誤差を考慮した位置匿名化手法の提案

著者	清 雄一, 大須賀 昭彦
雑誌名	電子情報通信学会論文誌. D, 情報・システム
巻	J97-D
号	5
ページ	964-974
発行年	2014-05-01
URL	http://id.nii.ac.jp/1438/00009034/

誤差を考慮した位置匿名化手法の提案

清　雄一[†a)]　　大須賀昭彦[†b)]

Location Anonymization on the Basis of Accuracy

Yuichi SEI[†a)] and Akihiko OHSUGA[†b)]

あらまし　年齢，年収，趣味等のユーザ属性と，ユーザの行動履歴とを関連付けてマイニングすることで，ユーザ属性や位置情報に応じた適切なマーケティングや広告配信をすることが可能となる．しかし，あるユーザの行動履歴の一部を知る攻撃者にこの情報がわたると，関連付けられたユーザ属性と個人を結び付けられるリスクがある．従来研究において，ユーザの行動履歴を知る攻撃者に対してもユーザ属性と個人を結び付けられることを防ぐため，k-匿名性等の指標に基づく匿名化手法が多数提案されている．しかし，ユーザの位置情報には誤差が含まれていることが考慮されておらず，誤差がある環境下では個人が特定されるリスクが増加する．また，匿名化後のデータの有効性指標にも誤差が考慮されていない．本論文では，位置情報には誤差があるという現実的な環境を想定し，新しいプライバシー指標，匿名化後のデータにおける有効性指標，及びこれら指標に基づいた匿名化アルゴリズムを提案する．シミュレーション評価を実施し，従来手法と比べて匿名化後のデータの有効性を向上させ，同時に，個人が特定されるリスクを低減することを示す．

キーワード　ユビキタスコンピューティング，プライバシー，位置情報，匿名化

1. まえがき

ユーザの位置情報の履歴と，ユーザの性別や年齢，年収，居住地等のユーザ属性とを関連付け，ユーザ属性ごとの消費行動を分析する研究が行われている [1]．このような分析を実施することで，より適切なマーケティングや広告配信を行うことができるようになる．

本論文では，ユーザ属性や位置情報を直接取得していない事業者が，他事業者からこれらの情報を受領し，マイニングを実施する環境を想定する．マイニングを実施する企業は氏名や住所等の情報を排除した後のデータ（位置情報の履歴と，対応するユーザ属性）のみを受け取れば良い．しかし，例えばユーザ Alice が時刻 t に位置 (x, y) にいたという事実を知る攻撃者が存在した場合，その時刻と位置に対応するユーザ属性が Alice の属性であることが判明してしまう．

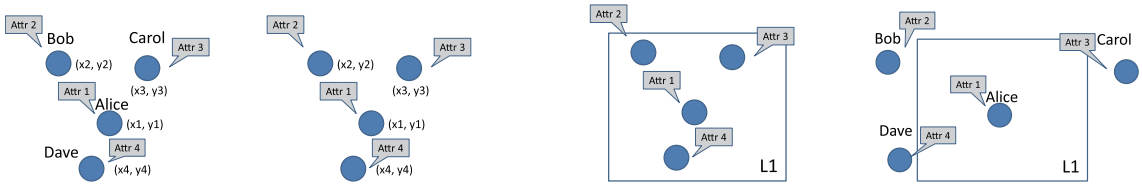
この問題を解決するため，各ユーザの位置情報を匿名化する研究が盛んに行われている．例えば，位置情報に関して k-匿名化 [2], [3] を行った場合，マイニング事業者へ提供する位置情報は点ではなくエリアで表現され，この匿名化エリアには k 人以上のユーザがいることが保証されるとされている．しかし，匿名化を行う事業者が把握している位置情報に誤差がある場合，匿名化エリアに実際には k 未満のユーザしかいない可能性がある．

また既存の k-匿名化手法におけるアルゴリズムが目標としていることは，匿名化エリアの最小化である．しかし，位置情報に誤差が含まれている場合，匿名化エリアにユーザが存在していない可能性もある．したがって，匿名化エリアの最小化だけではなく，そのエリアにユーザが実際に存在している確率についても評価指標の一つとすべきであると考えられる．

本論文ではこのように，位置情報に誤差がある環境の場合，k 人以上のユーザが存在すること保証するとしている既存研究の問題点を明らかにし，新しくプライバシー指標及び有効性指標を提案する．また，提案指標に基づいた，位置情報の誤差を考慮した匿名化手法の提案を行う．

本論文の構成は次のとおりである．**2.** では本論文が想定する環境や，解決を目指す問題について述べる．**3.** で関連研究を記述する．**4.** において本論文で提案

† 電気通信大学大学院情報システム学研究科，調布市
Graduate School of Information Systems, The University of Electro-Communications, Chofu-shi, 182–8585 Japan
a) E-mail: sei@is.euc.ac.jp
b) E-mail: ohsuga@euc.ac.jp



(a) Sensed data (b) Inadequate anonymization (c) Anonymization based on sensed data (d) True information

図 1 匿名化の例
Fig. 1 An example of anonymization.

する評価指標を，**5.** において匿名化手法を提案する．**6.** ではシミュレーション評価を実施する．考察を **7.** で行い，**8.** で本論文をまとめる．

2. 想定モデル及び問題定義

想定する位置情報の利用モデル，位置情報の誤差の扱いや，本論文で取り組む問題について述べる．

2.1 位置情報の利用モデル

どのような属性をもったユーザがどのような行動を取ったかに関するデータマイニングを行うことを想定する [1]．本論文における匿名化対象は，ある一時刻における位置情報である．各時刻において独立に位置情報を匿名化し，データマイニングに利用することを想定する．連続的な時間にわたる匿名化への拡張は将来課題とし，本論文では対象としない．

また，ユーザの位置情報や属性情報を保有する企業を**情報管理企業**，情報管理企業から匿名化したデータを受領する企業を**情報利用企業**と記述することにし，これらの企業が異なることを想定する．また，情報管理企業は信頼できるが，情報利用企業は信頼できないものと仮定する．

2.2 位置情報と誤差の表現

位置情報には誤差が含まれると想定する．モバイル端末の OS として広く利用されている iOS，Android，Windows Phone OS 等では，位置を緯度，経度で，誤差の大きさを円の半径で表している [4]～[6]．本論文でも同様に，情報管理企業は，中心座標 (x, y) 及び，位置情報の誤差として円の半径 r の情報を得られると想定する．ユーザ u の中心座標を x 座標及び y 座標で表し，(x_u, y_u) と表現する．また，誤差を表す円の半径を r_u とする．この円を各ユーザの**存在円**と呼ぶ．既存研究では，この誤差を無視して匿名化を行っている．また本論文では議論を簡単にするために，この存在円の中にユーザが必ず存在していると想定する．

2.3 取り組む問題の概要

情報管理企業が把握している情報が図 1 (a) であったとする．この情報をそのまま情報利用企業に提供すると，情報利用企業は，各ユーザの位置情報とユーザ属性を一意に特定できてしまう．この問題は，ユーザ名（ユーザ ID）を削除しただけでも十分な解決とはならない（図 1 (b)）．何故なら，提供された情報に「Alice」を示す識別子が無かったとしても，情報利用企業が仮に Alice が (x_1, y_1) に存在したことを知っている場合，その位置情報が Alice を示す識別子の役割を担うことになるためである．

3. に示す多くの研究では，図 1 (c) のように，k 人以上が存在する**匿名化エリア** $L1$ を導出し（図 1 (c) では $k = 4$），この k 人は $L1$ に存在する，ということだけを情報利用企業に提供する．結果，$L1$ に存在するユーザは必ず k 人以上存在するため，k 人のうちどれが Alice であるかを判断することができない．

しかし，位置情報に誤差がある場合，情報管理企業が信じている情報と，現実世界との間に乖離が生じる．例えば図 1 (d) に示すように，現実世界では Alice のみが $L1$ に存在する，ということもあり得る．

このとき情報利用企業が別の情報源から，エリア $L1$ には Alice しかいないという事実を把握している状況を考える．提供を受けた各ユーザのユーザ属性から，実際に $L1$ に存在する可能性が相対的に高いユーザ属性を推定できた場合，そのユーザ属性は Alice のユーザ属性である可能性が高いことが判明してしまう．

例えば，ユーザ属性に保有する電子機器の情報が含まれていて高性能な GPS 機器を保有するユーザ属性がある場合や，ユーザ属性に普段利用する Web サービス名が含まれていて頻繁に GPS 情報を利用することが推測されるユーザ属性がある場合，そのユーザ属性に紐付く位置情報は正しい可能性が高い．別の例として，ユーザ属性に趣味の情報が含まれており，特にその趣味の人が $L1$ に存在する店舗に行くことが多い

と考えられる場合は，このユーザ属性に紐付く位置情報は相対的に正しい可能性が高いと言える．

通常，k-匿名性を満たす匿名化が行われている場合，特定のユーザ Alice のユーザ属性は高々$1/k$ の確率でしか当てることはできないが，このような例の場合は，より高い確率で Alice のユーザ属性を推測できてしまう．本論文では，誤差を考慮した上で，匿名化エリアに一定の確率 w 以上で k 人以上が存在することを保証することでこの問題に対応する．

また，各匿名化エリアの面積は小さいほど有用な情報であると考えられるが，匿名化エリア $L1$ に存在するとされている各ユーザが，実際に $L1$ に存在するかどうかは誤差を考えると不明確であり，適切なデータマイニングを行うためには，実際に $L1$ に存在している可能性をできるだけ向上させる必要があると考えられる．

以上より本論文では，各匿名化エリアに確率 w 以上で k 人以上が存在することを保証し，各匿名化エリアの面積を小さくすると同時に，各匿名化エリアに存在するとされている各ユーザがそのエリアに実際に存在している確率を向上させることを目指す．より厳密には **4.1** で定義するプライバシー指標を満たした上で，**4.2** で定義する有効性をできるだけ向上させる．

3. 関連研究

ユーザを一意に特定できないように匿名化を行う指標の一つとして，k-匿名性が提案されている [7]．Alice の位置が (x, y) であるとき，$x_1 < x < x_2$，$y_1 < y < y_2$ を満たす x_1，x_2，y_1，y_2 を用意し，Alice が (x_1, x_2, y_1, y_2) の頂点によって表される矩形領域のエリアに存在するという情報のみを公開する．このとき，この領域に k 人以上のユーザが存在するようにエリアを構築する．領域内に k 人以上のユーザが存在するため，公開されるユーザ属性の中で，どのユーザ属性が Alice を表しているかを特定することができない．

k-匿名性を対象とする既存研究は，k 人以上のユーザが存在する匿名化エリアの面積を最小化することを目指している．この問題は NP 困難であることが示されているため [8]，計算量が少ないヒューリスティックなアルゴリズムが提案されている．位置情報に限定せず k-匿名化手法として Mondrian アルゴリズム [9] が広く利用されており，複数の位置匿名化手法においてもベースの手法として採用されている [10], [11]．

ユーザの位置情報を一時刻のみに限定せず，一定時間にわたる連続した位置の履歴を匿名化する研究も行われている [3], [12]．これらの研究は，複数の時刻 $t_1, t_2, \ldots$ における位置情報の履歴から，ある時刻 t における Alice の場所を知っている攻撃者に，t 以外の時刻に Ailce がどこにいたかを知られることを防ぐ．これらの研究ではある時刻における位置情報の k-匿名化を行っており，その匿名化に本論文で提案する手法を利用することができると考えられる．

ユーザ ID 及びそのユーザの位置情報を公開するシナリオを対象として，位置情報を匿名化する研究もある．このような研究では，他のユーザとの関係を考慮せずに位置情報を曖昧化することによって，攻撃者に正確な位置が伝わらないようにする方法も取られている [13]．Ardagna ら [14] のように，位置情報の取得誤差を考慮して匿名化を行っている研究もあるが，各ユーザ個別の位置情報に対する匿名化であって k-匿名化のように他ユーザとの関係が考慮されていない．k-匿名化を行うよう拡張することも可能であるが，この場合，本論文で指摘しているような通常の k-匿名化手法と同じ問題が生じる．

公開する位置情報から，そのユーザが学校にいたのか繁華街にいたのか等，セマンティックな情報が漏洩しないよう匿名化を行う研究もある [15]．各ユーザに対し，位置情報に基づいたサービスを提供する場合にはこのような匿名化が必要となる．しかし本研究では，各ユーザに対してサービスを提供するのではなく，ユーザを一意に特定しない状態でマイニングを行うシナリオを想定している．したがって，位置のセマンティックな匿名化は本論文のスコープ外とする．

4. 指標の提案

位置情報の誤差を考慮した場合に必要となる，プライバシー指標及び有効性指標を提案する．

4.1 プライバシー指標

プライバシー指標として (w, k)-匿名性を提案する．全ての匿名化エリアに対し，w 以上の確率で k 人以上が存在することを保証している状態を，(w, k)-匿名性が満たされていると定義する．

4.2 有効性指標

匿名化エリアを最小化すると有効性が向上するような指標が一般的に採用されている [2], [16]．本論文では誤差を考慮するため，そのエリア内に存在する確率も考慮する．

誤差を考慮した有効性指標は，ユーザ集合を U，ユー

ザ $u \in U$ が匿名化エリア $L(u)$ に分類され，ユーザ u がそのエリアに存在する確率を $p_{u,L(u)}$ としたとき，

$$Utility = \sum_{u \in U} \frac{(p_{u,L(u)})^{\alpha}}{|L(u)|} \quad (1)$$

と表すことができる．ここで $|L(u)|$ はエリア $L(u)$ の面積を表し，パラメータ α は存在確率を重視する度合いである．α は 0 以上の値を取り，値が大きいほど存在確率を重視する度合いが強まる．

5. 匿名化アルゴリズム

提案指標に基づく匿名化アルゴリズムを提案する．利用する主な変数やパラメータ名を表 1 に記す．

5.1 課　　題

本論文で提案する手法を構築するにあたり，考慮すべき課題を以下に挙げる．

（1） ユーザの存在密度に偏りがある

多くのユーザが存在している場所と，ほとんど存在していない場所が存在する．存在密度が高い部分を匿名化エリアの境目にしてしまうと，境界付近のユーザについて最悪の場合，存在確率を 25%にしてしまい，匿名化後のデータの有効性が低下する．

（2） 位置情報の誤差の大きさに偏りがある

位置情報の誤差が小さい人を優先して匿名化エリアを設定することにより，全体の有効性を向上させることができる可能性がある．

5.2 概　　要

提案手法は，エリア分割フェーズ及びエリア拡大フェーズを繰り返し，最後にエリア縮小フェーズを実施する．

エリア分割フェーズは，k-匿名化を行う手法として広く利用されている Mondrian アルゴリズムをベースとする．Mondrian アルゴリズムでは初期のエリアを最も抽象化されたエリアに設定し，k-匿名性が満たせなくなるまで分割を繰り返すトップダウン型のアプローチである．提案手法では，Mondrian アルゴリズムに基づいてエリアを分割する際に，それが (w,k)-匿名性を満たしているかを確認する．満たしている場合にのみ分割を行う．

表 1　Notation
Table 1　Notation.

N	全ユーザ数
$L(u)$	ユーザ u が存在する匿名化エリア
$\|L\|$	匿名化エリア L の面積
(x_u,y_u)	ユーザ u の位置情報の観測値における中心座標
C_u	ユーザ u が存在する可能性のある円（ユーザ u の存在円）
r_u	C_u の半径
U_L	匿名化エリア L に存在する確率が 0 より大きいユーザ集合
α	匿名化エリアに存在する確率を重視する度合い

通常の匿名化では匿名化エリアの面積最小化が目標となるが，本論文では各ユーザが匿名化エリアに実際に存在している確率も考慮するため，匿名化エリアを広げるほうが有効性向上につながる場合がある．エリアの境界に多くのユーザが存在していた場合がこれに該当する（前節の課題 1）．エリア分割フェーズ後にエリア拡大フェーズを設け，有効性が向上する場合には匿名化エリアを拡大する．

エリア分割フェーズとエリア拡大フェーズを繰り返し，それ以上分割できない状態になったとき，最後にエリア縮小フェーズを実施する．エリア縮小フェーズを実施しない場合，ある匿名化エリア $L1$ に存在するとされるユーザの中心座標（真の座標ではなく情報管理企業が把握している座標）は必ず $L1$ の範囲内にある．しかし，(w,k)-匿名性を満たしてさえいれば，ユーザの中心座標が当該エリア内にある必要はない．匿名化エリアに存在するユーザの位置情報の誤差の大きさに偏りがある場合は，誤差が小さいユーザ周辺に匿名化エリアを限定することで有効性が向上する可能性がある（前節の課題 2）．有効性が向上する場合は，ユーザの中心座標の位置にかかわらず，(w,k)-匿名性を満たす範囲内で匿名化エリアを縮小する．

以降で，これらのフェーズの詳細を記述する．

5.3 エリア分割フェーズ

x 座標と y 座標のうちより範囲が広いほうを対象とし，対象エリアに含まれるユーザの位置の中央値で分割を試みる（図 2 左）．図において，各ユーザの中心座標を黒点で，存在円を灰色の円で表している．分割後の各エリアに w 以上の確率で k 人以上のユーザが存在する場合にのみ分割を実行する．分割できない場合は，もう片方の座標について分割を試みる．分割されたエリアに対し，同様の処理を繰り返すことでトップダウン的に分割していく．

ユーザ u の中心座標を (x_u,y_u)，その誤差を r_u と表す．また，ユーザ u がエリア L に存在する確率を $P_1(L,u)$ とおく．$P_1(L,u)$ は，エリア L と，中心座標が (x_u,y_u) であり半径が r_u である円との重なり合っている面積から求めることができる．

エリア L に存在する確率が 0 より大きいユーザ集

合を U_L とすると，L に k 人以上のユーザが存在する確率 $P(L,k)$ は次式で表すことができる．

$$P(L,k) = 1 - \sum_{\{S|S\in\mathfrak{P}(U_L)\wedge|S|<k\}} \left[\prod_{u\in S} P_1(L,u)\cdot \prod_{u\notin S\wedge u\in U_L} (1-P_1(L,u))\right]. \quad (2)$$

$\mathfrak{P}(U_L)$ は，集合 U_L のべき集合を表す．式 (2) は，1 から，U_L に含まれるユーザのうち k 人未満だけが L に存在する確率を引いている．

ここで，集合 U_L の要素数が大きい場合，計算不可能なほど計算量が増大する場合がある．例えば，$|U_L| = 50$ で $k = 10$ である場合，$\{S \in \mathfrak{P}(U_L) \wedge |S| < k\}$ を満たす S の数は約 30 億通り存在する．この問題に対応するため，計算量を減らす方法を **5. 6** で述べる．

分割後の 2 エリアを $L1$，$L2$ としたとき，$P(L1,k) < w$ または $P(L2,k) < w$ の場合，分割を取りやめる．

5. 4　エリア拡大フェーズ

分割後に生じる二つの匿名化エリアのうち，片方のエリアを L と記す．以下の処理は両エリアに対して独立に実行する（図 2 右）．

分割して生じたもう片方のエリアと接する面を，境界面と呼ぶことにする（図 2 の A_0）．L の境界面を，有効性が最も向上する位置まで拡大することを目指す．匿名化エリアを縮小せず拡大するのみであるので，エリア分割フェーズで (w,k)-匿名性が満たされている場合は，エリア拡大フェーズ後でも必ず (w,k)-匿名性が満たされている．

まず，L に含まれる各ユーザ u の存在円 C_u を完全に含むよう，境界面を拡大する（図 2 の A_1）．拡大されたエリアを L_{max} とする．境界面をこれ以上拡大してもエリア内のユーザの存在確率は増加しないため，有効性が向上する可能性のあるエリアとしては L_{max} が最大のエリアである．元のエリア L と L_{max} の範囲内で有効性を最も向上させるエリアを求める．

ユーザの存在円は円形で表現されるため，L の境界面を拡大させることによって各ユーザの存在確率が増加する割合は徐々に減少する．したがって，L と L_{max} で挟まれる領域をパラメータとしたときの有効性は単峰関数となり，その極大値を求める問題に帰着できる．このような問題では黄金分割探索を用いることによって極大値を算出することができる [17], [18]．

L の境界面の座標を A_0，L_{max} の境界面の座標を A_1 とする．新たに二つの境界面 L_{n1} と L_{n2} の座標 A_{n1}，A_{n2} を次のように定義する．

$$A_{n1} = \frac{\phi\cdot A_0 + A_1}{\phi+1}, \quad A_{n2} = \frac{\phi\cdot A_1 + A_0}{\phi+1} \quad (3)$$

ここで，ϕ は黄金比であり，$\phi = (1+\sqrt{5})/2$ である．

L_{n1} 及び L_{n2} の各座標 A_{n1}，A_{n2} のうち，有効性が小さくなるほうの境界面が L_{n1} であったとする．このとき，匿名化エリア L の境界面を A_0 から A_{n1} に更新する．逆に，有効性が小さくなるほうの境界面が A_{n2} であった場合は，匿名化エリア L_{max} の境界面を A_1 から A_{n2} に更新する．

更新された L 及び L_{max} に対し，同様の分割探索処理を繰り返すことにより，有効性を最大化する境界面を導出することができる．

5. 5　エリア縮小フェーズ

エリア縮小フェーズの処理は，全ての匿名化エリアに対して独立に実施する．以下では，ある匿名化エリア L に対する処理を記述する（図 3）．

L 内の全ての存在円において，少なくとも一部が L に含まれるような最小エリアを L_{min} とする．L と L_{min} の間で最適なエリアを特定する．L を縮小する方向は上下左右の 4 方向がある．各方向について黄金分割を行うための 2 点を導出し，各点で分割したときの $P(L,k)$ 及び Utility をそれぞれ計算する．(w,k)-匿名性を満たした上で最も Utility が向上する方向にのみ分割を実施する．この処理を繰り返し，Utility を最大化するエリアを特定する．

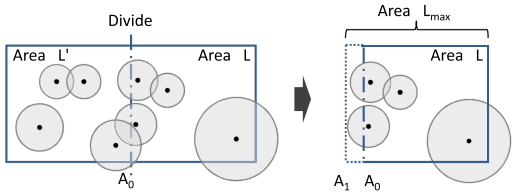


図 2　エリア分割とエリア拡大
Fig. 2　Area divide and expansion.

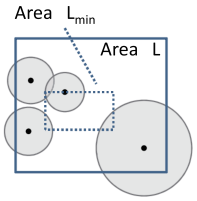


図 3　エリア縮小
Fig. 3　Area decrease.

5.6 準最適化アルゴリズム

式 (2) の計算は，**5.3** で述べたように場合によっては計算不可能なほど計算量が増大する．本節では，計算量を大幅に削減する手法を述べる．

まず，$|U_L| < k$ のときは明らかに，匿名化エリア L に k 人以上が存在する確率は 0 であるため，計算を行う必要はない．$|U_L| \geq k$ のときのみ計算を行う．

また，式 (2) を計算する際に，k の値をそのまま用いる必要がない場合がある．対象とするエリア L に存在する確率が 0 より大きいユーザ集合 U_L の中で，m 人の存在円が当該エリアに完全に包含されている場合，残り $|U_L| - m$ 人のうち $k-m$ 人以上が実際にエリア L に存在する確率を求めれば良い．

更に，次のように近似解を算出することができる．各ユーザが対象とするエリアに存在する確率を例えば 0 から 0.1，0.1 から 0.2，のように d 段階に分割する．例として $d = 10$ のとき，あるユーザが対象とするエリアに存在する確率が 0.25 だった場合，0.2 とみなす．存在確率を過小評価するため匿名化後の有効性は減少するが，(w, k)-匿名性は保証することができる．このような計算を行う場合の手順を以下に示す．各ユーザ u について，対象とするエリア L に存在する確率を $P_1(L, u)$ とする．変数 j $(j = 1, \ldots, d-1)$ に対し，$|U_L| - m$ 人の中で，$P_1(L, u)$ が j/d 以上 $(j+1)/d$ 未満であるユーザ数を c_j とすると，L に k 人以上のユーザが存在する確率 $P(L, k)$ に対し，

$$P(L,k) \geq 1 - \sum_{q=0}^{k-m-1} \sum_{i_1=\underline{\Lambda}(1)}^{\bar{\Lambda}(1)} \sum_{i_2=\underline{\Lambda}(2)}^{\bar{\Lambda}(2)} \cdots \sum_{i_{d-1}=\underline{\Lambda}(d-1)}^{\bar{\Lambda}(d-1)} \prod_{j=1}^{d-1} \left[{}_{c_j}C_{i_j} \cdot (\frac{j}{d})^{i_j} \cdot (1 - \frac{j}{d})^{c_j - i_j} \right], \quad (4)$$

where

$$\bar{\Lambda}(s) = \min(c_s, q - \sum_{j=1}^{s-1} i_j),$$

$$\underline{\Lambda}(s) = \max(0, q - \sum_{j=1}^{s-1} i_j - \sum_{j=s+1}^{d-1} c_j).$$

が成り立つ(注1)．

ここで，以下の定理が成り立つ．

(注1)：$\sum_{j=1}^{0} i_j = 0$ として計算する．

定理 1.

$|U_L| \geq k$ のとき，式 (4) 中の $\bar{\Lambda}(s)$ 及び $\underline{\Lambda}(s)$ について，常に，

$$\bar{\Lambda}(s) \geq \underline{\Lambda}(s)$$

が成り立つ．

証明を付録 1. に記す．

式 (4) において，$\prod_{j=1}$ の箇所は，各 c_j $(j = 1, \ldots, d-1)$ のうち，それぞれちょうど i_j 人が L に存在している確率を求めている．$\sum_{i_1} \sum_{i_2} \ldots \sum_{i_{d-1}}$ の箇所は，総和が q となるような i_j $(j = 1, \ldots, d-1)$ の全組合せを表しており，$\sum_{q=0}$ の箇所は，L に存在する人数を q とし，q を 0 から $k-m-1$ まで変動させている．

例として，10 人が匿名化エリア L に存在する確率が 0 より大きい状態であるとき，$P(L, k)$ を計算することを考える．例えば $k = 7$，$c_3 = 5$，$c_5 = 1$，$c_7 = 3$ であったとする．また，確率 1 で L に存在するユーザが 1 人いたとする．この状況を式 (4) にあてはめると，

$$P(L,7) \geq 1 - \sum_{q=0}^{7-1-1} \sum_{i_3=\max(0,q-(1+3))}^{\min(5,q)} \sum_{i_5=\max(0,q-i_3-3)}^{\min(1,q-i_3)} \sum_{i_7=\max(0,q-(i_3+i_5))}^{\min(3,q-(i_3+i_5))} \left[{}_5C_{i_3} (\frac{3}{10})^{i_3} (1-\frac{3}{10})^{5-i_3} \cdot {}_1C_{i_5} (\frac{5}{10})^{i_5} (1-\frac{5}{10})^{1-i_5} \cdot {}_3C_{i_7} (\frac{7}{10})^{i_7} (1-\frac{7}{10})^{3-i_7} \right]$$

である．結果，$P(L, 7) \geq 0.15$ が得られる．

5.7 解　　析

本節では匿名化アルゴリズムについて計算量の解析を行う．特に計算量が多いのは式 (4) である．$\prod_{j=1}$ の部分は $O(d)$ で表される数だけ乗算を行う．$\sum_{q=0}$ の部分は $O(k)$，$\sum_{i_1} \cdots \sum_{i_{d-1}}$ の部分は $O(k^d)$ で表される数だけ加算を行う．

式 (4) の計算は，最終的に出力される匿名化エリアの数におおむね比例した数だけ実行される．したがって最終的な計算量は $O(dk^d N)$ である．

6. 評　　価

提案手法の評価方法及び評価結果について述べる．

6.1 データセット

オープンソースの移動体シミュレータ Siafu [19] を利用し，ユーザを 100 人から 1000 人に設定してシミュ

レーションを行った．Siafu はコンテクスト情報を考慮したユーザの移動についてのシミュレータとして，多くの研究で利用されている．データセットの移動範囲を約 4.2km × 4.2km とし，5 分ごとの位置情報のデータを 4 時間分利用した．

Siafu における位置情報には誤差が含まれていないため，各ユーザに対し，誤差を 5m から 500m までの範囲でランダムに設定した．Siafu における元のデータを真の値とし，匿名化処理には誤差を加えた値を利用した．匿名化にあたっては，移動範囲を 1m 四方のセル状に分割した座標系を利用した．

またデフォルト値として，$k = 10$, $w = 0.99$, $\alpha = 1$, $N = 1000$ に設定した．比較対象として，通常の Mondrian アルゴリズムによる手法を選定した．

6.2 評価結果

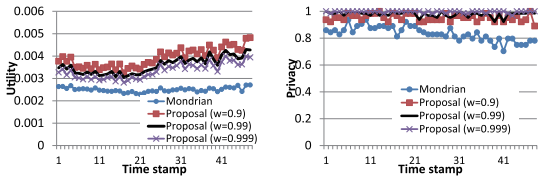
評価指標として，式 (1) の Utility 及び次式で示す Privacy を利用する．Privacy は，k-匿名化や (w, k)-匿名化を実施したとき，匿名化エリアに k 人以上のユーザが実際に存在している割合を表す．具体的には次のとおりである．匿名化エリアの総数を A_L，k 人以上のユーザが実際に存在している匿名化エリアの数を A'_L とおく．このとき Privacy を

$$Privacy = A'_L / A_L \tag{5}$$

と定義する．

匿名化エリア L に k 人以上のユーザが存在する確率 $P(L, k)$ の値は，厳密な値を算出する式 (2) ではなく，下限値を計算する式 (4) を $d = 10$ に設定して用いた．

5 分ごと 4 時間分のデータに対して Utility 及び Privacy の値を計測した結果を図 4 に示す．図 4 (a) から分かるように，時間ごとに変動はあるが，提案手法が既存手法よりも常に Utility の値が上回っている．また，Privacy に関しては，既存手法は 0.8 前後で推移しているのに対し，提案手法はおおむね w で設定した以上の値（0.9〜0.999）で推移していることが分かる（図 4 (b)）．これらより，Utility 及び Privacy いずれの指標においても，提案手法が既存手法を上回っていると言える．



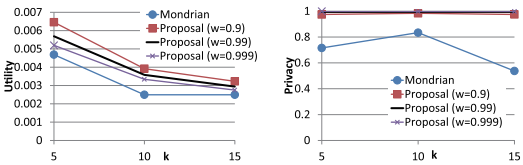
(a) Utility　(b) Privacy

図 4　5 分ごとの Utility 及び Privacy
Fig. 4　Utility and Privacy every 5 minutes.

Utility 及び Privacy に関し，4 時間分のデータの平均値を算出した結果を以下に示す．

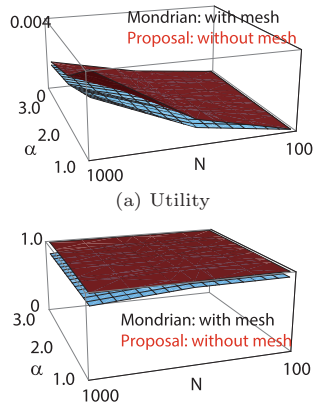
図 5 は，w 及び k の値を変動させて Privacy 及び Utility を評価した結果を表している．Privacy に関しては，どの k の値を取っても提案手法が既存手法を上回っている．また，w の設定値が 1 に近づくほど，Privacy の値も 1 に近づいている．Utility に関しても，提案手法が既存手法を上回っていることが分かる．w の値が増加するほど，匿名化エリアを広げる必要が生じるため Utility の値が減少していることが分かる．また k の値が大きくなるほど匿名化エリアの面積は増加するため，Utility は減少している．

ユーザ数を N とし，N を 100 から 1000 まで，α を 1.0 から 3.0 まで変化させたときの Utility 及び Privacy を計測した結果を図 6 に示す．Utility に関しては，ユーザ数が増加するほど提案手法と既存手法の値の差が拡大していることが分かる．また，α の値が増



(a) Utility　(b) Privacy

図 5　w 及び k を変動させたときの Utility 及び Privacy
Fig. 5　Utility and Privacy with varying w and k.



(a) Utility

(b) Privacy

図 6　N 及び α を変動させたときの Utility 及び Privacy
Fig. 6　Utility and Privacy with varying N and α.

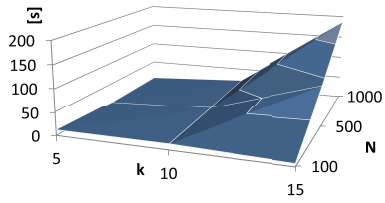


図 7　処 理 時 間
Fig. 7　Calculation time.

加するほど，提案手法も既存手法も Utility の値が減少している．Privacy に関しては，ユーザ数や α の値による影響はほとんど見られなかった．

最後に，匿名化に必要な時間を計測した．実験は，OS が Windows 7 Professional 64 bit，CPU が Intel Xeon CPU X5675 @ 3.07GHz 3.06 GHz，RAM が 12GB である PC を利用して行った．ユーザ数 N を 100 から 1000 まで，k を 5 から 15 まで変動させたときの計測結果を図 7 に示す．

k の増加に対し，処理時間が指数関数的に増加していることが分かる．これは式 (4) の計算量が，k の値に応じて指数関数的に増加するためである．したがって，k の値が大きい場合には提案手法は有効ではない．しかし，多くの既存研究で k は 3 から 20 程度の値が用いられており，この範囲内であれば有効であると考えられる．N の増加に対しては，処理時間は線形に増加していることが分かる．

一方，既存手法はパラメータ値にほとんど依存せず，処理時間は 1 秒程度であった．このように，処理時間に関しては既存手法のほうが優位である．しかしながら，既存手法は位置情報の誤差を考慮していない．一方で，Utility や Privacy は誤差を考慮した指標となっている．したがって，既存手法において，処理時間を増加させることを許容し，処理方法を工夫したとしても，位置情報の誤差が考慮されていないため，Utility や Privacy の値を大きく増加させることは難しいと考えられる．

7. 考　　察

想定環境の妥当性及びパラメータの設定について考察を述べる．

7.1　想定環境の妥当性

・ユーザの位置情報を知っている攻撃者が存在する

攻撃者が，Alice がエリア $L1$ にいたことを知っている背景として考えられる例として考えられるものを以下に挙げる．

（1）　Alice がある時刻 t_1 にエリア $L1$ に存在したことを物理的に観測した

（2）　他の位置情報管理事業者 B から，Alice がある時刻 t_1 にエリア $L1$ に存在したことを把握した

2 番目の例は，事業者 B の提供する情報を用いて Alice に何らかのサービスを提供することを想定している．したがって，事業者 B から受領する情報には Alice のユーザ属性は含まれないが，個人を特定できる識別子が含まれているという状況である．

特に，今後複数事業者間においてユーザ情報を共有することが多くなってくると，2 番目のように他の情報源と結びつけられる状況は増加すると考えられる．

・取得できる位置情報には誤差があり，誤差の大きさにばらつきがある

駅の改札通過時や店舗での購入時等では，ユーザの位置情報を正確に取得することが可能である．一方，GPS を用いた計測では，誤差は数 m 未満の場合もあるが数十 m を超えることもあり [20]，Wi-Fi のみを利用した場合，誤差が 500m を超えることもある．位置情報取得精度は将来的に向上するが，精度のばらつき自体は常に生じ得ると考えられる．

7.2　パラメータの設定

・k の値

複数の既存研究で k の値はおおむね 3〜20 程度に設定されており [10], [16], [21]，本論文でもその範囲内で k の値を変動させて評価を行った．適切な値の設定については将来課題である．

・α 及び w の値

本論文で導入した α や w の設定方法も将来課題である．また，w を多段階に設定可能なように拡張することも考えられる．例えば，ある定数 k' $(k' < k)$ を導入し，匿名化エリアに k' 人以上は 100%の確率で存在することを保証し，その上で，k 人以上が存在する確率を w に設定する，といったことが可能なようにプライバシー指標を拡張することができる．この場合，提案している匿名化アルゴリズムを修正する必要があるが，修正箇所はプライバシー条件が満たされているかどうかのチェック箇所だけであるため，単純な修正で対応できると考えられる．

8. む　す　び

ユーザ属性と行動履歴の情報を元に，どのような属性をもったユーザがどのような行動を取りやすいかに

ついてマイニングをすることを想定し，ある個人の位置情報を知る攻撃者に，その個人とユーザ属性を結び付けられないようにするための匿名化手法を提案した．

位置情報に誤差がある環境を想定すると，従来の匿名化手法では，保護されるとされているレベル以上にユーザ属性が漏洩するリスクがあることを示した．このリスクに対応するため，匿名化エリアに w 以上の確率で k 人のユーザが存在することを保証する，(w,k)-匿名性という新しい指標を提案した．更に，匿名化後のデータの有効性指標として，匿名化エリアのサイズだけでなく，匿名化エリアに実際に存在する確率を考慮する指標を提案した．

これらの新しい指標や想定環境に応じた匿名化手法を提案し，シミュレーション評価を実施した．提案指標の下では，従来手法よりプライバシー漏洩リスクを低減させると同時に，匿名化後のデータの有効性を向上させることができた．

将来課題として，実データを用いて大規模な評価を行う必要がある．更に，本論文では Mondrian アルゴリズムに対して拡張を行ったが，その他の匿名化手法に対して本手法を適用し，評価する必要があると考えている．また，連続的な時間にわたる k-匿名化への拡張を行う必要がある．

謝辞　本研究は JSPS 研究費 24300005，23500039，25730038 の助成を受けたものです．

本研究を遂行するにあたり，研究の機会と議論・研鑽の場を提供して頂き，ご指導頂いた国立情報学研究所/東京大学本位田真一教授をはじめ，活発な議論と貴重なご意見を頂いた研究グループの皆様に感謝致します．

文　献

[1] 飯尾　淳，吉田圭吾，小池亜弥，清水浩之，白井康之，桑山晃一，栗山桂一，小浪宏信，高山隼佑，“属性付き位置情報ログが示す行動特性と消費傾向の関係，” 情処学論，vol.52, no.7, pp.2256–2267, 2011.

[2] A. Gkoulalas-Divanis, P. Kalnis, and V.S. Verykios, “Providing K-anonymity in location based services,” ACM SIGKDD Explor. Newsl., vol.12, no.1, pp.3–10, 2010.

[3] 中西健一，高汐一紀，徳田英幸，“粒度の動的変更による位置匿名性についての考察，” 情処学論，vol.46, no.9, pp.2260–2268, Sept. 2005.

[4] Microsoft Inc., “Windows Phone Dev Center,” http://dev.windowsphone.com/en-us/develop

[5] Apple Inc., “iOS Developer Library,” http://developer.apple.com/library/ios

[6] Google Inc., “Android Developers,” http://developer.android.com/

[7] K. LeFevre, D. DeWitt, and R. Ramakrishnan, “Incognito: Efficient full-domain k-anonymity,” Proc. ACM SIGMOD, pp.49–60, 2005.

[8] A. Meyerson and R. Williams, “On the complexity of optimal K-anonymity,” Proc. PODS, pp.223–228, 2004.

[9] K. LeFevre, D.J. DeWitt, and R. Ramakrishnan, “Mondrian multidimensional K-anonymity,” Proc. IEEE ICDE, p.25, 2006.

[10] H. Hu, J. Xu, S.T. On, J. Du, and J.K.-Y. Ng, “Privacy-aware location data publishing,” ACM Trans. Database Systems, vol.35, no.3, pp.1–42, 2010.

[11] O. Abul, F. Bonchi, and M. Nanni, “Never walk alone: Uncertainty for anonymity in moving objects databases,” Proc. 24th IEEE ICDE, pp.376–385, 2008.

[12] M. Terrovitis and N. Mamoulis, “Privacy preservation in the publication of trajectories,” Proc. IEEE MDM, pp.65–72, 2008.

[13] M.E. Andrés, N.E. Bordenabe, K. Chatzikokolakis, and C. Palamidessi, “Geo-indistinguishability: Differential privacy for location-based systems,” CoRR, vol.abs/1212.1, Dec. 2012.

[14] C. Ardagna, M. Cremonini, S. De Capitani di Vimercati, and P. Samarati, “An obfuscation-based approach for protecting location privacy,” IEEE Trans. Dependable and Secure Computing, vol.8, no.1, pp.13–27, 2011.

[15] M. Xue, P. Kalnis, and H.K. Pung, “Location diversity: Enhanced privacy protection in location based services,” Proc. International Symposium on Location and Context Awareness, pp.70–87, Springer, 2009.

[16] 高橋　翼，宮川伸也，伊東直子，“移動軌跡ストリームに対するリアルタイム k 匿名化手法の提案，” 日本データベース学会論文誌，vol.10, no.1, pp.37–42, 2011.

[17] J. Kiefer, “Sequential minimax search for a maximum,” Proc. American Mathematical Society, vol.4, no.3, pp.502–506, March 1953.

[18] L. Nazareth and P. Tseng, “Gilding the lily: A variant of the Nelder-Mead algorithm based on golden-section search,” Computational Optimization and Applications, vol.22, no.1, pp.133–144, 2002.

[19] M. Martin and P. Nurmi, “A generic large scale simulator for Ubiquitous computing,” Proc. 3rd MobiQuitous, pp.1–3, IEEE, July 2006.

[20] N.M. Drawil, H.M. Amar, and O.A. Basir, “GPS localization accuracy classification: A context-based approach,” IEEE Trans. Intelligent Transportation Systems, vol.14, no.1, pp.262–273, 2013.

[21] L. Yao, G. Wu, J. Wang, F. Xia, C. Lin, and G. Wang, “A clustering K-anonymity scheme for location privacy preservation,” IEICE Trans. Inf. & Syst.,

vol.E95-D, no.1, pp.134–142, Jan. 2012.

付　　録

1. 定理 1 の証明

Proof. $|U_L| \geq k$ より，次式が導かれる．

$$m + \sum_{j=1}^{d-1} c_j >= k \tag{A·1}$$

以下で，数学的帰納法を用いて，式 (A·1) が満たされているとき定理 1 が成り立つことを証明する．

1) $s = 1$ のとき定理 1 が成り立つことを示す．

$s = 1$ のとき式 $\bar{\Lambda}(s)$ 及び $\underline{\Lambda}(s)$ は以下のようになる．

$$\begin{aligned} \bar{\Lambda}(1) &= \min(c_1, q) \\ \underline{\Lambda}(1) &= \max(0, q - \textstyle\sum_{j=2}^{d-1} c_j) \end{aligned} \tag{A·2}$$

ここで，$q \geq 0$ であり，かつ，任意の j ($j = 1, \ldots, d-1$) について，$c_j \geq 0$ であるから，

$$H_1 = q - \sum_{j=2}^{d-1} c_j \tag{A·3}$$

とおくと，

$$H_1 \leq c_1 \tag{A·4}$$

が常に成り立つことを示せば良い．

q の取り得る最大値は，式 (4) より，$q = k - m - 1$ である．したがって，式 (A·3) より

$$H_1 \leq k - m - 1 - \sum_{j=2}^{d-1} c_j \tag{A·5}$$

である．また，

$$\sum_{j=2}^{d-1} c_j = \sum_{j=1}^{d-1} c_j - c_1 \tag{A·6}$$

であるから，

$$H_1 \leq k - m - 1 - \sum_{j=1}^{d-1} c_j + c_1 \tag{A·7}$$

となる．式 (A·1)，(A·7) より

$$H_1 \leq c_1 - 1 \tag{A·8}$$

となるため，式 (A·4) が常に成り立つ．したがって，$s = 1$ のとき定理 1 が成り立つ．

2) $s = s'$ のとき定理 1 が成り立つと仮定し，$s = s' + 1$ のときも成り立つことを示す．

$s = s'$ のとき式 $\bar{\Lambda}(s)$ 及び $\underline{\Lambda}(s)$ は以下のようになる．

$$\begin{aligned} \bar{\Lambda}(s') &= \min(c_{s'}, q - \textstyle\sum_{j=1}^{s'-1} i_j) \\ \underline{\Lambda}(s') &= \max(0, q - \textstyle\sum_{j=1}^{s'-1} i_j - \sum_{j=s'+1}^{d-1} c_j) \end{aligned} \tag{A·9}$$

ここで，

$$H_{s'} = q - \sum_{j=1}^{s'-1} i_j - \sum_{j=s'+1}^{d-1} c_j \tag{A·10}$$

とおくと，$s = s'$ のとき定理 1 が成り立つと仮定していることから，常に，

$$H_{s'} \leq c_{s'} \tag{A·11}$$

が成り立つ．$s = s' + 1$ のとき式 $\bar{\Lambda}(s)$ 及び $\underline{\Lambda}(s)$ は以下のようになる．

$$\begin{aligned} \bar{\Lambda}(s'+1) &= \min(c_{s'+1}, q - \textstyle\sum_{j=1}^{s'} i_j) \\ \underline{\Lambda}(s'+1) &= \max(0, q - \textstyle\sum_{j=1}^{s'} i_j - \sum_{j=s'+2}^{d-1} c_j) \end{aligned} \tag{A·12}$$

ここで，$q \geq 0$ であり，かつ，任意の j ($j = 1, \ldots, d-1$) について，$c_j \geq 0$ であるから，

$$H_{s'+1} = q - \sum_{j=1}^{s'} i_j - \sum_{j=s'+2}^{d-1} c_j \tag{A·13}$$

とおくと，

$$H_{s'+1} \leq c_{s'+1} \tag{A·14}$$

が常に成り立つことを示せば良い．

式 (A·10)，(A·13) より，

$$H_{s'+1} = H_{s'} - i_{s'} + c_{s'+1} \tag{A·15}$$

となる．ここで，式 (A·11) より，

$$H_{s'+1} \leq c_{s'} - i_{s'} + c_{s'+1} \tag{A·16}$$

となり，式 (4) より $i_{s'} \leq c_{s'}$ であるから，

$$H_{s'+1} \leq c_{s'+1} \tag{A·17}$$

となるため，式 (A·14) が常に成り立つ．したがって定理 1 が成り立つ．

以上より，数学的帰納法により，定理 1 が成り立つ．□

（平成 25 年 7 月 28 日受付，10 月 21 日再受付）

清　　雄一　（正員）

1981 年生．2009 年東京大学大学院情報理工学系研究科博士後期課程修了．同年（株）三菱総合研究所入社．同社情報技術研究センター，金融ソリューション本部等に所属．2013 年より電気通信大学助教，現在に至る．分散コンピューティング，セキュリティ，プライバシー保護技術等の研究に従事．情報処理学会，電子情報通信学会，IEEE Computer Society 各会員．

大須賀昭彦　（正員）

1958 年生．1981 年上智大学理工学部数学科卒．同年 (株) 東芝入社．同社研究開発センター，ソフトウェア技術センター等に所属．1985〜1989 年 (財) 新世代コンピュータ技術開発機構 (ICOT) 出向．2007 年より，電気通信大学大学院情報システム学研究科教授．2012 年より，国立情報学研究所客員教授兼任．工学博士 (早稲田大学)．主としてソフトウェアのためのフォーマルメソッド，エージェント技術の研究に従事．1986 年度情報処理学会論文賞受賞．IEEE Computer Society Japan Chapter Chair，人工知能学会理事，日本ソフトウェア科学会理事を歴任．情報処理学会，電子情報通信学会，人工知能学会，日本ソフトウェア科学会，IEEE Computer Society 各会員．