Vol.2019-CSEC-86 No.7
Vol.2019-SPT-34 No.7
2019/7/23

対象者の人数と人間関係に制約のない 移動履歴とSNSアカウントの照合

大岡 拓斗[1,a)]　松本 瞬[1]　市野 将嗣[1]　緑川 耀一[2]　吉井 英樹[2]　吉浦 裕[1,b)]

概要：移動履歴のプライバシーリスクを評価するために，移動履歴と SNS アカウントの照合方式を提案，評価した．先行研究には，移動履歴と SNS から交友関係が観測可能であること，移動履歴と SNS の対象者群が同一であることという制約があった．提案手法は，個々の移動履歴と SNS アカウントを機械学習でモデル化して類似度を判定することで，これらの制約を解消した．53 人の被験者の移動履歴と SNS アカウントを用いた評価により，被験者の SNS アカウントを 10 万の不特定多数のアカウントに混ぜ込んでも，1 週間の移動履歴 53 件のうち 9 件について本人の SNS アカウントを 100 アカウント以内に絞り込めることを明らかにした．

キーワード：プライバシー，個人情報，移動履歴，SNS

1. はじめに

個人の位置を時系列で記録した移動履歴が産業や公共分野で有効利用されている．例えば，移動履歴を元に商品や飲食店などをレコメンドする広告最適化サービス [1] がある．しかし，移動履歴は機微なパーソナルデータであり，個人の自宅や通勤通学先，留守時間などが推定できる．そのため，プライバシーを保護しながら有効利用するための技術，制度が重要となる．これらの技術，制度を検討する前提として，移動履歴の機微の度合いを明確化する必要がある．なかでも，移動履歴に関わる最も大きなプライバシーリスクとして，移動履歴から個人が特定されるリスクを明確化する必要がある．

Srivatsa らは，移動履歴と同じ人物の SNS アカウントを特定し，移動履歴からの個人特定のリスクを示した [2]．Srivatsa の手法では複数人の移動履歴間の接触頻度 (同一時間帯，同一地域に滞在していた頻度) と，同じ複数人の SNS アカウント上の交友関係に類似性があることに注目した．この類似性に着目し，移動履歴間の接触関係と SNS アカウント間の交友関係の照合をすることで，個人特定のリスクを示した．

しかし，Srivatsa らの手法には以下の 4 つの制約が存在する．

(1) 対象者が互いに交友関係を持つ
(2) 対象者が物理世界と電子世界の両方で交友関係を持つ
(3) 物理世界と電子世界の交友関係に類似性が存在する
(4) 移動履歴群と SNS アカウント群の被験者が同一集団である

実際の攻撃ではこれらの制約が成立するとは限らない．本稿では，機械学習を用いることで Srivatsa の制約を解消しつつ移動履歴と SNS アカウントを照合する手法を提案し，評価実験を行った．

2. 関連研究

2.1 概要

移動履歴のプライバシーに関する研究は，プライバシーリスクの明確化の研究 [2][3][4][5][6] とプライバシー保護 (匿名化) の研究 [7][8] の 2 つに分けられる．

プライバシーリスクの明確化の研究は，個人が特定されていない移動履歴から該当する個人を特定する研究 (以下，個人の特定) [2][3][4][5] と移動履歴から年齢や性別を推定する研究 (以下，個人属性の推定) [6] に分けられ，このうち本稿に関連するのは個人の特定の研究である．

さらに，個人の特定の研究は移動履歴と SNS を照合する研究 [2] と，二つの移動履歴群の間で同一人物の移動履歴を照合する研究 [3][4][5] がある．本稿に最も関連するの

[1] 電気通信大学大学院情報理工学研究科，Graduate School of Informatics and Engineering, The University of Electro-Communications
[2] ソフトバンク株式会社，〒 105-7317 東京都港区東新橋 1-9-1 東京汐留ビルディング, SoftBank Corp.,Tokyo Shiodome Bldg., 1-9-1, Higashi-shimbashi, Minato-ku, Tokyo 105-7303,Japan.
[a)] oooka.takuto@uec.ac.jp
[b)] yoshiura@uec.ac.jp

は，移動履歴と SNS を照合する研究である．

2.2 移動履歴と SNS の照合

Srivatsa らは，移動履歴と同じ人物の SNS アカウントを特定する手法を示した [2]．文献 [2] では，移動履歴間の接触頻度と SNS 上の交友関係に類似性があることに注目した照合手法を提案した．

文献 [2] の照合手法は，被験者をノード，交友関係をリンクとした人物間の接触関係グラフを生成する．また，SNS の友人関係から SNS アカウントをノード，友人関係をリンクとした人物間の交友関係グラフを生成する．生成された 2 つの関係グラフの照合を行ったところ，被験者のうち 82.0%について本人の SNS アカウントが特定された．SNS アカウントは公開されることが多く，また SNS アカウントから個人を特定する技術があるため，文献 [2] は個人特定のリスクを示した．

しかし，文献 [2] の手法には以下の 4 つの制約が存在する．

(1) 対象者が互いに交友関係を持つ
(2) 対象者が物理世界と電子世界の両方で交友関係を持つ
(3) 物理世界と電子世界の交友関係に類似性が存在する
(4) 移動履歴群と SNS アカウント群の被験者が同一集団である

文献 [2] では，移動履歴群と無関係の集団 27 人を SNS アカウント群に追加し評価実験を行ったが，精度が 25%に下がった．

実際の攻撃では，これらの制約が成立するとは限らない．また，文献 [2] における評価実験はデータとして St Andrews 大学の学生 27 人の移動履歴と Facebook アカウントを用いており，小規模かつ偏りがあるという問題点が存在する．

2.3 移動履歴間の照合

Murakami は，SNS 上の投稿文や追尾によって得られる移動履歴に，欠損した位置情報が存在することを想定した際の照合手法を示した [5]．文献 [5] の手法では，テンソル分解とビタビアルゴリズム・Forward Filtering Backward Sampling アルゴリズムを活用することで，欠損した位置情報を補間しながら学習を行う．また，人物間に共通した特徴も活用しながら学習を行うことで，各人物の移動履歴に含まれる位置情報の件数が少ない場合にも対応できる．文献 [5] の手法は，5.1.2 節で述べる前処理を行い，SNS の投稿文を移動履歴に変換することで移動履歴と SNS の照合も可能になる．

3. 提案方式

3.1 節に提案方式が満たすべき要件を示し，3.2 節に提案方式の具体的なアルゴリズムを示す．

3.1 満たすべき要件

2.2 節で述べた文献 [2] の手法の 4 つの制約を解消するため，以下の要件を設定する．

要件 1: 照合の対象となる人物の間の関係
SNS の対象者同士の交友関係を利用しない．

要件 2: 照合およびその準備に利用可能なデータ
攻撃者は M 人の移動履歴と N 人の SNS アカウントを照合する．M と N の下限は各々 1 である．すなわち，$M:1, 1:N$，および M と N の値が異なる $M:N$ 照合を含む．攻撃者は，照合の準備にあたって，上記の M 個の移動履歴と N 個の SNS アカウント (投稿文を含む) を利用できるとする．

3.2 提案方式

機械学習を用いて，移動履歴と SNS アカウントのモデルを学習する．各モデルを用いて，移動履歴 i と SNS アカウント j が同一人物である確率 P_{ij} を算出する $(1 \leq i \leq M, 1 \leq j \leq N)$．それぞれの i について P_{ij} が最大となる j を求め，これを照合結果とする．

3.2.1 SNS アカウントから疑似的な移動履歴への変換

SNS アカウント j の投稿文から地名を抽出する．その後地名を緯度と経度に変換し，投稿時刻と合わせて疑似的な移動履歴とする．これを疑似移動履歴 j と呼ぶことにする．

3.2.2 学習

M 人の移動履歴のモデル $\mathcal{M}^i$ を学習する $(1 \leq i \leq M)$．モデル $\mathcal{M}^i$ の学習においては，移動履歴 i を正例，移動履歴 $i'(1 \leq i' \leq M, i' \neq i)$ を負例とする．特徴量やアルゴリズムを変えて，K 種類のモデルを学習する．したがって，KM 個のモデル $\mathcal{M}^{ik}$ を学習する $(1 \leq i \leq M, 1 \leq k \leq K, K$ はモデルの種類数)．$M = 1$ の場合は負例がないので学習しない．

同様に N 人の SNS アカウントから変換された疑似移動履歴について，モデル $\mathcal{N}^{jl}$ $(1 \leq j \leq N, 1 \leq l \leq L, L$ はモデルの種類数) を学習する．$N = 1$ の場合，一般に公開されている SNS アカウントを負例として $\mathcal{N}^{1l}$ を作成する．

3.2.3 モデル毎の確率算出

j 番目の疑似移動履歴を $\mathcal{M}^{ik}$ に入力し，i 番目の移動履歴と j 番目の SNS アカウントが同一人物である確率 P_{ijk} を算出する $(1 \leq i \leq M, 1 \leq j \leq N, 1 \leq k \leq K)$．なお，$M = 1$ の場合は，$\mathcal{M}^1$ を学習しないので，P_{ijk} は算出しない．一方，i 番目の移動履歴を $\mathcal{N}^{jl}$ に入力し，i 番目の移動履歴と j 番目の SNS アカウントが同一人物である確率 P'_{ijl} を算出する $(1 \leq i \leq M, 1 \leq j \leq N, 1 \leq l \leq L)$．

3.2.4 統合確率の算出

P_{ijk}，P'_{ijl} を統合して，i 番目の移動履歴と j 番目の SNS アカウントが同一人物である確率 $\mathcal{P}_{ij}$ を算出する．$M = 1$ の場合は，P'_{ijl} のみから $\mathcal{P}_{ij}$ を算出する．

3.2.5 SNS アカウントの特定

各 i について (1) 式により，移動履歴 i と同一人物の SNS アカウント $\hat{j}_i$ を選定する．

$$\hat{j}_i = \arg\max_{1 \le j \le N} \mathcal{P}_{ij} \quad (1)$$

3.3 提案方式の期待される性質

提案方式は，SNS アカウント N 人の所有者の間の交友関係を用いていないため，3.1 節の要件 1 を満たす．また，$M \ge 1, N \ge 1$ の任意の M と N において，$\mathcal{P}_{ij}$ を算出し，移動履歴 i と同一人物の SNS アカウントを選定することができるので，3.1 節の要件 2 を満たす．

4. データセット

4.1 移動履歴

電気通信大学の学生 24 人と一般の被験者 29 人，合わせて 53 人の被験者のスマートフォンの MAC アドレスを元に，Wi-Fi アクセスポイントへのプローブ要求から移動履歴を取得した[*1]．対象期間は 2015 年 1 月 29 日から 2016 年 4 月 17 日までとした．このデータには緯度，経度，時刻が含まれており，一日一人あたり平均 214 件のデータが存在する．

表 1　移動履歴の例

緯度	経度	時刻
35.33917	139.48697	2015/2/6 6:26
35.39559	139.46653	2015/3/10 19:00
35.6988	139.77228	2015/3/12 7:06
35.64999	139.54363	2015/3/19 16:53

4.2 SNS アカウント

移動履歴と同一の被験者 53 名の Twitter[9] アカウントの投稿文を収集した．この投稿は 2016 年 4 月 28 日を起点に最大で 3000 件遡り，地名を含むものだけを抽出したものである．地名を含む投稿数は一人一日あたり，1.972 件であった．また，同時期に国内の公開されている Twitter アカウントから無作為に選んだ 10 万人分のアカウントの投稿文を収集した．地名を含む投稿数は一人一日あたり，2.042 件であった．

5. 実装

5.1 前処理

移動履歴群，SNS アカウント群のデータセットに対してそれぞれ前処理を施す．

[*1] 被験者からは個別に明確な事前同意を得た上で移動履歴の収集を行った．また，本研究は電気通信大学の倫理委員会の審査を経て実施している．

5.1.1 移動履歴群

被験者が立ち止まった場合，同じ位置情報を大量に記録するため，照合精度が低下する恐れがある．そこで，収集した実験データに対して以下の 4 種類の前処理を施した．

- オリジナル： 前処理を施さない
- 10 分間隔： 時間を 10 分区画に区切り，10 分区間内に同じ位置のデータが複数存在するとき，それを一つにする
- 中央時刻： 一定時間同じ場所で計測された場合，1 回だけ計測されたとみなす
- 両端中央時刻： 一定時間同じ場所で計測された場合，3 回だけ計測されたとみなす

5.1.2 SNS アカウント群

SNS アカウントの投稿文は GeoNLP[10] を用いて地名を緯度，経度，時刻の情報に変換する．これを擬似的な移動履歴 (以下，擬似移動履歴) として扱う．また，地名を含む投稿文は 1 日あたり 1.972 件であり，移動履歴のデータ数 (1 日あたり 214 件) と比較すると 100 分の 1 である．よって，投稿文から擬似移動履歴に変換する際に，100 日分の投稿文を 1 日の投稿文として扱う．また，平日と休日は移動傾向が異なるため，平日休日を分けて処理した．

5.2 特徴量

移動履歴，擬似移動履歴のデータセットから 3 種の特徴量を生成した．

5.2.1 場所毎の訪問頻度 (訪問頻度)

移動履歴の分析に場所毎の訪問頻度を用いることは古くから行われている [6]．この手法を参考にし，以下の特徴量を生成した．

(1) 関東区域 1km メッシュ (訪問頻度 1km)
図 1 のように，関東地区を 1km 単位の格子状に区切り，126×126 のメッシュを生成する．移動履歴から，各メッシュへの訪問回数をカウントする．また，各メッシュ内について，訪問回数ではなく訪問有無 (0 または 1) を記録する方式も検討した．全てのメッシュへの訪問回数または訪問の有無の値を横一列に並べ次元数 $126 \times 126 = 15876$ のベクトルとし，これを特徴量として用いた．

(2) 日本全域 5km メッシュ (訪問頻度 5km)
日本地図全域を 5km 単位の格子状に区切り，474×428 のメッシュを生成する．その後，関東区域 1km メッシュと同様の操作を行い，次元数 $474 \times 428 = 202872$ のベクトルとし，これを特徴量として用いた．

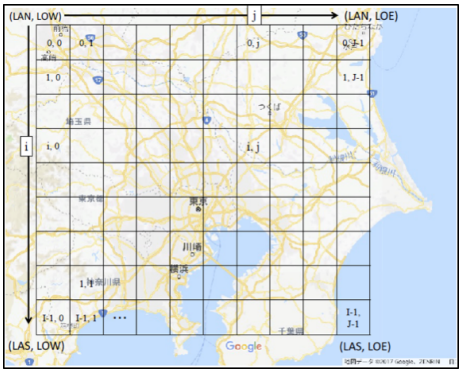

図 1　関東地区 1km メッシュ概要図

5.2.2　ユーザー間の接触 (接触頻度)

文献 [2] の手法を参考にした．2 人のペアが同一時間帯に近くにいた回数を特徴量にした．同一日における 2 人の位置情報を比較する．その時間差，距離差について表 2 の各欄に従い頻度をカウントし，$7 \times 5 = 35$ 次元の特徴ベクトルとした．この処理を (移動履歴の対象者) × (擬似移動履歴の対象者) の組み合わせで行った．

表 2　接触頻度の算出表

時間差 \ 距離差 (km)	0-1	1-2	2-4	4-8	8-16
0 分-10 分					
10 分-20 分					
20 分-30 分					
30 分-1 時間					
1 時間-2 時間					
2 時間-3 時間					
3 時間-6 時間					

5.2.3　特徴量の選定

前処理 4 種，特徴量 3 種，訪問の回数と有無の 2 種，学習アルゴリズム 2 種 (Logistic 回帰 [11]，XGBoost[12]) の全組み合わせについて予備実験を行ったところ，以下の組み合わせが有効であった．

- 方式 1： 前処理 2-関東区域 1km メッシュ-Logistic 回帰-訪問有無特徴量
- 方式 2： 前処理 2-日本全域 5km メッシュ-Logistic 回帰-訪問有無特徴量
- 方式 3： 前処理 1-ユーザ間接触頻度-XGBoost-回数表現特徴量

5.3　モデルの生成と評価

5.3.1　モデルの生成

(1) 方式 1 および 2

図 2 のように移動履歴 i を正例，移動履歴 $i'(1 \leq i' \leq M, i' \neq i)$ を負例として学習させ，モデル $\mathcal{M}^{iK}$ $(1 \leq i \leq M, K = 1, 2$，$K$ は方式の番号) を生成する．これを全ての移動履歴 i に対して行う．なお，$M = 1$ の場合は負例が存在しないので $\mathcal{M}^{1K}$ は学習しない．

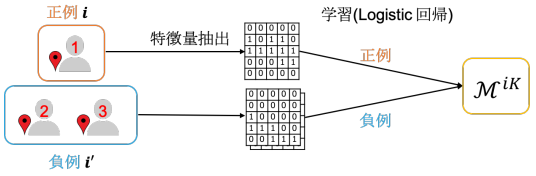

図 2　モデル $\mathcal{M}^{iK}$ の学習 $(K = 1, 2)$

同様に N 人の SNS アカウントから変換された疑似移動履歴について，図 3 のようにモデル $\mathcal{N}^{jL}(1 \leq j \leq N, L = 1, 2$，$L$ は方式の番号) を学習する．$N = 1$ の場合，一般に公開されている SNS アカウントを負例として $\mathcal{N}^{1L}$ を作成する．

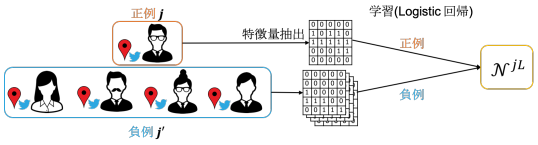

図 3　モデル $\mathcal{N}^{jL}$ の学習 $(K = 1, 2)$

(2) 方式 3

移動履歴 (i, i) のペアを正例，移動履歴 $(i, i')(1 \leq i' \leq M, i' \neq i)$ のペアを負例として，識別器 $\mathcal{M}^{i3}$ を学習する．このとき，特徴量は図 4 のように移動履歴同士の差分から生成する．これを全ての移動履歴 i に対して行う．なお，$M = 1$ の場合は負例が存在しないので $\mathcal{M}^{13}$ は学習しない．

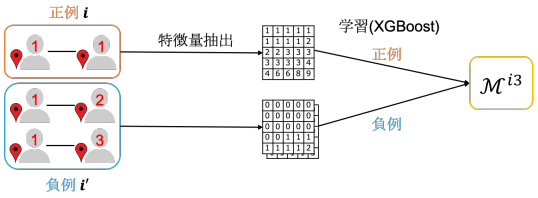

図 4　モデル $\mathcal{M}^{i3}$ の学習

同様に図 5 のように N 人の SNS アカウントから変換された疑似移動履歴 j について，モデル $\mathcal{N}^{j3}(1 \leq j \leq N)$ を学習する．

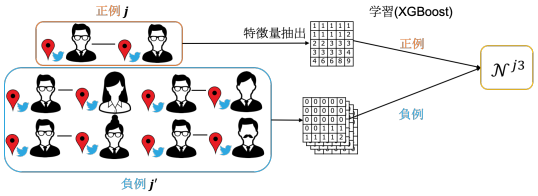

図 5　モデル $\mathcal{N}^{j3}$ の学習

5.3.2 モデルの評価

(1) 方式 1 および 2

3.2.3 節に従いモデル $\mathcal{M}^{iK}(K=1,2)$ に擬似移動履歴 j から生成した特徴量を入力し，同一人物である確率 P_{ijk} を算出する．同様にモデル $\mathcal{N}^{jL}(L=1,2)$ に移動履歴 i から生成した特徴量を入力し，同一人物である確率 P'_{ijl} を算出する．

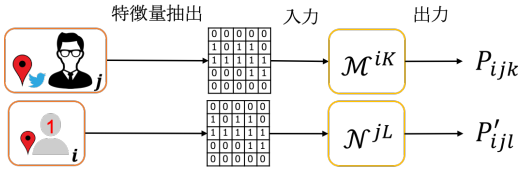

図 6 モデル $\mathcal{M}^{iK}, \mathcal{N}^{jL}$ の評価 $(K, L = 1, 2)$

(2) 方式 3

入力する特徴量は移動履歴 i と擬似移動履歴 j を用いて生成する．方式 1 および 2 と同様に，生成した特徴量をモデル $\mathcal{M}^{i3}, \mathcal{N}^{j3}$ に入力し，同一人物である確率 P_{ij3}, P'_{ij3} を算出する．

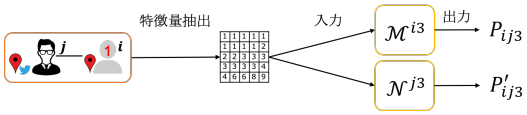

図 7 モデル $\mathcal{M}^{i3}, \mathcal{N}^{j3}$ の評価

算出した確率 P_{ijk}, P'_{ijl} を基に図 8 のような確率表を生成する．

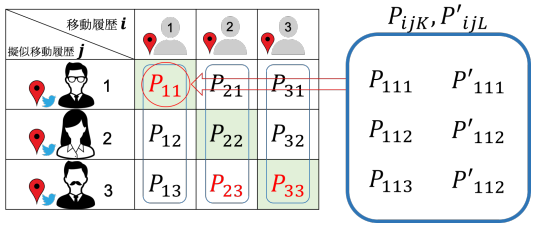

図 8 移動履歴 3 人，擬似移動履歴 3 人とした時の確率表

5.3.3 スコア統合

図 8 のように各モデルから算出された確率表に対して，図 9 のように確率を統合する．その際，方式 1，2，3 のスコアを揃えるため各確率に対して正規化を行った．統合の手法は以下の 2 つを検討した．

(1) 各モデルから算出された確率の平均値を用いる

(2) 各モデルから算出された確率のうち最も高い値を用いる

予備実験を行ったところ，(1) の手法の方が精度が高かった．以降のスコア統合では，(1) の手法を用いることとする．

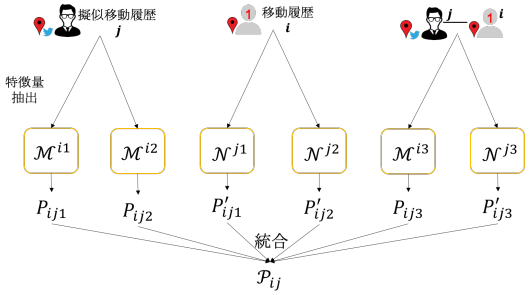

図 9 スコア統合

6. 評価結果

6.1 被験者 53 人を用いた評価

6.1.1 各モデルの評価結果

53 人の移動履歴とその同一人物の 53 人の SNS アカウントから得られた擬似移動履歴の照合を行ったところ，以下のような結果になった．表 3 の結果は図 8 のように確率表を算出し，照合を行った結果である．また，学習アルゴリズムにはランダム性があるため，評価実験は 5 回行いその平均を結果とした．

表 3 方式 1 の評価結果

モデル	一意特定	5 人以内	10 人以内	平均順位
$\mathcal{M}^{i1}$	18.0(34.0%)	33.6(63.4%)	38.2(72.1%)	11.2
$\mathcal{N}^{j1}$	17.2(32.5%)	32.4(61.1%)	38.0(71.7%)	9.3
$\mathcal{M}^{i1}$&$\mathcal{N}^{j1}$	25.2(47.6%)	35.8(67.6%)	41.0(77.4%)	7.6

表 3 の 1 行目は方式 1 を用いたモデルである $\mathcal{M}^{i1}$ の評価結果を示す．移動履歴に対して SNS アカウントが一意に特定できたのは 53 人中 18.0 人であり，5 人に絞り込めたのは 33.6 人，10 人に絞り込めたのは 38.2 人となった．また，SNS アカウントの本人らしさの順位の平均は 11.2 位となった．表 3 の 2 行目はモデル $\mathcal{N}^{j1}$ の評価結果を示す．SNS アカウントが一意に特定できたのは 53 人中 17.2 人であり，5 人に絞り込めたのは 32.4 人，10 人に絞り込めたのは 38.0 人となった．また，SNS アカウントの本人らしさの順位の平均は 9.3 位となった．表 3 の 3 行目はモデル $\mathcal{M}^{i1}$ とモデル $\mathcal{N}^{j1}$ を統合した評価結果を示す．SNS アカウントが一意に特定できたのは 53 人中 25.2 人であり，5 人に絞り込めたのは 35.8 人，10 人に絞り込めたのは 41.0 人となった．また，SNS アカウントの本人らしさの順位の平均は 7.6 位となった．

評価結果より，モデル $\mathcal{N}^{j1}$ よりモデル $\mathcal{M}^{i1}$ の方が全体的な精度が優れている．ただし，$\mathcal{M}^{i1}, \mathcal{N}^{j1}$ の統合をすることで精度が一意特定 25.2 人，5 人への絞り込みが 35.8 人と $\mathcal{M}^{i1}, \mathcal{N}^{j1}$ の評価結果と比べて向上しているため，$\mathcal{M}^{i1}$ では照合できなかった被験者が $\mathcal{N}^{j1}$ では照合に成功しているということが言える．

次に表 4 に方式 2 を用いたモデルである $\mathcal{M}^{i2}, \mathcal{N}^{j2}$ の評

価結果を示す．

表 4 方式 2 の評価結果

モデル	一意特定	5 位以内	10 位以内	平均順位
$\mathcal{M}^{i2}$	19.2(36.2%)	32.8(61.9%)	39.2(74.0%)	9.0
$\mathcal{N}^{j2}$	13.8(26.0%)	30.8(58.1%)	36.2(68.3%)	8.7
$\mathcal{M}^{i2}\&\mathcal{N}^{j2}$	22.6(42.6%)	37.6(70.9%)	42.4(80.0%)	6.2

表 4 より，モデル $\mathcal{N}^{j2}$ よりモデル $\mathcal{M}^{i2}$ の方が全体的な精度が優れており，その差が方式 1 と比べて顕著であることがわかる．また，方式 1 と同様にモデルを統合すると精度が向上するという性質が観測できた．方式 2 は 5km メッシュの情報を用いており，方式 1 は 1km メッシュの情報を用いている．よって，位置精度 5km の移動履歴でも個人の特定に繋がりうるということが言える．

次に表 3 と表 4 を比較した．$\mathcal{M}^{i1}$ の一意特定人数は 18.0 人，$\mathcal{M}^{i2}$ の一意特定人数は 19.2 人となっている．このことから，モデル $\mathcal{M}^{iK}$ は方式 2(全国区域 5km メッシュ) の方が方式 1(関東区域 1km メッシュ) よりも優れており，詳細な情報より粗い情報を特徴量に用いた方が優れた精度となることがわかる．しかし，$\mathcal{N}^{j1}$ の一意特定人数は 17.2 人，$\mathcal{N}^{j2}$ の一意特定人数は 13.8 人となっているため，$\mathcal{N}^{jL}$ は $\mathcal{M}^{iK}$ とは逆の性質を持つことがわかる．

次に表 5 に方式 3 を用いたモデルである $\mathcal{M}^{i3}$,$\mathcal{N}^{j3}$ の評価結果を示す．

表 5 方式 3 の評価結果

モデル	一意特定	5 位以内	10 位以内	平均順位
$\mathcal{M}^{i3}$	17.2(32.5%)	33.6(63.4%)	35.4(66.8%)	9.8
$\mathcal{N}^{j3}$	24.4(46.0%)	34.6(65.3%)	37.0(69.8%)	8.7
$\mathcal{M}^{i3}\&\mathcal{N}^{j3}$	23.4(44.2%)	38.0(71.7%)	39.0(73.6%)	7.4

表 5 より，一意特定の精度はモデル $\mathcal{N}^{j3}$ の方がモデル $\mathcal{M}^{i3}$ より優れており，一方で，一意特定以外の精度はモデル $\mathcal{M}^{i3}$ の方が優れていた．つまり，方式 3 においてモデル $\mathcal{M}^{i3}$ は全体的な精度を上げる性質を持ち，モデル $\mathcal{N}^{j3}$ は一部の被験者の精度を上げる性質を持つことがわかる．また，$\mathcal{N}^{j3}\&\mathcal{M}^{i3}$ を統合すると一意特定の精度がわずかに低下しており，全体的な精度は向上している．

次に方式 1，2，3 のスコア統合を行った．結果を表 6 に示す．

表 6 全方式統合の評価結果 $(K, L = 1, 2, 3)$

モデル	一意特定	5 位以内	10 位以内	平均順位
$\mathcal{M}^{iK}$	29.4(55.5%)	44.4(83.8%)	48.0(90.6%)	5.6
$\mathcal{N}^{jL}$	33.6(63.4%)	42.0(79.3%)	44.4(83.8%)	5.8
$\mathcal{M}^{iK}\&\mathcal{N}^{jL}$	39.2(74.0%)	45.8(86.4%)	47.0(88.7%)	3.1

表 6 より，一意特定の精度はモデル $\mathcal{M}^{iK}(K = 1, 2, 3)$ の統合の方がモデル $\mathcal{N}^{jL}(L = 1, 2, 3)$ の統合より高いことが観測できた．また，5 人，10 人にの絞り込みの精度はモデル $\mathcal{N}^{jL}(L = 1, 2, 3)$ の統合の方が高かった．これは方式 3 特有の性質であるため，$\mathcal{N}^{jL}(L = 1, 2, 3)$，$\mathcal{M}^{iK}(K = 1, 2, 3)$ をそれぞれ統合すると方式 3 の影響力が強くその性質が残るということが言える．また，モデルを全て統合した場合 $(\mathcal{M}^{iK}\&\mathcal{N}^{jL}, K, L = 1, 2, 3)$，どのモデルよりも精度が高いので，ある方式では一意特定できなかった被験者が別の方式では一意特定できていることが分かり，各方式の欠点を互いに補い合っているということが言える．

6.1.2 日数を短縮した場合の評価

6.1.1 節の評価実験は 90 日分の移動履歴を用いていたが，その日数を D 日まで削減して照合を行った．D の値は 1 ヶ月 $(D = 30)$ と，1 週間 $(D = 7)$，1 日 $(D = 1)$ とした．モデルは表 6 で最も精度が高かった $\mathcal{M}^{iK}\&\mathcal{N}^{jL}(K, L = 1, 2, 3)$ を用いた．

表 7 $M : N = 53 : 53$ の日数削減評価

	$D = 90$	$D = 30$	$D = 7$	$D = 1$
一意特定	39.2(74.0%)	27.9(52.6%)	22.5(42.5%)	7.3(13.8%)
5 人以内	45.8(86.4%)	38.6(72.8%)	34.9(65.9%)	12.6(23.8%)
10 人以内	47.0(88.7%)	43.7(82.5%)	40.1(75.7%)	19.9(37.6%)
平均順位	3.1	4.6	6.2	13.7

$D = 90$ の列は表 6 のモデル $\mathcal{M}^{iK}\&\mathcal{N}^{jL}(K, L = 1, 2, 3)$ の評価結果と同一である．$D = 30$ の場合，被験者 53 人のうち一意特定できたのは 27.9 人，5 人，10 人に絞り込めたのは 38.6 人，43.7 人となり，本人らしさの順位の平均は 4.6 位となった．$D = 7$ の場合，被験者 53 人のうち一意特定できたのは 22.5 人，5 人，10 人に絞り込めたのは 34.9 人，40.1 人となり，本人らしさの順位の平均は 6.2 位となった．$D = 1$ の場合，被験者 53 人のうち一意特定できたのは 7.3 人，5 人，10 人に絞り込めたのは 12.6 人，19.9 人となり，本人らしさの順位の平均は 13.7 位となった．

表 7 より，$D = 90$ から $D = 7$ にかけては大きな精度の低下は見られないが，$D = 7$ から $D = 1$ にかけて急激に精度が低下している．このことから，移動履歴が 7 日分あれば被験者の特定につながりうるということが言える．

6.2 大規模評価

不特定多数の Twitter アカウント R 人から擬似移動履歴を生成し，その中に被験者 53 人の擬似移動履歴を埋め込む．その後，6.1.1 節と同様に移動履歴 53 人と擬似移動履歴 $R+53$ 人の照合を行った結果を表 8 に示す．また，モデルは表 6 で最も精度が高かった $\mathcal{M}^{iK}\&\mathcal{N}^{jL}(K, L = 1, 2, 3)$ を用いた．この評価実験は CPU が Intel Core i9-7920X，メモリが 128GB の計算機を用いて行い，総計算時間は 104 時間 21 分 50 秒であった．

表 8 において，$R = 0$ は表 6 のモデル $\mathcal{M}^{iK}\&\mathcal{N}^{jL}(K, L = 1, 2, 3)$ の評価結果と同一である．$R = 1000$ の時，一意特定できたのは被験者 53 人のうち 8.0 人であり，5 人，10 人，100 人に絞り込めたのはそれぞれ 26.0 人，29.0 人，49.0 人となり，$R = 0$ の時と比較すると，全体的に精度が下がっ

ている．

表 8　大規模評価結果

	$R=0$	$R=1000$	$R=10000$	$R=100000$
一意特定	39.2(74.0%)	8.0(15.1%)	0.0(0.0%)	0.0(0.0%)
5 人以内	45.8(86.4%)	26.0(49.1%)	2.8(5.3%)	0.0(0.0%)
10 人以内	47.0(88.7%)	29.0(54.7%)	7.8(14.7%)	0.8(1.5%)
100 人以内	-	49.0(92.5%)	32.4(61.1%)	21.1(39.8%)
1000 人以内	-	-	48.4(91.3%)	34.8(65.7%)
平均順位	3.1	50.8	390.0	3780.2

$R=10000$ の時，一意特定できた被験者は存在しなかった．また，5 人，10 人，100 人，1000 人に絞り込めたのはそれぞれ 2.8 人，7.81 人，32.4 人，48.4 人となった．$R=100000$ の時，一意特定，および 5 人に絞り込めた被験者は存在しなかった．また，10 人，100 人，1000 人に絞り込めたのはそれぞれ 0.8 人，21.1 人，32.4 人となった．

6.2.1　M：N=53：1053 の場合の日数短縮の評価

6.1.2 節と同様の日数短縮の評価を行った．移動履歴 53 人と擬似移動履歴 1053 人の照合を行った結果を表 9 に示す．

表 9　$M:N=53:1053$ の日数削減評価

	$D=90$	$D=30$	$D=7$	$D=1$
一意特定	8.0(15.1%)	8.0(15.1%)	5.9(11.1%)	3.4(6.4%)
5 人以内	26.0(49.1%)	18.9(35.7%)	16.3(30.8%)	8.0(15.1%)
10 人以内	29.0(54.7%)	23.0(43.4%)	22.6(42.6%)	10.2(19.3%)
100 人以内	49.0(92.5%)	37.1(70.0%)	34.5(65.1%)	19.0(35.9%)
平均順位	50.8	135.7	145.7	270.6

表 9 より，6.1.2 節と同様の性質が観測できた．また，5 人，10 人，100 人と絞り込みの人数を増やすにつれ，その性質がより顕著に現れるということが言える．

6.2.2　M：N=53：1 万 53 の場合の日数短縮の評価

移動履歴 53 人と擬似移動履歴 1 万 53 人の照合を行った結果を表 10 に示す．

表 10　$M:N=53:1$ 万 53 の日数削減評価

	$D=90$	$D=30$	$D=7$	$D=1$
一意特定	0.0(0.0%)	1.6(3.0%)	0.6(1.1%)	0(0%)
5 人以内	2.8(5.3%)	5.9(11.1%)	4.3(8.1%)	1.3(2.5%)
10 人以内	7.8(14.7%)	7.8(14.7%)	6.7(12.6%)	2.1(4.0%)
100 人以内	32.4(61.1%)	23.5(44.3%)	22.6(42.6%)	10.3(19.4%)
1000 人以内	48.4(91.3%)	37.9(71.5%)	34.9(65.9%)	18.2(34.3%)
平均順位	390.0	1117.8	1361.7	2536.0

表 10 より，6.2.1 節と同様の性質が観測できた．また，一意特定の項目に注目すると $D=90$ から $D=30$ にかけて精度が向上している．よって，90 日の精度が低い場合，例外的にこのようなケースが観測できるということが言える．

6.2.3　M：N=53：10 万 53 の場合の日数短縮の評価

移動履歴 53 人と擬似移動履歴 10 万 53 人の照合を行った結果を表 11 に示す．表 11 より，6.1.2 節と同様の性質が現れている．また，5 人，10 人以内の項目に注目すると，日数を短縮することで精度が向上するという 6.2.2 節でも見られた例外的ケースが観測できた．

表 11　$M:N=53$:10 万 53 の日数削減評価

	$D=90$	$D=30$	$D=7$	$D=1$
一意特定	0.0(0.0%)	0.1(0.2%)	0(0%)	0(0%)
5 人以内	0.0(0.0%)	0.9(1.7%)	1.4(2.7%)	0(0%)
10 人以内	0.8(1.1%)	2(3.8%)	2.6(4.9%)	1.1(2.1%)
100 人以内	21.1(39.8%)	9.6(18.1%)	9.1(17.2%)	2.4(4.5%)
1000 人以内	34.8(65.7%)	24.3(45.9%)	25.2(47.6%)	11.9(22.5%)
平均順位	3780.2	10978.0	12698.6	25647.4

6.3　1 対多の照合

6.3.1　M：N=53：1 の照合

表 12 は $M:N$=53 : 1 の照合結果を表す．擬似移動履歴の数 N が 1 の時，モデル $\mathcal{N}^{jL}$ を生成することができない．よって，表 12 はモデル $\mathcal{M}^{i1}\&\mathcal{M}^{i2}\&\mathcal{M}^{i3}$ の統合を用いた．

表 12　$M:N$=53 : 1 の評価結果

	一意特定	5 位以内	10 位以内	平均順位
$M:N$=53 : 1	26.8(50.7%)	45.4(85.7%)	49.8(93.7%)	3.5

表 6 の $\mathcal{M}^{iK}(K=1,2,3)$ の精度と表 12 を比較すると，一意特定の特定率のみわずかに低下しており，それ以外の項目の精度はほぼ変わらないことが観測できた．

6.3.2　M：N=1：53 の照合

6.3.1 節と同様に $M:N$=1 : 53 の照合を行った結果を表 13 に示す．移動履歴の数 M が 1 の時，モデル $\mathcal{M}^{iK}$ を生成することができないため，$\mathcal{N}^{j1}\&\mathcal{N}^{j2}\&\mathcal{N}^{j3}$ の統合を用いた．

表 13　$M:N$=1 : 53 の評価結果

	一意特定	5 位以内	10 位以内	平均順位
$M:N$=1 : 53	23.8(44.9%)	28.2(53.2%)	37.0(69.8%)	9.5

表 6 の $\mathcal{M}^{jL}(L=1,2,3)$ の精度と表 13 を比較すると，全体的に精度が低下しているが，一意特定の精度の低下は見られなかった．

7.　まとめ

本稿は移動履歴と SNS アカウントを照合する手法を提案し，Wi-Fi 基地局から得られた移動履歴と Twitter アカウントを用いた評価実験を行った．文献 [2] では，移動履歴と SNS の交友関係に類似性があることに着目した手法を提案していたが，この手法には，物理世界と電子世界の交友関係に類似性があり，かつ移動履歴群と SNS アカウント群の被験者が同一集団であるという制約がある．提案手法

は，移動履歴毎，SNS アカウント毎に独立にモデルを生成し，類似度を判定するため文献 [2] の制約を解消している．また，文献 [2] の評価実験では，データとして St Andrews 大学の学生 27 人データを用いており，小規模かつ偏りが存在したが，本稿では電気通信大学生 29 名，一般人 23 名を合わせた 53 人のより多様な被験者で評価実験を行った．

移動履歴から生成するモデル，SNS アカウントの投稿文から生成するモデル，およびこの 2 つのモデルの平均値をとる手法を評価したところ，平均値をとる手法が最も精度が高かった．この手法を用いて，移動履歴の被験者数：SNS アカウントの被験者数=53：53 で照合を行ったところ，移動履歴の 74.0%について同一人物の SNS アカウントを一意特定できた．また，移動履歴の位置精度が 5km の場合でも，42.6%について一意特定できた．よって，位置精度を 5km に匿名化した移動履歴でも，個人特定のリスクが十分大きいと言える．上述の評価実験では 90 日分の移動履歴を用いたが，移動履歴の日数を 7 日に短縮し評価実験を行ったところ，被験者のうち 42.5%を一意特定できた．よって，公開される移動履歴が 1 週間以上あれば，個人特定のリスクが十分大きいと言える．

次に被験者の SNS アカウントを 10 万人の不特定多数の SNS アカウントに埋め込み，53：10 万 53 の照合を行ったところ，被験者のうち 39.8%が 100 人に絞り込めた．移動履歴の日数を 7 日に短縮したところ，被験者のうち 17.2%が 100 人に絞り込めた．よって，対象者の規模が 10 万人であっても，公開される移動履歴が 1 週間以上あれば，個人特定のリスクが存在すると言える．また，この評価実験では 10 万 53 人のモデルの学習およびテストを実行したが，CPU が Intel Core i9-7920X，メモリが 128GB の計算機を用いて計算をしたところ，かかった日数は約 4 日 (104 時間) であった．

さらに，文献 [2] では不可能であった 1：N の照合が可能となった．被験者 53 人の移動履歴群と SNS 群のセットに対し，1：53 の照合を 53 人分行ったところ，被験者のうち 44.9%を一意特定できた．53：53 の場合と比較しても，精度の大幅な低下は見られないため，公開される移動履歴が一人分でも個人特定のリスクが十分大きいと言える．

参考文献

[1] Google: GoogleAds, , available from ⟨https://ads.google.com⟩ (accessed 2019-04-11).
[2] Srivatsa, M. and Hicks, M.: Deanonymizing mobility traces: Using social network as a side-channel, *Proceedings of the 2012 ACM conference on Computer and communications security*, ACM, pp. 628–637 (2012).
[3] Ma, C. Y., Yau, D. K., Yip, N. K. and Rao, N. S.: Privacy vulnerability of published anonymous mobility traces, *IEEE/ACM transactions on networking (TON)*, Vol. 21, No. 3, pp. 720–733 (2013).
[4] Shokri, R., Theodorakopoulos, G., Le Boudec, J.-Y. and Hubaux, J.-P.: Quantifying location privacy, *2011 IEEE symposium on security and privacy*, IEEE, pp. 247–262 (2011).
[5] Murakami, T.: Expectation-maximization tensor factorization for practical location privacy attacks, *Proceedings on Privacy Enhancing Technologies*, Vol. 2017, No. 4, pp. 138–155 (2017).
[6] 松尾豊，岡崎直観，中村嘉志ほか：位置履歴からのユーザ属性の推定，情報処理学会論文誌，Vol. 48, No. 6, pp. 2106–2117 (2007).
[7] Oya, S., Troncoso, C. and Pérez-González, F.: Back to the drawing board: Revisiting the design of optimal location privacy-preserving mechanisms, *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, ACM, pp. 1959–1972 (2017).
[8] Chatzikokolakis, K., Elsalamouny, E. and Palamidessi, C.: Efficient utility improvement for location privacy, *Proceedings on Privacy Enhancing Technologies*, Vol. 2017, No. 4, pp. 308–328 (2017).
[9] Twitter: Twitter, , available from ⟨https://twitter.com⟩ (accessed 2019-04-11).
[10] 国立情報学研究所：GeoNLP-文章を自動的に地図化する地名情報処理システム, , 入手先 ⟨https://geonlp.ex.nii.ac.jp⟩ (参照 2019-04-11).
[11] Hosmer Jr, D. W., Lemeshow, S. and Sturdivant, R. X.: *Applied logistic regression*, Vol. 398, John Wiley & Sons (2013).
[12] Chen, T. and Guestrin, C.: XGBoost: A scalable tree boosting system, *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, ACM, pp. 785–794 (2016).