Time-aware multi-resolutional approach to re-identifying location histories by using social networks

Takuto Ohka
Graduate School of Informatics and Engineering
University of Electro-Communications
Tokyo, Japan
t.ohka@uec.ac.jp

Shun Matsumoto
Graduate School of Informatics and Engineering
University of Electro-Communications
Tokyo, Japan
s.matsumoto@uec.ac.jp

Masatsugu Ichino
Graduate School of Informatics and Engineering
University of Electro-Communications
Tokyo, Japan
ichino@inf.uec.ac.jp

Hiroshi Yoshiura
Graduate School of Informatics and Engineering
University of Electro-Communications
Tokyo, Japan
yoshiura@uec.ac.jp

***Abstract*—Identifying people from anonymous location histories is important for two purposes. i.e. to clarify privacy risks in using the location histories and to find evidence of who went where and when. Although linking with social network accounts is an excellent approach for such identification, previous methods need information about social relationships and have a limitation on the number of target data sets. Moreover, they make limited use of time information. We present models that overcome these problems by estimating the sameness and difference of people by using combinations of time and distance. Our proposed method uses these models along with multi-resolution models for both sides of linking, i.e. location histories and social network accounts. Evaluation using real data demonstrated the effectiveness of our method even when linking only one pseudonymized and obfuscated location history to 1 of 10,000 social network accounts without any information about social relationships.**

Keywords—Privacy, Re-identification, Location history, Social network

I. INTRODUCTION

Location histories of people, i.e. when and where people have been, are used for marketing, shop placement, and sophisticated advertising in the commercial sector and for tuning transportation systems and planning evacuations in the public sector. However, there is much concern about using location histories because malicious people can analyze them to estimate private information of individuals such as addresses, affiliations, human relationships, and unusual behaviors. Location histories are therefore often anonymized to hide the persons they represent [1, 2]. However, researchers have demonstrated that such hidden people can be "re-identified" by linking the location histories with other data so that the persons represented by the two sets of data are the same.

Location histories are also important for forensics in cyber and physical worlds. We can analyze them to estimate when and where people were. However, because location histories are usually anonymous, we cannot directly know who the people were. Study of re-identifying people from location histories is therefore important to clarify privacy risks as well as to use the location histories for forensic purposes.

Re-identification of location histories has been actively studied [3–11]. Srivatsa et al. proposed linking location histories to social network accounts [3], which is an effective approach because social network data are generally easy to obtain. Moreover, it is often easier to re-identify social network accounts than to re-identify location histories directly. However, Srivatsa et al.'s method requires two assumptions: (1) people represented by location histories and social networks have social relationships that are observable in both sets of data, and (2) the number of location histories is not too small and is not too different from the number of social network accounts. Murakami proposed a method for linking location histories to mobile profiles that correspond to individuals [4]. He asserted that his method can be used to link location histories and social network accounts by deriving mobile profiles from social network data. However, his method makes limited use of time-related information because a mobile profile in his method is represented by a set of probabilities of moving from one location to another. It thus uses only the order in which two locations were visited.

Using time information more effectively could improve the precision of re-identifying location histories. However, it is not easy to put this idea into practice because the combination of location and time does not affect re-identification uniformly—the effect depends on the person.

In this paper, we present a method that links location histories and social network accounts based on the content of these data instead of the relationships between them. Because the method selects a corresponding social network account for each location history independently, it works for any numbers of location histories and social network accounts. It improves the accuracy of re-identification by making use of time information in a novel way. Our contributions are summarized as follows.

978-1-7281-8915-4/20/$31.00 ©2020 IEEE
DOI 10.1109/QRS-C51114.2020.00038

- Our method does not assume any social relationship between people represented by the data and it works when the numbers of people represented by the two sets of data are greatly different, e.g. it works for the 1 vs. 10,000 case.
- How the combination of time and distance affects re-identification is learned for each (unknown) person from the location history and social network account data.
- Evaluation using 53 location histories and 100,053 social network accounts demonstrated the re-identification performance of our proposed method as well as its ability to search millions of social network accounts for the one that correspond to the given location history.

II. Related work

A. Overview of re-identifying location histories

Hereafter we use ***LH*** for location history and ***AC*** for social network account. Methods for re-identification have been investigated in various data domains such as LHs [3–11], social networks [12–15], Web browsing histories [16], purchase records [17], and databases [18]. A few methods have been developed for general data types [19, 20]. We focus on re-identification of LHs and social networks.

Re-identification of LHs. Re-identification of an LH has been generally approached by linking it to another data record so that both data records represent the same person. If the other data record is not anonymous, the LH is directly re-identified and, even if the data record is anonymous, the attacker obtains more clues to re-identify the LH.

Shokri et al. developed a method that re-identifies pseudonymized obfuscated LHs by linking them to mobile profiles that correspond to individuals [5]. Srivatsa et al. linked pseudonymized LHs to social network accounts (ACs) by matching a social graph generated from LHs to one generated from ACs [3]. The methods of Shokri et al. and Srivatsa et al. will be analyzed in detail in Section 2.2.

Ma et al. linked pseudonymized noisy LHs to the original LHs. They assumed probabilistic distributions of noises and performed maximum likelihood estimation on the basis of the probability of pseudonymized noisy LHs being generated from the original LHs [6]. Gambs et al. generated a transition matrix between two locations from each LH and identified pairs of LHs that represented the same person on the basis of the similarity between the two corresponding transition matrices [7]. Riederer et al. assumed a Poison distribution for the probability of a person visiting each location in the target area. They estimated the degree of two LHs representing the same person by calculating the likelihood of the LH amalgamated from the two LHs on the basis of the assumed Poison distribution [8]. Murakami enhanced Shokri et al.'s method to enable it to cope with mobile profiles that have little information, which will be detailed in Section 2.2 [4, 9]. Manousakas et al. used the topology of graphs in which a node represents a location and an edge represents the transition between locations [10].

Re-identification of social networks. Narayanan et al. developed two classical methods that use graph matching [12] and machine learning [13]. Most recent methods are based on one of these two methods [14] [15].

B. Re-identifying LHs using social networks

Srivatsa et al. re-identified LHs using real data from social networks [3]. Murakami asserted that his method can also re-identify LHs using social networks [4]. Since Murakami's method is based on that of Shokri et al. [5], we analyzed the methods of Srivatsa et al., Shokri et al., and Murakami.

Srivatsa et al.'s method transforms a set of LHs into a graph in which each node represents the LH of a person [3]. An edge between two nodes represents contact between the two corresponding people, meaning that they were in close proximity of each other for at least a specified period. A set of ACs are also transformed into a graph in which each node represents an AC, and an edge between two nodes represents a link (such as friendship link) between them. The two graphs are matched, and each pair of nodes in the two graphs is linked.

Srivatsa et al.'s approach is effective because social network data are generally easy to obtain. ACs are often accompanied with real names. Even if they are not, they may be easier to re-identify than to re-identify LHs directly. However, their method requires two assumptions, which are too strong in many real situations.

1. People represented by LHs and ACs have social relationships that are observable in both sets of data.
2. The number of LHs is not too small and is not too different from the number of ACs.

The first assumption is too strong because it does not hold if the people represented by the LHs are not socially related. It does not hold either if the LHs are obfuscated as is usual in real world use. For examples, with perturbation of location up to 100 m, we cannot know whether two persons were spatially close or 100-m distant and, with time perturbation, we cannot know whether they were spatially close or only visited the same place at different times.

The second assumption is also too strong. Assume, for example, three LHs that have a mutual relationship (i.e. the graph for them is a triangle) and three ACs that also have a mutual relationship. There are six equally possible mappings between the two graphs and we cannot re-identify at all. Assume we have 10 LHs to re-identify but have ACs of 1000 candidate people because we cannot narrow down the candidates. The graph of 10 LHs probably matches multiple subgraphs of the graph of 1000 ACs. Thus, we also cannot re-identify.

In Shokri et al.'s method [5] [11], knowledge about each person's mobility (e.g. location of home and workplace) is used to generate a model of the person, which is called a mobile profile. The model is represented by a matrix in which the elements represent the probability of transition between two locations. For each pair of a pseudonymized obfuscated LH and a transition matrix (i.e. the person's mobile profile), the

probability of the LH being produced from the matrix is calculated. The calculated probabilities are used to identify the mapping between the LHs and matrices.

Murakami enhanced Shokri et al.'s method by forming a tensor from the transition matrices of people (i.e. their mobile profiles) [4]. Elements of the profile are complemented by factorizing the tensor using information from other elements of the same profile as well as using information from other profiles. Murakami's method thus copes with incomplete knowledge about the mobility of people. He asserted that knowledge about mobility can be derived from social networks.

Although an LH and knowledge about a person's mobility have information of time as well as of locations, Shokri et al.'s and Murakami's methods use limited time information, i.e. the order in which two locations were visited, because the basic tool for representing a model is a transition matrix.

III. Problem Statement

A. Terminology

- Data item: Quadrant consisting of a pseudonym, latitude, longitude, and time, where pseudonym is consistent for each person.
- Location history (LH): Set of data items, as shown in Fig. 1.
- $S^{LH}=\{LH_1, LH_2, \cdots, LH_i, \cdots, LH_M\}$: Set of LHs that are the targets of re-identification, where $M=|S^{LH}|$. Note that S^{LH} is a set of sets because each element, i.e. each LH, is a set of data items.
- $S^{AC}=\{AC_1, AC_2, \cdots, AC_j, \cdots, AC_N\}$: Set of social network accounts (ACs) used to re-identify S^{LH}, where $N=|S^{AC}|$.
- S^{Person}: Set of people including those represented by S^{LH} and S^{AC}. It may also include other people.
- $Person^{LH}$: Mapping from S^{LH} to S^{Person} such that $Person^{LH}(LH)$ is the person represented by LH.
- $Person^{AC}$: Mapping from S^{AC} to S^{Person} such that $Person^{AC}(AC)$ is the person represented by AC.

$Person^{LH}(LH)$ and $Person^{AC}(LH)$ are abbreviated Person(LH) and Person(AC) when obvious. LH_i and AC_j are used to denote an arbitrary LH in S^{LH} and AC in S^{AC}, respectively. Fig. 2 illustrates the problem structure.

Pseudonym	Latitude	Longitude	Time	
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

Fig. 1. Example location history

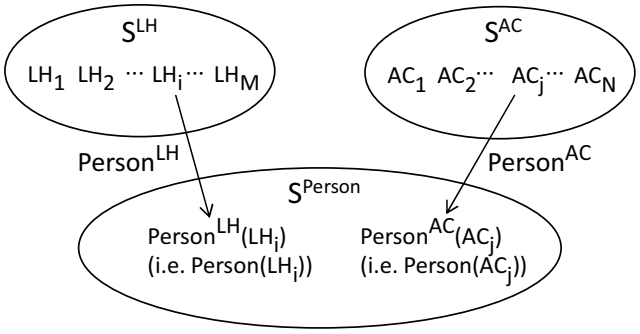

Fig. 2. Problem structure

- Same(i, j): Proposition that $Person(LH_i)$ and $Person(AC_j)$ are the same person.
- Model(<model descriptor>, AC_j): Model of $Person(AC_j)$'s mobility, generated from data of AC_j; used to calculate Score(i, j) for $1 \le i \le M$. Because multiple models are generated for the same AC, a model descriptor is used to discriminate them. Mobile profiles like those used in Shokri et al.'s and Murakami's methods are kinds of model.
- Model(<model descriptor>, LH_i): Data generated from LH_i; used to calculate Score(i, j) for $1 \le j \le N$.
- Score(<model>, i, j): Numerical value calculated using a model representing the likelihood of Same(i, j).
- Score(i, j): Numerical value representing the overall likelihood of Same(i, j), which is obtained by fusing Score(<model>, i, j)s from multiple models.

B. Our problem

Given $S^{LH}=\{LH_1, LH_2, \cdots, LH_i, \cdots, LH_M\}$ and $S^{AC}=\{AC_1, AC_2, \cdots, AC_j, \cdots, AC_N\}$, we attempt to estimate the correct mapping $\mathcal{M}^{correct}$ from S^{LH} to S^{AC}, which satisfies Same(i, j') for $AC_{j'}=\mathcal{M}^{correct}(LH_i)$ and $1 \le i \le M$. Note that we neither know $Person^{LH}$ nor any pair of LH and AC that represents the same person when we start this estimation. The estimation results in the answer, i.e. estimated mapping $\mathcal{M}^{est}$. The metric used for evaluating our answer is $|S^{correct}|/M$, where $S^{correct}=\{LH \mid LH \in S^{LH}, \mathcal{M}^{est}(LH)=\mathcal{M}^{correct}(LH)\}$.

To overcome the problems in previous methods, we aimed to achieve three goals.

1. The method should not require an assumption about any relationship among people represented by LHs and those represented by ACs.
2. The method should work with any number of LHs and ACs; specifically, it should work when there is only one LH and when the number of LHs is far less than that of ACs.
3. The method should use more time information than simply the order in which two locations were visited.

IV. DATASETS

A. Location history dataset

We obtained LHs for 53 volunteers in cooperation with a Wi-Fi service company. The company collected records of probe requests that included the volunteers' MAC addresses obtained at 400,000 nationwide Wi-Fi base stations in Japan[1]. The collection period was 90 days (January 25 through April 23, 2017). Data was not obtained for all 90 days for some volunteers due to stoppage of the Wi-Fi function on their device. The company provided the records after transforming them into a set of data items, each of which is a quadrant consisting of a pseudonym for the volunteer's MAC address, the latitude and longitude of the Wi-Fi base station, and the time of the probe request. The location of the probing device was within 100 m of the base station as that was the maximum distance at which probe requests could be received. Table shows the statistics for the volunteers' LH data, where the average and median were taken over the 53 volunteers.

B. Social network dataset

We obtained access to the Twitter accounts of the 53 volunteers, collected the tweets on each account at the end of April, 2017, and thereby obtained 3645.6 tweets per account.

We applied a tool for natural language analysis to these tweets to identify the tweets that may refer to places. We analyzed these identified tweets and found that "places" were mentioned in one of five ways: (1) the volunteer was at the place mentioned in the tweet at the time of the tweet (e.g. "I'm drinking in Shinjuku"), (2) the volunteer was at the place mentioned in the tweet at a time different from that of the tweet (e.g. "I drank in Shibuya last week"), (3) the volunteer was not at the mentioned place (e.g. "I wish I could live in Kyoto"), (4) there was ambiguity about whether the tweet belonged in (1) through (3) without additional information such as a photo, and (5) the "place" was not actually a place but another thing such as the name of a person or a company.

TABLE I. STATISTICS FOR VOLUNTEERS' LOCATION HISTORY DATA

	Average	Median
No. of data items	21,361.36	13,068
No. of data items per day	244.90	161.04

TABLE II. STATISTICS FOR VOLUNTEERS' TWITTER ACCOUNT DATA

	Average	Median
No. of times that place names were tweeted	67.54 116.58	13 50
No. of times that place names were tweeted per day	0.69 1.31	0.34 0.96

Type (1) tweets are useful for correct re-identification. Type (2) tweets are partially useful for an appropriate re-identification algorithm. Some type (4) tweets are potentially type (1) or (2) and therefore are useful or partially useful. Type (3) tweets are noises that lead to incorrect linking. Type (5) tweets are also noises that are inevitable when we automate our method by using a natural language analyzer.

Table shows the statistics for the volunteers' AC data reflecting this analysis. The top number in each cell is for type (1) tweets, and the bottom number is the total for types (1), (2), and (4). The two numbers respectively represent the infimum and supremum of the number of tweets useful for correct re-identification; the numbers are surprisingly small.

V. PROPOSED METHOD

A. Linking different types of data

Because we must not use any relationship between people represented by LHs and ACs, we based our method on the content of each LH and AC. The main contents of an AC are texts, pictures, and user profile. We use only the texts for our method. Because an LH is represented as a set of quadrants, as shown in Fig. 1, and texts are represented in natural language, we must first unify their data types to enable matching. We convert the texts into quadrants by extracting the place names from them, converting each place name into a coordinate value (i.e. a pair of latitude and longitude), and using the times of the posts. We use Google's Geocoding API [21] to convert the place names into coordinate values.

B. Models of people's mobility

For each account AC_j, we generate three models that represent the mobility of Person(AC_j). The first and second models represent the location-visiting pattern of Person(AC_j). They are based on the frequency of visiting each location (i.e. the number of times each location was visited) as represented by the data of AC_j. The difference between these two models is in area and resolution. The first model (the ***small-fine model***) uses detailed discrimination for locations in the region of Person(AC_j)'s daily life, and the second model (the ***large-coarse model***) uses rough discrimination for locations in the wider region. For each LH_i, the number of times each location was visited is counted, and the totals are input into the small-fine and large-coarse models of Person(AC_j) to calculate Score(small-fine AC_j, i, j) and Score(large-coarse AC_j, i, j).

The third model of Person(AC_j), the ***time-aware model***, utilizes the time information to estimate the sameness and difference of people. It reflects physical impossibility, i.e. a person cannot be at two locations at the same time and cannot move between distant locations in an impossibly short time. The time-aware model also reflects our intuition that the greater the number of pairs of data items from an LH_i and an AC_j that are close in location and time, the greater the probability of Same(i, j). The data items in LH_i and those in AC_j are compared to calculate time difference and distance between them. The

[1] Each volunteer gave written permission for us to obtain their location and Twitter data. The ethical committee of our university authorized this research.

calculated time difference and distance are input into the time-aware model of AC_j to obtain Score(time-aware AC_j, i, j). Details of generating and using these models are described in Section 6.

C. Dual models

We generate not only models of Person(AC_j) but also models of Person(LH_i) for each LH_i though the identify of Person(LH_i) is unknown. These models are used to calculate Score(small-fine LH_i, i, j), Score(large-coarse LH_i, i, j), and Score(time-aware LH_i, i, j). Thus, we generate and use three models for each side of the linking.

D. Basic flow

Perform preprocessing. Texts posted on each AC_j are transformed into a set of data items, each of which is a quadrant (pseudonym, latitude, longitude, time).

Generate models for ACs. The data items of each AC_j are transformed into three kinds of data for the small-fine, large-coarse, and time-aware models, from which three models, i.e. Model(small-fine, AC_j), Model(large-coarse, AC_j), and Model(time-aware, AC_j) are generated.

Generate models for LHs. Three models, i.e. Model(small-fine, LH_i), Model(large-coarse, LH_i), and Model(time-aware, LH_i), are similarly generated for each LH_i.

Use models to calculate scores. The three models for each AC and those for each LH are used to calculate six scores for each pair of LH_i and AC_j; the scores are fused into Score(i, j).

Link LHs and ACs. Each LH_i is linked to $AC_{j'}$, where $j' = Argmax_{(1 \le j \le N)}\big(Score(i,j)\big)$.

VI. Building and using models

The six models of people's mobility introduced in Section 5.2 are described in detail here along with how they are generated and used.

A. Frequency-based models

We use the term ***region*** for the entire geographical area considered for re-identification and the term ***cell*** for a sub-region, i.e. the unit used for representing a location.

Feature used for generating models. While the amount of data in social networks useful for re-identification is limited, as shown in Table □, the region we considered is wide, i.e. the whole of Japan, leading to very sparse data. Murakami demonstrated that, when data are sparse, a simple model based on the frequency of visiting each cell is better than one based on the frequency of moving between cells [9]. We therefore used the frequency of visiting each cell to generate the models.

Multiple resolutions for representing models. Because the volunteers lived in or around Tokyo, most locations in their LHs were in the Tokyo area. However, they sometimes travelled to distant locations. Such outliers are strong clues for re-identification. However, if the region considered is large to include distant locations, e.g. the whole country, we must consider many cells, making our data sparse with little AC information. We therefore use two models, i.e. a small-fine model with fine-grained cells for the region around Tokyo and a large-coarse model with coarse-grained cells for the whole country. In the evaluations described in Section 7, the region for the small-fine model was 126 km × 126 km centred around Tokyo with a cell size of 1 km^2. The region and cell size for the large-coarse model were 2370 km × 2140 km and 5 km^2.

Generating models. Texts posted on each AC_j ($1 \le j \le N$) are transformed into quadrants, from which feature vectors are generated for the small-fine and large-coarse models. The value of the k-th element in a feature vector is the number of times the k-th cell was visited. Models for AC_j are generated by using a machine learning algorithm with these feature vectors as positive and negative samples. The positive samples are feature vectors generated from AC_j data, and the negative samples are those generated from data from other ACs. Models for each LH_i are generated similarly.

Using models. Given LH_i, we generate feature vectors for the small-fine and large-coarse models in the same way as in the generation phase. These two sets of feature vectors are input into Model(small-fine, AC_j) and Model(large-coarse, AC_j) to obtain Score(small-fine AC_j, i, j) and Score(large-coarse AC_j, i, j), respectively. Given AC_j, we similarly generate two sets of feature vectors and input them into Model(small-fine, LH_i) and Model(large-coarse, LH_i) to obtain Score(small-fine LH_i, i, j) and Score(large-coarse LH_i, i, j), respectively. We thereby obtain four scores for each i and j.

B. Time-aware models

Dividing a model on the basis of time, as suggested by Shokri et al. [5], is not effective because it makes the sparse data even sparser. We therefore use time information based on physical impossibility. However, the determination of physical impossibility depends on the person. For example, it is unlikely that an elderly person can walk 1 km in 10 minutes while it is likely for a younger person. Thus, the probability of Same(i, j) for an elderly person is less than that for a younger person when a data item in LH_i has location L1 and time T1 while a data item derived from AC_j has L2 being 1 km from L1 and T2 being 10 minutes after T1. Because the attribute information (such as age and gender) of a person is not available for re-identification, a model of physical impossibility must be learned for each (unknown) person from the person's LH and AC data.

Moreover, we attempted to model our intuition that the greater the number of pairs of data items from an LH_i and an AC_j that are close in location and time, the greater the probability of Same(i, j). Because the validity of this intuition also depends on the person, a model reflecting this intuition must be learned for each person on the basis of his or her data.

Overview of learning and using models. A time-aware model that represents these two properties (i.e. physical impossibility and spatial-temporal closeness) is learned on the basis of the number of data items that satisfy each combination of time and distance interval. Fig. 3 illustrates the data structure, a ***time-distance matrix***, used to do this. Each row in the matrix

represents a time interval such that T_k represents the interval $[T_k, T_{k+1})$, and each column represents a distance interval such that D_l represents the interval $[D_l, D_{l+1})$, where $1 \leq k \leq K$, $1 \leq l \leq L$, and K and L are the number of time and distance intervals, respectively. The element in row T_k and column D_l represents the intersection of $[T_k, T_{k+1})$ and $[D_l, D_{l+1})$. A time-aware model is learned for each LH by filling each cell of the matrix with the number of times that pairs of data items satisfy the corresponding time-distance condition. The symbol c_{kl} represents the number in the cell(k, l). A time-aware model is similarly learned for each AC.

A time-aware model learned from pairs of data items for a person should have smaller values in the upper-right cells, which correspond to pairs of data items close in time but distant in location (i.e. representing physical impossibility). In contrast, it should have larger values in the upper-left cells, which correspond to pairs of data items close in location and time (i.e. representing spatial-temporal closeness).

In the time-aware-model usage phase, we assume that an LH and an AC represent the same person and the cells in the time-distance matrix are filled using pairs of data items from the LHi and ACj. The resultant matrix is checked to determine whether it has the same pattern as those of the time-aware models of LHi and ACj, i.e. whether the LH and AC truly represent the same person.

	D_1	…	D_l	…	D_L
T_1	c_{11}		c_{1l}		c_{1L}
…					
T_k	c_{k1}		c_{kl}		c_{kL}
…					
T_K	c_{K1}		c_{Kl}		c_{KL}

Fig. 3. Time-distance matrix for learning and using time-aware model

Details of learning models. The data items (i.e. quadrants) in each LH are divided into subsets under the condition that the period (i.e. start and end times) of each subset are common among LHs (Fig. 4 (a)). A feature vector for a positive example is generated from each subset of quadrants as follows. We consider all pairs of quadrants in the subset. For example, if the subset consists of three quadrants, q1, q2, and q3, we consider six pairs (Fig. 4 (b)). For each pair of quadrants, we calculate the distance between the two locations and the time difference. We then add frequency values to the corresponding cells in the time-distance matrix (shown in Fig. 3). The resultant matrix represents how far Person(LHi) moves in a specific time interval. It is transformed into one dimension, which is a feature vector of a positive example for Model(time-aware, LHi).

A negative example is generated from each subset α of quadrants in LHi. For each of the other LHs, we take a subset β of quadrants so that the periods of α and β coincide. We consider all pairs of a quadrant from α and one from β and, by using these pairs, we add frequency values to the time-distance matrix as we do for positive examples (Fig. 4(c)). The resultant matrix represents the distribution of the time differences and distances between Person(LHi) and persons represented by other LHs. These positive and negative examples are used to learn Model(time-aware, LHi), as shown in Fig. 4(d).

A time-aware model for each ACj is similarly learned following a pre-process that transforms text in ACj into quadrants.

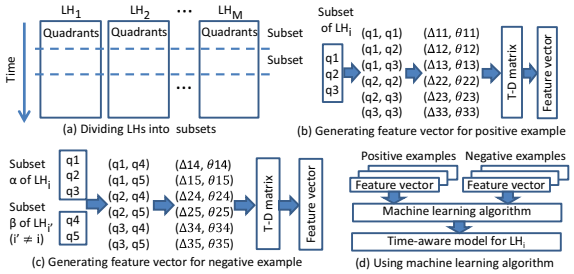

Fig. 4. Learning time-aware model for LHi (Δab and θab represent time difference and distance between quadrants qa and qb)

Details of using models. Given LHi and ACj, we divide the LHi and ACj data in the same way as in the learning phase. We consider all pairs of a subset of quadrants from LHi and one from ACj under the condition that the two subsets have the same period. For each pair of subsets, say γ and δ, we consider all pairs of a quadrant from γ and one from δ. For example, if γ has five quadrants and δ has two quadrants, we consider ten pairs. Using these pairs, we add frequency values to the time-distance matrix as in the learning phase. The resultant matrix represents the distribution of time differences and distances between Person(LHi) and Person(ACj). The matrix is then transformed into a feature vector. The feature vectors generated from LHi and ACj are input into Model(time-aware, ACj) and Model(time-aware, LHi) to obtain two score values for the time-aware models, i.e. Score(time-aware ACj, i, j) and Score(time-aware LHi, i, j), for each i and j.

VII. Evaluation

A. Implementation and preliminary evaluation

We fused the six scores into Score(i, j) by normalizing each of the scores and averaging them.

For the time-aware model, we instantiated the time-distance matrix with (0, 10, 20, 30, 60, 120, 180, 360) for time intervals and with (0, 1, 2, 4, 8, 16) for distance intervals, where the units for time and distance were minutes and kilometres, respectively. Evaluation using alternative parameters is left for future work.

For the time-aware models, the data for each LH and AC were divided into subsets with a period of one day because physical impossibility works more effectively for a pair of data items for which the times are close. For the two frequency-based models, the LH data were also day-wise divided into subsets whereas we tested several options for dividing the AC data. We used scikit-learn [22] for implementing machine learning and tested three machine learning algorithms, i.e. logistic regression using linear discrimination plane, SVM (support vector machine) with a radial basis function kernel using a non-linear discrimination plane, and XGBoost (eXtreme Gradient Boosting) using decision trees.

We evaluated the three machine learning algorithms for the time-aware models and found that XGBoost performed best. For the two frequency-based models, we evaluated all combinations of the three machine learning algorithms and the options for dividing the AC data. We found that logistic regression performed best with a period of 245 days for a subset of AC data[2]. The evaluation results presented below were obtained using these best combinations.

B. Evaluations with original LHs

Each evaluation was performed ten times because the machine learning algorithms used have randomness. The rates of correct linking shown hereafter are average values.

53 vs. 53 evaluation. The 53 LHs of the volunteers were linked to their ACs after the correspondence between the LHs and ACs was made unknown to the method. Table shows the results, where row A1 shows the number and rate (%) of LHs that were linked to the correct ACs using the small-fine AC models. Row A3 shows the number and rate using the small-fine and large-coarse AC models, and row A5 shows the number and rate using the small-fine, large-coarse, and time-aware AC models. Rows D1 through D5 show the results using dual models, e.g. D1 shows the results using the AC and LH small-fine models, and D5 shows the results using all six models. From Table , we can see that combining models with different resolutions (small-fine and large-coarse) was effective; i.e. the rates for A3/L3/D3 were better than those for A1/L1/D1 and those for A2/L2/D2 with only one exception (A3 was worse than A1). The time-aware models were also effective; i.e. the rates for A5/L5/D5 were better than those for A3/L3/D3. The dual-model approach (combinations of LH and AC models) was also effective; i.e. the rates for D1–D5 were better than those for A1–A5 and L1–L5. Table also shows that using all six models (D5) resulted in the best performance, and the rates shown hereafter are those obtained with this best mode unless otherwise noted.

TABLE III. RESULTS OF 53 VS. 53 EVALUATION

ID	Model	No. of LHs linked correctly (%)
A1	small-fine, AC	17.2 (32.5%)
A2	large-coarse, AC	13.8 (26.0%)
A3	A1 & A2	16.8 (31.7%)
A4	time-aware, AC	24.4 (46.0%)
A5	A1 & A2 &A4	33.6 (63.4%)
L1	small-fine, LH	18.0 (34.0%)
L2	large-coarse, LH	20.2 (38.1%)
L3	L1 & L2	23.2 (43.8%)
L4	time-aware, LH	17.2 (32.5%)
L5	L1 & L2 &L4	29.4 (55.5%)
D1	small-fine, Dual	23.4 (44.2%)
D2	large-coarse, Dual	22.6 (42.6%)
D3	D1 & D2	28.6 (54.0%)
D4	time-aware, Dual	24.2 (45.7%)
D5	D1 & D2 & D4	39.8 (75.1%)

53 vs. many evaluations. We hid the ACs of the 53 volunteers among 1000, 10,000, and 100,000 noise ACs by inserting each of the volunteers' ACs randomly into the list of the noise ACs. We then attempted to link the 53 LHs to the 53 ACs hidden in the noise ACs. Table shows the rate of an LH being correctly linked and the rate of an LH for which the correct AC was among the top 100 scored ACs. As shown, 15.1% of the LHs were correctly linked to ACs hidden in the 1000 noise ACs. Furthermore, if we assume that the linkability of 100 ACs can be checked more precisely with the human eye, 39.8% of the LHs might have been correctly linked to ACs hidden in the 100,000 noise ACs.

The total processing time for the 53 vs. 100,053 evaluation was 6.8 days, including 60 hours for obtaining data from ACs, 96 hours for generating 53 LH models and 100,053 AC models, and 8 hours for re-identification. The processing time is proportional to the number of ACs if we limit the number of other ACs for generating negative samples and limit the number of LHs. We used a workstation having a 12-core Intel i-9 CPU and 128-GB memory. It cost 3600 USD. We also used three PCs, each of which cost 450 USD, to obtain the AC data. Thus, it is practical to search millions of candidate ACs for ones to be linked with LHs.

TABLE IV. RESULTS OF 53 VS. MANY EVALUATION

No. of ACs	53	1053	10,053	100,053
Correct linkage (%)	75.1	15.1	0.0	0.0
Top 100 scores (%)	—	92.5	61.1	39.8

1 vs. 53 and more evaluations. An LH was linked to one of the 53 ACs either not hidden or hidden in 1000, 10,000, or 100,000 noise ACs. This trial was performed for each of 53 LHs. Because we used only one LH in each trial, we could not generate a model for the LH because negative examples from other LHs were not available. Thus, we used only the three AC models instead of using all six models. As shown in Table , the rates for 1 vs. 53 correct linking, 1 vs. 1053 top 100 scores, and 1 vs. 10,053 top 100 scores were almost the same as those for 53 vs. 53 or more evaluations (Table).

TABLE V. RESULTS OF 1 VS. 53 OR MORE EVALUATIONS

No. of ACs	53	1053	10,053	100,053
Correct linkage (%)	63.4	3.8	0.0	0.0
Top 100 scores (%)	—	84.5	58.5	9.4

C. Evaluations with obfuscated LHs

We used only parts of the LHs (data for one month, one week, and one day) instead of the data for all three months. Shortening the LHs corresponds to a type of obfuscation, i.e. frequently changing a person's pseudonym. We also used only

[2] In Tables and , the number of LH data items per day is roughly 245 while that of AC data items is roughly 1. A period of 245 days for a subset of ACs led to roughly the same number of data items included in the LH and AC data subsets.

the large-coarse models. This limited model usage corresponds to another kind of obfuscation, i.e. degrading the resolution of the location data from one second and 100 meters to one day and 5 km. (Note that a feature vector for a large-coarse model is generated from the number of times each 5-km^2 cell is visited in one day.)

As shown in Tables □ and Ⅶ, there was still the risk of re-identification with one-week LHs despite the short periods and low resolution in both the 53 vs. 53 or more case and the 1 vs. 53 or more case.

TABLE VI. RESULTS OF 53 VS. 53 OR MORE WITH SHORT AND LOW-RESOLUTION LHS

Target	No. of ACs	3 months	1 month	1 week	1 day
Correct linkage (%)	53	42.6	31.9	26.2	13.0
	1053	7.9	7.0	0.0	0.0
Top 100 scores (%)	10,053	47.5	32.5	30.6	17.7
	100,053	19.6	14.7	9.4	2.6

TABLE VII. RESULTS OF 1 VS. 53 OR MORE WITH SHORT AND LOW-RESOLUTION LHS

Target	No. of ACs	3 months	1 month	1 week	1 day
Correct linkage (%)	53	26.0	24.3	22.3	7.4
	1053	1.5	1.5	0.0	0.0
Top 100 scores (%)	10,053	33.6	24.9	21.7	10.6
	100,053	0.4	0.9	0.4	1.1

VIII. CONCLUSION

We identified problems in linking location histories with social networks, i.e. the need for information about social relationships, a limitation on the number of target data sets, and the insufficient use of time information. Our proposed models make better use of time information by enabling the sameness and difference of people to be estimated on the basis of both time and location. Our proposed method uses these models along with multi-resolution models for both sides of linking, i.e. location histories and social network accounts. Evaluation using real data demonstrated the effectiveness of our method even for 1 pseudonymized and obfuscated location history vs. 10,000 social network accounts without any information about social relationships.

Future work includes evaluating the method using different data sets and parameters, using meta-learning to fuse the scores, and using pictures and user profiles on social network accounts as well as texts. Another research direction is to study how to utilize location histories while alleviating the privacy risk such as the study of tradeoff between the utility and risk of location histories [23].

REFERENCES

[1] M. E. Nergiz, M. Atzori, Y. Saygin, and B. Guc, "Towards trajectory anonymization: A generalization-based approach", Transactions on Data Privacy 2(1), pp. 47-75, 2009.

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

[3] M. Srivatsa, and M. Hicks, "Deanonymizing mobility traces: using social networks as a side-channel." In: 19th ACM Conference on Computer and Communications security, pp. 628–637, 2012.

[4] T. Murakami, "Expectation-maximization tensor factorization for practical location privacy attacks.", In: 17th Privacy Enhancing Technologies Symposium, pp. 138-155, 2017.

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

[6] C. Y. Ma, D. K. Yau, N. K. Yip, and N. S. Rao, "Privacy vulnerability of published anonymous mobility traces." IEEE/ACM Transactions on Networking 21(3), 720-733, 2013.

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

[8] C. Riederer, Y. Kim, A. Chaintreau, N. Korula, and S. Lattanzi, "Linking users across domains with location data: theory and validation." In: 25th International Conference on World Wide Web, pp. 707-719, 2016.

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

[10] D. Manousakas, C. Mascolo, A. R. Beresford, D. Chan, and N Sharma, "Quantifying privacy loss of human mobility graph topology." In: 18th Privacy Enhancing Technologies Symposium, pp. 5-21, 2018.

[11] R. Shokri, G Theodorakapoulos, G Danezis, J.-P. Hubaux, and J.-Y. Le Boudec, "Quantifying location privacy: The case of sporadic location exposure." In: 11th Privacy Enhancing Technologies Symposium, pp. 57-76, 2011.

[12] A. Narayanan, and V. Shmatikov, "De-anonymizing social networks." In: 30th IEEE Symposium on Security and Privacy, pp. 173-187, 2009.

[13] A. Narayanan, H. Paskov, N. Z. Gong, J. Bethencourt, E. Stefanov, E. C. R. Shin, and D. Song, "On the feasibility of Internet-scale author identification." In: 33rd IEEE Symposium on Security and Privacy, pp. 300–314, 2012.

[14] R. Overdorf, and R. Greenstadt, "Blogs, Twitter feeds, and Reddit comments: cross-domain authorship attribution." In: 16th Privacy Enhancing Technologies Symposium, pp. 155-171, 2016.

[15] W.-H. Lee, C. Liu, S. Ji, P. Mittal, and R. B. Lee, "Blind De-anonymization attacks using social networks." In: 16th Workshop on Privacy in the Electronic Society, pp. 1-4, 2017.

[16] J. Su, A. Shukla, S. Goel, and A. Narayanan, "De-anonymizing Web browsing data with social networks." In: 26th International Conference on World Wide Web, pp. 1261-1269, 2017.

[17] T. Minkus, and K. Ross, "I know what you're buying: privacy breaches on eBay." In: 14th Privacy Enhancing Technologies Symposium, pp 164-183, 2014.

[18] S. K. K. Santu, V. Bindschadler, C. X. Zhai, and C. A. Gunter, "NRF: A Naive re-identification framework." In: 17th Workshop on Privacy in the Electronic Society, pp. 121-132, 2018.

[19] G. Danezis, and C. Troncoso, "You cannot hide for long: de-anonymization of real-world dynamic behavior." In: 12th Workshop on Privacy in the Electronic Society, pp. 49-59, 2013.

[20] S. Ji, W. Li, M. Srivatsa, and R. Beyah, "Structural data de-anonymization: quantification, practice, and implications." In: 21st ACM Conference on Computer and Communications Security, pp.1040-1053, 2014.

[21] Geocoding API, https://developers.google.com/maps/documentation/geocoding/intro, last accessed 2020/07/11.

[22] scikit-learn: machine learning in Python, https://scikit-learn.org, last accessed 2020/06/24. scikit-learn: machine learning in Python, https://scikit-learn.org, last accessed 2020/07/11.

[23] D. Calacci, A. Berke, K. Larson, and A. S. Pentland, "The tradeoff between the utility and risk of location data and implications for public good." http://arxiv-export-lb.library.cornell.edu/pdf/1905.09350, last accessed 2020/07/11.