない．そこで，公平な評価を行うために病院 B 側から見た $\delta_{min,B}$ と $\delta_{max,B}$ を設定せずに評価を行った．図 3 に $\theta$={1%, 3%, 5%, 10%} として提案手法と既存手法の平均相対誤差の値を計測した結果を示す．なお，その他のパラメータは 5.1 節と同じにした．

この結果が示すように，$\theta$ が 1〜3%の時は提案手法は既存手法よりも数%ほど誤差が大きい．これは，集中型の既存手法のほうがより多くの分割を行えるため，より小さい範囲のユーザ数のカウントであっても相対誤差を小さくできるからである．しかし，$\theta$ が 5〜10%の時は提案手法と既存手法の差はほぼ無い．この結果から，提案手法は集中型の既存手法と大きな差がなく，有効な匿名化が行えることがわかった．

## 5.3 データ相関がマイニングに与える影響

様々なマイニング結果の影響を調べるために，2 つの病院間の通院回数の相関分析を行うユースケースを想定した評価を行った．この評価では，病院 A と病院 B が持つテーブルとして，$T_A$(ID, 通院回数 A)，$T_B$(ID, 通院回数 B, 病名 B, 分類) というデータ形式のテーブルを生成し評価を行った．ここで「通院回数 A」と「通院回数 B」は病院 A,B における通院回数である．この評価では「通院回数 A」と「通院回数 B」を QI,「病名 B」と「分類」を SA とおいている．生成する結合匿名テーブルは $T^*$(通院回数 A, 通院回数 B, 病名 B, 分類) である．

表 2 に，匿名化前のデータに対して「通院回数 A」と「通院回数 B」の相関係数を求めた結果と，提案手法で既存手法で匿名化後のデータに対して相関係数を求めた結果を示す．なお，各種パラメータは 5.1 節と同じある．この結果が示すように，匿名化前の元データでの相関係数は-0.03 であり，相関は無い．しかし，既存手法で匿名化後のデータに対して相関係数を求めた場合は-0.40 と負の相関があるという結果になってしまう．それに対し提案手法の場合の相関係数は-0.02 であり，-0.03 とほぼ変わずに相関は無いという結果になった．

また，病院 A と相関がある病院 C(整形外科) についても同様の評価を行った．この場合も提案手法の方が既存手法よりも相関係数の誤差が小さい．しかし，相関が無かった病院 A,B 間とは違い，相関がある病院 A,C 間では相関係数の誤差が大きくなってしまっている (0.23 と 0.03)．これは，ダミーユーザを相関にそって割当てることができないため，QI に相関があるとダミーユーザが偏ってしまうためである．その結果，分割回数が減ってしまい，匿名化の精度が悪くなってしまう．これは今後の課題である．

表 2: 匿名化前後における病院間の相関係数

| 対象病院 | 匿名化前 | 提案手法 | 既存手法 |
|---|---|---|---|
| 病院 A, 病院 B | -0.03 | -0.02 | -0.40 |
| 病院 A, 病院 C | 0.23 | 0.09 | -0.48 |

# 6 計算量と通信量の評価

本章では提案手法の平均的な計算量と通信量のオーダーを算出する．ダミーユーザ手法では，Step2 のスコア値 $S$ を計算する処理が，各分割の各分割点候補について *secure set intersection* を実行するため，計算量と通信量が大きくなる．

まず，平均計算量を算出する．分割の 1 回目は分割対象のグループのサイズは全ユーザ数 $|U|$ となる (以後，$|U|$=N とおく)．本手法の分割では多少の偏りはあるが平均的に中央で分割されるので 2 回目以降のグループサイズは $\frac{N}{2}, \frac{N}{4}, \cdots$ となる．また，これらのグループの個数は $2, 4, \cdots$ となる．さらに，各グループにおける分割点候補は，グループのサイズとほぼ同じなので $N, \frac{N}{2}, \frac{N}{4}, \cdots$ となる．ここで，*secure set intersection* の計算量は，2 つの集合の要素数を両方とも M とおいたとき $O(MloglogM)$ となる [10]．よって，1 回目の分割での計算量は，サイズ $N$ の 1 つのグループに対する計算量を $N$ 個の分割点候補分行うので $O(NloglogN) \times 1 \times N$，2 回目の分割ではサイズ $\frac{N}{2}$ の 2 つのグループに対する計算量を $\frac{N}{2}$ 個の分割点候補分行うので $O(\frac{N}{2}loglog\frac{N}{2}) \times 2 \times \frac{N}{2}$，3 回目は $O(\frac{N}{4}loglog\frac{N}{4}) \times 4 \times \frac{N}{4}$ となる．そして，これらを足した値は以下の関係を満たす．

$$N^2loglogN + \frac{N^2}{2}loglog\frac{N}{2} + \frac{N^2}{4}loglog\frac{N}{4} \cdots$$
$$< (1 + \frac{1}{2} + \frac{1}{4} + \cdots)N^2loglogN < 2N^2loglogN$$

よって，平均計算量は $O(N^2loglogN)$ である．

続いて，平均通信量を算出する．*secure set intersection*[10] の通信量は，2 つの集合の要素数を両方とも M とおいたとき $O(M)$ である．1 回目の分割での通信量は，サイズ $N$ の 1 つのグループに対する通信を $N$ 個の分割点候補分行うので $O(N) \times 1 \times N$，2 回目は $O(\frac{N}{2}) \times 2 \times \frac{N}{2}$，3 回目は $O(\frac{N}{4}) \times 4 \times \frac{N}{4}$ となる．よって，先ほどと同様に平均通信量は $O(N^2)$ である．

# 7 まとめ

本論文では，分散匿名化において双方の機関のユーザ集合が異なる場合のユーザ存在の漏洩問題に対し，ユーザ存在が推測される可能性を示した *δ-site-presence* という指標を提案した．そして，この指標を満たすためのダミーユーザ手法を導入し，実際の患者のレセプトデータを用いて評価を行った．結果，患者が通院しているか否かを隠蔽しながらも相対誤差 15%以下でデータ分析が可能であることが確かめた．

## 謝辞

本研究の一部は，経産省の「平成 23 年度次世代高信頼・省エネ型ＩＴ基盤技術開発・実証事業 (レセプト情報等の利活用基盤の開発)」プロジェクトの成果である．

# 参考文献

[1] 内閣官房 IT 戦略本部，“新たな情報通信技術戦略 (平成 22 年 5 月 11 日決定)，” 2010.

[2] U.S. NARA, “Standards for privacy of individually identifiable health information,” Federal Register, vol.67, no.157, pp.53182–53273, 2002.

[3] N. Mohammed, B.C.M. Fung, K. Wang, and P.C.K. Hung, “Privacy-preserving data mashup,” Proc. EDBT’09, pp.228–239.

[4] P. Jurczyk and L. Xiong, “Distributed anonymization: Achieving privacy for both data subjects and data providers” Proc. DBSec’09, pp.191–207.

[5] T. Takenouchi, T. Kawamura, and A. Ohsuga, “Distributed data federation without disclosure of user existence” Proc. DBSec’12, pp.282–297.

[6] L. Sweeney, “k-anonymity: a model for protecting privacy,” Int. J. Uncertain. Fuzziness Knowl.-Based Syst., vol.10, pp.557–570, 2002.

[7] Y. Lindell and B. Pinkas, “Secure multiparty computation for privacy-preserving data mining,” J. Privacy and Confidentiality, vol.1, pp.59–98, 2009.

[8] O. Goldreich, Foundations of Cryptography: Volume 2, Basic Applications, 2004.

[9] K. LeFevre, D.J. DeWitt, and R. Ramakrishnan, “Mondrian multidimensional k-anonymity,” Proc. ICDE’06, p.25.

[10] M.J. Freedman, K. Nissim, and B. Pinkas, “Efficient private matching and set intersection,” Proc. EUROCRYPT’04, pp.1–19.

[11] X. Xiao and Y. Tao, “m-invariance: Towards privacy preserving re-publication of dynamic datasets,” Proc. SIGMOD’07, pp.689–700.

[12] M.E. Nergiz, M. Atzori, and C. Clifton, “Hiding the presence of individuals from shared databases,” Proc. SIGMOD’07, pp.665–676.