情報処理学会論文誌 Vol.61 No.12 1814–1830 (Dec. 2020)

対象者の人数と対象者間の関係に制約のない移動履歴とソーシャルネットワークの照合方式

松本 瞬[1,a)] 大岡 拓斗[1] 市野 将嗣[1] 吉浦 裕[1]

受付日 2020年3月9日, 採録日 2020年9月10日

概要：仮名化および曖昧化された移動履歴からの対象者の再特定は，移動履歴の利活用における重大な脅威であるため，そのリスクを明らかにするために再特定の方法が多数研究されている．その中で，ソーシャルネットワークとの照合を通じて移動履歴の対象者を再特定する方法は有望なアプローチであるが，従来方式は，移動履歴間の接触関係とソーシャルネットワークのリンク関係が利用できない場合，移動履歴数およびソーシャルネットワークアカウント数が小さい場合，両者が大きく異なる場合には対応できなかった．本論文では，接触関係およびリンク関係を利用することなく，移動履歴とソーシャルネットを照合する方式を提案し，実データを用いた評価により，移動履歴数およびソーシャルネットワークアカウント数が小さい場合，両者が大きく異なる場合に対応できることを明らかにする．提案方式は，時間情報の利用，移動履歴とソーシャルネットワークの両者のモデルの併用，粒度の異なる 2 つのモデルの併用の点にも新規性を有する．多数の候補者のアカウントの中から移動履歴と同一人物のアカウントを見つけるという新しい視点の評価も行い，その評価での提案方式の有効性も明らかにする．

キーワード：プライバシリスク，移動履歴，再特定，ソーシャルネットワーク

Linking Location Histories and Social Networks without Assuming Target Persons' Number and Relation

Shun Matsumoto[1,a)] Takuto Ooka[1] Masatsugu Ichino[1] Hiroshi Yoshiura[1]

Received: March 9, 2020, Accepted: September 10, 2020

Abstract: Re-identifying people from location histories has been actively studied to clarify the risk of using location histories. Though using social networks as side data for this purpose is a promising approach, a representative previous method of this approach works only under the conditions that contact among the people are observable in the location histories and links among corresponding social network accounts are observable, the numbers of location histories and of the social network accounts are not too small, and the two numbers are nearly the same. In this paper, we describe a method of re-identifying people from location histories without using contact and link relations and experimentally demonstrate its effectiveness when the conditions on which the previous method depends do not hold. The proposed method also has its originality in using time information, dual models of the location histories and social networks, and models with different granularities. We also show the effectiveness of our proposed method by a new type of evaluation where the persons represented by location histories are identified from a large number of candidates.

Keywords: privacy risk, location history, re-identification, social network

1. はじめに

個人の位置を時系列で記録した移動履歴は，商品や飲食店の広告の最適化，交通システムの改善，災害時の避難計画等の様々な分野で有効利用されている．しかし，移動履

[1] 電気通信大学
The University of Electro-Communications, Chofu, Tokyo 182–8585, Japan
[a)] s.matsumoto@uec.ac.jp

© 2020 Information Processing Society of Japan

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，人間関係等が推定できる．そのため，移動履歴の利用にあたっては，仮名化や曖昧化（obfuscation）等の加工を行う場合が多い [1], [2]．しかし，これらの加工を行った場合でも，移動履歴を同じ対象者の別のデータと照合することにより，対象者を再特定できることが知られている．

移動履歴からの対象者の再特定は 2010 年ごろから活発に研究されている [3], [4], [5], [6], [7], [8], [9], [10]．なかでも，Srivatsa らは，移動履歴と同じ人物のソーシャルネットワークアカウントを特定する方式を提案している [3]．ソーシャルネットワークのアカウントおよびデータは公開されることが多い．また，ソーシャルネットワークのアカウントは実名の場合があり，匿名の場合でもユーザを再特定する方式が多数提案されている [11], [12], [13], [14]．そのため，ソーシャルネットワークアカウントを用いた再特定は，移動履歴からの再特定のアプローチとして有望である．しかし，Srivatsa らの方式は下記の 3 つの前提を必要とする．

(1) 再特定の複数の対象者は，移動履歴から抽出可能な物理世界の接触関係を持ち，ソーシャルネットワークではリンクによる関係を持つ．
(2) 物理世界の接触関係とソーシャルネットワーク上のリンク関係は照合可能である．
(3) 移動履歴数とソーシャルネットワーク数が少なすぎず，かつほぼ等しい．

これらの前提が実際の攻撃の際に成立するとは限らない．

移動履歴からの再特定において，時間情報は有効な手がかりになると予想される．しかし，従来方式における時間情報の利用は，上記の Srivatsa らの方式における接触関係の抽出および，2 つの場所への訪問時刻の前後関係に基づく場所間遷移の推定 [4], [5], [7] に限られる．

本論文では，ソーシャルネットワークアカウントを用いた移動履歴からの再特定方式を提案する．提案方式は，アカウントの投稿文の内容を移動履歴と比較し，アカウントと移動履歴の類似性を定量化する．そして，移動履歴に最も類似したアカウントを，同じ人物のアカウントとして特定する．提案方式は，移動履歴の接触関係とソーシャルネットワークのリンク関係を照合せず，移動履歴ごとに独立して類似アカウントを選定するので，Srivatsa らの方式の 3 つの前提条件を不要化できる．さらに，時間情報の新たな利用方法により，再特定の精度を向上させる．

本論文の構成は以下のとおりである．2 章では先行研究とその問題点を述べる．3 章では提案方式の概要，4 章では提案方式の詳細を述べる．5 章ではデータセットと実験方法について述べる．6 章では実装と評価について述べる．7 章では提案方式の優位性について考察し，8 章では結論と今後の課題について述べる．

2. 先行研究

2.1 概要

移動履歴のプライバシへの攻撃には，再特定だけでなく，ある時刻にどこにいたかを加工された移動履歴から推定 [4], [5]，目的地の推定 [15], [16]，複数の人物が出会ったことの推定 [4]，対象者の属性の推定 [17] 等があるが，ここでは再特定について調査する．

移動履歴からの再特定方式は，グラフの照合による方式 [3], [10]，場所間の遷移に基づく方式 [4], [5], [7]，場所ごとの訪問回数に基づく方式 [8], [9]，ノイズの位置の確率分布に基づく方式 [6] に分類できる．

再特定は，移動履歴を同一人物の別のデータに対応付けることで行われる．この別データはサイドデータと呼ばれる．サイドデータの対象者は既知であり，その人物が移動履歴の対象者と判定される．各方式の用いるサイドデータには，ソーシャルネットワーク [3]，移動履歴 [5], [6], [8], [9], [10]，攻撃者の知識 [4], [7] がある．サイドデータが移動履歴である場合には，再特定の対象である（匿名の）移動履歴とサイドデータである（実名の）移動履歴の間の対応付け，すなわち移動履歴間の対応付けとなる．方式 [4], [7] では，サイドデータを攻撃者の知識として抽象的に扱っている．これらの方式では，攻撃者が再特定対象者の移動に関する知識（住所，勤務先，通勤ルート，買い物先等）を何らかの手段で取得することを前提とし，それを場所間の遷移等の形でモデル化する．従来方式の分類を**表 1** に示す．

以下では，方式の分類に沿って論じることとし，最初にグラフの照合による方式を述べる．Srivatsa らは，複数の

表 1 従来方式の分類
Table 1 Classification of conventional methods.

方式	サイドデータ	方式例
グラフの照合	ソーシャルネットワーク（アカウント間のリンク関係）	Srivatsa et al. [3]
	移動履歴（位置情報のない場所の仮名および時刻情報）	Manousakas et al. [10]
場所間の遷移の整合性を判定	攻撃者の知識（位置と時刻の情報）	Shokri et al, [4]
	移動履歴（位置と時刻の情報，欠損の多いものに対応）	Murakami [5]
	攻撃者の知識（位置と時刻の情報）	Gamps et al. [7]
場所毎の訪問回数の整合性を判定	移動履歴（位置情報を含む複数の情報サービスの利用履歴）	Riederer et al. [8]
	移動履歴（位置と時刻の情報，欠損の多いものに対応）	Murakami [9]
ノイズ位置の確率分布との整合性を判定	移動履歴（位置と時刻の情報，ノイズ付加前の移動履歴）	Ma et al. [6]

© 2020 Information Processing Society of Japan

移動履歴間の接触関係のグラフと複数のソーシャルネットワークアカウント間のリンク関係のグラフを照合することにより，移動履歴と同一人物のソーシャルネットワークアカウントを特定した [3]．しかし，Srivatsa らの方式は，1 章で述べた 3 つの前提を必要とする．この問題については，2.2 節で詳しく分析する．Manousakas らは，移動履歴から，場所をノードとし場所間の遷移をエッジとするグラフを生成し，グラフ間の照合により，同一人物の複数の移動履歴を対応付けた [10]．Manousakas らの方式の特徴は，場所が位置情報のない仮名であっても再特定できることにある．Manousakas らの方式以外は，場所は何らかの位置情報をともなうことを前提としている．

次に，場所間の遷移に基づく方式について述べる．Shokri らは，位置の低解像度化，ダミーデータの付加等の移動履歴の曖昧化（obfuscation）を想定し，曖昧化された移動履歴と人物に関する知識を対応付けることで，移動履歴の対象者を特定した [4]．具体的には，再特定対象者の位置や移動に関する知識（自宅や勤務先等）から，場所間の遷移確率行列の形で，移動の傾向（移動プロファイル）を生成する．各移動履歴と各人物の移動プロファイルのペアについて，移動プロファイルを条件とし，移動履歴の条件付き生起確率を算出することで，移動履歴と移動プロファイルを対応付けた．攻撃者の知識が乏しく遷移確率行列に欠落が多い場合には，Gibbs サンプリングによって行列を補った．しかし，攻撃者の知識が非常に乏しい場合には，Gibbs サンプリングによる補完を行っても，有効な遷移確率行列（移動プロファイル）を生成できなかった．Murakami は，欠落の多い移動履歴から場所間の遷移確率行列を生成可能にした．個人ごとの遷移確率行列をまとめたテンソルに対して，EM アルゴリズムを用いたテンソル分解を適用することで，他人の遷移確率行列の情報を利用して各人の行列を補完した [5]．Gambs らは，移動履歴の 2 つの集合について，各々の移動履歴から場所間の遷移確率行列を生成し，行列間の類似度に基づいて，2 つの集合から同一人物の移動履歴を特定した [7]．

場所ごとの訪問回数に基づく方式について述べる．Riederer らは，複数の位置情報サービスの利用履歴から得られた移動履歴の集合から，同じ人物の移動履歴を推定した [8]．自然な移動履歴では，場所ごとの訪問回数がポアソン分布に従うとし，移動履歴のペアが同一人物に属すると仮定した場合と異なる人物に属すると仮定した場合の生起確率の比に基づいて，同じ人物の移動履歴を推定した．Murakami は，背景知識が少ない場合の移動プロファイルとして，場所間の遷移確率の行列よりも，場所ごとの訪問確率のベクトルの方が有効であることを明らかにした [9]．

最後に，ノイズの位置の確率分布に基づく方式について述べる．Ma らは，ランダムノイズに沿って位置を変更した移動履歴を対象とし，元の移動履歴に対応付けた [6]．ノイズの確率分布を想定し，元の移動履歴からの曖昧化された移動履歴の生起確率が最大になるように対応付けた．

提案方式は表 1 のいずれの分類にも属さない．場所ごとの訪問回数に加えて，時間と距離の関係に基づく方式であり，サイドデータとしてソーシャルネットワークの投稿文を利用する．3 章以下で述べるように，ソーシャルネットワークの投稿文から地名を抽出し，地名ごとの訪問回数を推定し，移動履歴における場所ごとの訪問回数と照合することで，Srivatsa らの方式の 3 つの前提を不要化する．また，時間と距離の関係に基づき，時間情報を有効利用することで再特定の精度を向上させる．

2.2 ソーシャルネットワークとの照合

提案方式に最も近い Srivatsa らの方式 [3] を詳しく分析する．なお，以下ではソーシャルネットワークのアカウントを単にアカウントと呼ぶことにする．Srivatsa らの方式は，移動履歴の集合とアカウントの集合を対象とする．移動履歴の集合から，ノードが移動履歴の対象者，エッジが対象者間の接触関係であるようなグラフを生成する．2 人の移動履歴が，位置と時刻が両方とも近いデータを含む場合に接触関係があるとする．一方，アカウント集合からは，アカウント間のリンクのグラフを抽出する．2 つのグラフを照合し，対応付けられたノードのペアが同一人物の移動履歴とアカウントであると推定する．

1 章で述べたように，アカウントとの照合を通じて移動履歴を再特定するアプローチは有望であるが，Srivatsa らの方式は下記の前提条件を必要とし，これが実際の攻撃では大きな制約になると考えられる．

(1) 再特定の複数の対象者は，移動履歴から抽出可能な物理世界の接触関係を持ち，ソーシャルネットワークではリンクによる関係を持つ．
(2) 物理世界の接触関係とソーシャルネットワーク上のリンク関係は照合可能である．
(3) 移動履歴数とソーシャルネットワーク数が少なすぎず，かつほぼ等しい．

(1) について，再特定の対象者間に社会的な交流がない場合には，アカウント間のリンクがないので方式が成立しない．また，たとえば同じ大学の出身者が全国に分散している場合，ソーシャルネットワーク上でリンク関係があっても，物理的にはほとんど接触はない．逆に，職場の同僚の場合，物理的な交友関係は頻繁であるが，ソーシャルネットワーク上ではつながっていない場合がある．

(2) について，両親と子供たちが同居している場合，全員が相互に物理的接触関係を持つが，LINE 上では大人同士，子供同士がつながり，大人と子供はつながっていない場合もある（たとえば，大人は子供に見せたくない会話をするため）．さらに，対象者が実際には物理的な接触関係を持っている場合でも，位置観測システムの粒度によっては，

© 2020 Information Processing Society of Japan

移動履歴のデータからは接触関係の有無を判断できない場合がある．たとえば，携帯電話の基地局からユーザの位置を測定する場合，位置の誤差が 500 m 程度になる．そのため，2 つの携帯電話が同じ基地局圏内にあったとしても，携帯電話を所有する 2 人が接触関係にあったとは判断できない．さらに，曖昧化によって移動履歴の位置や時間が加工されている場合にも，接触関係の有無は判断できない．

(3) について説明する．再特定の対象者が 3 人であり互いに親友である場合を考える．その場合，移動履歴から生成される接触グラフとアカウントから生成されるリンクグラフは，両方とも，三角形になる．2 つの三角形グラフの間には 6 つの同型写像が同じ確率で可能であるため，移動履歴の人物とアカウントの人物の対応関係はまったく推定できない．また，5 人の対象者の移動履歴があり，これらの 5 人が 1,000 人の候補者に含まれる場合を考える．5 人の接触グラフは，1,000 人のリンクグラフの様々なサブグラフに一致するため，人物間の対応関係は特定できない．

2.3 時間情報の利用

移動履歴は，位置と時刻の系列で表現される．そのため，位置情報と時間情報を利用することで再特定の精度を向上させられる可能性がある．2.1 節で述べたように，従来方式は，グラフの照合による方式 [3], [10]，場所間の遷移に基づく方式 [4], [5], [7]，場所ごとの訪問回数に基づく方式 [8], [9]，ノイズの位置の確率分布に基づく方式 [6] に分類できる．グラフの照合のうち Srivatsa らの方式では，2 人の移動履歴が，位置と時刻が両方とも近い要素を含む場合に，2 人は接触関係があると推定している．場所間の遷移に基づく方式は，場所を訪問した時刻の前後関係を利用している．従来方式における時間情報の利用は，筆者らの知る限り，これらの時間と場所の近さによる接触関係の推定および時刻の前後関係による遷移の推定に限られる．

Shokri らは，時間情報の利用について，時間帯ごとに遷移確率行列を生成することを示唆している [4]．たとえば，同じ人物でも平日と休日では移動の傾向が異なるので，遷移確率行列を平日と休日に分けて生成することが考えられる．しかし，Shokri らは，具体的な方法は示していない．時間帯として，平日と休日，午前と午後，1 週間ごと，季節ごと等様々な可能性がある．また，時間帯ごとに遷移確率行列を生成すれば，個々の行列の生成に利用可能なデータ（攻撃者が移動プロファイルを生成するための情報）が少なくなるため，正確な行列が生成できない可能性がある．これらの問題のため，時間帯ごとの行列生成は容易ではない．また，Shokri らは，短時間に遠距離の移動は困難であることを利用して，照合対象を絞り込むことを示唆している．しかし，この点についても具体的な方法は示しておらず，また，具体化は容易ではない．

3. 提案方式の概要

3.1 目標

従来方式の問題点を解決するために以下を目標とした．

目標 1：再特定の対象者に関する物理世界の接触関係とソーシャルネットワークのリンク関係のうち，一方または両方の情報が利用できなくても対応できる．

目標 2：移動履歴数およびアカウント数に制約がない．すなわち，移動履歴数とアカウント数のうち一方または両方が小さい場合，両者が大きく異なる場合にも対応できる．

目標 3：時間と場所の近さによる接触関係の推定および場所の訪問時刻の前後関係による遷移の推定以外の時間情報の利用を可能とし，再特定の精度を向上させる．

3.2 構成と処理の概要

提案方式は，再特定の対象である 1 つ以上の移動履歴とサイドデータである 1 つ以上のアカウントデータを入力する．そして，各移動履歴について，同じ人物に属すると推定されるアカウントを出力する．なお，本方式の扱う移動履歴は，**図 1** のような 4 つ組（仮名，緯度，経度，時刻）の系列である．ここで，仮名は（未知の）人物の ID であり，同じ人物の 4 つ組には同じ仮名を用いる．本方式の構成と処理を**図 2** に示す．

仮名	緯度	経度	時刻	
6	35.65703	139.71451	2017/1/25	6:16:35
…	…	…	…	
6	35.33917	139.48697	2017/2/6	6:26:08
6	35.39559	139.46653	2017/2/6	6:27:42
6	35.6988	139.77228	2017/2/6	6:30:59
6	35.64999	139.54363	2017/3/19	16:53:02

図 1 移動履歴の例

Fig. 1 Example location history.

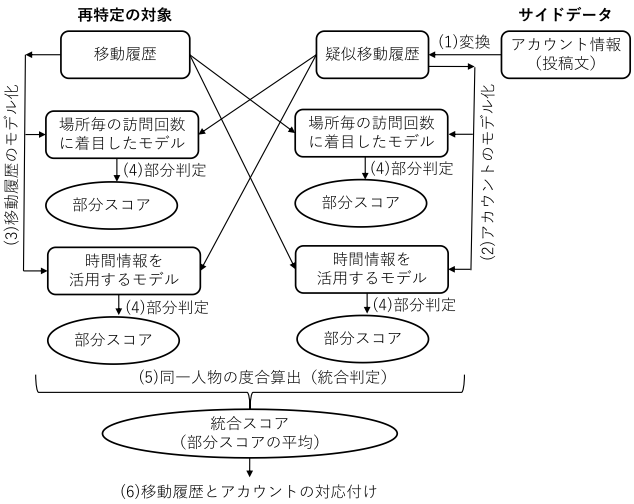

図 2 提案方式の構成と処理

Fig. 2 Structure and processes of proposed method.

（注）図中の (1)～(6) は本文中のステップ (1)～(6) を表す．

© 2020 Information Processing Society of Japan

(1) アカウント情報の疑似移動履歴への変換

各アカウントの投稿文に含まれる地名を位置情報（緯度と経度）に変換し，位置と投稿時刻の組合せからなる疑似移動履歴を生成する．このようにして，アカウント情報を疑似移動履歴に変換し，再特定の対象である移動履歴と照合する．地名から位置情報への変換には GeoNLP [18] を用いた．

(2) アカウントのモデル化

アカウントから生成した疑似移動履歴を，場所ごとの訪問回数に着目してモデル化する．このモデルは，ステップ (4) において，移動履歴とアカウントが同一人物に属するか判定するために用いる．加えて，時間情報を活用するためのモデルも生成する．このモデルは，① 時間が近く場所が遠い 2 つの移動履歴は同一人物ではない可能性が大きいこと，② 時間と場所の両方が近い 2 つの移動履歴は同一人物に属する可能性が大きいことの 2 点に着目し，移動履歴とアカウントが同一人物に属するか判定するために用いる．

(3) 移動履歴のモデル化

アカウントモデルの生成と同様の方法により，移動履歴から，場所ごとの訪問回数に基づくモデルおよび時間情報を活用するモデルを生成する．これらのモデルも，移動履歴とアカウント（疑似移動履歴）が同一人物に属するか判定するために用いる．

(4) 同一人物に属する度合の算出（部分判定）

移動履歴とアカウントのすべてのペアについて，移動履歴がアカウントのモデルに整合する度合および，アカウントが移動履歴のモデルに整合する度合を算出する．

(5) 同一人物に属する度合の算出（統合判定）

移動履歴とアカウントのすべてのペアについて，上記の複数の部分スコアを統合し，移動履歴とアカウントが同一人物に属する度合（統合スコア）を求める．複数の部分スコアから統合スコアを求める方法として平均を用いた．

(6) 移動履歴とアカウントの対応付け

各移動履歴に対して，上記の統合スコアが最大となるアカウントを対応付ける．

上記のステップ (2)，(3) で述べたように，提案方式では，物理世界の接触関係とアカウントのリンク関係のいずれも用いないので，目標 1 を満たし，1 章で述べた Srivatsa らの手法の前提 (1) と (2) を不要化することができる．また，接触関係とリンク関係を用いないことの帰結として，提案方式は，接触グラフとリンクグラフの照合を行わない．2.2 節で述べたように，移動履歴数とアカウント数が小さすぎず，かつほぼ等しいという制約はグラフの照合に起因するので，提案手法は目標 2 を満たし，Srivatsa らの方式の前提 (3) を不要化すると考える．さらに，上記 (2) の ①，② のように時間情報を活用することで，提案手法は目標 3 を満たし，Srivatsa らの手法，Shokri らの手法，Murakami の手法を含む既存手法の問題点（時間情報を利用が限定的であること）を解決すると考える．これらの効果については，6 章で評価する．

4. 提案方式の詳細

本章では，用語を定義した後，これらの用語を用いて，3.2 節の概要のうちのステップ (2)〜(5) の詳細を述べる．また，本方式の特徴の 1 つである両側モデルについて述べる．

4.1 用語の定義

- 移動履歴：位置と時刻からなる要素の系列．
- 移動履歴の要素（要素と略す）：(仮名，緯度，経度，時刻) の 4 つ組．同一人物の複数の要素は同じ仮名となる．仮名の明記が不要の場合は省略して，(緯度，経度，時刻) と略す．また，緯度と経度をまとめて位置とし，(仮名，位置，時刻) あるいは (位置，時刻) と略す場合がある．
- M，N：移動履歴数およびアカウント数．データセットの中に同一人物の複数の移動履歴および複数のアカウントはないことを前提とする．そのため，M，N は移動履歴およびアカウントの対象者数に等しい．
- 移動履歴 i，アカウント j：i 番目の移動履歴および j 番目のアカウント
- Same(i, j)：移動履歴 i の対象者とアカウント j の対象者が同一人物であること
- Score(i, j, <算出方法>)：Same(i, j) の確からしさを表す数値．大きいほど確からしい．提案方式では，複数の方法を用いて数値を算出するので，算出方法を区別するために <算出方法>) を付記する．
- Score(i, j)：複数の算出方法の結果である Score(i, j, <算出方法>) を統合して求めた最終的な確からしさ．

4.2 アカウントのモデル化

4.2.1 訪問回数に基づくモデルの生成

Shokri らの方式および Murakami の方式における人物ごとの移動プロファイルに相当するものを，アカウントごとの投稿文（疑似移動履歴）から生成する．Murakami は，攻撃に利用可能な位置情報が乏しい場合には，場所間の遷移確率よりも場所ごとの訪問確率の方が有効な移動プロファイルを作成できることを明らかにしている [9]．提案方式の場合には，5.2 節 (2) 項で述べるように，アカウントの投稿文に含まれる地名が少ないので，Murakami の結論を参考にして，場所ごとの訪問回数に基づいてモデルを生成することにした．Shokri らおよび Murakami を参考にして，モデルの対象となる矩形の地理的領域を想定し，その領域を一定サイズのメッシュに区切って，メッシュごとの訪問回数からモデルを生成する．

© 2020 Information Processing Society of Japan

(1) 多重解像度の利用

モデルの地理的な範囲としては，移動履歴中の場所のほとんどを含む領域（対象者の生活圏に相当）が考えられる．また，メッシュは細かい方が場所の相違を精度良く表すことができる．一方，移動履歴中に普段行く場所から大きく離れた場所がある．これらは，対象者の旅行先や出張先と考えられるが，訪問先で地名を含む投稿を行うことがある．このような遠隔地は他の対象者がほとんど訪問しないので，個人を特定する強い手がかりになると期待できる．しかし，モデルの対象領域に遠隔地を含めると，日本全体等の広い領域になるので，細かいメッシュを設けるとメッシュ数が膨大になる．一方で位置情報は少ないので，モデルの生成のためのデータがスパースになり，有効なモデルが生成できないことが懸念される．そこで，対象者の生活圏程度の狭い領域を細かいメッシュで区切った狭域細粒度モデルと，遠隔地を含む広い領域を大きいメッシュで区切った広域粗粒度モデルを併用することにした．異なる解像度のモデルを併用する点は著者らの新たな提案である．

(2) 機械学習の利用

訪問回数からモデルを生成する方法としては，訪問回数を正規化して確率として用いる，訪問回数を Gibbs サンプリングやテンソル分解，行列分解等により補完する [4], [5], [9]，何らかの確率モデルを仮定し実際の訪問回数にフィッティングすることで確率モデルのパラメータを最適化する [8] 等の方法がありうる．一方，近年，機械学習が急速に発展しており，今後も発展を続けることが期待される．また，機械学習アルゴリズムについてはネット上にライブラリが公開されており，誰もが利用可能である．そこで，機械学習を用いてモデルを生成すれば，提案方式を多くの技術者に利用可能とし，機械学習の精度向上にともなって提案方式の精度も向上すると考え，ここでは機械学習を用いることにした．

(3) モデル生成の詳細

3.2 節ステップ (1) で述べたように，モデルの生成に先立って，アカウントの投稿文は疑似移動履歴に変換されている．全アカウントの疑似移動履歴に含まれる位置情報から，4.2 節 (1) 項の狭域細粒度モデルおよび広域粗粒度モデルの考え方に沿って，各々の対象領域とメッシュサイズを決める．

各アカウント j からの狭域細粒度モデルの生成について述べる．アカウント j の疑似移動履歴の要素集合 S_j を投稿時刻によって部分集合 $S_{j1}, S_{j2}, \cdots, S_{jNj}$ に分割する．ここで Nj は部分集合の個数である．たとえば，1 日の疑似移動履歴を 1 つの部分集合とする．各部分集合 S_{ju} $(1 \leq u \leq Nj)$ の疑似移動履歴を用いて，メッシュごとの訪問回数をカウントし，特徴量とする．すなわち，部分集合 S_{ju} から生成される特徴ベクトルの α 番目の値は，対応する疑似移動履歴における α 番目のメッシュへの訪問回数である．これにより，アカウント j から Nj 個の特徴ベクトルが生成される．以上の処理を $1 \leq j \leq N$ について行う．アカウント j の Nj 個の特徴ベクトルを正例，j 以外のアカウントの $\sum_{1 \leq j' \leq N, j' \neq j} Nj'$ 個の特徴ベクトルを負例として，機械学習によりアカウント j の狭域細粒度モデルを生成する．

同様に各アカウント j から広域粗粒度モデルを生成する $(1 \leq j \leq N)$．

4.2.2 時間情報を活用するモデルの生成

Shokri ら [4] が示唆した時間帯ごとのモデル生成は，生成のための知識やデータを時間帯ごとに分割することになる．本方式の場合は，アカウントの投稿文に含まれる乏しい位置情報をさらに乏しくすることになるので適さない．そこで，同じく Shokri らが示唆した物理的な不可能性（短時間に長距離は移動できないので，2 つの移動履歴の時間が近く距離が遠い場合には，2 つの移動履歴の対象者は異なる可能性が高いということ）を取り上げ，具体化することにした．なお，2.3 節で述べたように，Shokri らは，具体的な方式は示していない．

物理的な不可能性は個人によって現れ方が異なる．たとえば，徒歩で 10 分間に 1 km 移動する確率は，若い人の場合は大きいが，高齢者の場合は小さい．そのため，移動履歴 i が要素（位置 1, 時刻 1）を含み，アカウント j[*1] が要素（位置 2, 時刻 2）を含み，位置 2 が位置 1 から 1 km，時刻 2 が時刻 1 から 10 分離れている場合，Score(i, j) は対象者が若い人の場合は大きく，高齢者の場合は小さい．そのため，物理的不可能性を表すモデルを個々の移動履歴あるいはアカウントについて（すなわち対象者ごとに）生成することにした．

物理的不可能性とは別に，移動履歴 i の要素とアカウント j の要素が，位置と時間の両方においてより近いほど Score(i, j) は大きいと考えられる．さらに，そのような位置と時間の近い要素のペアがより多いほど Score(i, j) は大きいと考えられる．これを時空間近接性と呼ぶことにする．時空間近接性の現れ方および時空間近接性と物理的不可能性の関係も個人によって異なるため，個々の移動履歴あるいはアカウントについて，そのモデルを生成することにした．

(1) 時間と距離の関係に基づくモデル化

図 3 を用いて，提案する時間情報活用モデルの概要を説明する．図 3 において，行 T_k は時間インターバル $[T_k, T_{k+1})$ を表す（$1 \leq k \leq K$，K は時間インターバルの数）．列 D_ℓ は距離のインターバル $[D_\ell, D_{\ell+1})$ を表す（$1 \leq \ell \leq L$，L は距離インターバルの数）．行が T_k で列が D_ℓ のセルは，$[T_k, T_{k+1})$ と $[D_\ell, D_{\ell+1})$ の両者を満たす時空間インターバルを表す．たとえば，あるセルは，[30 分，60 分) かつ [2 km, 4 km) の時空間インターバルを表す．$c_{k\ell}$ は，時空

[*1] 正確にはアカウント j から変換した疑似移動履歴

© 2020 Information Processing Society of Japan

	D_1	…	D_ℓ	…	D_L
T_1	c_{11}		$c_{1\ell}$		c_{L1}
…					
T_k	c_{k1}		$c_{k\ell}$		c_{kL}
…					
T_K	c_{K1}		$c_{K\ell}$		c_{KL}

図 3 時間情報活用モデルを生成するためのデータ
Fig. 3 Data for generating time-aware model.

間インターバル $[T_k, T_{k+1})$ かつ $[D_\ell, D_{\ell+1})$ のカウントを表す．

1 つのアカウントの疑似移動履歴が含む 2 つの要素（(位置 a，時刻 a）と（位置 b，時刻 b)）について，その時間差および距離がどの時空間インターバルに該当するかを計算し，該当する時空間インターバルのカウント $c_{k\ell}$ を 1 つ加算する．これを 1 つのアカウントのすべての要素のペアについて行うことで，そのアカウントの対象者に関して，時空間近接性と物理的不可能性を表すデータを生成し，このデータから当該アカウントの時間情報活用モデルを生成する．図 3 において，左上のカウントが大きくなることが時空間近接性，右上のカウントが小さくなることが物理的不可能性に相当する．

時間情報活用モデルの利用においては，移動履歴 i の要素とアカウント j の要素のすべてのペアに基づいて，図 3 の各セルのカウント $c_{k\ell}$ を求め，Same(i, j) を仮定した場合の時空間近接性と物理的不可能性を表すデータとする．このデータをアカウント j の時間情報活用モデルと比較することで，移動履歴 i がアカウント j と同一人物である確からしさ（Score(i, j, 時間情報活用モデル)）を求める．

(2) 機械学習の利用

時空間近接性と物理的不可能性を表すデータからモデルを生成する方法として，上述した訪問回数に基づくモデルの生成と同様の理由により，機械学習を用いる．

(3) モデル生成の詳細

各アカウント j からの時間情報活用モデルの生成について述べる（$1 \le j \le N$）．アカウント j から変換した疑似移動履歴 S_j を時刻によって部分集合 $S'_{j1}, S'_{j2}, \ldots, S'_{jN'j}$ に分割する（**図 4** (a)）．ここで N'j は，時間情報活用モデルのための部分集合の個数である．各部分集合 S'_{jv} から以下の手順で正例の特徴ベクトルを生成する（$1 \le v \le N'j$）．

(a) 部分集合 S'_{jv} の疑似移動履歴に含まれるすべての要素からペアを生成する（図 4 (b)）．各要素は (仮名，緯度，経度，時刻) の 4 つ組である．
(b) 各ペアについて，時間差と距離を算出し，図 3 の該当するセルのカウント $c_{k\ell}$ を加算する．
(c) 図 3 を 1 次元化し，特徴ベクトルとする．

以上の処理により，アカウント j から N'j 個の正例の特徴ベクトルを生成する．また，以下の手順 (a′)，(b′)，(c′) により負例の特徴ベクトルを生成する（$1 \le v \le N'j$，$1 \le j' \le N$，$j' \ne j$）．ここで，部分集合 S'_{jv} と部分集合 $S'_{j'v}$ は，同じ時間帯の疑似移動履歴とする．

(a′) 部分集合 S'_{jv} の疑似移動履歴に含まれるすべての要素と，部分集合 $S'_{j'v}$ の疑似移動履歴に含まれるすべての要素からペアを生成する（図 4 (c)）．

(b′)，(c′) は上記 (b)，(c) と同様である．

以上のように生成した正例および負例の特徴ベクトルを用いて，機械学習によりアカウント j の時間情報活用モデルを生成する．

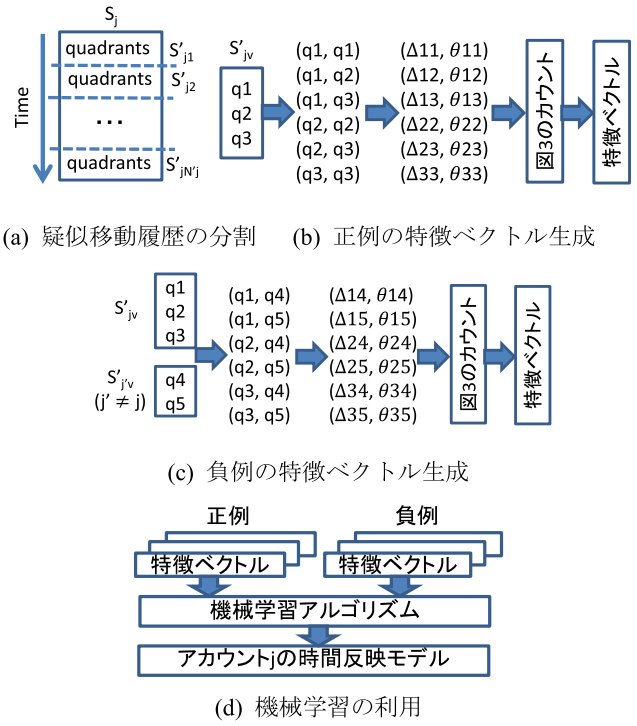

図 4 時間情報活用モデルの生成
Fig. 4 Generating time-information-usage model.
（注）q1～q5 は疑似移動履歴の要素（4 つ組），Δab and θab は 2 つの要素 qa，qb の時間差と距離

4.3 移動履歴のモデル化

上記の 4.2.1 項の処理を，疑似移動履歴の代わりに移動履歴に適用することにより，各移動履歴 i の訪問回数に基づくモデル（狭域細粒度モデルと広域粗粒度モデル）を生成する（$1 \le i \le M$）．また，4.2.2 項の処理を移動履歴に適用することにより，各移動履歴 i の時間情報活用モデルを生成する．

4.4 同一人物に属する度合の部分判定

移動履歴とアカウントが同一人物に属する度合の部分スコアの算出方法について述べる．

(1) 訪問回数に基づくモデルの利用

移動履歴 i とアカウント j のすべての組合せ（$1 \le i \le M$，

© 2020 Information Processing Society of Japan

$1 \leq j \leq N$）について，移動履歴 i の狭域細粒度モデルの特徴ベクトルをアカウント j の狭域細粒度モデルに入力し，Score(i, j, アカウント j の狭域細粒度モデル) を算出する．同様に，Score(i, j, アカウント j の広域粗粒度モデル) を算出する．また，アカウント j の狭域細粒度モデルの特徴ベクトルを移動履歴 i の狭域細粒度モデルに入力し，Score(i, j, 移動履歴 i の狭域細粒度モデル) を算出する．同様に Score(i, j, 移動履歴 i の広域粗粒度モデル) を算出する．以上により，i と j の 1 つのペアに対して，4 つの部分スコアを得る．

(2) 時間情報活用モデルの利用

すべての i，j について，共通の時間境界を用い，移動履歴 i とアカウント j の疑似移動履歴を部分集合に分割する．i と j のすべての組合せについて，以下の手順で特徴ベクトルを生成する．

- 移動履歴の部分集合とアカウントの疑似移動履歴の部分集合であって，同じ時間帯である 2 つの部分集合 A と B のすべてについて以下を行う．
- A の全要素と B の全要素のペアを算出する．たとえば A の要素が 3 個，B の要素が 5 個の場合，15 ペアとなる．
- 各ペアについて，時間差と距離を算出し，図 3 の該当するカウント $c_{k\ell}$ を加算することで特徴ベクトルを生成する．

上記により，i と j の組合せごとに，$N''_{i,j}$ 個の特徴ベクトルを生成する．ここで，$N''_{i,j}$ は移動履歴 i の部分集合とアカウント j の疑似移動履歴の部分集合のうち，時間帯の合致する部分集合の個数である．これらの特徴ベクトルは，Same(i, j) を仮定したときの対象者の時空間における移動パターンを表現したものである．$N''_{i,j}$ 個の特徴ベクトルを，移動履歴 i の時間情報活用モデルおよびアカウント j の時間情報活用モデルに入力し，Score(i, j, アカウント j の時間情報活用モデル) および Score(i, j, 移動履歴 i の時間情報活用モデル) を得る．

4.5 同一人物に属する度合の総合判定

上記 4.4 節で述べたように，すべての i と j の組合せ（$1 \leq i \leq M, 1 \leq j \leq N$）について，訪問回数に基づくモデルを利用した 4 つの部分スコアが算出され，時間情報活用モデルを利用した 2 つの部分スコアが算出される．これらの 6 つのスコアを統合して，最終的な Score(i, j) を算出する．スコア統合の方法としては平均を用いた．メタ学習を用いる高度な方法も考えられるが，その検討は今後の課題とした．

4.6 両側モデルの利用

3.2 節のステップ (2) ではアカウントのモデルを生成し，(3) では移動履歴のモデルを生成する．(4) で，移動履歴とアカウントモデルの整合性，アカウントと移動履歴モデルの整合性をスコア化し，(5) で部分スコアから統合スコアを算出する．すなわち，再特定対象である移動履歴とサイドデータであるアカウントの両側のモデルを併用する（図 2）．既存方式では，サイドデータのモデルのみ利用していた．たとえば，Shokri らの方式 [4] では，再特定対象者に関する攻撃者の知識（自宅や勤務先等の情報）がサイドデータである．これを移動プロファイル（遷移確率行列）としてモデル化し，移動履歴と移動プロファイルの整合性をスコア化していた．両側のモデルを併用して統合スコアを算出する点は著者らの新しい提案である．

5. 実験の説明

5.1 データセット

(1) 移動履歴

Wi-Fi 事業者の協力を得て，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が Wi-Fi アクセスポイントに発したプローブ要求の記録から移動履歴を取得した[*2]．対象期間は 2017 年 1 月 25 日から 2017 年 4 月 23 日までの 90 日間とした．しかし，スマートフォンの電源や Wi-Fi 機能を一時的にオフにする被験者がいたので，すべての被験者について 90 日分の移動履歴を取得してはいない．Wi-Fi 事業者は，スマートフォンの MAC アドレスを仮名化するとともに，Wi-Fi アクセスポイントを緯度と経度の情報に変換した後，プローブの時刻を付して著者らに渡した．そのため，1 人分の移動履歴は，4.1 節で定義した要素すなわち 4 つ組（仮名，緯度，経度，時刻）の系列である．プローブ要求は約 100 m 以内の基地局に無線で到達するため，（緯度，経度）の位置には 100 m 程度の誤差がある．時刻の粒度は秒である．

(2) ソーシャルネットワークアカウント

移動履歴と同一の被験者 53 人の Twitter アカウントのつぶやきを，2017 年 4 月 28 日を起点に過去に遡って，1 アカウントあたり 3645.6 件収集した．アカウントのリンク情報は，被験者の意向により収集できなかった．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．

5.2 データセットの分析

(1) 移動履歴

取得した移動履歴における 1 人あたりの要素数（4 つ組数）および 1 人・1 日あたりの要素数に関する統計を**表 2** に示す．平均および中央値に対して標準偏差が非常に大きく，個人によるデータ量の差が大きいことが分かる．

表 3 は，移動履歴のうち自宅のある市区町村，通学先の

[*2] 移動履歴の取得にあたって，被験者から個別かつ明確な許諾を得ている．また本研究は電気通信大学の研究倫理審査委員会の承認を得ている．

© 2020 Information Processing Society of Japan

表 2 移動履歴の統計
Table 2 Statistics of location histories.

項目	平均	中央値	標準偏差
4 つ組／人	21,361.36	13,068	24,862.88
4 つ組／人・日	244.90	161.04	258.25

表 3 訪問先の分類（%）
Table 3 Classification of visited locations (%).

分類	自宅付近	通学先付近	東京近郊	それ以外
一般	0	0	94.11	5.89
電気通信大学	33.57	24.48	89.15	10.85

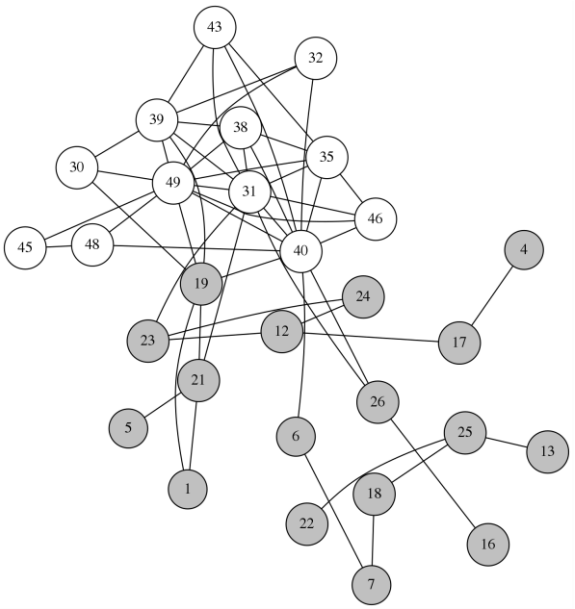

図 5 移動履歴間の接触グラフ
Fig. 5 Contact graph of location histories.

市区町村，東京近郊（東京，神奈川，埼玉，千葉，茨城），それ以外の場所である要素の割合を示す．被験者によって要素の総数が大きく異なるため，被験者ごとに割合を求め，被験者間で割合を平均した．一般被験者の場合は，自宅住所と通勤先住所の情報を提供していただけなかったので，東京近郊とそれ以外の割合のみ示す．電気通信大学被験者の場合，自宅の市区町村および通学先の市区町村の要素数は，東京近郊の要素数と重複している．また，自宅の市区町村と通学先の市区町村が同じ被験者もいる．

移動履歴間の接触グラフを**図 5** に示す．ここでは，Srivatsa らの実験 [3] と同様に，2 つの移動履歴が同じ基地局で 10 分以上にわたってともに観測されている場合に接触とした．図 5 のノードが移動履歴，エッジが接触関係を示す．灰色のノードは一般被験者，白色ノードは電気通信大学の被験者であり，ノード中の番号は被験者番号である．電気通信大学被験者 24 人のうち 12 人，一般被験者 29 人のうち 12 人は他の被験者と接触がなく，孤立していたので，図 5 から省いている．

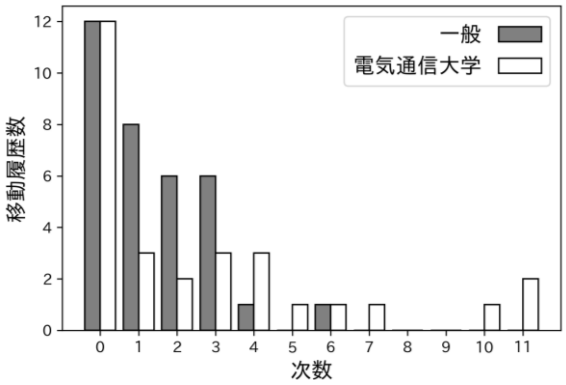

図 6 移動履歴の次数分布
Fig. 6 Degree distribution of location histories.

接触グラフにおけるノードの次数（ノードのエッジ数すなわち接触している移動履歴の数）の分布を**図 6** に示す．一般被験者の場合は次数の増加にともなって度数が減少するが，電気通信大学被験者の場合は次数の大きい被験者が存在する．電気通信大学の被験者は相互に友人関係があり，次数の大きい被験者はハブになっていると考えられる．

(2) ソーシャルネットワーク

形態素解析を用いて投稿文から地名の含まれるものを抽出し，それらを人手で分析した．その結果，形態素解析が地名と認識した名詞は以下の 5 種類に分類できることが判明した[*3]．

(a) 被験者は，投稿時刻に当該地名の場所にいた．たとえば「新宿で飲んでいる」．
(b) 被験者は，投稿時刻以外の時刻に当該地名の場所にいた．例「渋谷で飲んだ」．
(c) 被験者は，当該地名の場所にはいなかった．例「京都に住みたい」．
(d) (a)〜(c) のどれに該当するのか，判定できない．例「今渋谷にいるの？」は，被験者自身が現在渋谷にいて，友人と会いたがっている可能性を示唆するが，不明確である．
(e) 形態素解析の誤りにより，地名ではない単語（人名や会社名等）が地名と誤検知されている．

これらのうち (a) は移動履歴との照合に直接利用できる．(b) は方式によっては照合に利用できる．(d) の一部は，照合において (a) または (b) と同様の効果をもたらす．(c) と (e) はノイズとなる．(a)〜(e) の個数および割合は，3580 (41.9%)，1145 (13.4%)，938 (11.0%)，1454 (17.0%)，1425 (16.7%) であった．

以上の分析をふまえて，被験者のアカウントに関する統

[*3] (a)〜(e) の厳密な区別は困難である．たとえば (c) の例として「京都に住みたい」を取り上げたが，京都が好きなので旅行する可能性が大きく，文脈によっては (c) ではなく (b) に含めるべき場合もある．したがって，上記の分析は厳密ではないが，利用可能な位置情報が少ないという定性的な結論は得られる．

© 2020 Information Processing Society of Japan

表 4 アカウント投稿文の統計
Table 4 Statistics of texts posted on accounts.

項目	平均	中央値	標準偏差
地名の出現数／人	67.54	13	123.80
	89.15	30	138.07
	116.58	50	155.58
地名の出現数／人・日	0.69	0.34	1.07
	0.95	0.58	1.16
	1.31	0.96	1.27

計を**表 4** に示す．表 4 の各セルのなかには，上記 (a) に該当する地名の出現数，(a) または (b) に該当する地名の出現数，(a) または (b) または (d) に該当する地名の出現数の 3 つの数値を記載している．(a) の数値は照合に利用可能な位置情報数の下限，(a) または (b) または (d) は上限である．表 4 から，被験者のアカウントにおいて利用可能な位置情報は非常に少ないことが分かる．

上記の利用可能な地名（(a) または (b) または (d)）の解像度を調べたところ，市区町村より細かい地名（たとえば原宿は渋谷区の一部の地名）が 23.1%，市区町村名が 50.0%，都道府県名が 25.4%，都道府県より大きい国内の地名（関東等）が 1.4%，海外の地名が 0.1%であった．市区町村は数 km の範囲，都道府県は数 10 km の範囲に及ぶため，ソーシャルネットワークの投稿文における地名の解像度は非常に低い[*4]．

このように投稿文のうち照合に使える位置情報が少なく，しかも位置情報の解像度が低いことから，移動履歴との照合は容易ではないことが予想される．一方，投稿文の位置情報の解像度が非常に低いことから，移動履歴の解像度を落としても再特定の精度は顕著には低下しないことが予想される．

5.3 実験方法

提案方式を評価するための実験およびその目的は以下のとおりである．

(1) 53 vs. 53 の照合

移動履歴とサイドデータが同数である場合の再特定精度の評価は，すべての既存研究に共通した基本評価である．同一人物の移動履歴とアカウントが正しく対応付けられた割合を様々な条件下で評価し，以下を明らかにする．

- 提案方式はアカウント間のリンク関係を利用しない．そこで提案方式の精度を評価することで，目標 1 の一部（リンク関係が利用できなくても再特定可能）の達成度を明らかにする．
- 提案方式のうちアカウントのモデルだけ利用した場合の精度を評価することで，目標 1 全体（リンク関係と接触関係の両者が利用できなくても再特定可能）の達成度を明らかにする．
- 時間情報活用モデルを利用する場合としない場合の比較により，目標 3（時間情報の有効利用）の達成度を明らかにする．
- 狭域細粒度モデルと広域疎粒度モデルを別々に利用する場合と併用する場合の比較により，多重解像度の効果を明らかにする．
- アカウントのモデルと移動履歴のモデルを別々に利用する場合と併用する場合の比較により，移動履歴とサイドデータの両者のモデルの併用効果を明らかにする．

(2) 53 vs. 多数の照合

被験者 53 人のアカウントに不特定多数のアカウントを加えて多数のアカウントを設け，被験者 53 人の移動履歴と多数アカウントを照合して再特定精度を評価する．目標 2 の一部（移動履歴数とアカウント数が大きく異なる場合の対応）の達成度を明らかにする．

(3) 1 vs. 多数の照合

上記 (2) と同様に多数のアカウントを設け，被験者 1 人ずつの移動履歴と多数アカウントを照合して再特定精度を評価する．また，逆に被験者 53 人の移動履歴と被験者 1 人ずつのアカウントを照合して精度を評価する．これらにより，目標 2 全体（移動履歴数またはアカウント数が少ない場合，両者が大きく異なる場合への対応）の達成度を明らかにする．

(4) 曖昧化された移動履歴からの再特定

被験者 53 人の移動履歴を曖昧化した後に，53 人のアカウントと照合する．

- 解像度を低下させた移動履歴からは接触関係を判定することができないため，Srivatsa らの方式は対応できない．この状況で提案方式の再特定精度を評価し，提案方式の優位性を明らかにする．
- データの間引きおよび位置のランダム変更は提案方式の目標とは直接関係しないが，移動履歴に対する典型的な曖昧化方法であるため [2], [3], [4], [5], [6]，それらへの耐性を評価する．

6. 実装と評価

6.1 実装

5 章で述べたデータセットの被験者は，東京，神奈川，埼玉，千葉，茨城に在住しており，移動履歴中の位置情報のほとんどはその周辺にある．遠隔地を訪問することもあるが，データ取得期間内の位置情報は日本国内に限られていた．そこで，狭域細粒度モデルおよび広域粗粒度モデルの対象領域は，関東地方南部（126 km × 126 km）および日本全域（2,370 km × 2,140 km）とした．メッシュサイズは 1 km × 1 km および 5 km × 5 km とした結果，特徴ベクトルは 15,876（126 × 126）次元および 202,872（474 × 428）

[*4] 著者らが地名を位置情報に変換するために用いている GeoNLP [18] では，市区町村は市役所等の位置，都道府県は県庁等の位置に変換している．

© 2020 Information Processing Society of Japan

次元となった．

図 3 の時間情報活用モデルにおける時間インターバルおよび距離インターバルの例として，(0, 10, 20, 30, 60, 120, 180, 360) および (0, 1, 2, 4, 8, 16) を用いた．なお，時間インターバルの単位は分，距離インターバルの単位は km である．

移動履歴およびアカウントの疑似移動履歴を部分集合に分割する際，時間情報活用モデルについては，1 日単位のデータに分割した．狭域細粒度モデルおよび広域粗粒度モデルについては，移動履歴は 1 日単位のデータに分割した．アカウントデータの疑似移動履歴の分割は 3 通りを検討した．1 番目は移動履歴に合わせて 1 日単位とした．しかし，表 4 に示すように，疑似移動履歴に含まれる要素は 1 日あたり 1 個程度である．そのため，平均的には，上述した 15,876 次元および 202,872 次元の特徴ベクトルのうち 1 カ所に 1 が入り，他のすべてが 0 となるので，照合可能なパターンが特徴ベクトルに現れない可能性がある．そこで，2 番目として，移動履歴中の要素は約 245 個/日，疑似移動履歴中の要素は約 1 個/日であることから，要素数を合わせて，アカウントデータは 245 日単位に分割した．すなわち 245 日の疑似移動履歴から 1 つの特徴ベクトルを生成した．3 番目は，データを分割せず，1 アカウントの移動履歴全体から 1 つの特徴ベクトルを生成した．これらの 3 つの分割方法を用いて予備評価を行った結果，2 番目の分割方法を用いた再特定精度が最も高かったので，以後の評価では 2 番目を用いることにした．

機械学習のアルゴリズムとしては，ロジスティック回帰，Support Vector Machine（RBF カーネル利用），XGBoost（Extreme Gradient Boosting）の 3 例を取り上げ，scikit-learn のライブラリ [19] を用いて実装した．

移動履歴に対して，以下の前処理を加える場合と加えない場合を評価した．移動履歴は携帯端末が Wi-Fi 基地局と交信したプローブパケットの記録である．プローブパケットの交信はバースト的に短時間に多数回行われる場合がある．その場合，特定のメッシュの場所の訪問回数（すなわち狭域細粒度モデルおよび広域粗粒度モデルの特徴ベクトルの特定のセルの値）が極端に大きくなり，移動傾向を正確に捕捉できなくなる可能性がある．そこで，狭域細粒度モデルおよび広域粗粒度モデルについては，移動履歴のデータの時刻を 10 分間隔で量子化し，当該 10 分の間にプローブパケットが何回観測されても 1 回と数える前処理を検討した．アカウントの疑似移動履歴もそれに合わせて 10 分間隔に量子化した．一方，時間情報活用モデルについては，絶対時刻の差が重要になる．10 分間隔の量子化は時刻差に 20 分の誤差をもたらすので，前処理は行わなかった．

特徴ベクトルの各セルの値は，カウント（たとえば該当するセルへの訪問回数）である．しかし，機械学習では特徴量を 2 値化する（1 以上のカウントをすべて 1 とする）ことが多いので，2 値化しない場合とする場合を検討した．

6.2 53 vs. 53 の照合

被験者 53 人の移動履歴とアカウントを照合し，再特定率（同一人物のアカウントに正しく対応付けられた率）を評価した．狭域細粒度モデルおよび広域粗粒度モデルにおける前処理の有無，狭域細粒度モデルおよび広域粗粒度モデルにおける特徴ベクトルの 2 値化の有無，時間情報活用モデルにおける特徴ベクトルの 2 値化の有無，狭域細粒度モデルおよび広域粗粒度モデルにおける 3 つの機械学習アルゴリズム（ロジスティック回帰，SVM，XGBoost），時間情報活用モデルにおける 3 つの機械学習アルゴリズムの組合せである 72 通りのパラメータセットを評価した．機械学習アルゴリズムはランダム性を有するので各々 10 回評価し，平均を求めた．

その結果，前処理あり，狭域細粒度モデルおよび広域粗粒度モデルの特徴ベクトルは 2 値化あり，時間情報活用モデルの特徴ベクトルは 2 値化なし，狭域細粒度モデルおよび広域粗粒度モデルはロジスティック回帰，時間情報活用モデルは XGBoost のパラメータセットが，**表 5** に示す 15 通りの評価すべてにおいて最高の再特定率を示した．表 5 には，このパラメータセットにおける再特定率を記載している．

表 5 のア 1 の行は，アカウントの狭域細粒度モデルを用いた結果，53 個の移動履歴のうち 17.2 個（32.5%）が同一人物のアカウントに正しく対応付けられたことを示す．ア 3 の行は，アカウントの狭域細粒度モデルと広域粗粒度モデルを併用した結果，ア 5 は，アカウントの狭域細粒度モデル，広域粗粒度モデル，時間情報活用モデルを併用した結果を示す．移 1，移 3，移 5 は移動履歴の同様のモデルを用いた結果を示す．双 1 は，アカウントと移動履歴の狭域細粒度モデルを併用した結果，双 3 は，アカウントと移動履歴の狭域細粒度モデルと広域粗粒度モデル（すなわち

表 5 最良パラメータにおける再特定数と率
Table 5 Performance with best parameters.

種別	使用モデル	特定数(%)	種別	使用モデル	特定数(%)
ア1	ア，狭細	17.2 (32.5)	双1	双，狭細	23.4 (44.2)
ア2	ア，広粗	13.8 (26.0)	双2	双，広粗	22.6 (42.6)
ア3	ア1&ア2	16.8 (31.7)	双3	双1&双2	28.6 (54.0)
ア4	ア，時間	24.4 (46.0)	双4	双，時間	24.2 (45.7)
ア5	ア1&ア2&ア4	33.6 (63.4)	双5	双1&双2&双4	39.8 (75.1)
移1	移，狭細	18.0 (34.0)			
移2	移，広粗	20.2 (38.1)			
移3	移1&移2	23.2 (43.8)			
移4	移，時間	17.2 (32.5)			
移5	移1&移2&移4	29.4 (55.5)			

© 2020 Information Processing Society of Japan

4 モデル）を併用した結果，双 5 は，アカウントと移動履歴の狭域細粒度モデル，広域粗粒度モデル，時間情報活用モデル（すなわち 6 モデル）を併用した結果を示す．

表 5 から以下を観察することができる．

(1) 提案方式はソーシャルネットワークのリンク関係を利用していないが，表 5 の双 5（6 つのモデルすべての併用）は 75.1%の精度で再特定ができる．そのため，目標 1 の一部（リンク情報の不要化）を満たしている．次に物理世界の接触関係の情報が利用できない場合を考える．表 5 のア 1〜5 の結果は，アカウントのモデルだけを用いており，移動履歴のモデルは用いていない．再特定対象の移動履歴 i と各アカウント j（$1 \leq j \leq N$）について，Score(i, j) の N 個の値を算出して，Score(i, j) が最大となるアカウント j を選択している．この処理を各移動履歴 i について独立に行っているので，移動履歴間の関係は用いていない．そのため，物理世界の接触関係の情報が利用できなくても，表 5 のア 5 の精度（63.4%）は得られる．これは，最も精度が高かった双 5（75.1%）に比べて 84.4%の精度である．以上から，提案方式は，目標 1 全体（リンク関係と接触関係の両者の不要化）をおおむね達成できている．

(2) 表 5 のア 5/移 5/双 5 は，ア 3/移 3/双 3 よりも高精度である．双 5 の特定精度は双 3 の特定精度よりも 21.1%高い．また，ア 5/移 5 の特定精度は，ア 3/移 3 の特定精度よりも 31.7%/11.7%高い．このことから，訪問回数に基づくモデル（狭域細粒度モデルおよび広域粗粒度モデル）と時間情報活用モデルを併用すると，訪問回数に基づくモデルのみ用いる場合よりも再特定率は向上する．一方，時間情報活用モデルは，4.2.2 項で述べたように，接触関係とも前後関係とも異なる時間情報を含んでいる．以上から，提案方式は，接触関係および前後関係以外の時間情報を利用して再特定率を向上させているので，目標 3 を満たす．

(3) ア 3/移 3/双 3 は，ア 1/移 1/双 1 およびア 2/移 2/双 2 よりも高精度である（1 カ所例外があり，ア 3 はア 1 にわずかに劣る）．このことから，狭域細粒度モデルと広域粗粒度モデルの併用すなわち多重解像度の利用は有効であると考えられる．

(4) 双 1〜5 は，移 1〜5 およびア 1〜5 よりも高精度であることから，アカウントのモデルと移動履歴のモデルの併用は有効であると考えられる．

なお，以下の節では，ソーシャルネットワークのリンク関係の利用不可に対応できる双 5 のパラメータおよび，リンク関係と接触関係の両者の利用不可に対応できるア 5 のパラメータを用いて評価する．

6.3 53 vs. 多数の照合

本節では，多数の候補者のアカウントの中から移動履歴と同一人物のアカウントを見つけるという状況を想定する．被験者の 53 アカウントを不特定多数の β 個のアカウントに混ぜ込んでシャッフルし，被験者の 53 個の移動履歴を $(\beta + 53)$ 個のアカウントと照合した．**表 6** において，53 の移動履歴のうち同一人物のアカウントに正しく対応付けられた場合を「正解」とした．また，移動履歴 i と同一人物のアカウントを j としたときに，Score(i, j) が $(\beta + 53)$ 個のスコア Score(i, j')（$1 \leq j' \leq \beta + 53$）のうち上位 100 個に含まれた割合を「100 位以内」とした．表 6 では，正解および 100 位以内の割合を%で示している．

表 6 53 vs. 多数の再特定率（%）
Table 6 Re-identification rate in 53 vs. many evaluations.

(a) リンク関係の利用不可に対応する場合（双 5）

アカウント数	53	1,053	10,053	100,053
正解	75.1	15.1	0.0	0.0
100 位以内	-	92.5	61.1	39.8

(b) リンク関係と接触関係が利用不可の場合（ア 5）

アカウント数	53	1,053	10,053	100,053
正解	63.4	3.8	0	0
100 位以内	-	84.5	58.5	9.4

以下では，53 vs. 多数の照合の再特定率が，53 vs. 53 照合の再特定率よりも低くなる理由を考察する．6.2 節で示したように，53 vs. 53 照合の場合でも正解率は 75.1%であり，移動履歴に対して 24.9%の割合で他人のアカウントが対応付けられていた．このことから，i（$1 \leq i \leq M$）と j（$1 \leq j \leq N$，$j \neq i$）をランダムに選択したときに，ある確率 P で Score(i, j) < Score(i, i) になると考えられる．53 vs. 53 照合の誤り確率（同一人物のアカウントのスコアが最高にならない確率）は $1 - (1 - P)^{52}$ となる．1 vs. 多数の照合たとえば 1 vs. 1,053 照合の場合の誤り確率は，単純には $1 - (1 - P)^{1052}$ となる．被験者の 53 アカウントと不特定多数の 1,000 アカウントの性質に相違があるとしても，誤り確率は $1 - (1 - P)^{52}(1 - P')^{1000}$ となり，53 vs. 53 照合の誤り確率よりは大きい．また不特定多数のアカウントが増えるほど誤り確率は大きくなる．なお，P' は i（$1 \leq i \leq M$）と j'（j' は不特定多数アカウントの番号）をランダムに選択したときに，Score(i, j') < Score(i, i) になる確率である．同一人物のアカウントのスコアが 100 位以内にならない確率を仮定すれば，上記と同様の分析により，不特定多数のアカウントが増えるほど本人のアカウントが 100 位より下になる確率が高まることが説明できる．

一方，100 アカウントまでは，人手で詳細な調査が可能であるとすれば，100 位以内に絞り込めれば，人手調査が可能になる．たとえば，53 個の移動履歴の対象者を 100,053 個のアカウントから探す場合でも，双 5 のパラメータを用いる場合には，約 40%の割合で人手調査に持ち込むことができる．また，ア 5 の場合でも，10,053 個アカウントから約 60%の割合で人手調査に持ち込むことができる．人手調

© 2020 Information Processing Society of Japan

査では，たとえば，移動履歴の場所から勤務先が大学，移動時間から職種が教員であると推定し，アカウントの投稿文に「学生が頑張っているので楽しみ」とあれば，移動履歴とアカウントが同一人物である可能性が高いと判断できる．このような高度な判断により，100 アカウントからさらに絞り込める可能性がある．以上から，提案方式は目標 2 の一部（移動履歴数がアカウント数よりも極端に小さい場合の対応）をある程度は満たしていると考える．

なお，移動履歴の対象者の候補を事前に絞り込めるとは限らないため，本評価のように移動履歴の対象者を数千，数万の候補の中から見つける攻撃は，実際に行われる可能性がある．しかし，従来の移動履歴からの再特定の研究では，対象者を多数の候補者の中から見つけるという評価を行った例はなく，本研究が初めてである．

6.4 1 vs. 多数の照合

1 つの移動履歴に対して，$(\beta + 53)$ 個のアカウントから同一人物のアカウントを特定した．これは，攻撃者が移動履歴を 1 つだけ入手し，対応するアカウントを探す状況を想定している．したがって，攻撃者は移動履歴の負例を利用できず，移動履歴のモデルを生成できないので，表 5 のア 5 のモデルを用いた．対象とする 1 つの移動履歴を 53 の移動履歴から 1 つずつ選択し，53 回の評価を行って平均を求めた．なお，前述したように，機械学習はランダム性を有するので，1 回の評価結果として 10 回の平均値を用いている．評価結果を**表 7** に示す．

1 vs. 多数の照合における 100 位以内の率は，アカウント数が 10,053 の場合でも 60%に近い．このことから人手による照合との併用を仮定すれば，提案方式は，目標 2（移動履歴数やアカウント数が少数の場合，移動履歴数とアカウント数が大きく異なる場合の対応）をある程度は満たしていると考える．

なお，逆方向の評価として，1 つのアカウントに対して，53 の移動履歴の中から同一人物の移動履歴を特定した．この場合，アカウントのモデルが利用できないため，正解率は，表 5 の移 5 における 55.5%であった．

6.5 曖昧化された移動履歴からの再特定

曖昧化の例として，解像度の低下，データの間引き，位置のランダム変更を取り上げ，パラメータを変えて，53 vs. 53 における再特定の精度を評価した．

表 7 1 vs. 多数の再特定率（ア 5，%）
Table 7 Re-identification rate in 1 vs. many evaluations.

アカウント数	53	1,053	10,053	100,053
正解(%)	63.4	3.8	0.0	0.0
100 位以内(%)	-	84.5	58.5	9.4

(1) 解像度の低下

5 章のデータセットは，位置の誤差が 100 m，時間粒度は秒であるが，メッシュサイズを 5 km × 5 km，時間粒度を 1 日とした．広域粗粒度モデルと狭域細粒度モデルは，1 日幅でデータを分割しているので，時間粒度は 1 日になっている．広域粗粒度モデルのメッシュサイズは 5 km × 5 km である．狭域細粒度モデルのメッシュサイズは 1 km × 1 km であるが，これを 5 km × 5 km にすると，広域粗粒度モデルの関東地方の部分と等価になる．時間情報活用モデルは時間インターバルの最大値が 360 分であるため，時間粒度を 1 日にすると利用できない．以上から，提案方式においてメッシュサイズを 5 km × 5 km，時間粒度を 1 日とする自然な方法として，広域粗粒度モデルのみ利用することが考えられる．その場合の 53 vs. 53 の再特定精度は，表 5 の双 2 における 42.6%であり，位置誤差 100 m，時間粒度 1 秒の場合の精度（75.1%）に比べて，56.7%であった．

(2) データの間引き

各々の移動履歴の要素集合から，指定された割合をランダムに間引きした後，アカウントと照合した．ランダムな間引きは 10 回行い，再特定率の平均をとった．間引き率と再特定精度の関係を**表 8** に示す．双 5 の場合，99%の間引きの場合に再特定率は 50%に近かった．表 2 を見ると，1 人の移動履歴の要素数の中央値は 13,000 程度であるので，130 件程度の要素があれば 50%近くの再特定のリスクがあると考えられる．ア 5 の場合は，95%の間引き（650 件程度の要素）において，50%程度の再特定リスクがあると考えられる．

(3) 位置のランダム変更

各々の移動履歴の要素集合から，指定された割合をランダムに選択し，選択した要素について，位置を指定距離内でランダムに変更した．選択率と指定距離と再特定率の関係を**表 9** に示す．表 9 (a) および (b) を，データの間引きの結果（表 8 (a) および (b)）と比較すると，位置の変更範囲が 8 km 以上，間引き率と変更率が 99%以内では，間引きよりもランダム変更による精度低下が大きかった．理由としては，位置変更の指定距離が 8 km 以上になると，場所ごとの訪問回数に基づくモデルにおいて要素が異なるメッシュに移る可能性が高くなり，時間情報活用モデルにおいても表 4 の異なるセルに移る可能性が高くなる．その結

表 8 間引きされた移動履歴からの再特定率（%）
Table 8 Re-identification from subsampled location histories.

(a) リンク関係の利用不可に対応する場合（双 5）

間引き率	0	0.5	0.9	0.95	0.99	0.995	0.999
正解(%)	75.1	60.4	57.4	54.7	46.8	33.0	3.8

(b) リンク関係と接触関係が利用不可の場合（ア 5）

間引き率	0	0.5	0.9	0.95	0.99	0.995	0.999
正解(%)	63.4	51.7	50.9	49.6	38.3	27.2	3.8

© 2020 Information Processing Society of Japan

表 9 位置をランダム変更した移動履歴からの再特定率（%）
Table 9 Performance with randomly changed locations.

(a) リンク関係の利用不可に対応する場合（双 5）

指定距離＼変更率	0	0.5	0.9	0.95	0.99	0.995	0.999
2.0	75.1	62.4	52.8	51	50.4	49.1	41.2
4.0	75.1	57.9	49.5	48.9	49.7	46.3	40.1
6.0	75.1	54.4	44.2	44.8	41.7	39.8	32
8.0	75.1	52.7	44.6	42.7	39.8	35.9	29.1
10.0	75.1	50.9	38.5	35.8	32.2	30.3	18.5

(b) リンク関係と接触関係が利用不可の場合（ア 5）

指定距離＼変更率	0	0.5	0.9	0.95	0.99	0.995	0.999
2.0	63.4	57.5	52.5	51.2	47	44.6	26.8
4.0	63.4	54.5	49.1	49.4	44.9	38.4	22.3
6.0	63.4	52.8	50.2	48.9	42.7	37	23
8.0	63.4	51.1	48.2	48.1	39.4	35.7	19.3
10.0	63.4	49.7	43.4	33.8	31.5	30	15.9

果，再特定のための有用情報が減少するだけでなく，移動した要素がノイズとなるので，単純な間引きよりも精度の低下が大きいと考えられる．一方，位置変更の指定距離が 4 km 以内，間引き率と変更率が 99%以内では，間引きとランダム変更の影響は同程度であった．理由としては，指定範囲が 4 km 以下の場合は，訪問回数に基づくモデルにおいて要素が同じメッシュにとどまる可能性が高くなり，時間情報活用モデルにおいても図 3 の同じセルにとどまる可能性が高くなる．その結果，有用情報が残存する効果とノイズとなる効果が相殺されるため，単純な間引きと同等の精度になったと考えられる．さらに，間引き率と変更率が 99.5%以上の場合には，位置変更の指定距離にかかわらず，間引きよりもランダム変更による精度低下が小さかった．理由としては，99.5%以上の間引きでは，再特定のための有用情報がほとんどなくなるが，ランダム変更の場合は，一部の要素が同じメッシュおよびセルにとどまり有用情報が残る．そのため，間引きに比べると，ランダム変更の精度低下は小さいと考えらえる．

6.6 再特定しやすかった人の分析

53 vs. 1053 照合（表 6）で再特定された 8 人（15.1%）は，1,000 以上のアカウントの中から一意に再特定されている．そこで，これらの 8 人を再特定しやすい人の代表例として，移動履歴と投稿文の両面から分析した．

移動履歴中の場所を下記のように分類した．

- 遠隔地：関東南部以外．たとえば，広島，山形等．
- 他人があまり行かない都内/東京近郊の場所：鷲谷/春日部等．
- 他人が比較的よく行く都内/東京近郊の場所：西日暮里，四谷/大宮，横浜駅等．
- 他人が非常によく行く場所：被験者の多くが頻繁に行く場所．新宿，池袋等．

表 10 再特定しやすかった人の例
Table 10 Example people who were easy to re-identify.

タイプ	被験者仮 ID	移動履歴	投稿文
1	11	遠隔地 A に数回, 他人があまり行かない都内の場所 B に数回, 他人が非常によく行く場所 C に数 10 回.	左記の遠隔地 A に 100 回近く言及． B と C に各々数回言及．B は市区町村より細かい粒度.
	12	他人があまり行かない東京近郊の 2 か所 D, E のうち D に約 20 回，E に 1 回，他人がよく行く東京近郊の F に数 10 回.	D に 20 回程度，E に 1 回, F に 100 回近く言及.
2	21	他人があまり行かない都内の G に約 20 回，他人がよく行く都内の H に数 10 回.	G に数回，H に 10 数回言及. 2 か所とも市区町村より細かい粒度.
	22	他人があまり行かない都内の I，J のうち I に 10 回程度，J に数回.	I は 10 回程度, J は数回言及. 2 か所とも市区町村より細かい粒度.
3	31	遠隔地 K に数日滞在, 他人があまり行かない東京近郊の L に数日滞在.	K に数回言及, L は 1 回言及.

「他人がよく行く場所」「他人が非常によく行く場所」は再特定の手がかりとして弱いが，多数回行った場合は手がかりとなる．それ以外の場所は手がかりとなる．一方，投稿文中の場所は，5.2 節 (2) で述べたように市区町村名，都道府県名が多いが，市区町村より細かい粒度の地名（駅名や球場名を含む）は強い手がかりになると考えられる．

以上の観点から 8 人の挙動を分析した結果，8 人を下記の 3 つのタイプに分類することができた．各タイプの例を**表 10** に示す．

タイプ 1：移動履歴では，遠隔地，他人があまり行かない場所，他人がよく行く場所（多数回），他人が非常によく行く場所（多数回）の手がかりが 3 つ以上揃っており，そのいずれも投稿文で言及されている（被験者 4 人）．

タイプ 2：移動履歴の手がかりが 2 つ揃っており，そのいずれも投稿文で市区町村より細かい粒度で言及されている（3 人）．

タイプ 3：移動履歴の手がかりが 2 つ揃っており，両方の場所に数日滞在している．両方について投稿文で言及して

© 2020 Information Processing Society of Japan

いる（1 人）．

7. 提案方式の優位性に関する考察

7.1 目標達成のまとめ

6 章の実験結果をまとめると下記のとおりである．

- リンク情報を用いない場合の再特定精度は 75.1%であり，接触情報とリンク情報のいずれも用いない場合の精度は 63.4%であり，目標 1（接触情報とリンク情報の一方または両方を不要化）をおおむね達成している．これにより，既存方式の前提 (1) を不要化できる．
- 53 vs. 多数および 1 vs. 多数の照合において，アカウント数が 10,053 の場合でも同一人物のアカウントが上位 100 位以内になる確率が 60%程度であることから，目標 2（移動履歴数やアカウント数が少数の場合，移動履歴数とアカウント数が大きく異なる場合の対応）をある程度は満たしている．この点については，7.2 節 (2) および (3) でより詳しく分析する．
- 訪問回数に基づくモデルのみを利用する場合に比べて，時間情報活用モデルを併用することで，再特定精度を 29.4%向上させることができた．時間情報活用モデルは接触関係とも前後関係とも異なる情報を含むので，目標 3（接触関係，前後関係以外の時間情報の有効利用）を達成している．これにより，既存方式にはなかった時間情報の利用を可能にした．

7.2 Srivatsa らの方式との比較

5 章で述べたデータセット（以下では本データセットと呼ぶ）と 6 章の評価に基づき，提案方式と Srivatsa らの方式を比較する．必要に応じて，Srivatsa らの示した実験結果 [3] を引用する．Srivatsa らは 3 つのデータセット St Andrews（SA），Smallblue（SB），Infocom06（IC）を用いて実験したが，SA は Wi-Fi 基地局のデータであり，6 章のデータと同種である．SB と IC は Bluetooth のデータであるが，接触グラフに変換した後の処理は SA と同じなので参考にする．また，提案方式と Srivatsa らの方式の組合せについても検討する．

(1) 基本性能

移動履歴とアカウントが同数で曖昧化を加えない場合の精度を比較すると，SA/SB/IC を用いた Srivatsa らの方式の再特定率は 90%/80%/80% [*5]，本データセットを用いた提案方式の再特定率は 75.1%である．しかし，SA と SB では，移動履歴のどれもが 2 つ以上の他の移動履歴と接触（接触グラフのノードの次数が 2 以上），IC ではわずかな移動履歴が他の 1 または 2 個の移動履歴と接触し，大部分は 3 個以上の移動履歴と接触している．これに対し，本データセットでは，53 人の移動履歴のうち 24 人の移動履歴は他の移動履歴と接触せず孤立しており，接触グラフに含まれない．そのため，Srivatsa らの方式を本データセットに適用した場合の再特定率はたかだか 54.7%（すなわち (53 − 24)/53）である．このことから，Srivatsa らの方式の前提 (1) のうち物理世界の接触関係の成立は，データに大きく依存すると考えられる．移動履歴に接触関係が多い場合には Srivatsa らの方式の再特定精度は高い．しかし，接触関係が少ない場合には，Srivatsa らの方式の精度は低く，提案方式の方が精度が高いと考えられる．このように Srivatsa らの方式と提案方式は補完関係にあり，移動履歴の接触関係の多寡によって使い分けることが考えられる．

(2) 移動履歴やアカウントが少ない場合の性能

6.4 節で述べたように，提案方式は 1 移動履歴 vs. 53 アカウントの照合では 63.4%，1 アカウント vs. 53 移動履歴の照合では 55.5%の精度である．2 vs. 53 の実験は行っていないが，2 つの移動履歴あるいはアカウントの各々について 1 vs. 53 の照合を行えば，平均して 63.4%あるいは 55.5%の精度になる．3 vs. 53 の場合も同様である．一方，移動履歴あるいはアカウントが 1 つの場合はグラフを形成できないので，Srivatsa らの方式はまったく再特定できない．移動履歴が 2 つの場合は，アカウントのリンクグラフのすべてのエッジに同確率で照合するので再特定できない．3 つの場合も再特定は困難である．しかし，たとえば重要施設に侵入した移動履歴からの侵入者の再特定や，自社に毎日通っている社員らしい人物が気になる場所を訪問している場合の当該社員の再特定等，対象となる移動履歴が 1〜3 件程度で，候補が数十〜数百である状況は現実に想定できる．現実に想定可能な状況において，提案方式は Srivatsa らの方式よりも優れる．

(3) 移動履歴数とアカウント数が大きく異なる場合

Srivatsa らの方式は SB および IC データにおいて，アカウント数が移動履歴の 2 倍になると，アカウント数と移動履歴が等しい場合の再特定精度 90%/80%が 30%/20%に低下する．低下率は 66.7%/75.0%である．6.3 節の実験を拡張し，提案方式の 53 vs. 53 の再特定率（75.1%）を基準にして，53 移動履歴 vs. 106 アカウント（アカウント数が移動履歴の 2 倍）の再特定率を求めたところ 36.8%であり，低下率は 51.0% であった．このように，Srivatsa らの方式と提案方式では，アカウント数が移動履歴数の 2 倍になったときの再特定率の低下に顕著な差がみられなかった．データセットが異なることから，低下率の小さな差に有意性を見出すことはできないため，提案方式の優位性は確認できなかった．

一方，Srivatsa らの方式は SA/SB/IC において，アカウント数が移動履歴数の 1/2 になると，再特定精度が 90/80/80%から 30/20/20%に低下し，低下率は 66.7/75.0/75.0%である．提案方式において，53 移動履歴 vs. 26.5 アカウントの再特定精度を求めたところ 70.4%であり，低下率は 6.3%で

[*5] Srivatsa らの論文では再特定率をグラフで表しており，厳密な数値は記載していない．

© 2020 Information Processing Society of Japan

あった[*6]．このように，Srivatsa らの方式と提案方式の間には，アカウント数が移動履歴数の 1/2 になったときの再特定率の低下に顕著な差がみられた．以上から，移動履歴数に比べてアカウント数が小さい場合に，提案方式は Srivatsa らの方式よりも優れていると考える．

(4) 曖昧化処理への耐性

Srivatsa らは，接触およびリンクグラフの変更と，位置の変更について評価している．提案方式はグラフを用いないので，グラフの変更の影響は受けない．Srivatsa らは位置の変更の大きさを記載していないので比較できない．Srivatsa らは解像度の低下について評価していないが，分析可能であるため，提案方式と比較する．SA データは，位置（交信した基地局）の履歴であり，SB および IC データは端末間の交信の履歴である．移動履歴の多くは，SA のように位置の履歴である [2], [4], [5], [6], [7], [8], [9]．SA では Wi-Fi 基地局からの電波到達は 100 m 程度であり，2 つの移動履歴が同じ基地局で 10 分以上にわたってともに観測された場合に接触としている．提案方式は位置解像度 1 km，時間解像度 1 日の再特定率は 54.0%（表 5 の双 3），位置解像度 5 km，時間解像度 1 日の再特定率は 42.6%（表 5 の双 2）である．一方，Srivatsa らの方式は，これらの低解像度において接触関係を判定できないため再特定はできない．以上から，解像度低下への耐性に関し，提案方式は Srivatsa らの方式より優れると考える．

8. おわりに

ソーシャルネットワークアカウントとの照合を通じて移動履歴の対象者を再特定することは，移動履歴への攻撃として有望である．しかし，従来方式は，移動履歴間の接触関係とソーシャルネットワークのリンク関係が利用できない場合，移動履歴数およびソーシャルネットワークアカウント数が小さい場合，両者が大きく異なる場合には対応できなかった．また，時間情報の利用が限られていた．

そこで，接触関係およびリンク関係を利用しない方式を提案し，実データを用いた評価により，移動履歴数およびアカウント数が小さい場合，両者が大きく異なる場合の有効性を明らかにした．また，移動履歴とソーシャルネットワークの地名投稿との時間差および距離に基づいて，本人らしさと他人らしさを定量化し，再特定の精度を向上させることができることを明らかにした．多数の候補者のアカウントの中から移動履歴と同一人物のアカウントを見つけるという新しい視点の評価を行い，その評価でも提案方式の有効性を明らかにした．また，移動履歴とサイドデータの両者のモデルの併用，および狭い領域の細粒度モデルと広い領域の粗粒度モデルの併用の有効性を明らかにした．

今後の課題としては，地理的領域を分割するメッシュのサイズ，時間インターバル，距離インターバル等のパラメータを変えて評価を行うことに加え，再特定の容易な移動履歴と困難な移動履歴の特徴に基づいて，移動履歴からのプライバシ漏洩の防止方法を考察することがあげられる．

[*6] 26 アカウントの場合と 27 アカウントの場合の平均を求めた．

参考文献

[1] Nergiz, M.E., Atzori, M., Saygin, Y. and Guc, B.: Towards trajectory anonymization: A generalization-based approach, *Transactions on Data Privacy*, Vol.2, No.1, pp.47–75 (2009).

[2] Shokri, R., Freudiger, J. and Hubaux, J.-P.: A unified framework for location privacy, *Proc. 3rd Hot Topics in Privacy Enhancing Technologies* (2010).

[3] Srivatsa, M. and Hicks, M.: Deanonymizing mobility Traces: Using Social Networks as a Side-Channel, *Proc. 19th ACM Conference on Computer and Communications Security*, pp.628–637 (2012).

[4] Shokri, R., Theodorakopoulos, G., Le Boudec, J.-Y. and Hubaux, J.-P.: Quantifying location privacy, *Proc. 32nd IEEE Symposium on Security and Privacy*, pp.247–262 (2011).

[5] Murakami, T.: Expectation-maximization tensor factorization for practical location privacy attacks, *Proc. 17th Privacy Enhancing Technologies Symposium*, pp.138–155 (2017).

[6] Ma, C.Y., Yau, D.K., Yip, N.K. and Rao, N.S.: Privacy vulnerability of published anonymous mobility traces, *IEEE/ACM Trans. Networking*, Vol.21, No.3, pp.720–733 (2013).

[7] Gambs, S., Killijian, M.-O. and Cortez, M.N.: De-anonymization attack on geolocated data, *Proc. 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications*, pp.789–797 (2013).

[8] Riederer, C. et al: Linking users across domains with location data: theory and validation, *Proc. 25th International Conference on World Wide Web*, pp.707–719 (2016).

[9] Murakami, T.: A succinct model for re-identification of mobility traces based on small training data, *Proc. 15th International Symposium on Information Theory and Its Applications*, pp.164–168 (2018).

[10] Manousakas, D. et al.: Quantifying privacy loss of human mobility graph topology, *Proc. 18th Privacy Enhancing Technologies Symposium*, pp.5–21 (2018).

[11] Narayanan, A. and Shmatikov, V.: De-anonymizing social networks, *Proc. 30th IEEE Symposium on Security and Privacy*, pp.173–187 (2009).

[12] Narayanan, A. et al.: On the feasibility of Internet-scale author identification, *Proc. 33rd IEEE Symposium on Security and Privacy*, pp.300–314 (2012).

[13] Overdorf, R. et al.: Blogs, Twitter feeds, and Reddit comments: cross-domain authorship attribution, *Proc. 16th Privacy Enhancing Technologies Symposium*, pp.155–171 (2016).

[14] Lee, W. et al.: Blind de-anonymization attacks using social networks, *Proc. 16th Workshop on Privacy in the Electronic Society*, pp.1–4 (2017).

[15] Monreale, A., Pinelli, F., Trasarti, R. and Giannotti, F.: WhereNext: A location predictor on trajectory pattern mining, *Proc. 15th ACM SIGKDD International Conference on Knowledge Discovery and Data mining*, pp.637–646 (2009).

© 2020 Information Processing Society of Japan

[16] Xue, A. et al.: Destination prediction by sub-trajectory synthesis and privacy protection against such prediction, *Proc. 29th IEEE International Conference on Data Engineering*, pp.254–265 (2013).
[17] Matsuo, Y., Okazaki, N., Izumi, K., Nakamura, Y., Nishimura, T., Hasida, K., and Nakashima, H.: Inferring long-term user properties based on users' location history, *Proc. 20th International Joint Conference on Artificial Intelligence*, pp.2159–2165 (2007).
[18] 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，国立情報学研究所（オンライン），入手先 ⟨https://geonlp.ex.nii.ac.jp/⟩（参照 2020-02-18）.
[19] scikit-learn: machine learning in Python, available from ⟨https://scikit-lern.org⟩ (accessed 2020-02-18).

松本 瞬

2019 年電気通信大学情報理工学部総合情報学科卒業．同大学大学院情報理工学系研究科情報学専攻博士前期課程在学中．

大岡 拓斗

2019 年電気通信大学情報理工学部総合情報学科卒業．同大学大学院情報理工学系研究科情報学専攻博士前期課程在学中．

市野 将嗣 （正会員）

2003 年早稲田大学理工学部電子・情報通信学科卒業．2008 年同大学大学院理工学研究科博士課程修了．2007 年日本学術振興会特別研究員．2009 年早稲田大学大学院基幹理工学研究科研究助手．2010 年同大学メディアネットワークセンター助手．2011 年電気通信大学大学院情報理工学研究科助教．2016 年同大学院情報理工学研究科准教授．バイオメトリクス，ネットワークセキュリティに関する研究に従事．博士（工学）．電子情報通信学会会員．

吉浦 裕 （正会員）

1981 年東京大学理学部情報科学科卒業．日立製作所を経て，2003 年電気通信大学勤務．現在，同大学大学院情報理工学研究科教授．情報セキュリティ，プライバシ保護の研究に従事．博士（理学）．日立製作所社長技術賞（2000 年），情報処理学会論文賞（2005 年，2011 年），日本セキュリティ・マネジメント学会論文賞（2010 年，2016 年，2018 年），システム制御情報学会産業技術賞（2005 年），IEEE IIH-MSP best paper award（2006 年），IFIP I3E best paper award（2016 年）等受賞．電子情報通信学会，日本セキュリティ・マネジメント学会，人工知能学会，システム制御情報学会，IEEE 各会員．本会フェロー．

© 2020 Information Processing Society of Japan