# 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

高橋　翼† 宮川　伸也† 伊東　直子†

† 日本電気株式会社サービスプラットフォーム研究所　〒 211–8666 神奈川県川崎市中原区下沼部 1753
E-mail: t-takahashi@nk.jp.nec.com, s-miyakawa@ce.jp.nec.com, naoko@cj.jp.nec.com

**あらまし**　本稿では，移動軌跡をリアルタイムに匿名化する手法を提案する．移動軌跡は，測位対象ユーザの移動や生活の動線を表すプライバシ性の高い情報である．移動軌跡を匿名化することで，ユーザと緯度軌跡との対応付けを困難にし，移動や滞在の詳細を隠蔽する．既存の蓄積された移動軌跡を $k$ 匿名化する手法は，時々刻々と位置情報が測位される環境において，情報損失が大きくなるという問題や，$k$ 匿名性を担保できないという問題がある．提案手法では，位置情報の追加の度にリアルタイムに匿名化を実施する．移動の変遷に合わせて匿名移動軌跡を動的に再構成することで情報損失を抑制し，過去の移動軌跡を考慮した差分型の匿名化によって $k$ 匿名性を保証する．評価実験を通して，提案手法の有効性を検証する．

**キーワード**　匿名化，プライバシ保護，移動軌跡，位置情報，リアルタイム処理

# Real-time $k$-anonymization for Trajectory Stream

Tsubasa TAKAHASHI†, Shinya MIYAKAWA†, and Naoko ITO†

† Service Platforms Research Laboratories, NEC Corporation
1753, Shimonumabe, Nakahara-Ku, Kawasaki, Kanagawa, 211–8666 Japan
E-mail: t-takahashi@nk.jp.nec.com, s-miyakawa@ce.jp.nec.com, naoko@cj.jp.nec.com

## 1. はじめに

近年，携帯端末や自動車に搭載された GPS や無線 LAN 等によって計測された，位置情報の測位データを利用する場面が増加している．さらに，位置情報を定期的に取得し，携帯端末や自動車の移動軌跡や行動履歴を記録するサービスが増加している．位置情報は，自宅や勤務先，通学先等，ユーザ特有の位置を表わす情報である場合や，趣味や通院先等の他人に知られたくない位置を表わす情報である場合がある．そのため，位置情報はユーザの位置や位置に紐づく情報を特定したり，センシティブな機密情報であったりするため，プライバシ性が高い情報である．また，位置情報の時系列情報である移動軌跡は，ユーザの生活の導線をトレース可能な情報である．多くの人々の移動軌跡データセットを用いることで，人々の移動経路の解析や，移動に伴う情報や感染病等の伝搬経路の解析が可能となる．ただし，移動軌跡はユーザと一意に対応付けることが可能であり，非常にプライバシ性が高い．移動軌跡は，プライバシ性の高い場所への経路を含み，自宅等の滞在，不在も簡単に把握できる．さらに，サービス提供者やデータ解析者による移動軌跡のリアルタイムな利用によって，ユーザは常に追跡や監視といった脅威に晒される．そこで，このようなプライバシ情報をサービス提供者やデータ解析者に提供する際には，匿名化によって匿名性を確保することが求められる．本稿では，移動軌跡をリアルタイムに匿名化に関する問題を扱う．

匿名化とは，ユーザを特定できないようにプライバシ情報を加工する処理である．ユーザを特定できない度合を示す指標を匿名性指標と呼び，$k$ 匿名性（$k$-anonymity [1]）がよく知られている．プライバシ情報の中で，ユーザを必ずしも一意に識別することが可能な識別子ではないが，背景知識等を考慮するとユーザを識別する可能性がある情報は，準識別子（間接識別子）と呼ばれる．また，ユーザが他人に知られたくない情報はセンシティブ情報と呼ばれる．$k$ 匿名性は，匿名化したデータセットにおいて，同じ準識別子を持つセンシティブ情報が $k$ 個以上存在することを保証する指標である．$k$ 匿名性を保証することで，ユーザが特定される可能性が $1/k$ 以下となり，ユーザの特定を困難にすることができる．匿名性の指標としては，$k$ 匿名性以外にも，$\ell$ 多様性 ($\ell$-diversity [2]) や $t$ 近接性 (t-closeness [3]), $m$ 不変性 (m-invariance [4]) 等が知られている．

ユーザと移動軌跡が紐づけられることで，そのユーザにとってプライバシ性の高い場所を含むすべての行き先・滞在先が露わになる．また，行動の監視や追跡をされる危険性がある．そのため，移動軌跡のプライバシ性は単独の位置情報や，単なる

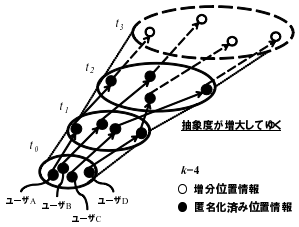


図 1　移動軌跡のリアルタイム匿名化

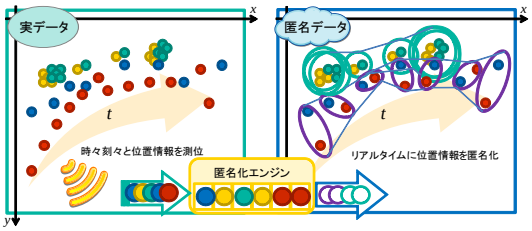


図 2　移動軌跡ストリームに対するリアルタイム匿名化

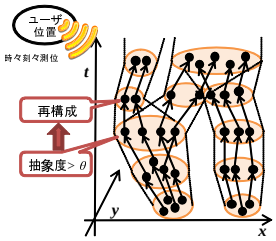


図 3　提案匿名化手法の概念図

複数の位置情報の組合せと比べて，極めて高いプライバシ性を有する．さらに，移動軌跡中のいくつかの位置情報の測位データが露わになるだけで，ユーザの特定が可能な場合がある．例えば，自宅の場所や勤務先，最寄り駅等の情報は，同僚や友人に知られていることが多い．そのため，移動軌跡の中にこれらの事前知識である場所の測位データが含まれていると，移動軌跡に紐づくユーザが特定され，事前知識以外の通院先や嗜好に関する場所への滞在情報が流出してしまうおそれがある．

そこで，ユーザと移動軌跡の紐づけを困難にするために移動軌跡の匿名化が必要とされる．蓄積された静的な移動軌跡を匿名化する手法がいくつか提案されている [5] [6] [7]．Abul ら [5] は，$(k, \delta)$-anonymity という匿名性指標を用い，データベースに蓄積された静的な移動軌跡をチューブ状に汎化する匿名化手法を提案している．この手法では，移動軌跡の始点と終点が明らかなデータに対して，移動軌跡間の距離が近いものをグループ化して抽象化することで匿名化を行う．このとき，始点ー終点間で一貫して $k$ 個以上の移動軌跡が同じグループ（匿名移動軌跡）に含まれることが保証される．これによって，ユーザの移動軌跡中のある 1 つの滞在地点が分かったとしても，どの移動軌跡であるかを特定することはできない．

Abul らの手法は，静的に蓄積された移動軌跡に対する匿名化には有効であるが，時々刻々と新たな位置情報を測位し，リアルタイムに伸長する移動軌跡の匿名化を想定していない．Abul らの手法は，始点と終点間の全経路において匿名性を満たすように一度に匿名化を行う巨視的なものであり，移動軌跡の増分のように局所的な位置情報を対象とするものではない．Abul らの手法で匿名化した移動軌跡に，対応する増分情報を接続し匿名性を維持し続けようとすると，時間の経過と共に，移動軌跡間の地理的類似性が薄くなり，過剰に抽象化した匿名情報となることが想定される (図 1)．また，過去に匿名化した移動軌跡を考慮せずに，異なる時間間隔の移動軌跡をそれぞれ独立して匿名化を行うと匿名性を担保できない．

本稿では，新たに測位された移動軌跡の増分位置情報を，既に匿名化された匿名移動軌跡を考慮しながらリアルタイムかつ連続的に匿名化する手法 CMOA (Continuous Moving Objects Anonymization) を提案する (図 2)．既に匿名化した情報を考慮することで，一貫した $k$ 匿名性を保証する．また，時々刻々と変化する人々の移動に合わせて匿名移動軌跡を構成する移動軌跡の組み合わせを動的に再構成することで，過度な抽象化を抑制する (図 3)．さらに，匿名移動軌跡の初期構成を密集度に応じて構築し，トレーサビリティを高く保つ．

本稿の以降の構成は以下の通りである．2 章では，本稿が対象とする問題を定義し，解決する課題を明確にする．3 章では，提案手法について詳細な説明を記載する．4 章では，提案手法の有効性を評価するための評価指標を導入する．5 章では，提案手法の有効性を評価実験で検証する．最後に 6 章にて本稿をまとめ，今後の検討課題についても示唆する．

## 2. 問 題 定 義

本稿では，特定のユーザの滞在先を表す位置情報として測位データ $(tid, x, y, t)$ が定期的に測位され，信頼できるプラットフォーム上に蓄積されることを想定する．ここで，$tid$ は移動軌跡 (ユーザ) を識別するための移動軌跡識別子である．単体の位置情報とは異なり，$tid$ を参照することで，ユーザの移動の軌跡を辿ることができる．$x$ は測位時刻 $t$ における測位データの水平方向の値 (経度) を表し，$y$ は垂直方向の値 (緯度) を表す．また，測位された位置情報は，第三者にリアルタイムに提供され利活用されるものとする．この第三者の信頼性は不明であり，敵対者である可能性もある．

### 2.1 移 動 軌 跡

移動軌跡 $\tau$ は以下のような測位データ $(tid, x, y, t)$ の測位時刻 $t$ に関して全順序のシーケンスからなる．

$$\tau = \{(tid, x_i, y_i, t_i) | \forall i, j \in \mathbb{Z},\ i > j,\ t_i > t_j\}$$

$x$, $y$ は GPS の経度，緯度に対応する．また，移動軌跡に対して，プライバシ性の高いセンシティブ情報 SA が付随する場合 $\tau = \{(tid, x_i, y_i, t_i, SA_i)\}$ も想定される．センシティブ情報には病状などが該当する．センシティブ情報を含む移動軌跡を分析することで，感染病の感染経路分析等が可能になる．本稿では，移動軌跡はリアルタイムに逐次測位されていることを想定する．特に，毎分すべての利用者の位置情報が測位され，移動軌跡のデータセット $T(tid, x, y, t, SA)$ が信頼できる事業者に蓄積され，第三者であるサービス事業者にリアルタイムに提供され，利用される．

### 2.2 脅　　威

敵対者 ($Adversary$) として，あるユーザの移動軌跡 $\tau$ のうち，いくつかの $(x, y, t)$ を知っているものを想定する．このとき，Adversary の知識 $A = \{(x, y, t)\}$ は，$A \subset \pi_{x,y,t}(\tau)$ であると仮定する．これは，あるユーザの自宅や勤務先をなんらかの形で知っている人々や，偶然出会いある一定時間，時間や空間を共にした人々などが該当する．このような人々の中には，ストーカー等の悪意のあるものが含まれている可能性がある．

このような想定の上で，移動軌跡 $\tau$ が漏えいし，ユーザが特定されたとき，以下のような脅威が想定される．

（1） ユーザの現在位置が露わになる．自宅への在宅/不在や，特定の場所への滞在等の状況が知られてしまう．

（2） ユーザの生活の動線が露わになる．本来知られるはずのない人に日常生活の一部を知られてしまう．

（3） ユーザの興味や嗜好，行動パターンが分析される．隠蔽しておきたい秘密の行動を隠すことができない．

（4） ユーザへの追跡や移動・行動軌跡の監視をされる．

脅威 (1)，(2)，(3) は，移動軌跡に対して，個人の特定を困難にするように準識別子を曖昧にする形での匿名化 (従来の $k$ 匿名化のような手法) を施すことで，対処できる．本稿は，この問題を扱う．脅威 (4) に対しては，個人の特定を困難にすることに加えて，移動軌跡として考え得るパターンを多様化する必要性がある．本稿では，準識別子を抽象化する形態の匿名化手法の実現を目的とする．$\ell$ 多様化のような形態での匿名化手法は将来の課題とする．

### 2.3 移動軌跡の $k$ 匿名化

本節では，移動軌跡に対する $k$ 匿名性について説明する．移動軌跡に対する $k$ 匿名性は，位置情報や統計情報などのある時刻におけるスナップショットデータに対する $k$ 匿名性とは性質が異なる．移動軌跡は，連続するデータが意味を持つシーケンスデータである．シーケンスデータに対する $k$ 匿名性は，シーケンスの開始から終了までの全区間において，常に同じ $k$ 個のメンバからなるグループを維持し続けることで実現できる．移動軌跡のデータセット $T$ を匿名化したデータセットを $T'$ とする．移動軌跡 $\tau$ を匿名化した匿名移動軌跡 $\tau'$ は以下のように表す．

$$\tau' = \{(tid', r_i, t_i, SA_i) | \forall i, j \in \mathbb{Z},\ i > j,\ t_i > t_j\}$$

$tid'$ は $tid$ と対応付く識別子であり，ランダムに割り振った識別子を用いる．$R$ は，$x$ と $y$ を汎化によってぼやかした領域を表す属性である．匿名移動軌跡の準識別子は，$(R, t)$ を想定し，ある場所にいつ滞在したかという情報から，特定の個人と匿名移動軌跡の対応付けを困難にする．同じ準識別子を持つ匿名移動軌跡の集合を，匿名チューブと呼ぶ．匿名チューブ $tube$ は，同じ準識別子 $(R, t)$ を持った匿名グループが時系列に連なった集合である．$\tau'$ から成る $T'$ は，以下の性質を満たすような $k$ 匿名化を実施する．

> ある匿名移動軌跡 $\tau'$ が存在する全区間 $[t_0, t_n]$ において，常に同じ準識別子 $(R, t)$ を持った匿名移動軌跡が少なくとも $k-1$ 個存在する．$(|\cap_{t \in [t_0, t_n]} tube_i[t]| \geqq k)$

これを移動軌跡の $k$ 匿名性と呼ぶ．$t_n$ は匿名移動軌跡 $\tau'$ の終端時刻，もしくは最新の測位時刻である．$tube$ の各測位時刻 $t$ の断面には，匿名グループに共通の準識別子 $(R, t)$ が含まれる．ここで，時刻 $t$ において，匿名チューブの領域を $R(tube[t])$，領域の面積を $S(tube[t])$，構成する移動軌跡の数を $N(tube[t])$，領域 $R$ の重心ベクトルを $Centroid(tube[t])$ とする．

### 2.4 移動軌跡のリアルタイム $k$ 匿名化

蓄積された静的な移動軌跡のデータセットに対して，移動軌跡の $k$ 匿名性を充足させる手法として，Abul らの研究がある．本稿で想定する環境では，移動軌跡の増分位置情報が時々刻々と測位される．

静的な移動軌跡に対する匿名化手法をリアルタイム環境で適用には，以下のような適用方法と問題点が考えられる．

（1） 既に匿名化した移動軌跡を始点，新たに測位した位置情報を終点とした匿名化を実施する．このとき，既に匿名化した構成の $k$ 匿名性を保った組み替えができないため，抽象度 (匿名化チューブの断面積 $S$) が大きくなる．

（2） 既に匿名化した移動軌跡とは無関係に，特定の時間間隔ごとに匿名化を実施する．過去と現在の匿名化に関連がないため，全データセットで見ると $k$ 匿名性を満たしていない．また，過去から現在に至るまでの移動を全くトレースできない．

本稿では，リアルタイム環境における移動軌跡の匿名化に関して以下の課題に取り組む．

- リアルタイムな連続的 $k$ 匿名化手法の実現
- 解像度を高く (抽象度を低く) 保った匿名移動軌跡の実現
- 可能な限り長時間トレース可能な匿名移動軌跡の実現

匿名化の対象は，移動軌跡 $\tau$ 全体ではなく，その最新の測位データ $(tid, x_n, y_n, t_n)$ である．最新の測位データを増分位置情報と呼ぶこととする．

後述の提案手法では，移動軌跡の $k$ 匿名性を差分的な匿名化によって充足させ，動的再構成によって解像度が高い匿名移動軌跡を生成する．新たな増分位置情報が測位されるまでに匿名化を完了し，リアルタイムに匿名化した増分位置情報を提供可能にする．

## 3. 移動軌跡のリアルタイム匿名化手法

### 3.1 概　　要

移動軌跡をリアルタイムに $k$ 匿名化する手法 CMOA (Continuous Moving Objects Anonymization) を提案する．提案匿名化手法 CMOA は，移動軌跡の $k$ 匿名性を差分的な匿名化によって充足させ，動的再構成によって解像度が高い匿名移動軌跡を生成する．CMOA は，匿名化によってぼやかされた範囲の面積を情報の損失 ($ILoss$)，同じ匿名グループに存在する移動軌跡の数をプライバシ利得 ($PGain$) と捉え，この二つの指標を考慮した匿名グループの密度 $\delta = \frac{PGain}{ILoss}$ が向上するように匿名グループの構成を動的に再構成する．

提案匿名化手法 CMOA は，初期匿名化，差分匿名化，動的再構成の 3 つの構成要素から成る．

第一の構成要素である初期匿名化では，単一の測位データ ($|\tau| = 1$ の移動軌跡) の汎化を行う．初期匿名化では，$ILoss$(領域のサイズ) の閾値 $\theta$ の範囲内で，$PGain$ 最大の (可能な限りたくさんのメンバから成る) 匿名チューブをクラスタリングによって生成する．このとき，$k$ 匿名性の充足を最も重視し，$k$ 匿名性を満たさない場合は $\theta$ を超えてもよい．

第二の構成要素は差分匿名化である．差分匿名化では，新たな増分位置情報に対して，既に匿名化された匿名移動軌跡の

**Algorithm 1** bisection(C, O)

```
if |C| ≧ k then
  if |C| ≧ 2k AND C.Area ≧ θ then
    {C1, C2} ← 2 − means(C)
    if |C1| ≧ k AND |C2| ≧ k then
      for all Ci ∈ {C1, C2} do
        bisection(Ci, O)
      end for
    end if
  end if
  O ← O ∪ C
end if
```

データセットに基づき，前回汎化した際と同一の匿名チューブの構成で汎化を実施する．

第三の構成要素は，動的再構成である．差分匿名化によって生成された匿名チューブの $ILoss$ と $PGain$ を評価して，基準を超えた匿名チューブを再構成する．再構成の結果，$k$ 匿名性を充足できない匿名移動軌跡が生じる．このような匿名移動軌跡は，移動軌跡の識別子 $tid$ を変更して，新たに匿名化を実施し直す．ただし，匿名移動軌跡の識別子の変更は，これまでの移動情報を完全に損失する．これを可能な限り行わずに匿名チューブを維持し続けるような匿名化手法の実現も本稿の課題の一つである．

匿名化結果として出力される情報は，匿名化を実施した時刻 $t_n$ における匿名移動軌跡 $\tau'$ の部分匿名移動軌跡である．

### 3.2 初期匿名化

初期匿名化では，クラスタリングによってユーザ毎の単一の位置情報をグループ化し，匿名チューブの始点を形成する．このとき，動的再構成における再構成の自由度を高く保つために，できるだけ多くの位置情報から匿名チューブを構成する．これは，$PGain$ を高めることにもなる．

ただし，あまりに大きな匿名チューブは情報損失 $ILoss$ が大きく，位置情報，移動情報としての価値が乏しい．そこで，$ILoss$(匿名移動軌跡が包含される領域のサイズ) に閾値 $\theta$ を与え，閾値の範囲内でできるだけ多くの位置情報を包含し，$PGain$ が最大となる匿名グループを構成する．この匿名グループを匿名チューブの始点とする．このとき，$k$ 匿名性は必ず充足し，$k$ 匿名性の充足に必要であれば閾値 $\theta$ の範囲内でなくてもよいものとする．

空間情報をクラスタリングする手法として，k-means 法がよく知られている．ここで，$k$ 匿名性の $k$ との混乱を避けるために，k-means のクラスタ数を表す変数として $\gamma$ を用いる．k-means 法では，クラスタの重心 (Centroid) とクラスタに属するデータとの距離が最小となるようにクラスタが形成される．よって，データが密集しているところでは，多くのメンバを持つ小さなサイズのクラスタを形成できる．本稿では，k-means 法を発展させた，k-means++法 [8] を利用する．k-means++法を利用して，段階的にデータセットをパーティショニングすることで匿名チューブを構成する．

移動軌跡を構成する属性のうち，空間座標を表す $(x, y)$ の二次元ベクトルをクラスタリングの対象とする．クラスタリングの対象となるデータの測位時刻 $t$ はすべて同一のものとする．

初期匿名化におけるクラスタリングでは，$\gamma = 2$ の k-means++法を再帰的に繰り返し行うことで所望のクラスタを得る．まずデータセットを一つの大きなクラスタとする．クラスタのメンバ数が $2k$ 以上であり，かつ領域のサイズが閾値 $\theta$ を超える場合には，$\gamma = 2$ の k-means++クラスタリングによって，二分割 ($bisection$) を行う．さらに，分割後のクラスタのメンバ数が $2k$ 以上であり，かつ領域のサイズが閾値 $\theta$ を超える場合には，さらなる $bisection$ を行う．$bisection$ を行った結果，クラスタのメンバ数が $k$ 未満になるクラスタが現れたら，当該 $bisection$ を取り消す．すべてのクラスタの領域のサイズが閾値 $\theta$ を下回る，もしくは一度以上 $bisection$ を取り消される ($k$ 匿名性の制約上分割が不可能となる) まで繰り返す．詳細なアルゴリズムを Algorithm 1 に示す．

Algorithm 1 によって，各移動軌跡 $\tau$ はいずれかのクラスタに割り当てられる．このクラスタを匿名チューブの始点となるスナップショット $tube[t_0]$ とする．各移動軌跡 $\tau$ には，匿名チューブの $R(tube[t_0])$ に対応する準識別子が与えらて匿名移動軌跡 $\tau'$ が生成される．以上によって，各匿名移動軌跡は，$T'$ 中に同じ準識別子 $(R, t)$ を持った $k-1$ 個以上の匿名移動軌跡が存在し，$k$ 匿名化される．

### 3.3 差分匿名化

差分匿名化では，新たな増分位置情報に対して，既に匿名化された匿名移動軌跡のデータセットに基づき，前回匿名化した際の匿名チューブとまったく同じ構成で汎化を実施する．言い換えると，最新の測位時刻 $t_n$ における匿名化は，測位時刻 $t_{n-1}$ における匿名チューブの断面 $tube[t_{n-1}]$ を構成する匿名移動軌跡の組み合わせと同じ組み合わせで $tube[t_n]$ を構成する．よって，差分匿名化を実施した際には，以下が成り立つ．

$$TID'(tube[t_n]) - TID'(tube[t_{n-1}]) = \phi$$

$TID'(tube[t])$ は時刻 $t$ における tube を構成する $\tau'$ の $tid'$ である．このように $t_{n-1}$ に $k$ 匿名化された匿名チューブと $t_n$ でもまったく同じメンバ構成で匿名化を行うことで，$k$ 匿名性を継続することができる．本提案では，$ILoss$ が閾値 $\theta$ を超えない間は，常にこの差分匿名化のみで匿名化を行う．これによって小さなコストで匿名化を実施することができる．

### 3.4 動的再構成

差分匿名化で形成された匿名チューブをより最適な構成へ変更する．動的再構成の対象となる匿名チューブは，匿名チューブのエリアのサイズが閾値 $\theta$ を超え，エリアのサイズが増加傾向にある匿名チューブである．動的再構成の条件は以下のように式で表わされる．

（1） $S(tube[t]), S(tube[t-1]) > \theta$

（2） $\frac{dS}{dt} = S(tube[t]) - S(tube[t-1]) > 0$

条件を満たす匿名チューブに対して，分割処理，合成処理を実施した匿名チューブが $PGain/ILoss$ を改善可能な場合に，動的再構成前の匿名チューブに代わる新たな匿名チューブを出力

**Algorithm 2** split($tube_o[t_n]$)

$\{tube'_1[t_n], tube'_2[t_n]\} \leftarrow 2 - means(tube_o[t_n])$
**if** $N_k(tube'_1[t_n])\delta(tube'_1[t_n]) + N_k(tube'_2[t_n])\delta(tube'_2[t_n])$
$> N_k(tube[t_n])\delta(tube[t_n])$ **then**
  **return** $\{tube'_1[t_n], tube'_2[t_n]\}$
**else**
  **return** $tube_o[t_n]$
**end if**

**Algorithm 3** combine($tube_o$, $Tube'_c$)

**for all** $tube_c \in Tube'_c$ **do**
  $tube'_o \leftarrow tube_o \cup tube_c$
  **if** $N_k(tube'_o[t_n])\delta_o(tube'_o[t_n]) > N_k(tube_o[t_n])\delta(tube_o[t_n]) +$
  $N_k(tube_c[t_n])\delta(tube_c[t_n])$ **then**
    $tube_o \leftarrow tube'_o$
  **end if**
**end for**

する．ここで，$PGain/ILoss$ は以下のように表わされる．

$$PGain/ILoss = \delta(tube[t]) = \frac{N(tube[t])}{S(tube[t])}$$

以降，$PGain/ILoss$ は $\delta$ で表わす．

**3.4.1** 分割処理

分割処理では，$ILoss$ が閾値 $\theta$ を超えた匿名チューブの分割を行う．このとき，分割後のそれぞれの匿名チューブに含まれる移動軌跡の数が $k$ 個以上になるように分割を行う．分割可能な匿名チューブは，含まれる移動軌跡の数が $2k$ 以上である必要がある．

差分匿名化によって生成された最新の匿名チューブの断面 ($tube[t_n]$) を構成する匿名移動軌跡の集合に対してクラスタリングによる分割を行う．分割処理の対象とする匿名チューブを $tube_o$ とする．$tube_o$ に対して，初期匿名化と同様に，2-means クラスタリングを実施する (Algorithm 2)．クラスタリングによって，生成された匿名チューブの集合を $Tube'$ とする．クラスタリング後の匿名チューブの集合が，分割前の匿名チューブよりも $\delta$ を改善できているかを評価する．

このとき，k-means 法を用いたクラスタリングでは，必ずしも分割後の匿名チューブ全てがメンバ数 $k$ 以上になるわけではない．提案匿名化手法 CMOA では，メンバ数 $k$ 未満の匿名チューブの生成は $k$ 匿名性に違反するため許されない．ただし，メンバ数 $k$ 以上の匿名チューブとメンバ数 $k$ 未満の匿名チューブとにクラスタリングされた場合に，メンバ数 $k$ 未満の匿名チューブを切り落とすことで，全体として $\delta$ を向上できる場合がある．$\delta$ を向上可能な場合には，メンバ数 $k$ 未満のクラスタを切り落とす．切り落とされた匿名移動軌跡は，匿名化結果として出力しない．

これらを考慮するために以下の式を評価し，満足する場合は，$tube'_i \in Tube'$ を匿名化結果として $tube_o$ と置き換える．

$$\sum_{tube'_i \in Tube'} N_k(tube'_i[t_n])\delta(tube'_i[t_n]) > N_k(tube_o[t_n])\delta(tube_o[t_n])$$

ここで，$N_k$ は匿名チューブのメンバ数 $N$ が $k$ 以上のときは，$N_k = N$ であり，$k$ 未満のときは，$N_k = 0$ となる関数である．

**3.4.2** 合成処理

もう一つが合成処理である．合成処理では，これ以上分割不可能な匿名チューブ同士を一つの匿名チューブに集約する処理を行う．これによって，匿名チューブの $PGain$ を向上できる．また，再び分割処理が可能になり，匿名チューブのメンバ構成に大きな自由度を与えることができ，エリアのサイズの小さな匿名チューブの継続可能性を高めることができる．

処理対象の匿名チューブの集合 $Tube_o$ は，メンバ数が $2k$ 未満かつ $ILoss$ が閾値 $\theta$ を超えた匿名チューブ $tube_o$ とする．$Tube_o$ と Suppression されている匿名移動軌跡を結合候補の匿名チューブ $Tube_c$ とする．$Tube_c$ の中で，時刻 $t_{n-m}$ と $t_n$ の両方で，$R(tube_o)$ に包含される重心を持つ $tube_c$ を結合する (図 4)．最終的な結合候補 $Tube'_c$ は以下のような匿名チューブである．

$$Tube'_c = \{tube_c | tube_c \in Tube_c \wedge Centroid(tube_c[t_n]) \subset R(tube_o[t_n]) \wedge Centroid(tube_c[t_{n-m}]) \subset R(tube_o[t_{n-m}])\}$$

結合後，以下の式を満足する場合には，結合後の匿名チューブへと更新する．

$$N_k(tube'_o[t_n])\delta_o(tube'_o[t_n]) > N_k(tube_o[t_n])\delta(tube_o[t_n]) + N_k(tube_c[t_n])\delta(tube_c[t_n])$$

合成処理の詳細を Algorithm3 に示す．$S(tube_o[t])$ が大きい $tube_o$ から順にこの処理を行う．

**3.4.3** 移動軌跡識別子の再割り当て

分割によって，匿名チューブのメンバ数が $k$ 未満になる場合と，合成処理後に，分割処理を実施した場合に $k$ 匿名性への違反が生じる．

前者の場合は，前述の通り切り落としを行い，合成処理によって $k$ 匿名性を充足するまで匿名化の結果 $\tau'$ として出力せず，内部でのみ匿名化の対象として処理を続ける．

後者の場合は，移動軌跡識別子を割り当て直す．特に本稿では，再割り当ての対象となる匿名移動軌跡の識別子の集合 $TID'(tube[t])$ の中から $tid'$ をランダム抽出し，新たな匿名移動軌跡の識別子として割り当てる．これによって，過去の移動軌跡とその後の移動軌跡との関連性はなくなる．ただし，ある方向から来たユーザの群れが $\ell$ 通りの方向へ移動していったことだけは判る．

## 4. 評価指標

提案匿名化手法 CMOA によって生成された匿名移動軌跡データセットの有用性と，プライバシ保護の度合いを評価する．本稿では，特定オブジェクトの移動を詳細にかつ長時間に渡って追跡できる移動軌跡ほど有用性が高いと考える．そこで，匿名移動軌跡の特定時刻における詳細さを測る指標として，解像度指標を導入する．また，匿名移動軌跡におけるトレーサビ

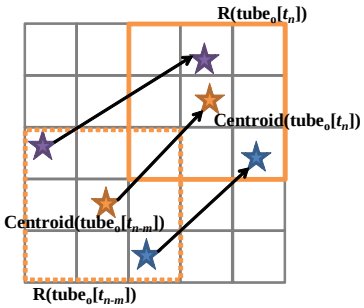


図 4　合成処理の候補

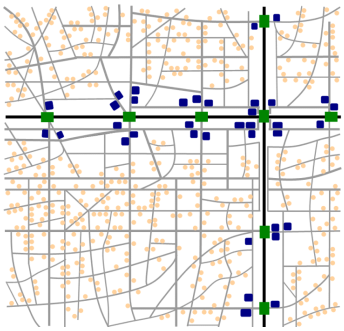

図 5　シミュレーションマップ

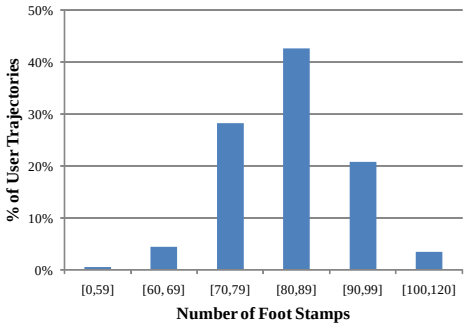


図 6　評価データセットのユーザ移動回数

リティを測る指標として最大継続時間を導入する．プライバシ保護の度合いは，Bayardo らによって提案された識別性指標 (Discernibility Metric [9]) によって評価する．

評価実験では，本節で取り上げた指標に基づいて評価を行う．

### 4.1　解像度指標 (RM)

匿名移動軌跡のデータセットの有用性，情報の損失度合いを評価する指標として，解像度指標 (*Resolution Metric*, RM) を定義する．移動軌跡に対して匿名化を実施すると，移動軌跡のスナップショットである位置情報が曖昧に抽象化される．これによって，ユーザの特定時刻における滞在地点の解像度が失われる．解像度指標 RM は，匿名化後のデータセットがどの程度の解像度を持っているかを測る指標である．RM は，元データの緯度経度で表わされるピンポイントな位置情報の解像度を 1 とする．匿名移動軌跡のスナップショット $\tau'[t]$ の解像度 $rm(\tau'[t])$ 以下のような式で表わす．

$$rm(\tau'[t]) = \frac{1}{S(\tau'[t])} = \frac{1}{ILoss(\tau'[t])}$$

ここで，切り落としによって匿名化結果から省かれたものは $S(\tau'[t]) = \infty$ とし，$rm(\tau'[t]) = 0$ とする．よって，全匿名化データの時刻 $t$ における RM は，以下のように表わされる．

$$RM[t] = \sum_{\tau' \in T'} ILoss(\tau'[t])^{-1}$$

### 4.2　最大継続時間 (MD)

移動軌跡は連続した位置情報から成り，ユーザの移動を時系列にトレースするための情報と言える．提案匿名化手法では，トレーサビリティを犠牲にすることで $k$ 匿名性を実現している．ここでは，匿名化結果である匿名移動軌跡がどの程度のトレーサビリティを保っているかを評価する．この指標を最大継続時間 (*Maximum Duration*, MD) と呼び，以下の式で表わす．

$$MD(tid) = max_{tid' \in TID'(tid)} Dur(tid')$$

ここで，$Dur(tid')$ は，同じ $tid'$ で移動し続けた時間を表わす．

### 4.3　識別性指標 (DM)

匿名化グループのサイズに基づいた匿名化データセットの品質を測る指標として *Discernibility Metric* (DM) [9] が提案されている．DM は，特定レコードの識別の困難さを表わしており，DM の高いデータセットほどプライバシの保護度合が高いと言える．DM は以下のような式によって定義する．

$$DM(D) = \sum_{i=1}^{n-1} |p_i|^2 + |p_n||D|$$

ここで，$p_i$ は各匿名化グループを，$p_n$ は切り落とされて匿名化グループから漏れた集合，$D$ は匿名化前のデータセットである．DM を時刻 $t$ について導入する場合，$PGain$ を用いて以下のように記述する．

$$DM[t] = \sum_{tube[t]} PGain(tube[t])^2 + |sup[t]||D|$$

ここで，$|sup[t]|$ は時刻 $t$ において切り落とされた移動軌跡の数を表わす．

## 5.　評 価 実 験

提案匿名化手法 CMOA の有効性を検証するために評価実験を行った．本節では，実験環境や評価結果を示し，考察を行う．

### 5.1　データセット

提案手法の有効性を検証するために，人々の移動の流れを表すデータセットを用いる．利用するデータセットは，1000，2000，5000，10000 人のユーザの 1 分おきの位置情報の人口データを用いる．データセットは，移動体シミュレータ Siafu [11] を利用して，徒歩と電車を利用した通勤のシナリオを作成した．データセットが移動するシミュレーションマップは水平方向：約 4.6km，垂直方向：約 3.2km の地図である．シミュレーションマップを図 5 に示す．生成したデータセットにおけるユーザの移動回数の分布を図 6 に示す．評価実験には，2 時間分の移動データを用いる．

シミュレーションマップを移動するユーザは以下のような特徴を持つ．

- 各ユーザには，自宅と勤務先がランダムに設定される．
- ユーザは，マップ上に敷かれた道路，線路上を移動する．
- ユーザはランダムに設定された時刻に自宅を出発し，勤務先を目指して移動を開始する．
- 最寄駅よりも勤務先の方が自宅から近い場合は，勤務先へ直接移動する．
- 最寄駅の方が自宅から近い場合は，最寄駅へ移動する．
- 電車は一定時間置きに駅に到着し，ユーザは電車が到着するまで駅に滞在する．

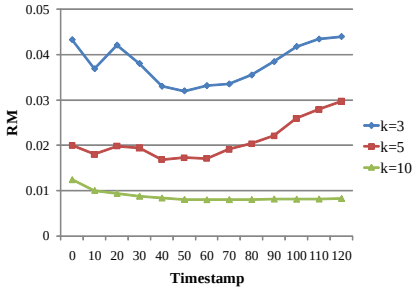


図 7　解像度指標 (RM)

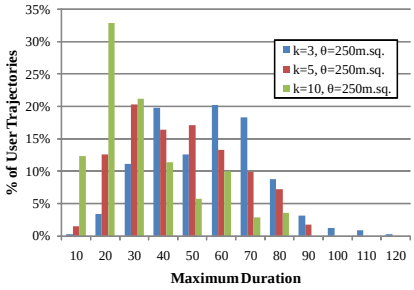


図 8　最大継続時間 (匿名度合いによる変化)

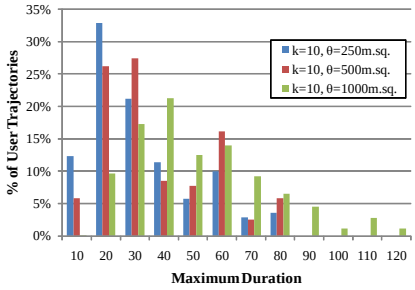


図 9　最大継続時間 (抽象化許容量による変化)

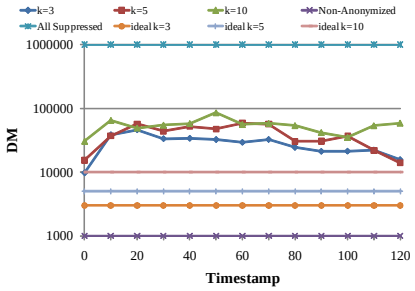


図 10　識別性指標 (DM)

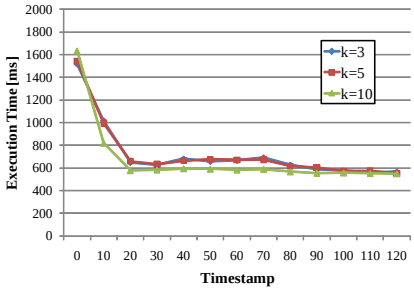


図 11　匿名化実行時間 ($k = 3, 5, 10$)

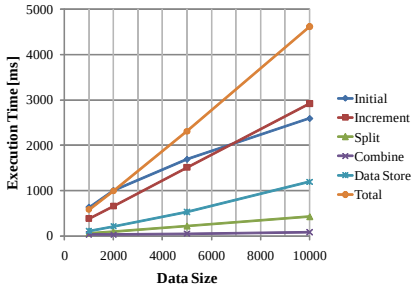


図 12　匿名化実行時間 (データサイズによる比較)

### 5.2　実験環境

以下のような計算機を用いて，評価実験を行う．

- CPU： Intel Xeon X3350 2.66GHz
- 主記憶容量： 8GB
- 外部記憶装置： HDD 500GB
- OS： CentOS 5.5

提案手法 CMOA を実装した匿名化エンジンの開発を行った．匿名化エンジンは以下のような環境を利用して開発した．

- プログラミング言語：Java 1.6.0_17
- DBMS：Apache Cassandra 0.6.6

匿名化エンジン中の各匿名化処理における距離計算では，計算を簡略化するために，シミュレーションマップをセル状に分割した座標系を利用した．セルのサイズには，総務省の統計局が発表している 8 分の 1 地域メッシュのサイズ [10] である 125m 四方を設定した．

また，匿名化エンジンに位置情報を定期的に送信するモジュール (位置情報送信モジュール) の開発を行った．以下のようなプログラミング言語，DBMS を利用した．

- プログラミング言語：Java 1.6.0_17
- DBMS：PostgreSQL 8.4.5

評価実験では，DB に格納されたすべてのユーザの特定時刻 ($t$) の位置情報を 1 分置きに匿名化エンジンに送信する．匿名化エンジンは，新たな位置情報が到着する度に匿名化を実施する．

### 5.3　評価結果

#### 5.3.1　解像度指標 RM の評価

解像度指標 RM は，特定時刻における情報の損失度合いを測る指標である．RM の時間変化を図 7 に示す．図 7 では，$[t_{n-9}, t_n]$ の区間における平均値を $t_n$ 上にプロットしている．$k = 3, 5, 10$ のときの $\theta$ はすべて 250m 四方とした．

$k$ の値が異なる何れの匿名化結果も，時間経過に伴って解像度の増減が観測できる (図 7)．また，$k$ が大きいほど，解像度が減少している．これは，$k$ 匿名性を満たすために，多くの移動軌跡を包含する必要があるためである．[11, 20] 分経過時には，[1, 10] 分経過時よりも解像度の改善が見られる．特に $k = 5$ においては，RM が改善し続け，初期状態よりも高い解像度を得ている．このように，動的再構成による密度の高い匿名チューブへの再構成の効果が表れていると言える．

#### 5.3.2　最大継続時間 MD の評価

移動軌跡は連続した位置情報を時系列に追うことでユーザの移動をトレースするための情報である．提案匿名化手法では，トレーサビリティを犠牲にすることで $k$ 匿名性を実現している．ここでは，匿名化結果である匿名移動軌跡がどの程度のトレーサビリティを保っているかを評価する．$tid'$ を変更せずに継続した時間の最大値 (MD) をユーザ毎に算出した．MD の分布を図 8 と図 9 に示す．図 8 は $k$ の値を変化させ，$\theta = 250m$ 四方に固定した場合である．図 9 は，$k = 10$ に固定し，$\theta = 250, 500, 1000m$ 四方に変化させた場合である．

最大継続時間 MD は，$k$ の値が増加すると，継続時間が短くなっていることが分かる (図 8)．特に，$k = 10$ のときは，最大でも 20 分程度しか継続できておらず，トレーサビリティは低い．$k$ の値を高めたときに同様の継続時間を保つためには，$\theta$ を増加させる必要があることが分かる．図 9 では，$\theta$ を増加させるに伴って MD が向上し，$\theta = 1000m$ 四方のときには，平均で 40 分前後と高いトレーサビリティを得ている．これは，$k = 3, \theta = 250m$ 四方とほぼ同等である．

以上より，$\theta$ は $k$ の値に依存することが分かる．ただし，データセットの性質にも依るところ大きいと推察できる．$\theta$ の自動的な決定や非パラメータ化が今後の課題の一つである．

**5.3.3** 識別性指標 DM の評価

識別性指標 DM は，匿名化データセットにおける特定レコードの識別の困難さを表わす指標であり，匿名度合いを表わす．DM をタイムスタンプ毎に導出し，$[t_{n-9}, t_n]$ の区間における平均値を図 10 の $t_n$ 上に示した．ideal k は，$k$ 匿名性を満たすために，すべての移動軌跡が切り落とされず，ちょうどサイズ $k$ にクラスタリングされた場合の DM の値を示している．Non-Anonymized は，全く匿名化を実施していない場合，All Suppressed は，完全に秘匿して全く情報を開示しない場合である．

識別性指標 DM は，$k$ の値が大きくなると，DM の値も大きくなる場合が多い．$k = 3$ では，時刻 $[10, 20]$ の範囲で $k = 5, 10$ と同程度の DM 値になっている．これは，匿名移動軌跡の継続性を高めるために，匿名チューブのメンバ数を初期匿名化や合成処理によって大きくしているためだと考えられる．また，全く匿名化を行っていない場合よりも DM が高く，完全に秘匿にした場合よりも DM が低い．各 ideal k の場合よりも DM の値が高いため匿名性は高いと言えるが，より情報の損失は大きいと言える．以上より，匿名化データセットからユーザを特定することは困難になっており，プライバシを保護可能であることが示唆できる．

**5.3.4** 性 能 評 価

提案手法を実装した匿名化エンジンの性能評価のために，匿名化処理に要する処理時間の計測を行った．図 11 に $k$ の値を変更した際の処理時間の計測結果を示す．図 12 にデータセットのサイズ $|D|$ を変更した際の処理時間の計測結果を示す．処理時間の計測は，各計測対象に対して 3 回ずつ行った．各図には，全計測結果の平均値を示している．

図 11 は，$k = 3, 5, 10$，$\theta = 250m$ 四方, $|D| = 1000$ における 1 回の匿名化にかかる処理時間の平均値を示す．

図 12 は，$k = 3$，$\theta = 250m$ 四方, $|D|$ =1000, 2000, 5000, 10000 における処理時間を匿名化処理の構成要素毎に計測した結果の平均値を示す．*Initial* は初期匿名化，*Increment* は差分匿名化，*Split* は分割処理 (移動軌跡識別子の再設定を含む)，*Combine* は合成処理，*Data Store* は匿名化結果の保存処理，*Total* は匿名化処理全体の時間である．

図 11 より，$k$ の値が増加しても，処理時間が増加しないことが分かる．また，図 12 から，データサイズ $|D|$ の増加に線形に比例するような処理時間であることが推測できる．1 万人分のデータセットを 5 秒以内で処理可能なことから，10 万人分のデータセットを 1 分以内に処理できる可能性がある．ただし，差分匿名化はデータセットサイズの増加に伴って急激に処理時間が増加している．既に匿名化したデータと新たに受信したデータとの対応付けに時間を費やしてしまっており，改善すべき点である．また，現在の合成処理は処理時間が非常に短い，これは高速な処理を実現するために，処理対象を限定し，簡略化を行っているためである．故に，トレーサビリティと解像度といった有用性向上のために，合成処理をより高度にすることも可能であることが示唆できる．

以上より，人の移動のようなデータセットに対して，時々刻々とリアルタイムに匿名化を実施可能であると言える．また，$k = 3$ の場合には，40 分程度のトレーサビリティを持った匿名移動軌跡を提供できる．評価に用いたデータセットは，1 分間隔に区切られたデータである．

## 6. ま と め

本稿では，移動軌跡に対して，リアルタイムに $k$ 匿名化する手法 CMOA の提案を行った．評価実験を通して，本提案手法が移動軌跡データセットをリアルタイムかつ連続的に匿名化できることを示した．また，匿名性を保証したまま，ある程度のトレーサビリティを持った匿名移動軌跡を提供できることを示した．ただし，匿名化の度合いを強くすると，トレーサビリティは失われていく．どの程度の匿名性を満たすべきかに関する技術的考察が今後の課題である．評価実験では，いくつかの指標で評価を行ったが，それらの指標による複合的な評価も今後の課題である．加えて，他のデータセットへの適用や，静的な移動軌跡の匿名化手法との比較も今後の課題である．

**謝辞** 本研究の一部は，総務省の委託研究「大規模仮想化サーバ環境における情報セキュリティ対策技術の研究開発」プロジェクトの成果である．

## 文 献

[1] L. Sweeney. k-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), pp. 557-570, 2002.

[2] A. Machanavajjhala, D. Kifer, J. Gehrke and M. Venkitasubramaniam. $\ell$-Diversity: Privacy Beyond k-Anonymity. ACM Transactions on Knowledge Discovery from Data (TKDD), 2007

[3] N. Li, T. Li, and S. Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and $\ell$-diversity. ICDE 2007, pp. 106–115, 2007.

[4] X. Xiao and Y. Tao. m-invariance: towards privacy preserving re-publication of dynamic datasets. SIGMOD2007, pp. 689–700, 2007.

[5] O. Abul, F. Bonchi, and M. Nanni. *Never Walk Alone* : Uncertainty for Anonymity in Moving Objects Databases. ICDE2008, pp. 376–385. 2008.

[6] M. E. Nergiz, M. Atozori, Y. Saygin and B. Güç. Towards Trajectory Anonymization: a Generalization-Based Approach. Transactions on Data Privacy, 2, pp. 47-75, 2009.

[7] A. Monreale, G. Andrienko, N. Andrienko, F. Giannotti, D. Pedreschi, S. Rinzivillo, and S. Wrobel. Movement Data Anonymity through Generalization. Transactions on Data Privacy, 3, pp. 91-121, 2010.

[8] D. Arthur, and S. Vassilvitskii. k-means++: the advantages of careful seeding. SIAM2007, pp. 1027-1035, 2007.

[9] R. Bayardo and R. Agrawal. Data privacy through optimal k-anonymity. ICDE2005, 2005.

[10] 総務省統計局．地域メッシュ統計．`http://www.stat.go.jp/data/mesh/`

[11] NEC Europe. Siafu, An Open Source Context Simulator. `http://siafusimulator.sourceforge.net/`