「マルチメディア，分散，協調とモバイル
(DICOMO2013)シンポジウム」 平成25年7月

不正確さを考慮した位置匿名化手法の提案

清 雄一[1]　大須賀 昭彦[1]

概要：性別や年齢等のユーザ属性と，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元来不正確さが含まれていることが考慮されていないため，位置情報の誤差を単純に扱おうとした場合に個人が特定されてしてしまうリスクが存在する．本論文では位置情報の不確実性に起因する漏洩リスクを導入及び定式化し，新しい匿名性指標，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．

Location Anonymization Considering Accuracy

Yuichi Sei[1]　Akihiko Ohsuga[1]

1. はじめに

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，どのような属性を持ったユーザがどのような場所へ行くのかについてマイニングをする研究が行われている [30]．ユーザ属性ごとの消費行動を分析することが可能となるため，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．

しかし，位置情報には誤差が含まれることが多く，マイニングを難しいものとしている．誤差なく正確な位置情報を取得できれば，どの店舗にユーザが入店したか，という事実まで把握することが可能である．一方，誤差が大きければ，駅単位や市区町村単位レベルでしかユーザの位置を把握できない．位置情報における誤差はばらつきが大きいため，精度よくマイニングを行うためには，各位置情報の誤差情報も利用するほうが良いと考えられ，本論文ではそのようなマイニングが行われることを想定する．

また本論文では，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．多くの既存研究においても，このような想定がなされている．マイニングを行うためには個人を特定する必要はないため，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，たとえばユーザ 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分かってしまうという問題がある．

各ユーザの位置情報を匿名化することによって，この問題を解決する研究が盛んに行われている．たとえば，k-匿名化 [9][13] 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．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．このとき，3 章で述べるように，ユーザ属性と個人を結び付けられる恐れがある．

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，実際の位置情報には誤差が含まれる可能性があるため，匿

[1] 電気通信大学
1-5-1 Chofugaoka, Chofu, Tokyo 182–8585, Japan

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．

本研究では位置情報の精度に大きなばらつきがある環境の場合，従来の匿名化手法をそのまま用いるとプライバシ漏洩リスクが増大する危険性があることを明らかにし，位置情報の精度を考慮した匿名化手法の提案を行う．従来手法よりもプライバシ漏洩リスクを低減できることをシミュレーションによって示す．また，匿名化エリアへのユーザの存在確率という新しい指標を導入し，既存の評価指標と組み合わせた場合に，既存研究よりも有効性の高い匿名化ができることを示す．

本論文の構成は次の通りである．2 章で関連研究とその課題について述べる．3 章では本論文が解く新しい問題を定義する．4 章において本論文で提案する評価指標を，5 章においてその評価指標に基づく匿名化手法を提案する．6 章では提案手法についてシミュレーション評価を実施する．考察を 7 章で行い，8 章で本論文をまとめる．

2. 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 [9][13] ．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，公開されるテーブルの中で，どのレコードが Alice を表しているかを一意に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エリア面積を最小化することをめざしている．この問題は NP 困難であることが示されているため [19][2]，より計算量が少なく匿名化後のデータの有効性が高まるよう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 [14] が広く利用されており，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 [1][11].

既存の k-匿名化手法は共通して，3 章で指摘するような攻撃モデルへの考慮がない．位置情報に誤差が含まれており，かつ，位置情報の精度情報も公開する場合においては，匿名化後のテーブルから個人が特定されてしまうリスクがある．

ユーザの位置情報を一時点のみに限定せず，一定時間にわたる位置情報の履歴を匿名化する研究も多数行われている [7][23]．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_i における Alice の場所を知っている攻撃者に，t_i 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とを目指している．これらの研究では各時刻における 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．

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 [3]．Ardagna ら [5] 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．

また，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 [26]．この匿名化には l-多様性 [17][24] 等，別の匿名化指標が利用される．各ユーザに対し，位置情報に基づくロケーションアウェア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データマイニングを行うシナリオを想定している．したがって，位置情報のセマンティックな匿名化は本論文のスコープ外である．

3. 問題定義

本章では，位置情報の利用モデルや攻撃モデルについて述べる．

3.1 位置情報の利用モデル

どのような属性を持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目的とする．このようなデータマイニングは，[30] 等で実際に行われている．

近年，複数事業者間が保有するデータを提供し合い，新しいビジネスやサービス構築に向けてマイニングを実施するモデルが提案されている．このように，位置情報を取得する事業者，ユーザ属性を保持する事業者，マイニングを実施する事業者がそれぞれ異なる場合がある．位置情報やユーザ属性は，他事業者に対して個人を特定されないよう管理する必要がある．したがって，事業者間で連携する場合，必要十分な匿名化を実施することが要求される [21], [29].

3.2 匿名化対象のデータ

本論文では，“位置情報” と “ユーザ属性” の 2 つを取り扱う．

ユーザ属性を匿名化しない場合は次のようなリスクがあ

る．攻撃者が Alice のユーザ属性を知っていると想定する．攻撃者は，匿名化前のテーブル T から，Alice のユーザ属性に該当するレコードを探し出すことにより，Alice の位置情報を取得することが可能である．本論文では，ユーザ属性を対象として通常の k-匿名化や l-多様化を行うことによってユーザ属性はすでに匿名化されており，ユーザ属性から位置情報を特定することは困難な状態になっていることを前提とする．

位置情報を匿名化しない場合は後述するように，あるユーザの位置情報を知る攻撃者が当該ユーザの属性を取得することが可能である．本論文では位置情報を匿名化することにより，この脅威に対する手法を提案する．

以下では，匿名化前のテーブルを T，匿名化後のテーブルを T^* と表現する．

3.3 位置情報と誤差の表現

位置情報には誤差が含まれることを想定する．モバイル端末の OS として広く利用されている iOS，Android 及び Windows Phone OS 等では，位置を緯度，経度で，精度を円の半径で表している [4], [10], [20]．本論文でも同様に，ユーザの位置情報を直接取得する事業者は，中心座標 (x, y) 及び円の半径 r の情報を得られると想定する．ユーザ i の中心座標を (x_i, y_i)，精度を表す円の半径を r_i とする．この円を各ユーザの**存在円**と呼ぶ．

匿名化後のテーブル T^* の各レコードには，各ユーザについて以下の情報を設定する．

- 存在する可能性のある匿名化エリア
- 匿名化エリアに実際に存在している確率
- ユーザの個人属性

ユーザの存在円が匿名化エリアに含まれないことを許容することもできるが，その場合は当該ユーザのデータを除外することと等しい．ほとんどの既存研究において，データを除外した匿名化は行っていないため，本論文でも同様に存在円の少なくとも一部が匿名化エリアに含まれるよう制約を設ける．

3.4 攻撃例

図 1，表 1 及び 2 を用いて攻撃例を説明する．位置情報を管理する事業者が図 1(a) の情報を得たとする．位置情報の精度，各ユーザ属性とともにテーブル化したものが表 1 である．表 1 に対し，項目 Location を対象として通常の k-匿名化を用いて匿名化（$k = 4$）されたテーブル T^* は表 2 で表されている．

攻撃者は T^* を閲覧できると想定する．また，Alice が時刻 t_1 においてエリア L_1 にほぼ確実に存在していたことを攻撃者は知っていると仮定する．攻撃者はこの情報を用いることにより，T^* の Accuracy=95%に該当する Info_1 が，Alice のユーザ属性である可能性が高いと判断できる．実際，現実には Alice 以外の 3 人が L_1 に全くいないということも起こり得る（図 1(b)）．

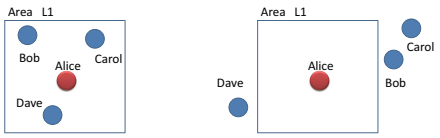
攻撃者が，Alice が時刻 t_1 においてエリア L_1 にいたことを知っている背景として考えられる例については，7 章において考察する．位置情報に誤差が存在せず，表 2 において Accuracy の項目が必要無ければ，Alice の個人情報は Info_1 から Info_4 までのいずれかであるか判断は全くできない．しかし，位置情報に誤差が含まれており，かつその誤差の程度を匿名化後のデータに含める必要がある場合，このような攻撃モデルが成立する．

表 1 時刻 t_1 における位置情報，精度，個人属性
Table 1 Location data, accuracy, and user attributes at time t_1

User	Location	Accuracy	Personal Information
Alice	(x1, y1)	95%	Info_1
Bob	(x2, y2)	55%	Info_2
Carol	(x3, y3)	30%	Info_3
Dave	(x4, y4)	25%	Info_4

表 2 位置情報に関して 4-匿名化されたテーブル
Table 2 4-anonymized table toward to location data

Location	Accuracy	Personal Information
L1	95%	Info_1
L1	55%	Info_2
L1	30%	Info_3
L1	25%	Info_4



(a) Sensed Data　　(b) True Information

図 1 4-匿名化
Fig. 1 4-Anonymization

4. 指標の提案

4.1 匿名性指標及びプライバシ指標

匿名性指標として (w, k)-匿名性を提案する．w 以上の確率で，匿名化エリアに k 人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．

またプライバシ指標として，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導入し，

$$Privacy = \frac{\text{No. of areas where } k \text{ or more users exist}}{\text{No. of areas}} \quad (1)$$

と定義する．

表 3 Notation

$L(i)$	ユーザ i が存在する匿名化エリア
$\|L\|$	匿名化エリア L の面積
(x_i, y_i)	ユーザ i の位置情報の観測値における中心座標
C_i	ユーザ i が存在する可能性のある円
r_i	C_i 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

4.2 有効性指標

T^* の有効性指標としては，匿名化エリアを最小化することが一般的である [28][9]．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必要がある．

有効性指標は，ユーザ集合を U，ユーザ $i \in U$ が匿名化エリア $L(i)$ に分類され，ユーザ i がそのエリアに存在する確率を $p_{i,L(i)}$ としたとき，

$$Utility = \sum_{i \in U} \frac{(p_{i,L(i)})^\alpha}{|L(i)|} \tag{2}$$

と表すことができる．ここで $|L(i)|$ はエリア $L(i)$ 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

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) ユーザの位置情報の精度に偏りがある
　精度が高い人と低い人がいる場合，高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．

5.2 概要

利用する変数やパラメータ名を表 3 に定義する．

4 章で提案した匿名性指標を満たし，前節で挙げた課題を考慮して有効性指標を最大化することを目的とする匿名化アルゴリズムを提案する．

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 k-匿名性だけでなく，同時に (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．

通常の匿名化では匿名化エリアの面積最小化が目的となるが，提案手法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．対象とするエリア内に存在するとされている各ユーザの存在円の少なくとも一部を全て含む最小サイズのエリアを導出し，対象エリアと最小エリアとの間で (w,k)-匿名性を満たした上で最も有効性の高いエリアを導出する．

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．分割後のエリアに k 人以上のユーザの中心座標が存在し，かつ，w 以上の確率で実際に 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 i の中心座標を (x_i, y_i), その精度を r_i とし, エリア L における 4 つの頂点の座標を (x_0, y_0), (x_0, y_1), (x_1, y_0), (x_1, y_1) とする．中心座標がエリア内に含まれている場合，ユーザ i がエリア L に存在する確率を $P_1(L, i)$ とすると，証明は省くが次式で表すことができる．

$$P_1(L,i) = \sum_{j=0}^{1} \sum_{k=0}^{1} Sub(|x_i - x_j|, |y_i - y_k|, r_i) \tag{3}$$

ここで，$a = |x_i - x_j|$，$b = |y_i - y_k|$，$r = r_i$ とおくと，$a = 0$ または $b = 0$ のとき

$$Sub(a,b,r) = 0$$

$a \neq 0$ かつ $b \neq 0$ のとき，

$$Sub(a,b,r) = \frac{\acute{a}\sqrt{\acute{r}^2 - \acute{a}^2} + \acute{b}\sqrt{\acute{r}^2 - \acute{b}^2} + \acute{r}^2\left(\frac{\pi}{2} - \arccos\frac{\acute{a}}{\acute{r}} - \arccos\frac{\acute{b}}{\acute{r}}\right)}{2\pi r^2},$$

where

$$\acute{a} = \min(a, r), \acute{b} = \min(b, r), \acute{r} = \min(r, \sqrt{a^2 + b^2})$$

である．

また，中心座標はエリアに含まれていないが存在円とエリアで重なる部分がある場合，ユーザ i がエリア L に存在する確率は，証明は省くが次式で表すことができる．

$$\begin{aligned} P_1(L,i) &= Sub'(x_0 - x_i, y_0 - y_i, x_1 - x_i, y_1 - y_i, r_i) \\ &+ Sub'(x_i - x_1, y_0 - y_i, x_i - x_0, y_1 - y_i, r_i) \\ &+ Sub'(x_0 - x_i, y_i - y_1, x_1 - x_i, y_i - y_0, r_i) \\ &+ Sub'(x_i - x_1, y_i - y_1, x_i - x_0, y_i - y_0, r_i) \end{aligned} \tag{4}$$

where

$$\begin{aligned} &Sub'(x_0, y_0, x_1, y_1, r) = [(r^2\theta - r^2 \sin\theta + a \cdot b)/2 \\ &+ (x'' - \max(0, x_0)) \cdot b + a \cdot (y'' - \max(0, y_0)) \\ &+ (x' - a - \max(0, x_0)) \cdot (y' - b - \max(0, y_0))]/(\pi r^2), \\ &x'' = \max(0, \min(x_1, \max(0, x_0, \sqrt{r^2 - \min(r, \max(y_1, 0))^2}))), \\ &x' = \max(0, \min(r, x_1, \sqrt{r^2 - \min(r, \max(y_0, 0))^2})), \\ &y'' = \max(0, \min(y_1, \max(0, y_0, \sqrt{r^2 - \min(r, \max(x_1, 0))^2}))), \\ &y' = \max(0, \min(r, y_1, \sqrt{r^2 - \min(r, \max(x_0, 0))^2})), \\ &a = x' - x'', \quad b = y' - y'', \quad \theta = 2 \cdot \arcsin(\sqrt{a^2 + b^2}/(2r)). \end{aligned}$$

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．

$$P(L,k) = 1 - \sum_{\{S|S \in \mathfrak{P}(U_L) \wedge |S| < k\}} \left[\prod_{i \in S} P_1(L,i) \cdot \prod_{j \notin S \& j \in U_L} (1 - P_1(L,j)) \right]. \tag{5}$$

$\mathfrak{P}(U_L)$ は，集合 U_L のべき集合を表す．ここで，集合 U_L の要素数が大きい場合，k の値によっては計算不可能なほど計算量が増大する．$P_1(L,i)$ を近似することで計算量を減らす方法を 5.6 で述べる．

全ての匿名化エリアについて $P(L,k) \geq w$ が満たされている場合，(w,k)-匿名性が満たされている．

5.4 エリア拡大フェーズ

分割後に生じる 2 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）．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めざす．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，エリア分割フェーズ後でも必ず (w,k)-匿名性が満たされている．

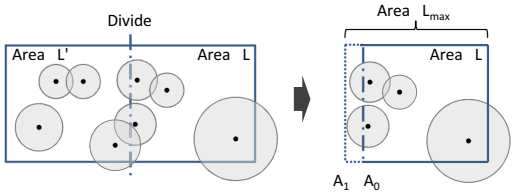


図 2 エリア分割とエリア拡大
Fig. 2 Area divide and expansion

まず，L に含まれる各ユーザ i の存在円 C_i ができるだけ含まれるよう，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向上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 [12][22]．

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 2 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．

$$A_{n1} = \frac{\phi \cdot A_0 + A_1}{\phi + 1}, \qquad A_{n2} = \frac{\phi \cdot A_1 + A_0}{\phi + 1} \tag{6}$$

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．

また，分割された 2 つのエリアの両方に対してエリア拡大フェーズを実施するため，両エリアとも拡大される場合は，重複部分が生じる．

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 に含まれるようにする最小エリアを 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

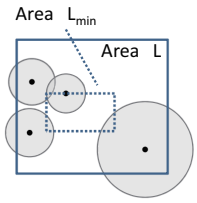


図 3 エリア縮小
Fig. 3 Area decrease

割を行うための 2 点を導出し，当該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．いずれの分割点も (w,k)-匿名性を満たさない場合，$P(L,k)$ が最も大きい点で分割する．このようにすることで，精度が高いユーザが多い方向に優先的にエリアを縮小することができる．この処理を繰り返し，Utility を最大化するエリアを特定する．

5.6 準最適化アルゴリズム

式 5 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．

さらに，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たとえば 0 から 0.1，0.1 から 0.2，のように n 段階に分割する．例として $n=10$ 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)-匿名性は満たすことができる．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 i について，対象とするエリア L に存在する確率を $P_1(L,i)$ とする．$|U_L|-m$ 人の中で，$P_1(L,i)$ が j/n 以上 $(j+1)/n$ 未満であるユーザ数を c_j とすると，次式が成り立つ．

$$P(L,k) \geq 1 - \sum_{q=0}^{k-m-1} \sum_{i_n=\underline{\Lambda}(n)}^{\bar{\Lambda}(n)} \cdots \sum_{i_1=\underline{\Lambda}(1)}^{\bar{\Lambda}(1)} \prod_{j=1}^{n} \left[(\frac{j}{n})^{i_j} \cdot (1-\frac{j}{n})^{c_j - i_j} \right], \tag{7}$$

where

$$\bar{\Lambda}(s) = \min(c_s, q - \sum_{j=1}^{s-1} i_j), \tag{8}$$

$$\underline{\Lambda}(s) = \max(0, q - \sum_{j=1}^{s-1} i_j - \sum_{j=s+1}^{n} c_j). \tag{9}$$

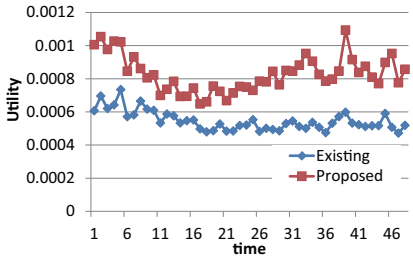
6. 評価

6.1 データセット

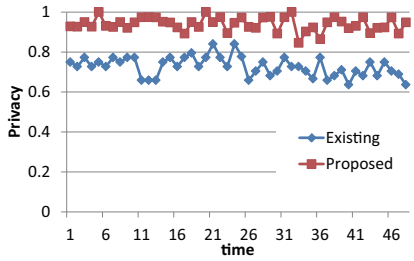
オープンソースの移動体シミュレータ Siafu[18] を利用し，ユーザを 100，300，500 人にそれぞれ設定してシミュレーションを行った．Siafu はコンテキスト情報を考慮したユーザの移動についてのシミュレータとして，多くの研究で利用されている [6][25]．データセットの移動範囲を約 4.2km × 4.2km の範囲とし，5 分ごとの位置情報のデータを 4 時間分利用した．Siafu では，各ユーザにランダムに家や会社が割り当てられ，一定の範囲でランダムに設定された時刻に，起床，出勤等の活動を行う．

Siafu における位置情報には誤差情報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，シミュレーションマップを 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，$k=5$，$w=0.9$，$\alpha=1$ に設定した．



(a) Utility



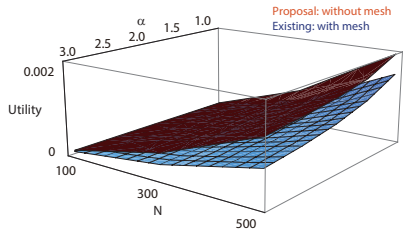
(b) Privacy

図 4 5 分毎の Utility 及び Privacy
Fig. 4 Utility and Privacy every 5 minutes

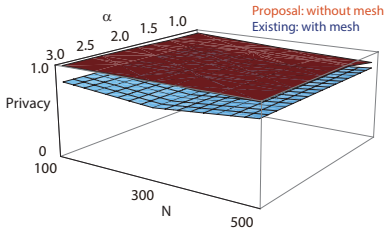
6.2 評価結果

評価指標として，式 2 の Utility 及び式 1 の Privacy を利用する．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 5 ではなく，下限値を計算する式 7 を $n=10$ に設定して用いた．

5 分間を 1 タイムスロットとし，48 タイムスロットに対して 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，時間毎に変動はあるが，提案手法が既存手法よりも常に Utility の値が上回っている．



(a) Utility



(b) Privacy

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

また，Privacy に関しては，既存手法は 0.7 から 0.8 程度で推移しているのに対し，提案手法は概ね 0.9 で推移していることが分かる（図 4(b)）．これらより，有効性及びプライバシ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．

Utility 及び Privacy に関し，48 タイムスロットの平均値を算出した結果を図 5 に示す．いずれの k 及び α の値に対しても，提案手法が有効性指標及びプライバシ指標において既存手法を上回っている．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．また，α の値が大きくなるほど，式 2 における値が小さくなるため，同じく Utility は減少する．図より，減少の割合は提案手法と既存手法であまり差がないことが分かる．また，Privacy の目標値 w を 0.9 に設定しているが，k 以上のユーザの存在円を匿名化エリアに含める必要があるため，ある一定のサイズ以上には匿名化エリアを縮小できない．したがって，提案手法における Privacy のほとんどの値は，0.9 を上回っている．

ユーザ数を N とし，N を 100 から 500 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．Privacy に関しては，ユーザ数が増加するほど，目標値である w に近い値が実現されている．

最後に，匿名化に必要な時間を計測した．表 4 に実験環境を示す．

ユーザ数 N を 100 から 500 まで，k を 2 から 10 まで変動させたときの計測結果を図 7 に示す．

表 4 実験環境

OS	Windows 7 Professional 64 bit
CPU	Intel Xeon CPU X5675 @ 3.07GHz 3.06 GHz
RAM	12.0 GB

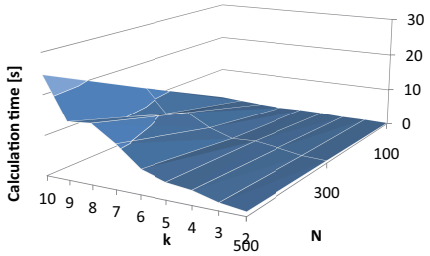


図 7 処理時間
Fig. 7 Calculation time

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 7 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 10 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言える．N の増加に対しては，処理時間はほぼ $O(N\log(N))$ で増加している．これはベースとしている Modrian アルゴリズムの計算量が $O(N\log(N))$ であることに依存している．

7. 考察

本研究で前提としている想定とその妥当性について議論する．

ユーザの位置情報を知っている攻撃者が存在する

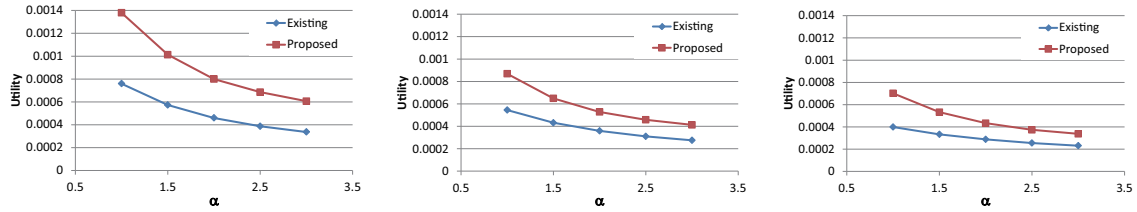
攻撃者が，Alice が時刻 t_1 においてエリア L_1 にいたことを知っている背景として考えられる例として考えられるものを以下に挙げる．

(1) Alice が時刻 t_1 にエリア L_1 に存在したことを物理的に観測した
(2) POS 情報にアクセス可能であり，Alice が時刻 t_1 にエリア L_1 内の店舗で商品を購入した記録を把握した
(3) 他の位置情報管理事業者 B から，Alice が時刻 t_1 にエリア L_1 に存在したことを把握した

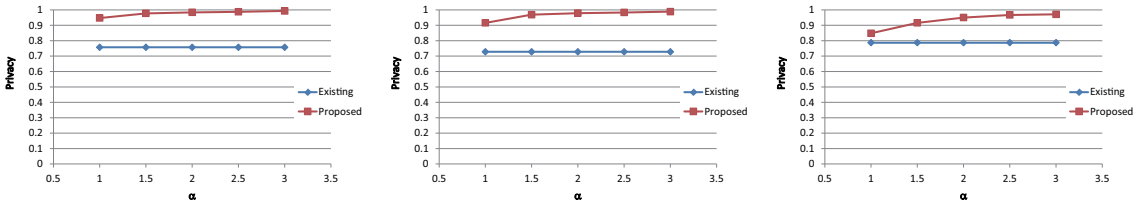
3 番目の例は，事業者 B の提供する情報を用いて Alice にロケーションアウェアサービスを提供することを想定している．したがって，事業者 B から受領する情報には Alice の個人属性は含まれないが，個人を特定できる識別子が含まれているという状況である．

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，例の 2 番目及び 3 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，取得精度にばらつきがある



(a) Utility ($k = 3$)　(b) Utility ($k = 5$)　(c) Utility ($k = 7$)



(d) Privacy ($k = 3$)　(e) Privacy ($k = 5$)　(f) Privacy ($k = 7$)

図 5　Utility 及び Privacy の平均値

Fig. 5　Average values of Utility and Privacy

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 [8]，Wifi を利用した位置情報取得については，誤差が 500m を超える場合もある [15]．

また，携帯端末等で常にユーザの位置情報を取得するには電池使用量が大きく，位置情報の取得頻度は一定以下に抑える必要がある [15]．取得できない時間帯においてユーザの位置を把握するためには，何らかのアルゴリズムを用いて推測しなければならない．推測手法に関してはいくつか提案されており [27][16]，これらの推測結果も匿名化テーブルへの入力として利用することが可能である．この場合，推測された位置情報の誤差はさらに大きくなる．

位置情報取得精度や推測精度は将来的に向上すると考えられるが，精度のばらつき自体は常に生じ得ると考えられる．

8. おわりに

ユーザ属性と行動履歴の情報を元に，どのような属性を持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を行うシナリオを想定した．

マイニングをするにあたって，位置情報について匿名化された後のテーブルに，匿名化エリアへの存在確率も含めることを前提とすると，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．さらに，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．提案する手法はシンプルであるが，従来手法と比べてプライバシ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，匿名化エリアに k 人以上が w 以上の確率で存在しているかどうかを確認する計算の処理高速化が挙げられる．また，シミュレーションではなく実データを用いてより大規模な評価を行う必要がある．

参考文献

[1] Abul, O., Bonchi, F. and Nanni, M.: Never Walk Alone : Uncertainty for Anonymity in Moving Objects Databases, *Proc. 24th IEEE ICDE*, pp. 376–385 (2008).

[2] Aggarwal, G., Feder, T., Kenthapadi, K., Motwani, R., Panigrahy, R., Thomas, D. and Zhu, A.: Anonymizing Tables, *Database Theory - ICDT*, Lecture Notes in Computer Science, Vol. 3363, Springer Berlin Heidelberg, pp. 246–258 (2005).

[3] Andrés, M. E., Bordenabe, N. E., Chatzikokolakis, K. and Palamidessi, C.: Geo-Indistinguishability: Differential Privacy for Location-Based Systems, *CoRR*, Vol. abs/1212.1 (2012).

[4] Apple Inc.: iOS Developer Library, `http://developer.apple.com/library/ios`.

[5] Ardagna, C., Cremonini, M., De Capitani di Vimercati, S. and Samarati, P.: An Obfuscation-Based Approach

for Protecting Location Privacy, *IEEE Trans. Dependable and Secure Computing*, Vol. 8, No. 1, pp. 13–27 (2011).

[6] Brgulja, N., Kusber, R., David, K. and Baumgarten, M.: Measuring the Probability of Correctness of Contextual Information in Context Aware Systems, *Proc. 8th IEEE International Conference on Dependable, Autonomic and Secure Computing*, pp. 246–253 (2009).

[7] Chen, R., Fung, B. C., Desai, B. C. and Sossou, N. M.: Differentially private transit data publication, *Proc. 18th ACM KDD*, pp. 213–221 (2012).

[8] Drawil, N. M., Amar, H. M. and Basir, O. A.: GPS Localization Accuracy Classification: A Context-Based Approach, *IEEE Trans. Intelligent Transportation Systems*, Vol. 14, No. 1, pp. 262–273 (2013).

[9] Gkoulalas-Divanis, A., Kalnis, P. and Verykios, V. S.: Providing K-Anonymity in location based services, *ACM SIGKDD Explorations Newsletter*, Vol. 12, No. 1, p. 3 (2010).

[10] Google Inc.: Android Developers, `http://developer.android.com/`.

[11] Hu, H., Xu, J., On, S. T., Du, J. and Ng, J. K.-Y.: Privacy-aware location data publishing, *ACM Trans. Database Systems*, Vol. 35, No. 3, pp. 1–42 (2010).

[12] Kiefer, J.: Sequential minimax search for a maximum, *Proceedings of the American Mathematical Society*, Vol. 4, No. 3, pp. 502–506 (1953).

[13] LeFevre, K., DeWitt, D. and Ramakrishnan, R.: Incognito: Efficient full-domain k-anonymity, *Proc. ACM SIGMOD*, pp. 49–60 (2005).

[14] LeFevre, K., DeWitt, D. and Ramakrishnan, R.: Mondrian Multidimensional K-Anonymity, *Proc. IEEE ICDE*, pp. 25–25 (2006).

[15] Lin, K., Kansal, A., Lymberopoulos, D. and Zhao, F.: Energy-accuracy trade-off for continuous mobile device location, *Proc. 8th MobiSys*, ACM Press, pp. 285–298 (2010).

[16] Lou, Y., Zhang, C., Zheng, Y., Xie, X., Wang, W. and Huang, Y.: Map-matching for low-sampling-rate GPS trajectories, *Proc. 17th ACM SIGSPATIAL GIS*, ACM Press, pp. 352–361 (2009).

[17] Machanavajjhala, A., Kifer, D., Gehrke, J. and Venkitasubramaniam, M.: l-diversity: Privacy beyond k-anonymity, *ACM TKDD*, Vol. 1, No. 1, pp. 3–es (2007).

[18] Martin, M. and Nurmi, P.: A Generic Large Scale Simulator for Ubiquitous Computing, *Proc. 3rd MobiQuitous*, IEEE, pp. 1–3 (2006).

[19] Meyerson, A. and Williams, R.: On the complexity of optimal K-anonymity, *Proc. ACM PODS*, pp. 223–228 (2004).

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

[21] Mohammed, N., Fung, B. C. M., Wang, K. and Hung, P. C. K.: Privacy-preserving data mashup, *Proc. EDBT*, ACM Press, pp. 228–239 (2009).

[22] Nazareth, L. and Tseng, P.: 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).

[23] Nergiz, M. E. and Atzori, M.: Towards Trajectory Anonymization : a Generalization-Based Approach, Vol. 2, No. 106, pp. 47–75 (2009).

[24] Xiao, X., Yi, K. and Tao, Y.: The hardness and approximation algorithms for l-diversity, *Proc. 13th EDBT*, ACM Press, pp. 135–146 (2010).

[25] Xu, C., Ma, X., Cao, C. and Lu, J.: Minimizing the Side Effect of Context Inconsistency Resolution for Ubiquitous Computing, *Proc. 8th MobiQuitous*, Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, Vol. 104, Springer Berlin Heidelberg, pp. 285–297 (2012).

[26] Xue, M., Kalnis, P. and Pung, H. K.: Location Diversity : Enhanced Privacy Protection in Location Based Services, *Proc. 4th International Symposium on Location and Context Awareness*, Springer-Verlag, pp. 70–87 (2009).

[27] Zheng, K., Zheng, Y., Xie, X. and Zhou, X.: Reducing Uncertainty of Low-Sampling-Rate Trajectories, *Proc. 28th IEEE ICDE*, pp. 1144–1155 (2012).

[28] 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).

[29] 竹之内隆夫，川村隆浩，大須賀昭彦：ユーザ存在の特定を困難にした分散匿名化の提案〜２診療機関のレセプトデータを用いた有効性の評価〜，電子情報通信学会論文誌，Vol. J96-D, No. 3, pp. 596–610 (2013).

[30] 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，情報処理学会論文誌， Vol. 52, No. 7, pp. 2256–2267 (2011).