軌跡データ部分公開によるプライバシリスクに関する研究

小久保彰博† 吉川 正俊†

† 京都大学大学院情報学研究科 〒 606-8501 京都市左京区吉田本町 36-1
E-mail: †kokubo@db.soc.i.kyoto-u.ac.jp, ††yoshikawa@i.kyoto-u.ac.jp

あらまし 近年では携帯電話やスマートフォンの普及と共に，GPS 機能を利用した位置情報サービスは我々の生活にかかせないものとなっており，これらのサービスは LBS(Location-based Service) と呼ばれる．LBS によって収集された軌跡データは有用性が高い一方で個人のプライバシを強く反映したデータであるため，LPPM(Location privacy-preserving mechanisms) よって匿名化加工を施す研究がなされている．LBS ユーザには，それぞれ保護したい情報やその保護したい強度などプライバシ選好があると考えられるが，我々が知る限り既存の LPPM はユーザのプライバシ選好を考慮していない．そこで本研究では，プライバシ選好を考慮した LPPM 提案に先駆けて，プライバシ選好を $Preference$ と定義した．また，攻撃者の $Preference$ に関する知識を確率分布によって定義し，攻撃者の知識の誤差を評価する誤差関数も定義した．これらを実際的な状況に沿って適用することを目的として実験を行った．

キーワード 軌跡データ, プライバシリスク

1 はじめに

1.1 軌跡データ

軌跡データは，位置情報の系列である．GPS 機能を搭載したモバイル端末（携帯電話，スマートフォン）やカーナビゲーション装置がある時刻ごとにその位置情報を記録したものを時系列順に並べたものは，その端末を持ったユーザやカーナビゲーション装置を搭載した車の移動した軌跡を表すデータとなる．これが軌跡データである．

近年では携帯電話やスマートフォンの普及と共に，GPS 機能を利用したサービスも我々の生活では広く一般的なものとなっている．例えば，近くのレストランを検索できたり，そのレストランまでの道のりを検索できたり，道順をリアルタイムで提供してくれる Google マップや，友達同士の位置情報をリアルタイムで共有するサービスである Zenly などのアプリケーションがある．これらのサービスはメインの機能に位置情報を利用するため，LBS(Location-based Service) と呼ばれる．

LBS によって収集されたユーザの位置情報のデータの集合すなわち軌跡データは社会的に応用性・有用性の高いデータである．例えば，都市計画，ビジネス分析，輸送最適化，疫学分析，健康保険などの用途があげられる [6].

一方で，軌跡データは個人のプライバシを強く反映したデータであるため，簡単に公開することはできないデータである．軌跡データそのものを公開せず，統計データを公開したとしても個人を特定できてしまうという研究がある [7].

また，近年では個人のプライバシ保護への関心は高まっており，位置情報に関するプライバシ意識は特に高いという調査結果もある [8]．ヨーロッパでは GDPR(General Data Protection Regulation：一般データ保護規則) [9] が 2016 年 4 月に制定され，2018 年 5 月に施行された．

1.2 軌跡データのプライバシ保護

LBS プロバイダが LBS によって入手したユーザの軌跡データを第三者に提供する場合，なんらかの方法でユーザが特定されないようにデータを匿名加工しなければならない．例えば，1 人のユーザについてランダムな文字列を id を振り直す（仮名化）方法が考えられるが，これは十分な加工ではない．ユーザのプライバシを保護するためには，仮名化をした上でさらにデータを加工する必要がある．

ユーザのプライバシ保護のために軌跡データを加工するタイミングは 2 つ考えられる．ユーザが端末から LBS に対して現在の位置情報を送るタイミングと，LBS プロバイダが第三者に軌跡データベースを提供するタイミイングである．端末でユーザの位置情報に何らかの匿名加工を施すメカニズムは LPPM(Location privacy-preserving mechanisms) と呼ばれる．また，LBS プロバイダーが持つ軌跡データベースを何らかの匿名加工をして第三者に公開することは PPDP(Privacy Preserving Data Publishing) と呼ばれる．

LPPM を用いて匿名化することは，LBS プロバイダーが信頼できない可能性を考慮している点で，より強力な仮定を置いており，ユーザの端末側でデータの匿名化加工をする方が適切であると言える．

1.3 プライバシの要件

プライバシの保護の目標をどのように設計するかという問題は今のところ普遍的な正解がない問題である．すなわち，このプライバシ基準を満たせば軌跡データを公開してもユーザのプライバシを保護しているといえるというプライバシ基準は未だ確立されていない，研究されるべき領域である．

どの程度自分のプライバシを保護したいかというのは，各ユーザによって様々であり，各ユーザが自分の好みによって設定できるようにするべきである．また，同じように自分の軌跡データのうちどの部分を保護したいかは各ユーザによって様々

である．すなわち，自宅の位置情報を知られたくないと考えるユーザもいれば，自宅の位置情報を知られるのは構わないが，病院に通院しているのは知られたくないと考えるユーザもいる．

ところが，現在ユーザがプライバシ保護の程度を指定できるような UI は見られない．例えば，iPhone では，各アプリケーションについて，位置情報を利用するかしないかの二つしか選択肢がない．

ユーザごとに保護したいプライバシは様々であると考えられ，それらを保護できるように軌跡データを加工する必要がある．現在軌跡データのプライバシ保護でなされる研究では，軌跡データそのものがどれだけ推測されているかという観点からプライバシを保護の程度を評価する傾向があるが，その視点だけでは足りないと考えられる．

1.4 論文の貢献

この論文では，ユーザの保護したいと考える軌跡データの部分を一般的に定義できる選好 (preference) を導入し，それがどの程度攻撃者に推測されているかという尺度でプライバシリスクを定量化する．これによって多岐にわたるユーザの秘匿したい情報を一つの尺度でプライバシ保護の程度を測定することが可能となる．

2 関 連 研 究

既存研究の全体像は表 1 のようにまとめることができる．

表 1　PPDP の研究と LPPM の研究の比較

	PPDP の研究	LPPM の研究
注目するプライバシ	軌跡データのプライバシ	位置データのプライバシ
データ量	大きい	小さい
生データの保存場所	サーバー	クライアント
プライバシ保護の処理	オフライン	オンライン
想定している攻撃者	データを利用する第三者	サービスプロバイダ
プライバシ保護するタイミング	データベースを提供するとき	LBS にリクエストを送るとき

2.1 プライバシ保護型データ出版 (privacy preserving data publishing) の研究

LBS プロバイダを信頼し，LBS が提供するデータを利用する第三者を信頼しないというシナリオの元，軌跡データベースをプライバシを保護しつつ公開するプライバシ保護型データ出版 (privacy preserving data publishing) の研究は盛んに行われている．この分野の研究を Fiore らは調査し結果をまとめている [1]．ここでの軌跡データのデータベースとは，識別子 (名前や電話番号)，日時，位置情報 (緯度・経度)，属性 (性別・職種・収入) などのタプルからなるデータベースのことである．プライバシ保護型データ出版について，攻撃と匿名化の二つの観点で研究を分類している．

まず，公開された軌跡データベースへの攻撃に関する研究についてであるが，文献 [1] は以下の 3 つの側面から研究を分類している．

- 攻撃者の目的 (レコード連結・属性連結・確率)
- サイド情報の形式 (時空間点・特徴量・補足情報)
- サイド情報のソース (内部情報，別のデータベース)

次に，公開された軌跡データベースの匿名化に関する研究についてであるが，文献 [1] は以下の 3 つのプライバシ原理を用いて分類している．

- プライバシ原理なし
- 識別不可能性
- 情報無価値性

プライバシ原理なしとは，曖昧化という手法を用いて，真の位置情報に何かしらノイズを加えることでプライバシリスクを軽減するも，特に厳密なプライバシの定義はなく，結果としてプライバシという面では何も保証していない手法であることを示す．具体的な手法に mix-zones [10] [11] がある．識別不能性とは複数の軌跡データ同士で区別できない性質のことで，プライバシ基準を k-匿名性として，プライバシ保護を実現する方法がある．具体的には，軌跡データの時刻や位置情報の粒度を大きくすることによって k-匿名性を満たすようにする手法がある．情報無価値性とは，ある情報を受け取った前後で知識が変化しないという性質で，プライバシ基準を差分プライバシ [12] とする手法がある．差分プライバシは「位置データプライバシ」の点では情報無価値性をうまく担保できる．すなわち，ある一瞬一瞬の位置情報のプライバシは守れる．その一方で，軌跡データのプライバシを守るのは今の所困難な状況にある．最近の傾向として，軌跡データにノイズを加える手法から，データの特徴を捉えたままプライバシを保護して新しいデータを生成するという手法が多くなっている．

このサーベイによる PPDP への匿名化手法への結論は以下にまとめられる．軌跡データの匿名化は非常に複雑で困難な問題である．プライバシ原理なしの曖昧化手法 (mix-zones など) は単純に機能していない．時空間的な解像度を下げても，軌跡を短くしたり絡め合わせても，攻撃のリスクは未だに残っており，複雑な手法も完璧には程遠い．識別不可能性の k-匿名化技術は，かなり成熟しており，匿名加工したデータも十分な有用性を持っている一方で，k をスケールできていない．情報無価値性の差分プライバシやその拡張は，軌跡データに適用するのが非常に難しい．情報無価値性という原則に則った解決法では，差分プライベートなモデルを生成し，人工的な軌跡を作るという手法をとるものが多い．この手法により匿名化したデータが全体的な特徴しか維持していないことは明らかで，実在の個人を追跡することで可能な分析はしにくくなっている．すなわち，モデルを作る途中で失った特徴が，出力するデータベースに反映されているという保証はない．

差分プライバシを用いて軌跡データベースを生成する研究は近年盛んに行われている．Chen らはモントリオールの交通系 IC カードシステムのデータを差分プライバシモデルを使って公開するための手法を提案した [2]．hybrid-granularity prefix tree 構造をもとにして，データに有用性を残しつつ，差分プライベートな交通データに加工する．前処理として prefix tree に固有の性質を用いて特徴を推論することで，データの有用性を高める．差分プライバシを保証して，大量の系列データを公開する実用的な方法をはじめて提案したものである．

同じく Chen らは，Differential Privacy を sequential data

に適用するのが難しい原因が，データに潜在的な連続性と高次元という性質が含まれるためということを踏まえ上記の問題に取り組んでいる [3]．具体的には，可変長 n-gram によって，系列データベースの本質的な情報を抽出し，探索木構造と Markov assumption をもとにした斬新な手法を用いて，加えるノイズを小さくしている．

He らは，ε-差分プライバシを保証する DPT という軌跡生成システムを提案している [4]．

Gursoy らは，DP-Star というモデルを使い生成した軌跡データベースの有用性をさらに改善している [5]．

3 問題設定

3.1 前　　提

本研究では，より強力な仮定である LBS が攻撃者になりうると仮定をおき，実際の状況では，1 人のユーザは複数の LBS を利用しているかもしれないが，ここでは 1 人のユーザと一つの LBS に注目する．ユーザが複数の LBS を利用していても，端末側が各 LBS について，それぞれユーザの守りたい部分を保護するように動作すれば良いだけである．さらに，端末はユーザの軌跡データを常時記録しているものとする．すなわち，ユーザの生の軌跡データは全て端末側に保存されており，端末は，そのデータを踏まえて各 LBS に位置情報を提供するごとにどれほどノイズをかけて出すのか，それともそもそも提供しないのかを判断して出すような状況を前提とする．

3.2 プライバシ保護の選好

軌跡データのプライバシ保護の研究分野では，ユーザごとに保護したい部分は様々でかつ，保護したい部分ごとに保護したい強度も異なる可能性があるという点はあまり考慮されてこなかった．ユーザごとに保護したい部分や強度が異なるとは，例えば,「自宅の場所は守りたいと思っているが，旅行先の移動経路はあまり守りたいと思っていない.」や,「自宅の場所は強さ 1 で守りたいと思っているが，通院先の病院の場所は強さ 2 で守りたい」などである．これまでの研究ではプライバシは k-匿名化や差分プライバシなど厳格な基準で正解データを数学的な手法で保護していたものの，軌跡データ中の重要な部分（ユーザが強く保護したい部分）と重要でない部分（ユーザは保護したいと思っていない部分）を考慮せずに軌跡データの任意の部分を同じ重要度で保護している．本研究では，実際の状況では各ユーザごとに保護したいと思っている情報やその強度は異なり，それに合わせたプライバシ保護が必要であるという観点から，ユーザの保護したい部分をプライバシ選好とし，それを定義する．

以下でユーザの軌跡を保護するとはどういうことかを考えるためにユーザのプライバシ選好の具体例をあげる．

3.3 プライバシ選好の具体例

具体的にユーザが LBS に対して非公開にしたい情報の例を考える．以下の 4 つの場合についてそれぞれ考える．

（1） ある日のお昼にあるレストランに行ったことを隠したい

（2） 病院に通院したことは全て隠したい

（3） 家の場所を隠したい

1 の場合　ある日のお昼にあるレストランに行ったことを隠したい

まず，最初の例を考える．これはすなわち，ある時刻 t_1（ある日のお昼）において，ある場所 l_1（レストラン）に存在したということを非公開にしたいというユーザのニーズである．攻撃者はこれまでユーザが LBS を使うことによって得たユーザの位置情報をもとにユーザに関する知識（例えば時刻 t_1 でユーザはどこにいるのか）を獲得する．この知識は，時刻 t_1 にユーザは A 地点に 20 %，B 地点に 30 %，C 地点に 50 %で存在するであろうという場所に関する確率分布と考えることができる．すなわち，攻撃者はあるユーザについて時刻 t ごとに，地図上の各地点についてその地点にいる確率をそれぞれ持っているとできる．この攻撃者の知識を $P^{(\alpha)}(t)$ とする．このとき，ユーザのプライバシリスクは，時刻 t でのユーザの真の位置と攻撃者の知識である確率分布との距離と逆相関の関係にある．今の場合で言えば，時刻 t_1 におけるユーザの真の位置 l_1 と攻撃者のユーザに関する知識すなわち場所に関する確率分布 $P^{(\alpha)}(t_1)$ との距離が大きければ大きいほど攻撃者はユーザの真の位置を分かっておらずプライバシリスクは低く，距離が小さいほど攻撃者はユーザの真の位置を正確に分かっており，プライバシリスクは高いと言える．例えば，攻撃者が確率 1 でユーザは時刻 t_1 に場所 l_1 にいると断言した場合，ユーザのプライバシは完全に露呈していると言える．さて，ユーザの保護したい情報のことを $preference$ とすると，この時のユーザの $preference$ は「時刻 t_1 において場所 l_1 にいること」となる．

2 の場合　病院に通院したことは全て隠したい

病院の場所を l_2 とする．どんな時刻についても l_2 にいたことを攻撃者に推測されることを防ぎたいというユーザのニーズであることから，このときのユーザの $preference$ は,「任意の時刻 t について，場所 l_2 にいること」となる．

3 の場合　家の場所を隠したい

家の場所を l_3 とする．LBS 上で自分の家の位置情報を家というラベルと共に登録しない限りは，攻撃者はユーザの家を家と断定することはできない．しかし，毎日夜から朝（一般的に人が家にいる時間帯）にかけて l_3 にユーザがいることがわかれば，毎日夜から朝にかけている場所は家に違いないという推論のもと，攻撃者はユーザの家の位置情報を特定することが可能となる．病院との違いは，病院に行ったことを隠すには一度でも病院に訪れたことを推測されればプライバシを保護できていないと言えるのに対し，家の場所を隠すというのは家にいるような時間帯に大抵家にいるということを推測されればプライバシを保護できていないと言えるところである．すなわち，一度だけユーザが家にいることを攻撃者に公開したとしても，攻撃者はその位置情報がユーザの家であるということは言えないのに対し，一度だけ病院にいることを攻撃者に公開することは，ユーザが病院にいたことを攻撃者に知らせることになる．さて，ユーザの保護したい情報のことを $preference$ とし，一般的に人が家にいる時間帯（平日の夜から朝にかけてと土日祝日）を

T とすると，このときのユーザの $preference$ は，「時刻 $t \in T$ において，ユーザの分布が l_3 であること」となる．

4 プライバシ選好

以下ではユーザの守りたい情報である $Preference$ を定義する．

4.1 仮　　定

時間 T は本来ならば連続的な集合であるが，ここでは時間 T は離散な時刻集合 $T = \{t_1, t_2, ...\}$ であるとする．t_i と t_{i+1} の間隔は i によらず一定である．また，ユーザ側（端末）と攻撃者側（LBS）は時間 T を共有しているものと仮定する．

4.2 定　　義

［定義 1］ **地図**
地図を m 個の節点の集合 V とそれらを結ぶ枝の集合 E からなるグラフ $G(V, E)$ で表現するとする．このとき，地図上の位置は $v_i \in V (i \in [1, m])$ と表せる．

［定義 2］ **軌跡**
ユーザの軌跡を l として，時刻 t における位置情報を $l(t)$ とする．$l(t) \in V$ であるため，以下の式が成り立つ．

$$\forall t, \exists i \text{ s.t. } v_i = l(t) \tag{1}$$

［定義 3］ **位置情報の確率分布**
位置情報の確率分布 P を以下のように定義する．

$$P = (p_1, p_2, ..., p_m) \tag{2}$$

$$\sum_{i=1}^{m} p_i = 1 \tag{3}$$

ただし，p_i は v_i である確率．

［定義 4］ **ユーザの位置情報の確率分布**
ユーザの位置情報の確率分布を $P^{(u)}$ として，時刻 t におけるユーザの位置情報の確率分布を $P^{(u)}(t)$ とする．このとき，$l(t) = v_k$ とすると $P^{(u)}(t)$ は以下のように表せる．

$$P^{(u)}(t) = (p_1^{(u)}, p_2^{(u)}, ..., p_k^{(u)}, ..., p_m^{(u)}) \tag{4}$$

$$p_i^{(u)} = \begin{cases} 1 & (i = k) \\ 0 & (i \neq k) \end{cases} \tag{5}$$

［定義 5］ **Preference**
ユーザが保護したい情報をプライバシ選好とする．

$$Preference = \{preference_1, preference_2, ..., preference_n\} \tag{6}$$

$$preference_i = (V_i, T_i, MinimumError_i) \tag{7}$$

ただし，V_i は V の部分集合で $V_i \in V$，T_i は T の部分集合で $T_i \in T$ であるとする．$MinimumError_i$ とは，攻撃者の $preference_i$ に関する知識の誤りの下限であり，ユーザが $preference_i$ についてどれほどの強さで守りたいかを表す．

［定義 6］ **Preference に関するユーザの確率分布**
$preference_i$ に関するユーザの確率分布を $P^{(u)}(V_i, T_i)$ とする．$t \in T_i$ かつ $l(t) \in V_i$ なる t の集合を T'，集合 T' のサイズを $|T'| = k$ とする．このとき，

$$P^{(u)}(V_i, T_i) = \frac{1}{k} \sum_{t \in T'} P^{(u)}(t) \tag{8}$$

［定義 7］ **攻撃者の Preference に関する知識**
攻撃者の $preference_i$ に関する知識を位置情報の確率分布によって定義する．まず，ユーザの時刻 t における位置情報に関する攻撃者の知識を $P^{(a)}(t)$ とすると，$P^{(a)}(t)$ は以下のように表すことができる．

$$P^{(a)}(t) = \{p_1^{(a)}, p_2^{(a)}, ..., p_m^{(a)}\} \tag{9}$$

$$\sum_{i=1}^{m} p_i^{(a)} = 1 \tag{10}$$

ただし，$p_i^{(a)}$ は攻撃者が時刻 t においてユーザが地点 v_i に存在すると考える確率である．このとき，$preference_i = (V_i, T_i)$ に関する攻撃者の知識 $P^{(a)}(V_i, T_i)$ は，$t \in T_i$ かつ $l(t) \in V_i$ なる t の集合を T' として，$k = |T'|$ とすると，以下のように定義できる．

$$P^{(a)}(V_i, T_i) = \frac{1}{k} \sum_{t \in T'} P^{(a)}(t) \tag{11}$$

［定義 8］ **位置情報と位置情報の確率分布の距離**
位置情報 v と位置情報の確率分布 P の距離 $error(v, P)$ は以下のように定義できる．ただし，$distance(v, v_i)$ はノード v と v_i とのグラフ G 上の最短距離であるとする．

$$error(v, P) = \sum_{i=1}^{m} distance(v, v_i) p_i \tag{12}$$

これを用いて，例えば時刻 t におけるユーザの位置 $l(t)$ に関して，攻撃者の知識を $P^{(a)}(t)$ とすれば，その距離 $error(l(t), P^{(a)}(t))$ は以下のようにできる．

$$error(l(t), P^{(a)}(t)) = \sum_{i=1}^{m} distance(l(t), v_i) p_i^{(a)}(t) \tag{13}$$

［定義 9］ **位置情報の確率分布の距離**
位置情報の確率分布 P と P' の距離 $error(P, P')$ は以下のように定義できる．

$$error(P, P') = \sum_{i=1}^{m} \sum_{j=1}^{m} distance(v_i, v_j) p_i p'_j \tag{14}$$

これを用いて，時刻 t におけるユーザの位置情報の確率分布を $P^{(u)}(t)$，それに関する攻撃者の知識を $P^{(a)}(t)$ とすると，その距離 $error(P^{(u)}(t), P^{(a)}(t))$ は以下のようにできる．

$$error(P^{(u)}(t), P^{(a)}(t)) = \sum_{i=1}^{m} \sum_{j=1}^{m} distance(v_i, v_j) p_i^{(u)}(t) p_j^{(a)}(t) \tag{15}$$

5 実　　験

5.1 実験の目的

上記で定めたユーザの $Preference$ に対する攻撃者の知識を具体的に推定し，その誤差を実際の地図上で具体的な状況に沿って測定することを目的に実験を行う．具体的には，ある範囲のユーザの軌跡中の点を公開した時の，別の点のプライバシリスクを実際の地図データを用いて測定する．軌跡データは本来時刻の増加にともなって永遠に続く系列のデータであるが，ここでは特定の範囲を切り取りその範囲のみを考慮して実験を行う．また，具体的な error 関数を複数試すことによって，それぞれの error 関数の性質やグラフの特徴による影響を検証する．

5.2 実験の概要

ごく単純な設定により実験を行う．一つの LBS と 1 人のユーザがいるとする．ユーザはスタート地点からゴール地点までの経路を移動する．この経路はスタート地点からゴール地点までの最短経路であるとする．時刻 $t = t_s$ にスタート地点を出発したことは攻撃者（LBS）は知っているとする．ユーザは時刻 $t = t_g$ にゴール地点にいたことを隠したいと思っているとする．この時スタート地点からゴール地点までの経路中のある一つの中間地点を攻撃者（LBS）に公開することが，どれほどゴール地点のプライバシに影響するかを測定する．すなわち，ユーザの軌跡データを $l(t)$，$Preference$ を $\{preference_1\}, preference_1 = \{S_1 = \{v_g\}, T_1 = \{t_g\}\}$ とすると，時刻 t_s から時刻 t_g の範囲の経路 $l(t)$ に注目して，攻撃者に $l(t_s)$ と $l(t_m)$ を公開したときの，ユーザの $Preference$ に関するプライバシリスクを検証するために，$error(l(t_g), P^{(a)}(t_g))$ を測定する．ただし，error 関数に関しては定義しておらず，今回の実験で複数の選択肢を比較する．また，error 関数がグラフの性質によってどのような影響を受けるのかを検証するために複数の都市について実験を行う．

ここで実験の設定を単純にするために，いくつか仮定を置く．

仮定 1　ユーザは等速で移動しており，その速さは時速 4km/h である．

仮定 2　ユーザはスタート地点からゴール地点まで最短経路上を移動する．

仮定 3　攻撃者は，ユーザが等速で移動していると想定するとする．

仮定 4　攻撃者は，ユーザの移動が最短経路であると想定するとする．

仮定 5　ユーザの移動は 30 分であるとする（$t_g - t_s = 30$ 分）．

また，攻撃者はユーザの 2 つの情報 $l(t_s)$ と $l(t_m)$ を得ることによって，時刻 $t = t_g$ でのユーザの位置 $l(t_g)$ を推測するが，攻撃者が想定するユーザの移動について 2 つのケースを考え実験を行う．すなわち，中間地点以降もユーザが移動し続けると攻撃者が想定する場合と，中間地点以降にユーザが移動をやめることもあると想定する場合の二つの場合をそれぞれ実験する．

場合 1　中間地点以降もユーザが移動し続けると攻撃者は想定する

場合 2　中間地点以降にユーザが移動をやめることもあると攻撃者は想定する

それぞれに場合についても，攻撃者は，ユーザのスタート地点と中間地点の位置情報と時刻からユーザの移動速度を推定する．中間地点 v_m を得ることによって，時刻 $t = t_g$ でユーザが存在すると攻撃者が推測するノードの集合を V_m とし，攻撃者は V_m 中の各ノードについて等確率でユーザが存在すると推測するとする．よって V_m に含まれるノードの数を k とすると，攻撃者がもつユーザの $Preference$ に関する知識，すなわち，時刻 $t = t_g$ におけるユーザの位置情報の確率分布は，V_m に含まれるノードは確率 $1/k$ でそのほかのノードは確率 0 になるような位置情報の確率分布となる．

さて，攻撃者が想定するユーザの移動について，中間地点以降もユーザが移動し続けると攻撃者が想定する場合のときの V_m は，以下のようになる．攻撃者はユーザの経路のスタート地点と中間地点を得ることによってユーザの速度を計算し，それ以降も等速で移動すると想定することから，V_m に含まれるノードはスタート地点からの最短経路長が 2km の地点である．また，攻撃者はユーザの移動が最短経路であると想定することから，スタート地点と V_m に含まれるノードまでの最短経路には中間地点が含まれる．よって，V_m は，スタート地点からの最短経路に中間地点を含みかつ，その経路長が 2km であるようなノードの集合となる．

次に，攻撃者が想定するユーザの移動について，中間地点以降にユーザが移動をやめることもあると攻撃者が想定する場合の V_m は，以下のようになる．攻撃者はユーザの経路のスタート地点と中間地点を得ることによってユーザの速度を計算し，それ以降も等速で移動するまたは途中で移動をやめると想定することから，V_m に含まれるノードはスタート地点からの最短経路長が 2km 以内の地点である．また，攻撃者はユーザの移動が最短経路であると想定することから，スタート地点と V_m に含まれるノードまでの最短経路には中間地点が含まれる．よって，V_m は，スタート地点からの最短経路に中間地点を含みかつ，その経路長が 2km 以内のノードの集合である．

5.3 実験の方法

以下が大まかな実験の手順である．

（1） 適切な範囲の地図データを用意

（2） ユーザの経路を生成

（3） 軌跡データ中の各点を中間点としたとして，以下を繰り返す

（4） 攻撃（攻撃者によるゴール地点の推測）

（5） 評価（攻撃者のゴール地点に関する知識の $error$ を計測）

以下で，経路の生成，攻撃，評価について説明する．

5.3.1 経路の生成

本来ならば軌跡データすなわち位置情報と時刻のタプルの系列をもとに実験を行う想定であるが，仮定により，長さ 2km の経路すなわち位置情報の系列があれば今回の実験は行うことが

できる．よって，用意した地図データの中心をスタート地点とし，スタート地点からの最短経路長が 2km であるようなノードを探索し，条件に合うノードをゴール地点とすることで経路を生成して，実験に用いる．

5.3.2 攻　　撃

スタート地点，中間地点から，攻撃者が時刻 $t = t_g$ でユーザが存在すると推測するノードの集合 V_m を求める．V_m は，中間地点以降もユーザが移動し続けると想定する場合（場合 1）は，スタート地点からの最短経路に中間地点を含みかつその経路長が 2km であるようなノードの集合であり，中間地点以降にユーザは移動をやめることもあると想定する場合（場合 2）は，スタート地点からの最短経路に中間地点を含みかつ，その経路長が 2km 以内のノードの集合である．場合 1, 場合 2 のそれぞれのときの攻撃を attack1, attack2 とする．

さて，地図データは道路ネットワークのグラフデータであり，このグラフの各ノードは道路の交差点を意味する．このとき，必ずしもスタート地点からの最短経路長が 2km のところにノードが存在するとは限らない．よって attack1 について，スタート地点からの最短経路に中間地点を含みかつその経路長が 1.8km から 2km であるようなノードの集合を V_m であるとする．

5.3.3 評　　価

攻撃者の知識（時刻 $t = t_g$ でのユーザの位置を推定した確率分布）とゴール地点（時刻 $t = t_g$ での実際のユーザの位置）との $error$ を測定する．

また，error 関数に加えて，総延長距離（TRL: Total Road Length）という概念により攻撃者の知識とユーザの真の位置との誤差を測定することも行う．総延長距離とは，道路ネットワーク上のあるエリアに対して定義される距離であり、エリア内の全ての道路の長さの和である。すなわち、グラフ $G(V, E)$ と重み関数 $w : E \to \mathbb{R}$ について，グラフ G の部分グラフ $G'(V', E')$ の総延長距離とは，E' に含まれる枝 e の重みの総和である．

［定義 10］ **TRL(Total Road Length)**

グラフ G(V, E)，枝の重み関数 $w : E \to \mathbb{R}$ について，グラフ G の部分グラフ $G'(V', E')$ の総延長距離（TRL）を以下のように定義する．

$$TRL(G') = \sum_{e \in E'} w(e) \qquad (16)$$

今回の実験では攻撃者はユーザのいそうなノード集合を推測するにとどまっているため，攻撃者はユーザのいそうなエリアを推測しているとみなすこともできる．このとき，ユーザのいる場所はノードとは限らずグラフの枝上にいることもあるため，攻撃者が推測するエリアの総延長距離が長いほどユーザがどこにいるか分かっていないとみなせる．

5.4 デ　ー　タ

本論文はグラフ上で $Preference$ とそれに関する攻撃者の知識の $error$ を定義している．したがって公開されている道路ネットワークとして，Open Street Map Japan[1] より，図 1〜3 に示すような道路データを取得した．

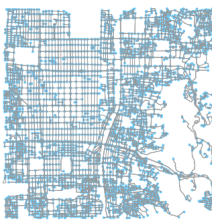

図 1　京　　都

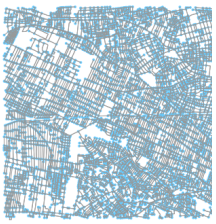

図 2　東　　京

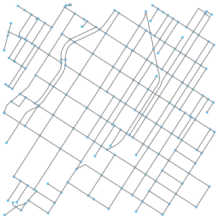

図 3　北　海　道

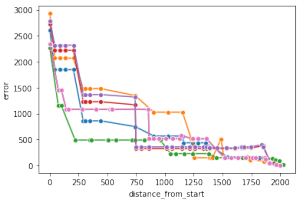

図 4　場合 1・京都での error の変化

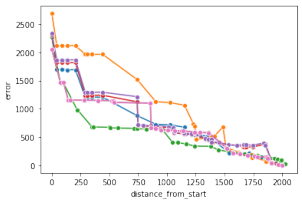

図 5　場合 2・京都での error の変化

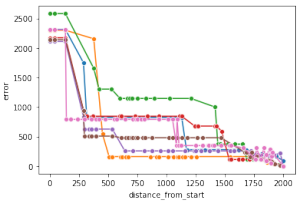

図 6　場合 1・東京での error の変化

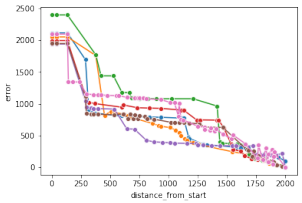

図 7　場合 2・東京での error の変化

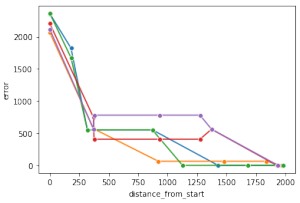

図 8　場合 1・北海道での error の変化

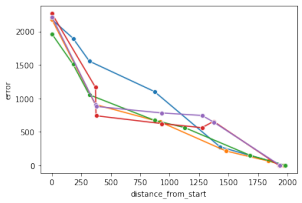

図 9　場合 2・北海道での error の変化

5.5 結　　果

京都，東京，北海道の三箇所について同様の実験を行った．それぞれの場所での 4.8km 四方の地図は図 1, 図 2, 図 3 のようになった．三箇所について場合 1, 場合 2 としたときの，複数の経路で中間地点をスタート地点からゴール地点まで変化させたときの error と TRL の変化の図を以下に示す．各図について，横軸の distance_from_start とは，中間地点 v_m とスタート地点との距離（最短経路長）で単位はメートル，同じく縦軸も単位はメートルである．

図 4，図 6，図 8 に注目する．初期値が京都，東京，北海道の順で大きいという違いはあるが，数値，グラフ全体の傾向ともに大きな違いはない．３つともについて，中間地点のスター

1 : https://openstreetmap.jp/

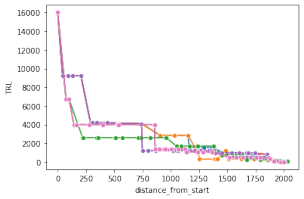

図 10　場合 1・京都での TRL の変化

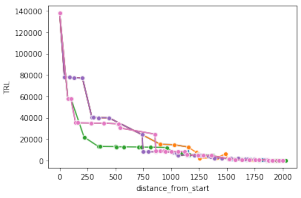

図 11　場合 2・京都での TRL の変化

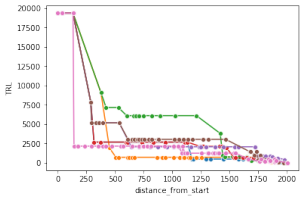

図 12　場合 1・東京での TRL の変化

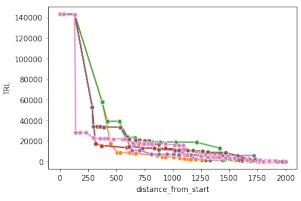

図 13　場合 2・東京での TRL の変化

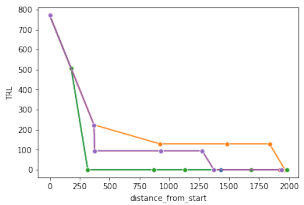

図 14　場合 1・北海道での TRL の変化

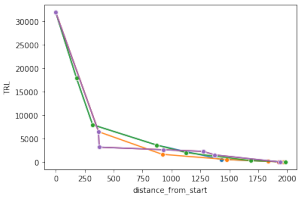

図 15　場合 2・北海道での TRL の変化

ト地点からの距離が 250 メートルを超えたあたりで大きく下がるという特徴がある．また，*error* の値の変化はゆるやかではなく，あるノードからノードへの小さい移動が大きく *error* の値を下げる点が各グラフ，各経路に見られる．

図 5，図 7，図 9 に注目する．初期値についても，中間地点のスタート地点からの距離が大きくなるにつれて *error* が小さくなる様子も，京都，東京，北海道それぞれについて大きな差はみられない．これらについても，*error* の値の変化はゆるやかではなく，あるノードから次のノードへの小さい移動が大きく *error* の値を下げる変化が各経路に見られる．

次に，同じ場所での場合を変えたときの *error* の変化に注目する．すなわち，図 4 と図 5，図 6 と図 7，図 8 と図 9 を比較する．全体的に場合 2 の方がスタート地点から中間地点までの距離が大きくなるにしたがって減る *error* の値はゆるやかである．

図 10，図 12，図 14 に注目する．まず京都・東京と北海道では TRL の大きさが大きく異なり，京都・東京での TRL は北海道での TRL のおよそ 20 倍ほど大きい．グラフの傾向として，中間地点のスタート地点からの距離が 250 メートルを超えたあたりから TRL の値が大きく下がるという特徴は同じである．また，中間地点のスタート地点からの距離が大きくなっているのにも関わらず TRL が変化しないという点は，格場所・各ルートにみられる．

図 11，図 13，図 15 に注目する．場合 1 のときと同様に，京都・東京と北海道では TRl の大きさが大きく異なる．一方でグラフの傾向として中間地点のスタート地点からの距離が 250 メートルを超えたあたりから大きく TRL の値が下がる点は共通している．加えて，中間地点のスタート地点からの距離が大きくなっているのにも関わらず TRL が変化しない特徴が各ルートにみられる．

次に，同じ場所での場合を変えたときの *error* の変化に注目する．すなわち，図 4 と図 5，図 6 と図 7，図 8 と図 9 を比較する．全体的に場合 2 の方がスタート地点から中間地点までの距離が大きくなるにしたがって減る *error* の値はゆるやかである．

5.6　考　　察

実験の設定から，中間地点がよりスタート地点に近いすなわち隠したいと思っているゴール地点から遠いほど攻撃者の知識に誤りが大きく *error* は大きくなりプライバシは守られるであろうということは定性的に明らかである．注目すべきは，中間地点がよりゴール地点に近くなっているのにも関わらず *error* が全く変化しないことがあるという点であり，必ずしも *error* の値は distance_from_start に関して滑らかに変化する訳ではないということ．例えば，図 4 のピンク色の経路に注目すると distance_from_start が 175 メートルあたりから 850 メートルあたりまで *error* の値は常に 1000 メートルほどであり，これはスタートからの距離が 175 メートルの中間地点を攻撃者に公開することは，スタートからの距離が 850 メートルの中間地点を攻撃者に公開することと同じ程度のプライバシリスクがあることを意味する．これについては TRL についても同様の特徴がみられる．

次に注目すべき点は，各場所について *error* の値に大きな変化がみられない点である．京都・東京は北海道に比べ，同じ広さのエリア内のノード数が大幅に大きいのにも関わらず，*error* の値がほとんど同じ程度の大きさで変化することは，図 4，図 6，図 8 および，図 5，図 7，図 9 から分かる．これは，例えば半径 1 の円があり，ユーザの真の位置が円の中心，攻撃者はこの円の円周上のいくつかの点にユーザがおり，ノードごとの確率は等しいと推測するとき，そのノードの数に寄らず *error* は常に 1 となることから，ノードが増えたとしてもノードごとの重みが小さくなることで全体の *error* はあまり変わらないということから説明がつく．一方で，これは直感に反する結果でもある．すなわち，V_m の範囲の面積が同じくらいであったとしても，そのエリア内のノードの個数や枝の総延長距離が大きい方が攻撃者はユーザの位置を特定しにくく，よりプライバシリスクは小さいすなわち *error* は大きくなるであろうと考えられるものの，現在の *error* の定義では，同じくらいのプライバシリスクであると言える．

6　今後の課題

今後の課題について，以下にまとめる．

- error 関数の再検討

現在の *error* の定義は，二次元平面上での定義としては上手く機能している一方で，実際の地図を考慮できていないと言える

ことは，今回の実験からわかった．すなわち，北海道と東京など，同じ面積のエリア内の総延長距離が大きく異なる二つの場所で，似たような *error* の値を得るのは，*error* 関数が適切でないからと言える．

- 既存のメカニズム（LPPM）の $Preference$ の観点からの分析

すでにいくつか提案されている LPPM（Location privacy-preserving mechanisms）がユーザの $Preference$ をどれほど保護しているかを検証し，それが十分なプライバシ保護を達成しているのかを分析する必要がある．同時に，どれほどの保護が実用上適切なのかも検討する必要がある．

- ユーザの $Preference$ を保護するメカニズムの提案

端末側でユーザの $Preference$ を LBS の使用に応じて動的に保護し，位置情報を公開，匿名加工を行うメカニズムを提案することは今後の課題である．

7 おわりに

本研究では，LBS を使用するユーザの実際的な状況やプライバシニーズに即して，ユーザが守りたい情報を $Preference$ として定義した．これは，ユーザの保護したい情報すなわちプライバシ選好は様々で，個々のユーザに即したプライバシ保護が必要であるという視点を背景としている．加えて，本研究で定義した $Preference$ がユーザの様々なプライバシニーズに対応することを具体例を踏まえながら示した．また，単純な設定で実験を行い，攻撃者の $Preference$ に対する知識の誤差を測定した．今後の研究として，本研究で定義した $Preference$ を保護する LPPM の検討，およびどれほど強く $Preference$ を守れば実用的に個人をプライバシを保護していると言えるのかの検証などが考えられる．

謝辞 本研究は JSPS 科研費基盤研究 (S) No. 17H06099, (A) No. 18H04093 の助成を受けたものです。

文　献

[1] Fiore, Marco, Panagiota Katsikouli, Elli Zavou, Mathieu Cunche, Franccoise Fessant, Dominique Le Hello, Ulrich Matchi Aivodji, Baptiste Olivier, Tony Quertier and Razvan Stanica.“Privacy in trajectory micro-data publishing : a survey.”(2019).

[2] Chen, Rui, Benjamin C. M. Fung, Bipin C. Desai and Nériah M. Sossou.“Differentially private transit data publication: a case study on the montreal transportation system.”KDD (2012).

[3] Chen, Rui, Gergely Ács and Claude Castelluccia.“Differentially private sequential data publication via variable-length n-grams.”CCS ’12 (2012).

[4] He, Xi, Graham Cormode, Ashwin Machanavajjhala, Cecilia M. Procopiuc and Divesh Srivastava.“DPT: Differentially Private Trajectory Synthesis Using Hierarchical Reference Systems.”PVLDB 8 (2015): 1154-1165.

[5] Gursoy, Mehmet Emre, Ling Liu, Stacey Truex and Lei Yu.“Differentially Private and Utility Preserving Publication of Trajectory Data.”IEEE Transactions on Mobile Computing 18 (2019): 2315-2329.

[6] Kanza, Yaron and Hanan Samet.“An online marketplace for geosocial data.”GIS ’15 (2015).

[7] Xu, Fengli, Zhen Tu, Yong Li, Pengyu Zhang, Xiaoming Fu and Depeng Jin.“Trajectory Recovery From Ash: User Privacy Is NOT Preserved in Aggregated Mobility Data.”WWW (2017).

[8] NTT データ経営研究所.“パーソナルデータに関する一般消費者の意識調査”http://www.keieiken.co.jp/aboutus/newsrelease/161122/supplementing01.html (2016)

[9] 個人情報保護委員会.“GDPR（General Data Protection Regulation：一般データ保護規則）”https://www.ppc.go.jp/enforcement/infoprovision/laws/GDPR/ (2019)

[10] Beresford, Alastair R. and Frank Stajano.“Location Privacy in Pervasive Computing.”IEEE Pervasive Computing 2 (2003): 46-55.

[11] Beresford, Alastair R. and Frank Stajano.“Mix zones: user privacy in location-aware services.”IEEE Annual Conference on Pervasive Computing and Communications Workshops, 2004. Proceedings of the Second (2004): 127-131.

[12] Dwork, Cynthia.“Differential Privacy.”ICALP (2006).