プライバシを考慮した移動系列情報解析のための安全性の提案

著者	川本 淳平, 福地 一斗, 照屋 唯紀, 佐久間 淳
内容記述	SCIS 2013 The 30th Symposium on Cryptography and Information Security Kyoto, Japan, Jan. 22 - 25, 2013 The Institute of Electronics, Information and Communication Engineers
雑誌名	SCIS 2013
発行年	2013-01
URL	http://hdl.handle.net/2241/119467

SCIS 2013 The 30th Symposium on
Cryptography and Information Security
Kyoto, Japan, Jan. 22 - 25, 2013
The Institute of Electronics,
Information and Communication Engineers

プライバシを考慮した移動系列情報解析のための安全性の提案
A Privacy Model for Movement Sequence Mining

川本 淳平 * 福地 一斗 † 照屋 唯紀 * 佐久間 淳 *‡
Junpei Kawamoto Kazuto Fukuchi Tadanori Teruya Jun Sakuma

あらまし 本論文ではプライバシを考慮した移動系列情報解析のための安全性を提案する．モバイル端末やカーナビゲーションシステムから収集できる移動系列情報を解析することで，交通事故や自然災害を早期に検出することが期待されている．一方で，個々人の移動系列情報はプライバシ情報を含んでおり，解析に用いられたデータや解析結果から個人が特定されるようなことがあってはならない．移動系列情報は，時間と共に増大していく情報であるため時間が経つにつれプライバシを保護することが難しくなる情報である．本研究では，各時点で駅や交差点など POI に滞在している人数を公開するユースケースを考え，各 POI 間の移動確率がマルコフモデルで既知となっている場合に個人行動の追跡を防ぐ新しいプライバシモデルを提案する．

キーワード プライバシ保護データ出版，移動系列情報，プライバシモデル

1 はじめに

近年，スマートフォンやカーナビゲーションシステムの普及により，それらモバイル端末に搭載されている GPS から大人数の移動系列情報をリアルタイムに収集することが可能になった．この大規模な移動系列情報は，事故や渋滞の早期発見や自然災害発生時における人の流れなど様々な解析に利用が期待されている [1]．一方で，通信事業者やナビゲーションサービス提供者が顧客の端末から得た移動系列情報を解析機関に提供する場合，プライバシの問題を解決する必要が生じる．個人の移動系列情報には，人には知られたくない行動が記録されている可能性があるためである．それゆえ，移動系列情報をリアルタイムに提供する場合には，個人が特定できないことや個人の移動が追跡されないことを保証する必要がある．

この移動系列情報を第三者へリアルタイムに提供する場合におけるプライバシ保護にはいくつかの課題がある．先ず，個々人の移動系列情報は記録している時間が長ければ長いほど他人とは異なる系列になるという性質への対処である．すなわち，移動系列情報に対して k-匿名性 [2] や ℓ-多様性 [3] といった，他人の系列との区別困難性を基にした安全性を保証することは系列情報が長い場合は短い場合に比べて難しい．Fung らの LKC プライバシ [4] では，攻撃者が入手できる情報を制限することでこの問題に対処している．具体的には，攻撃者は，人々の移動系列のうち長さ L 以下の部分系列のみを取得できると仮定している．このように，移動系列の長さを制限することで他人との区別困難性を基にした安全性を保証することはできる．しかし，本稿で考えている移動系列情報のリアルタイム提供では，攻撃者が提供された情報全てを入手することは容易であると考えられる．そのため，攻撃者についての仮定は極力置かずにプライバシを定義する必要がある．

また，情報のリアルタイム提供に関しては，情報を複数回提供する場合のプライバシ問題が存在する [5, 6]．この問題は，情報を一定時間おきに複数回提供する場合において，各提供時に k-匿名性や ℓ-多様性といった既存のプライバシ定義を満足していても，別の時間に提供された情報と合わせることでプライバシ要件が破られてしまう問題である．Xiao らは，ある時刻にプライバシ要件を満足するために重要だった人の情報は，以後の時間においてその人の情報が削除されていたとしても提供情報には残しておく必要がある場合について議論している [5]．

このように，既存研究の多くでは，移動系列情報に対するプライバシと複数回提供時のプライバシを独立に

* 筑波大学大学院システム情報工学研究科，茨城県つくば市天王台 1-1-1，Graduate School of Systems and Information Engineering, Information and Systems, University of Tsukuba, 1-1-1, Tennodai, Tsukuba, Ibaraki, {junpei, tadanori}@mdl.cs.tsukuba.ac.jp, jun@cs.tsukuba.ac.jp
† 筑波大学情報学群情報科学類, 茨城県つくば市天王台 1-1-1, College of Information Science, University of Tsukuba, 1-1-1, Tennodai, Tsukuba, Ibaraki, kazuto@mdl.cs.tsukuba.ac.jp
‡ 科学技術振興機構さきがけ, Japan Science and Technology Agency

表 1: 記号のまとめ.

記号	意味
N	総人数
t	時刻（$t \in \{1, 2, \cdots, T\}$）
l	POI （$\ell \in \{\ell_1, \ell_2, \cdots, \ell_L\}$）
S_1	マルコフモデル M_1 における状態集合
$\boldsymbol{P}$	マルコフモデルにおける遷移確率行列
$\boldsymbol{\pi}(t)$	時刻 t における出力ヒストグラムベクトル
$\boldsymbol{\kappa}(t)$	時刻 t におけるターゲットの状態ベクトル
$\tilde{\boldsymbol{\kappa}}(t+s)$	時刻 $t+s$ における攻撃者の状態推測値

議論している．さらに，移動系列情報に代表される位置情報に対し，一般化階層構造を用いた匿名性メカニズム [7, 8] を適用すると，提供情報に対する解析結果が悪くなる，すなわち利便性が低くなる問題もある [9]．一方で，前述の解析用途に人々の移動系列情報を用いる場合，個々人の系列すべてを利用しなくても十分な場合もある．本稿では，あらかじめ定められた地点 (POI; point of interest) に滞在している人数のみをヒストグラムという形で提供することを考えプライバシを定義する．本稿で提案するプライバシ定義では，人々の行動がマルコフモデルに従うことを仮定している．また，時刻 t に攻撃対象の状態，すなわちどの POI にいたのかという情報を知り得た攻撃者が，提供されたヒストグラムとマルコフ遷移確率を用いて s 時間後の攻撃対象の状態を推測する攻撃を考える．

その上で，出力ヒストグラムが攻撃対象の状態推測に対する確信度に与える寄与のプライバシ侵害度合いは，攻撃者が対象の状態を観測してからの経過時間 s によって変化するという **経過時間とプライバシ侵害度合いの関係** を導入する．これは，出力ヒストグラムによって，小さい経過時間 s と大きい経過時間 s' があるとき，s' における対象の状態が漏洩する方がプライバシ侵害度合いは大きいということを意味している．この経過時間とプライバシ侵害度合いの関係を仮定し，出力ヒストグラムに対するプライバシを定義する．

2 基本事項

本稿では，GPS などによる移動系列情報を提供する **人々**，人々から得た移動系列情報を集約し解析機関へ提供する **システム**，そして，解析機関へ提供された情報を盗み見てある人の行動情報を得ようとする **攻撃者**，及び **攻撃対象** の四つのステークホルダを考える．以降では，これらステークホルダについて定式化する．なお，以降で導入する記号のまとめを表 1 に示す．

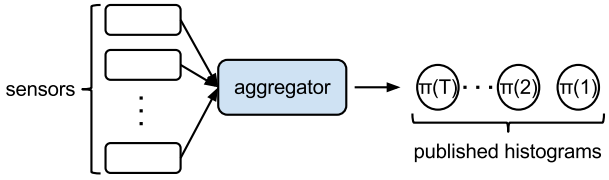


図 1: 移動情報の集約及び提供システム.

2.1 移動系列情報を提供する人々

本稿では，人々から提供される移動系列情報のうち，予め定めた移動手段による同じく予め定めた POI 間の移動系列にのみに着目する．例えば，移動手段が鉄道利用で POI が鉄道駅からなる場合や，移動手段が高速道路の走行で POI がサービスエリアやインターチェンジからなる場合などである．つまり，解析機関がある交通ネットワーク内における人々の移動系列情報を解析することを考えている．そのため，POI は任意の POI ペアが 1 ステップで直接到達可能であるか否かの関係を基にグラフ (***POI* グラフ**) として表現できる．今，与えられた POI の集合を $\{\ell_1, \ell_2, \cdots, \ell_L\}$ とすると，POI グラフにおける頂点集合 V は POI の集合に等しく $V = \{\ell_1, \ell_2, \cdots, \ell_L\}$ となる．また，直接到達可能な POI 間に無向枝を張り，その枝集合を E と書くと，POI グラフ G は形式的に $G(V, E)$ と表すことができる．

次に，この POI にいる人は次にどの POI に向かう傾向にあるという，人々の一般的な行動パタンをモデル化する．本稿では，マルコフモデルによって一般的な行動パタンを表現する．先ず，簡単なモデル化として POI グラフ G における頂点，すなわち各 POI を状態とするマルコフモデル M_1 を考える．よって，このモデルにおける状態集合を S_1 とすると $S_1 = V$ となる．このモデルにおける遷移確率 P は，POI から POI への遷移確率 $P : S_1 \times S_1 \to [0, 1]$ となる．ただし $\forall \ell \in S_1, \sum_{\ell' \in S_1} P(\ell, \ell') = 1$ である．この遷移確率の行列表現を行確率行列 $\boldsymbol{P}$ と書くことにする．なお，POI グラフ G 上の人々の行動のマルコフモデルは M_1 以外にも考えられる．4 節ではそうした複雑なモデル化について議論する．

2.2 移動系列情報の集約システム

次に，人々の携帯電話やカーナビゲーションシステムに内蔵される GPS から緯度・経度情報を収集し，それらを解析機関へ提供するシステムを導入する．このシステムは，自明なプライバシ問題の観点から収集した情報そのものを解析機関へ提供するのではなく，先ず駅名や交差点名など予め与えられた POI に滞在している人数のみを 5 分間隔，10 分間隔といった定められた時間間隔ごとに集計し出力することを考える．従ってこのシス

テムは，時間間隔ごとに各 POI に滞在している人数の分布，すなわちヒストグラムを出力する．図 1 は，このシステムの概要を記したものである．このシステムは個々人の GPS（センサー）から緯度・経度を受け取ると，緯度・経度と POI を対応付けるクリーニング処理を行う．その後，各 POI における滞在人数を定められた時間間隔ごとに集計し出力する．なお，このシステムは N 人の緯度・経度情報を収集するとする．

本稿では，ヒストグラムが出力される時刻を t で表し，全部で T 時刻まであると考えることにする．すなわち，$t \in \{1, 2, \cdots, T\}$ となる．本稿では，ある時刻にヒストグラムを出力する場合に，そのヒストグラムをそのまま公開して良いのか否かについて議論する．つまり，過去に出力されたヒストグラムを併せて用いることで，今公開しようとしているヒストグラムからプライバシを侵害することが起きないようにヒストグラムの出力方法を考えたい．このとき，過去に出力された無限個のヒストグラムを用いた攻撃は現実的ではないと考える．そのため，本稿で対象とする時間は有限であると仮定する．

時刻 t における出力ヒストグラムは L 次元のベクトル $\boldsymbol{\pi}(t)$ として表し，各 POI に滞在している人数の割合を意味する．よって，$\boldsymbol{\pi}(t)$ の i 番目の要素を $\pi_i(t)$ と書くと，これは時刻 t において POI ℓ_i に滞在している人数の割合を表し，$\sum_{i=1}^{L} \pi_i(t) = 1$ となる．

2.3 攻撃者と攻撃対象

攻撃者によって行動記録が暴かれようとしている対象の行動を定式化する．なお，攻撃対象は一人とする．時刻 t における対象の状態を確率ベクトル $\boldsymbol{\kappa}(t)$ と書く．この $\boldsymbol{\kappa}(t)$ は，一カ所のみ 1 でその他が 0 である状態ベクトルである．例えば，時刻 t で対象が POI ℓ_k に滞在していたとすると，

$$\boldsymbol{\kappa}(t) = (0, 0, \cdots, 0, \overset{k}{\check{1}}, 0, \cdots, 0)^t$$

となる[1]．

攻撃者は，この対象の時刻 t における状態 $\boldsymbol{\kappa}(t)$ を知ることができるとする．これは，実際に対象をある POI で見かけた場合やその他の外部情報を使うことで実現可能である．そして，攻撃者はこの背景知識を基に s 時間後の対象の状態 $\boldsymbol{\kappa}(t+s)$ を推測する．つまり，攻撃者は時刻 t に対象が POI ℓ_k にいたという情報を基に，時刻 $t+s$ における滞在位置を推測する攻撃を行う．この推測値は確率ベクトルであるとし $\tilde{\boldsymbol{\kappa}}(t+s)$ と書く．よって，$\tilde{\boldsymbol{\kappa}}(t+s)$ の i 番目の要素を $\tilde{\kappa}_i(t+s)$ と書くと，$\forall i, \tilde{\kappa}_i(t+s) \in [0,1]$ かつ $\sum_i \tilde{\kappa}_i(t+s) = 1$ である．なお，この間，時刻 $t+1$ から $t+s$ の間は直接対象の状態を知ることはできないとする．その他攻撃者が利用できる情報は，システムが出力するヒストグラム $\boldsymbol{\pi}(t), \boldsymbol{\pi}(t+1), \cdots, \boldsymbol{\pi}(t+s)$ とマルコフモデルにおける遷移確率 P とする．攻撃者が遷移確率 P を入手できることは現実的ではない．実際，遷移確率 P 人々の行動記録全体を知り得たとしても推定しか行えない．しかし，本稿では，遷移確率 P さえ知っている強い攻撃者に対して安全性を議論するため，攻撃者は P を知っていると仮定する．

[1] ベクトル $\boldsymbol{v}$ の転置ベクトルを $\boldsymbol{v}^t$ と書く．

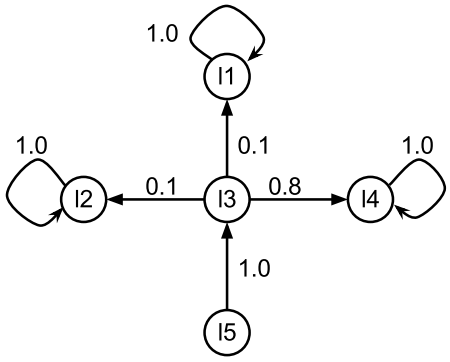


図 2: 行動パタンが自明な場合．

3 安全性定義

本節では，図 1 に示したシステムが出力するヒストグラム $\boldsymbol{\pi}(1), \boldsymbol{\pi}(2), \cdots, \boldsymbol{\pi}(T)$ に対する安全性を定義する．提案の安全性は，人々の一般的な行動パタンがマルコフモデル M_1 として既知の場合に，マルコフモデルのみから推測される行動とヒストグラムも併せて推測される行動の比を用いて定義する．安全性は，上記の比ではなく，マルコフモデルとヒストグラムから推測される攻撃者の確信度のみを用いて定義することも可能ではある．しかし，図 2 に示すように行動パタンが自明であるような場合，ヒストグラムがプライバシを侵害していないにも関わらず攻撃者の確信度が高くなってしまう問題がある．図 2 は，五つの状態 $\{\ell_1, \ell_2, \cdots, \ell_5\}$ からなるマルコフモデルの遷移確率を表している．遷移確率より，ℓ_5 にいる人の多くが $\ell_5 \to \ell_3 \to \ell_4$ と行動することが分かる．この場合，出力ヒストグラムの有無によらず高い確信度で攻撃者は対象の行動を推定できる．したがって，マルコフモデルとヒストグラムから推測される攻撃者の確信度が高いからと言って出力ヒストグラムがプライバシを侵害しているとは言い切れない．そこで，マルコフモデルのみからの推測とヒストグラムも併せた推測との比を用いて自明でない行動に対する推定の度合いを評価している．すなわち，この二つの差が小さければ出力ヒストグラムが攻撃者に有利な情報を与えていることにはならずヒストグラムはプライバシを侵害しないと言える．

本節では，まず出力ヒストグラムが攻撃者の確信度に与えるゲイン定義した後，安全性を定義する．その後，

プライバシリスクの評価方法について述べた後，例を用いて実際に計算を行う．

3.1 攻撃者へのゲインと安全性定義

本稿で想定する攻撃者は，時刻 t に知り得た攻撃対象の状態 $\boldsymbol{\kappa}(t)$ を基に，時刻 $t+s$ における対象の状態を推測する．この攻撃者が利用できる情報は，マルコフモデルにおける遷移確率 P と図 1 のシステムが出力するヒストグラム $\boldsymbol{\pi}(t), \boldsymbol{\pi}(t+1), \cdots, \boldsymbol{\pi}(t+s)$ である．攻撃者が推測した，時刻 $t+s$ における対象の状態 $\tilde{\boldsymbol{\kappa}}(t+s)$ の i 番目の要素 $\tilde{\kappa}_i(t+s)$ は，時刻 $t+s$ に対象が POI ℓ_i に滞在している確信度を表している．ここで，攻撃者が出力ヒストグラムを用いずにこの推測を行った場合と，ヒストグラムを用いた場合の二種類を考える．ヒストグラムを用いない場合，

$$\tilde{\kappa}_i(t+s) = p(\ell_i|\boldsymbol{\kappa}(t); \boldsymbol{P})$$

と言える．またヒストグラムを用いた場合は，

$$\tilde{\kappa}_i(t+s) = p(\ell_i|\boldsymbol{\kappa}(t), \boldsymbol{\pi}(t), \boldsymbol{\pi}(t+1), \cdots, \boldsymbol{\pi}(t+s); \boldsymbol{P})$$

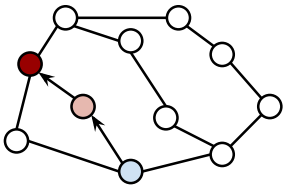
と書ける．この二つの比は，出力ヒストグラムが攻撃者に与えるゲインと考えられ，比が大きくなれば攻撃者は高い確信度で攻撃対象の状態を推測できる．この考えを基に，出力ヒストグラム $\boldsymbol{\pi}(t), \boldsymbol{\pi}(t+1), \cdots, \boldsymbol{\pi}(t+s)$ による攻撃者へのゲインを定義する．

定義 3.1 (出力ヒストグラムによる攻撃者へのゲイン) 与えられた攻撃対象の時刻 t における状態 $\boldsymbol{\kappa}(t)$ と遷移確率行列 $\boldsymbol{P}$ の基で，出力ヒストグラム $\boldsymbol{\pi}(t), \boldsymbol{\pi}(t+1), \cdots, \boldsymbol{\pi}(t+s)$ から時刻 $t+s$ に攻撃対象が $POI\ \ell_i$ に滞在していると推測する攻撃者の確信度へのゲインを

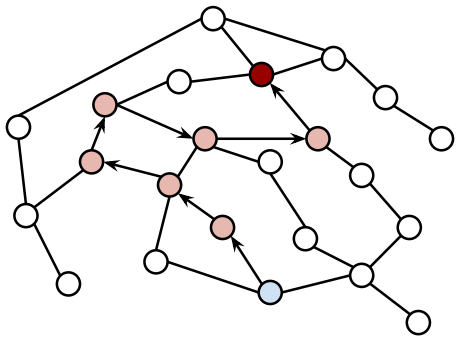
$$\begin{aligned}&\mathrm{Gain}(\boldsymbol{\pi}(t), \boldsymbol{\pi}(t+1), \cdots, \boldsymbol{\pi}(t+s); \ell_i, \boldsymbol{\kappa}(t), \boldsymbol{P}) \\ &\quad = \frac{p(\ell_i|\boldsymbol{\kappa}(t), \boldsymbol{\pi}(t), \boldsymbol{\pi}(t+1), \cdots, \boldsymbol{\pi}(t+s); \boldsymbol{P})}{p(\ell_i|\boldsymbol{\kappa}(t); \boldsymbol{P})}\end{aligned}$$

と定める．

さらに，本研究では，**経過時間とプライバシ侵害度合いの関係** を導入する．これは，出力ヒストグラムが同じゲインを攻撃者に与える場合，経過時間 s が大きい場合に与えるゲインの方が，s が小さい場合に与えるゲインよりプライバシの侵害度合いが大きいことを言う．なぜなら，攻撃対象が図 3 のような移動を行ったとし，どちらの場合も出力ヒストグラムが同じゲインを攻撃者に与えたとする．経過時間 s が小さい時，攻撃対象が到達可能な POI は少なく対象の滞在 POI を推測することは容易であるため，出力ヒストグラムによるゲインのプライバシ侵害度合いは大きくないと考えられる．一方，十分時間が経った後，すなわち経過時間 s が大きい時，攻撃対象が到達可能な POI は多く滞在 POI の推測は容易ではない．それにも関わらず，出力ヒストグラムが攻撃者の確信度にゲインを与える場合，プライバシの侵害度合いは大きいと考えられるからである．



(a) 小さい経過時間 ($s = 2$).



(b) 大きい経過時間 ($s = 7$).

図 3: 経過時間とプライバシ侵害度合いの関係.

この経過時間とプライバシ侵害度合いの関係を仮定し出力ヒストグラムのプライバシを定義する．

定義 3.2 (出力ヒストグラムのプライバシ) s 時間で攻撃対象の到達可能な状態集合を $S_1' \subseteq S_1$ とすると，出力ヒストグラム $\boldsymbol{\pi}(t), \boldsymbol{\pi}(t+1), \cdots, \boldsymbol{\pi}(t+s)$ が，任意の $\ell_i \in S_1'$ に対して次の式を満足するとき単調減少関数 $\epsilon : \mathbf{Z}_+ \to \mathbf{R}_+$ の基で $\boldsymbol{\kappa}(t)$ を持つ攻撃者に対して安全であると言う．

$$\mathrm{Gain}(\boldsymbol{\pi}(t), \boldsymbol{\pi}(t+1), \cdots, \boldsymbol{\pi}(t+s); \ell_i, \boldsymbol{\kappa}(t), \boldsymbol{P}) < \epsilon(s) \tag{1}$$

3.2 攻撃者へのゲインの評価

先ず，$s = 1$ の場合について式 (1) を評価する．出力ヒストグラムを用いない場合の推測 $p(\ell_i|\boldsymbol{\kappa}(t); \boldsymbol{P})$ は，時刻 t における状態と遷移確率から求めることができる．すなわち，

$$p(\ell_i|\boldsymbol{\kappa}(t); \boldsymbol{P}) = [\boldsymbol{\kappa}^t(t)\boldsymbol{P}]_i$$

である[2]．出力ヒストグラムを用いた場合の推測 $p(\ell_i|\boldsymbol{\kappa}(t), \boldsymbol{\pi}(t), \boldsymbol{\pi}(t+1); \boldsymbol{P})$ は，

$$\begin{aligned}&p(\ell_i|\boldsymbol{\kappa}(t), \boldsymbol{\pi}(t), \boldsymbol{\pi}(t+1); \boldsymbol{P}) \\ &\quad = \frac{p(\boldsymbol{\pi}(t), \boldsymbol{\pi}(t+1)|\ell_i, \boldsymbol{\kappa}(t); \boldsymbol{P}) p(\ell_i, \boldsymbol{\kappa}(t); \boldsymbol{P})}{p(\boldsymbol{\kappa}(t), \boldsymbol{\pi}(t), \boldsymbol{\pi}(t+1); \boldsymbol{P})}\end{aligned} \tag{2}$$

[2] ベクトル $\boldsymbol{v}$ の i 番目の要素を $[\boldsymbol{v}]_i$ と表す.

と計算できる．この式の，$p(\boldsymbol{\pi}(t), \boldsymbol{\pi}(t+1)|\ell_i, \boldsymbol{\kappa}(t); \boldsymbol{P})$ は，時刻 t における状態 $\boldsymbol{\kappa}(t)$ から時刻 $t+1$ に状態 ℓ_i 移動した場合の，ヒストグラム $\boldsymbol{\pi}(t)$ 及び $\boldsymbol{\pi}(t+1)$ の尤度である．今，時刻 t における対象の滞在 POI が ℓ_k であるとする．すなわち，$\boldsymbol{\kappa}(t)$ が，

$$\boldsymbol{\kappa}(t) = (0, 0, \cdots, 0, \overset{k}{\check{1}}, 0, \cdots, 0)^t$$

という形であるとする．これは，対象は ℓ_k から ℓ_i へ移動したとことを表している．このとき，残り $N-1$ 人の行動に関して，ある一人が時刻 t から $t+1$ の間に ℓ_s から ℓ_e に移動したとする．ただし，$\ell_s, \ell_e \in S_1$ である．この人の行動を $a = (\ell_s \to \ell_e)$ と書くことにする．この行動 a が起きる確率は，遷移確率行列 $\boldsymbol{P}$ の i, j 成分を $P_{i,j}$ と書くと $P_{s,e}$ である．出力ヒストグラム $\boldsymbol{\pi}(t)$ と $\boldsymbol{\pi}(t+1)$ に矛盾しない $N-1$ 人分の可能な行動にはさまざまな組み合わせが存在するが，そのうちの一つを $A = \{a_1, a_2, \cdots, a_{N-1}\}$ と書くことにする．この行動集合 A の尤度 $p(A|\ell_i, \boldsymbol{\kappa}(t); \boldsymbol{P})$ は，

$$p(A|\ell_i, \boldsymbol{\kappa}(t); \boldsymbol{P}) = \prod_{(\ell_s \to \ell_e) \in A} P_{s,e}$$

となる．この尤度が $p(\boldsymbol{\pi}(t), \boldsymbol{\pi}(t+1)|\ell_i, \boldsymbol{\kappa}(t); \boldsymbol{P})$ となる．また，$p(\ell_i, \boldsymbol{\kappa}(t); \boldsymbol{P}) = [\boldsymbol{\kappa}^t(t)\boldsymbol{P}]_i$ である．よってある行動集合 A に対する式 (2) は，

$$p_A(\ell_i|\boldsymbol{\kappa}(t), \boldsymbol{\pi}(t), \boldsymbol{\pi}(t+1); \boldsymbol{P}) = \frac{p(A|\ell_i, \boldsymbol{\kappa}(t); \boldsymbol{P}) p(\ell_i, \boldsymbol{\kappa}(t); \boldsymbol{P})}{\sum_{\ell_i, \boldsymbol{\kappa}(t)} p(A|\ell_i, \boldsymbol{\kappa}(t); \boldsymbol{P}) p(\ell_i, \boldsymbol{\kappa}(t); \boldsymbol{P})} \quad (3)$$

となる．従って，ある行動集合 A に対する出力ヒストグラム $\boldsymbol{\pi}(t)$，$\boldsymbol{\pi}(t+1)$ が攻撃者へ与えるゲインは，

$$\mathrm{Gain}_A(\boldsymbol{\pi}(t), \boldsymbol{\pi}(t+1); \ell_i, \boldsymbol{\kappa}(t), \boldsymbol{P}) = \frac{p(A|\ell_i, \boldsymbol{\kappa}(t); \boldsymbol{P}) p(\ell_i, \boldsymbol{\kappa}(t); \boldsymbol{P})}{p(\ell_i|\boldsymbol{\kappa}(t), \boldsymbol{P}) \sum_{\ell_i, \boldsymbol{\kappa}(t)} p(A|\ell_i, \boldsymbol{\kappa}(t); \boldsymbol{P}) p(\ell_i, \boldsymbol{\kappa}(t); \boldsymbol{P})}$$

となる．この値が最も大きい場合であっても式 (1) を満たせば，出力されたヒストグラム $\boldsymbol{\pi}(t)$, $\boldsymbol{\pi}(t+1)$ は安全である．よって $s=1$ の場合，定義 3.2 の条件は，出力ヒストグラム $\boldsymbol{\pi}(t)$ 及び $\boldsymbol{\pi}(t+1)$ が任意の $\ell_i \in S_1'$ と単調減少関数 $\epsilon : \mathbf{Z}_+ \to \mathbf{R}_+$ に対して，

$$\max_A \mathrm{Gain}_A(\boldsymbol{\pi}(t), \boldsymbol{\pi}(t+1); \ell_i, \boldsymbol{\kappa}(t), \boldsymbol{P}) < \epsilon(1)$$

を満たすことの必要条件となる．以上の議論は $s > 1$ の場合にも拡張でき，次の定理を導く．

定理 3.1 ヒストグラムの集合 $\{\boldsymbol{\pi}(t), \boldsymbol{\pi}(t+1), \cdots, \boldsymbol{\pi}(t+s)\}$ と $\boldsymbol{\kappa}(t)$，攻撃対象の時刻 $t+s$ における予測状態 $l_i \in S_1'$ があるとする．出力ヒストグラム集合に矛盾しない攻撃対象以外の $N-1$ 人の可能な行動の集合 A の集合族を $\mathcal{A}$ と書くと，任意の $\ell_i \in S_1'$ に対してある単調減少関数 $\epsilon : \mathbf{Z}_+ \to \mathbf{R}_+$ が，

$$\max_{A \in \mathcal{A}} \mathrm{Gain}_A(\boldsymbol{\pi}(t), \boldsymbol{\pi}(t+1), \cdots, \boldsymbol{\pi}(t+s); \ell_i, \boldsymbol{\kappa}(t), \boldsymbol{P}) < \epsilon(s)$$

を満足すれば，出力ヒストグラム集合 $\{\boldsymbol{\pi}(t), \boldsymbol{\pi}(t+1), \cdots, \boldsymbol{\pi}(t+s)\}$ は $\epsilon(s)$ の基で定義 *3.2* のプライバシを満たす．

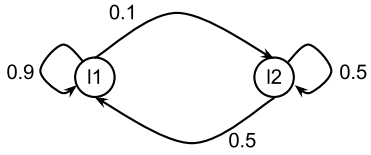


図 4: マルコフモデルの例．

3.3 攻撃者へのゲインの計算例

POI 集合は $\{\ell_1, \ell_2\}$ に対して図 4 に示すようなマルコフモデル M_1 があるとする．このとき，遷移確率行列 $\boldsymbol{P}$ は，

$$\boldsymbol{P} = \begin{pmatrix} 0.9 & 0.1 \\ 0.5 & 0.5 \end{pmatrix}$$

である．また，時刻 t における攻撃対象の状態 $\boldsymbol{\kappa}(t) = (1, 0)^t$ であるとする．よって，出力ヒストグラムを用いない攻撃者の推測 $p(\ell|\boldsymbol{\kappa}(t); \boldsymbol{P})$ $(\ell = \ell_1, \ell_2)$ は，

$$\begin{aligned} p(\ell_1|\boldsymbol{\kappa}(t); \boldsymbol{P}) &= [\boldsymbol{\kappa}^t(t)\boldsymbol{P}]_1 = 1 \times 0.9 + 0 \times 0.5 = 0.9 \\ p(\ell_2|\boldsymbol{\kappa}(t); \boldsymbol{P}) &= [\boldsymbol{\kappa}^t(t)\boldsymbol{P}]_2 = 1 \times 0.1 + 0 \times 0.5 = 0.1 \end{aligned}$$

となる．

今，時刻 t 及び $t+1$ における出力ヒストグラム $\boldsymbol{\pi}(t)$, $\boldsymbol{\pi}(t+1)$ が，

$$\boldsymbol{\pi}(t) = \begin{pmatrix} 2 \\ 1 \end{pmatrix}, \qquad \boldsymbol{\pi}(t+1) = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$$

であったとする．なお，このヒストグラムは，時刻 t において ℓ_1 にいた人のどちらかあるいは両方が時刻 $t+1$ において ℓ_2 に移動したことを表している．これは，遷移確率行列から見て稀な場合であり，出力ヒストグラムによるプライバシリスクは大きくなると予想される．

次に，攻撃対象者が時刻 $t+1$ で ℓ_1 に滞在していると仮定する．このとき，攻撃対象者以外の可能な行動は対象性を考慮すると，$A_1 = \{(\ell_1 \to \ell_2), (\ell_2 \to \ell_2)\}$ のみであり，尤度は，

$$p(A_1|\ell_1, \boldsymbol{\kappa}(t); \boldsymbol{P}) = \prod_{(\ell_s \to \ell_e) \in A_1} P_{s,e} = 0.05$$

となる．同様に，攻撃対象者が時刻 $t+1$ で ℓ_2 に滞在していると仮定すると，攻撃対象以外の可能な行動は，

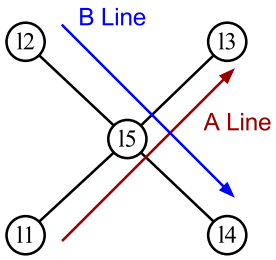


図 5: 滞在地点が前時刻の滞在地点に依存する例.

$A_2 = \{(\ell_1 \to \ell_1), (\ell_2 \to \ell_2)\}$ と $A_3 = \{(\ell_1 \to \ell_2), (\ell_2 \to \ell_1)\}$ の二種類が考えられる．尤度はそれぞれ，

$$p(A_2|\ell_2, \boldsymbol{\kappa}(t); \boldsymbol{P}) = 0.45, \quad p(A_3|\ell_2, \boldsymbol{\kappa}(t); \boldsymbol{P}) = 0.05$$

である．従って，式 (3) の分母は，

$$\sum_{\ell_i, \boldsymbol{\kappa}(t)} p(A|\ell_i, \boldsymbol{\kappa}(t); \boldsymbol{P}) p(\ell_i, \boldsymbol{\kappa}(t); \boldsymbol{P}) = 0.095$$

となり，

$$\mathrm{Gain}_{A_1}(\boldsymbol{\pi}(t), \boldsymbol{\pi}(t+1); \ell_1, \boldsymbol{\kappa}(t), \boldsymbol{P}) = \frac{0.05 \times 0.9}{0.9 \times 0.095} \approx 0.5263$$

$$\mathrm{Gain}_{A_2}(\boldsymbol{\pi}(t), \boldsymbol{\pi}(t+1); \ell_2, \boldsymbol{\kappa}(t), \boldsymbol{P}) \approx 4.7368$$
$$\mathrm{Gain}_{A_3}(\boldsymbol{\pi}(t), \boldsymbol{\pi}(t+1); \ell_3, \boldsymbol{\kappa}(t), \boldsymbol{P}) \approx 0.5263$$

となる．定理 3.1 より，$\epsilon(1) > 4.7368$ であればこの出力ヒストグラム $\boldsymbol{\pi}(t)$，$\boldsymbol{\pi}(t+1)$ は ϵ の基で安全である．逆に $\epsilon(1) \leq 4.7368$ であれば，システムはヒストグラム $\boldsymbol{\pi}(t)$，$\boldsymbol{\pi}(t+1)$ をそのまま出力してはならず，何らかのメカニズムを用いてヒストグラムを書き換えることになる．このメカニズムの考案は今後の課題である．

4 複雑なマルコフモデルへの拡張

本節では，2 節で導入した M_1 モデルを拡張したモデルについて議論する．M_1 モデルでは，状態を POI グラフの頂点としていた．つまり，M_1 モデルにおける状態集合 S_1 は，POI グラフ G の頂点集合 V と等しかった．このモデルは，ある POI に居た人は次の時刻にどの POI に向かいやすいのかを表している．一方で，これだけでは捉えきれない状況も存在する．例えば，POI が駅である場合を考える．この時，ある駅に滞在している人が次の時刻にどの駅に向かいやすいかが，どの路線を使ってその駅に到達したのかに影響される場合が考えられる．図 5 は，そのような状況の例を示したもので，五つの駅が POI $\{\ell_1, \ell_2, \cdots, \ell_5\}$ であり，(ℓ_1, ℓ_5, ℓ_3) と (ℓ_2, ℓ_5, ℓ_4) がそれぞれ別の路線である A 線と B 線の駅とする．このとき，駅 ℓ_5 での乗り換えは通過する人に比べて少ないと仮定すると，時刻 t で ℓ_5 に滞在している人が時刻 $t+1$ にどの駅に向かいやすいかは，どちらの路線で ℓ_5 に到達したのかに依存する．

このような状態を取り扱うためには，各時刻に滞在している POI だけではなく，その前の時刻に滞在していた POI も併せた POI のペアを状態として持つモデルが必要になる．このモデルを M_2 とする．M_2 モデルにおける状態集合 S_2 は，POI グラフ G における頂点ペアの部分集合であり $S_2 \subseteq V \times V$ と書ける．2 節で議論したように POI グラフでは，直接到達可能では無い頂点間に枝は無い．それゆえ S_2 は $V \times V$ の部分集合となる．また，M_2 モデルにおける遷移確率 P_2 は，$P_2 : S_2 \times S_2 \to [0, 1]$ である．

さらに，M_2 モデルでも取り扱えない状況も考えられる．M_2 モデルにおける状態は POI のペアであり，現在の滞在 POI だけではなく，一つ前の時刻に滞在していた POI も併せて状態とすることで経路を基に遷移確率を考えることができた．しかし，同じ経路であっても移動速度を含めて遷移確率を考えたい場合もある．例えば，同じ鉄道路線を移動中であっても各駅停車の電車に乗車中と特急列車に乗車中ではある駅において次に向かう路線に差がある場合などである．この場合は，一つ前の時刻だけではなく二つ前の時刻まで考える必要がある．このようにして M_3 モデルが考えられる．同様にして，n 個前の時刻まで考慮した M_n モデルを考えることができる．M_n モデルにおける状態集合 S_n は $S_n \subset V^n$ となる[3]．また，遷移確率 P_n は，$P_n : S_n \times S_n \to [0, 1]$ と定義できる．

このように，M_1 モデルを基にさらに複雑なモデルを考えることができる．一方で，3 節で議論した安全性モデルは，一般的なマルコフモデル上であれば同様に定義できる．すなわち，これらの複雑なモデルに対しても 3 節の議論は有効である．

5 関連研究

人々の行動がマルコフモデルに従うと仮定しプライバシを議論している研究として，Götz らの MaskIt がある [10]．MaskIt では，携帯電話が収集する位置情報や，誰と通話しているのかという利用者のコンテキストをアプリケーションに提供して良いのか否かを判断するプライバシ基準について論じている．ここでは，コンテキスト間の遷移をマルコフモデルによって表現しており，利用者が予め秘匿したいと指定したコンテキストすなわち状態を攻撃者によって推測される確率を基にプライバシを定義している．MaskIt では，単一の携帯端末利用者の状態を秘匿することを目的としているため，利用者が

[3] V^1 を V と定義し，$V \times V^{n-1}$ を V^n と書く．

秘匿すべき状態を指定するという方法を取っている．一方，本稿で扱う問題では，大人数の人々から収集する情報が対象であり，個々人に秘匿すべき状態を提示してもらう方法は現実的ではない．それゆえ，本稿では利用者個々人に依存せず普遍的な安全性を議論している．

プライバシを考慮しつつヒストグラムを提供するという問題に関しては，公開されたヒストグラムが差分プライバシ (differential privacy) [11, 12] を満足するように摂動を加える手法が提案されている [13]．しかし，Xu らの手法は，ヒストグラムを複数回提供する場合を考慮していない．1 節で議論したように，ある時間に提供した情報がプライバシ要件を満たしていても，他の時間に提供された情報と合わせることでプライバシ要件が破られてしまう問題があるため，この手法を単純に我々の問題へ適用することはできないと考える．

最後に，センサーから収集される情報に対するプライバシ保護を集約者が保証するのではなく，センサーが個別に保証するアプローチの研究が行われている [14, 15]．Rastogi らは，センサーが各時刻に数値を出力する場合において，集約結果が差分プライバシを満足する方法を提案している [14]．提案手法では，マルチパーティ秘密計算 [16] を用いている．そのため，センサーと集約者の間で秘密計算を行った結果．最終的な値が差分プライバシを満足することになる．Shi らも同様にセンサーからの集約演算に秘密計算を用いており，各センサーが実際の値を集約者に開示すること無しに集約を行っている [15]．本稿で想定している問題では，GPS 情報の集約者は通信事業者やナビゲーションサービス提供者である．それゆえ，人々がそれら事業者に対して GPS 情報を秘匿したまま解析者に集約結果を提供する状況は稀である．むしろ，集約者と解析者の間で情報提供する段階におけるプライバシ保護が必要である．したがって，本稿では集約者が保証すべき安全性について議論している．

6　まとめと今後の課題

本稿では，通信事業者やナビゲーションサービスプロバイダが顧客のスマートフォンやカーナビゲーションシステムに搭載されている GPS などから収集した移動系列情報を，解析の為に第三者機関に提供する場合におけるプライバシ定義を提案した．提案プライバシは，移動系列情報を集約し各時間，各 POI における滞在人数ヒストグラムを出力するシステムを想定し，人々の行動がマルコフモデルで表せられることを仮定している．また，攻撃者として時刻 t に攻撃対象の状態，すなわちどの POI にいたのかという情報を基に，システムが出力したヒストグラムとマルコフ遷移確率を用いて s 時間後の攻撃対象の状態を推測する存在を考えている．その上で，出力ヒストグラムが同じゲインを攻撃者に与える場合，経過時間 s が大きい場合に与えるゲインの方が，s が小さい場合に与えるゲインよりプライバシの侵害度合いが大きいという，経過時間とプライバシ侵害度合いの関係を導入した．そして，この関係を仮定しプライバシを定義した．また，提案のプライバシ定義は人々の状態をマルコフモデルによって表せていれば，さまざまなモデル化に対しても有効である．実際, 4 節において複雑なマルコフモデルに対しての議論を行った．

今後の課題は，提案のプライバシ要件を満足する出力ヒストグラム $\boldsymbol{\pi}(1), \boldsymbol{\pi}(2), \cdots, \boldsymbol{\pi}(T)$ を計算するメカニズムの考案である．

謝辞

本研究は，最先端研究開発プログラム「超巨大データベース時代に向けた最高速データベースエンジンの開発と当該エンジンを核とする戦略的社会サービスの実証・評価」の助成を受けました．

参考文献

[1] Shashi Shekhar, Viswanath Gunturi, Michael R. Evans, and KwangSoo Yang. Spatial Big-Data Challenges Intersecting Mobility and Cloud Computing. In *Proc. of the Eleventh ACM International Workshop on Data Engineering for Wireless and Mobile Access*, pp. 1–6, Scottsdale, AZ, USA, May 2012. ACM Press.

[2] Latanya Sweeney. k-anonymity: a model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, Vol. 10, No. 5, pp. 1–14, 2002.

[3] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. L-diversity: Privacy Beyond k-Anonymity. *ACM Transactions on Knowledge Discovery from Data*, Vol. 1, No. 1, pp. 1556–4681, March 2007.

[4] Benjamin C. M. Fung, Ming Cao, Bipin C. Desai, and Heng Xu. Privacy Protection for RFID Data Categories and Subject Descriptors. In *Proc. of the 24th ACM Symposium on Applied Computing*, pp. 1528–1535, Honolulu, HI, USA, 2009. ACM Press.

[5] Xiaokui Xiao and Yufei Tao. m-Invariance: Towards Privacy Preserving Re-publication of Dynamic Datasets. In *Proc. of the ACM SIGMOD*

International Conference on Management of Data, pp. 689–700, Beijing, China, 2007. ACM Press.

[6] Raymond Chi-Wing Wong, Ada Wai-Chee Fu, Jia Liu, Ke Wang, and Yabo Xu. Global Privacy Guarantee in Serial Data Publishing. In *Proc. of the 26th International Conference on Data Engineering*, pp. 956–959, Long Beach, CA, USA, 2010. IEEE Computer Society.

[7] Kristen LeFevre, David J. DeWitt, and Raghu Ramakrishnan. Incognito: Efficient Full-domain k-Anonymity. In *Proc of the ACM SIGMOD International Conference on Management of Data*, pp. 49–60, Baltimore, MD, USA, June 2005. ACM Press.

[8] Yeye He and Jeffrey F. Naughton. Anonymization of set-valued data via top-down, local generalization. *Proc. of the International Conference on Very Large Databases*, Vol. 2, No. 1, pp. 934–945, August 2009.

[9] 川本 淳平, 佐久間 淳. 位置情報解析のためのプライバシ保護手法. 2012 年度 全国共同利用研究発表大会「CSIS DAYS 2012」, D08, 千葉, 2012.

[10] Michaela Götz, Suman Nath, and Johannes Gehrke. MaskIt: Privately Releasing User Context Streams for Personalized Mobile Applications. In *Proc. of the International Conference on Management of Data*, pp. 289–300, Scottsdale, AZ, USA, May 2012. ACM Press.

[11] Cynthia Dwork, Frank Mcsherry, Kobbi Nissim, and Adam Smith. Calibrating Noise to Sensitivity in Private Data Analysis. In *Proc. of the third Theory of Cryptography Conference*, pp. 265–284, New York, NY, USA, 2006. Springer.

[12] Noman Mohammed, Rui Chen, Benjamin C M Fung, and Philip S. Yu. Differentially Private Data Release for Data Mining. In *Proc. of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 493–501, San Diego, CA, USA, 2011. ACM Press.

[13] Jia Xu, Zhenjie Zhang, Xiaokui Xiao, Yin Yang, and Ge Yu. Differentially Private Histogram Publication. In *Proc. of the 28th IEEE International Conference on Data Engineering*, pp. 32–43, Washington, DC, USA, 2012. IEEE Computer Society.

[14] Vibhor Rastogi and Suman Nath. Differentially Private Aggregation of Distributed Time-Series with Transformation and Encryption. In *Proc. of the 2010 ACM SIGMOD International Conference on Management of Data*, pp. 735–746, Indianapolis, IN, USA, 2010. ACM Press.

[15] Elaine Shi, T-H. Hubert Chan, Eleanor Rieffel, Richard Chow, and Dawn Song. Privacy-Preserving Aggregation of Time-Series Data. In *Proc. of the Network and Distributed System Security Symposium*, San Diego, CA, USA, 2011.

[16] Benny Pinkas Yehuda Lindell. Secure Multiparty Computation for Privacy-Preserving Data Mining. *The Journal of Privacy and Confidentiality*, Vol. 1, No. 1, pp. 59–98, 2009.