Linking User Identities Across Social Networks via Frequency Domain Analysis

Hui Xue*†, Bo Sun‡§, Weixuan Mao‡
*Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
†School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
‡National Computer Network Emergency Response Technical Team/Coordination Center of China, Beijing, China
Emails: xuehui@iie.ac.cn, {sunbo, maoweixuan}@cert.org.cn

***Abstract*—User identity linkage refers to linking different social accounts belonging to the same natural person. Now user identity linkage across social networks based on spatiotemporal data has attracted more and more attention. However, the existing methods have some problems, such as trajectory processing is not suitable for sparse data, and grid processing leads to information loss and abnormality. Because of the above problems, we propose an accurate and efficient method of user identity linkage via wavelet transform, WTLink, which expresses the user identity in the form of several critical points obtained through a novel wavelet transform application mode. Then the user identities are linked by calculating the similarity between their representations with a proposed metric. We compare this method with several existing user identity linkage methods based on spatiotemporal data on real datasets. The results show that this method exceeds the baseline methods in terms of effectiveness and efficiency.**

***Index Terms*—spatiotemporal data, user identity link, social network, location, frequency domain**

I. Introduction

With the popularity of GPS devices such as cars, mobile phones and smart bracelets, the availability of spatiotemporal data is increasing. Recently, many social networks have generated more spatiotemporal data, such as Foursquare, Twitter and Instagram. Many users have registered accounts on these platforms and published information with geographic locations. This information builds a bridge between the real world and the virtual world, provides an unprecedented opportunity to analyze users' real-world behaviour, and has the potential to improve various location-based services. Researchers use similar spatiotemporal data to connect user accounts of different social network platforms, that is, user identity linkage (UIL) of the same natural person from different social networks, as shown in Fig. 1. By linking user accounts across social networks and integrating complementary information sources, more comprehensive user information can be obtained, which can better promote the development of cross-domain recommendation, personalized advertising and other fields. Therefore, user identity linkage across social networks based on spatiotemporal data has attracted more and more attention. However, with the deepening of the research, some inevitable problems have brought significant challenges to this work.

§Bo Sun is the corresponding author.
ISBN 978-3-903176-57-7

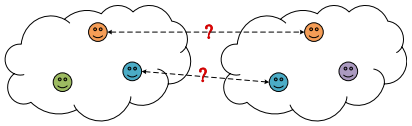

Fig. 1. User identity linkage is a task to find out user identities from different social networks belonging to the same natural person.

Spatiotemporal data in social networks have the following characteristics:

- **Sparsity.** The traditional GPS data set records the user's location automatically through GPS equipment, and the time interval between two continuous points is usually very short. However, in social networks, the generation of spatiotemporal data is user-driven. That is, users, decide whether to publish and when to publish information with spatiotemporal data. For privacy or personal will, users often publish less spatiotemporal data. The period and space span of adjacent data points may be huge, some even more than one year [1].
- **Heterogeneity.** Due to the business characteristics of different social networks, their spatiotemporal data often have different structural characteristics, such as time interval, period, and space span. This situation brings difficulties to the user identity linkage across social networks.
- **Imbalance.** Users have different preferences and use frequencies for different social networks, which leads to a significant gap in the amount of data from different social networks.
- **Incompleteness.** The integrity of spatiotemporal data from different social networks is different. Some of them only have longitude and latitude positions, and some of them also have semantic tags.
- **Noise.** Due to the error of the GPS itself, the subjective will of users, the difference in business characteristics of social networks and other reasons, some spatiotemporal data will be lost or offset, resulting in the generation of noise data.

The existing user identity linkage methods based on spatiotemporal data are effective to a certain extent, but there are still some problems to be solved.

- **Trajectory processing is not suitable for sparse data.** Some methods [2]–[4] connect spatiotemporal data points into lines in time order, which are called trajectories, and then link users' identities by calculating the similarity between trajectories. This method is only suitable for dense spatiotemporal data with low time intervals, such as automobile GPS data. However, this method is not applicable for sparse spatiotemporal data, such as social network spatiotemporal data, because the time and space interval between adjacent spatiotemporal data points is significant.
- **Grid processing leads to information loss and exceptions.** Existing methods often use gridding to process spatiotemporal data [1]. This method can effectively reduce the computational complexity, but there may be information loss and anomalies at the edge of the grid [5], which will affect the accuracy of user identity linkage.

Because of the above problems, we propose a new method **WTLink** (Wavelet Transform-based user identity Linkage) for user identity linkage through frequency domain analysis to solve the problem of cross-social network user identity linkage based on spatiotemporal data. According to the characteristics of human mobile activities, there are often only a few places each person often moves to, such as home, school and workplace. Therefore, many closer records often appear in the user's spatiotemporal data records. Inspired by this, firstly, we decompose and reconstruct the original spatiotemporal data records through the frequency domain analysis technology (precisely, Discrete Wavelet Transform, DWT [6]–[8]). Secondly, we extract several key points from the original spatiotemporal data records according to the reconstructed signals used as the user identity representation. Finally, we link the user identities according to the similarity between user identity representations. Unlike the traditional DWT reconstruction method, we proposed a new method to generate a shorter signal. We further proposed a boundary box filter method to improve identity linkage accuracy and efficiency when selecting user identity pair candidates. Our experiments on real datasets show that this method can achieve user identity linking well and exceeds the existing methods in terms of effectiveness and efficiency. Our contributions mainly include the following aspects.

- We proposed a new method, WTLink, for cross-social network user identity linkage through frequency domain analysis. The method extracts the user identity representation from the user's spatiotemporal data with a new application mode of discrete wavelet transform (DWT). Then it uses the proposed fuzzy Ochiai coefficient to measure the similarity of user identities and links the user identities. We are the first to link user identities via frequency domain analysis.
- The traditional application mode of discrete wavelet transform uses the approximate coefficients obtained after the DWT decomposition of the original signal to carry out DWT reconstruction to obtain a smoothed signal with the same length as the original signal. In contrast, we propose a new application mode for DWT. Firstly, it decomposes the original signal to obtain detail coefficients. Then, it reconstructs the signal with the detail coefficients. Lastly, a key point sequence shorter than the original signal is selected from the original signal by analyzing the reconstructed signal. The short key point sequence is considered to be the representation of the original signal, which is beneficial for improving the efficiency of user identity linkage.
- A boundary box filter method is proposed and embedded into the user identity linkage framework, improving user identity linkage accuracy and efficiency.
- Experiments on two real datasets show that WTLink exceeded the comparative baseline methods in effectiveness and efficiency. Compared with the SOTA, WTLink improves F1 scores on the FS-TW and IG-TW datasets by 2.6% and 12.9%, with a reduction in data size by 15.9% and 3.9%, respectively.

The rest of this paper is organized as follows. We put forward the related work in the second section and our method in the third section. The fourth section gives the experimental results, and the fifth summarizes the paper.

II. Related Work

With the development of social networks, more and more spatiotemporal data is being posted by users. Consequently, many researchers have utilized this spatiotemporal data on social networks to link users' identities and have achieved positive results.

Han et al. [9] used the spatiotemporal data generated by users in social networks to solve the problem of user identity linking in an unsupervised way. They proposed a new user identity linking framework, which uses the spatiotemporal collaborative clustering model under the spectral division paradigm to find account groups with very similar spatiotemporal trajectories. Riederer et al. [10] divided location and time into bins of corresponding geographic regions or time intervals. First, the similarity of all user pairs was calculated, then the maximum bipartite graph matching was performed, and finally, the matching user was obtained. Chen et al. [5] proposed the STUL model. Firstly, spatial features were extracted by the density-based clustering method, temporal features were extracted by the Gaussian mixture model, and different weights were assigned to features. Then, a new method to measure user similarity was proposed based on these features. Users whose similarity is higher than the threshold are considered to be linked. Wang et al. [11] first used a correlation graph to capture the co-occurrence locations of all user IDs across multiple services. Then, based on this graph, a set matching algorithm was proposed to discover candidate ID sets, and Bayesian reasoning was used to generate the confidence score of candidate ranking. Chen et al. [1] proposed a solution based on kernel density estimation to realize the user identity connection between location-based social networks, which alleviates the data sparsity problem in user similarity

measurement and improves measurement accuracy. In order to improve search efficiency, they developed a grid-based location data structure to simplify the search space.

In recent years, Chen et al. [12] propose a multi-platform UIL model ULMP based on location data. Firstly, it prunes the search space by searching the first k candidate user accounts. Then, it designs a matching score based on kernel density estimation to search the linked user accounts by combining the spatial and temporal information. Feng et al. [2] proposes a framework DPLink based on deep end-to-end learning. It first uses the deep learning tool to obtain the potential representation of the motion trajectory, then calculates the similarity between the two vectors as the similarity of the trajectory, and finally completes the user identity linkage task. This framework is the first time deep learning technology has been applied to the user identity linkage problem of heterogeneous mobile trajectories based on different qualities of service.

The existing methods often process the spatiotemporal data in the way of trajectories or gridding, and then link the user identity. Trajectories are only applicable to dense spatiotemporal data, but not to sparse spatiotemporal data, such as spatiotemporal data in social networks. Grid processing can improve the efficiency of user identity linkage, but information loss and abnormality may occur at the edge of the grid, which will affect the accuracy of user identity linkage.

III. Method

A. Definition

The problem of user identity linkage across social networks based on spatiotemporal data can be formally defined as follows:

Symbol indicators: A spatiotemporal data record r can be expressed as (u, l, t), where u represents the user identity, l represents the geographic location in the form of latitude and longitude, and t represents the timestamp. Let $G = \{V, R\}$ denote a social network platform, where $V = \{v_1, v_2, \ldots, v_{n_u}\}$ represents all n_u users in the social network platform; R is a n_u-dimensional vector, and vector elements are sets of spatiotemporal data records for each user.

Input: A source social network platform $G^s = \{V^s, R^s\}$, and a target social network platform $G^t = \{V^t, R^t\}$.

Output: A user pair set $P \subseteq V^s \times V^t$, which represents the user identities that have been linked (which means that they belong to the same natural person) from the source social network platform to the target social network platform.

B. Motivation

Fourier Transform (FT), Continuous Wavelet Transform (CWT) and Discrete Wavelet Transform (DWT) are three common frequency domain analysis techniques. DWT overcomes FT's shortcomings in decomposing non-stationary and abrupt signals and reduces the redundancy and calculation time of CWT. Therefore, we choose DWT to process spatiotemporal data records. DWT decomposes the original signal into a group of approximate coefficients (cA) and a group of detail coefficients (cD) through the wavelet function, which respectively represents the low-frequency approximate characteristics and high-frequency detail characteristics of the original signal.

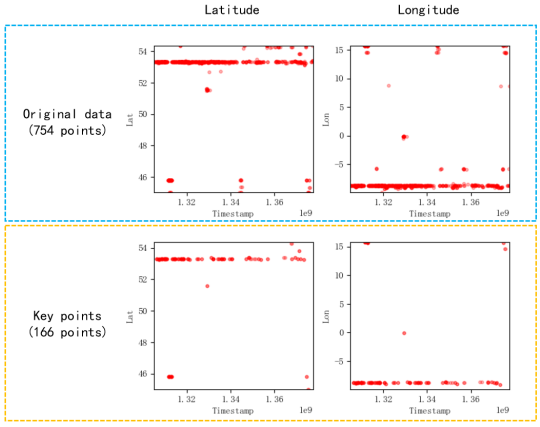

Fig. 2. An example of extracting key points via DWT

According to the characteristics of human mobile activities, there are often only a few places where each person often stays, such as home, school and workplace. Therefore, many records with similar positions often appear in the user's spatiotemporal data records, which correspond to DWT detail coefficients (cD). Inspired by this, different from the trajectory and grid modes, we propose a user identity linkage method WTLink based on DWT, which extracts some key points as user identity representation from original data points via DWT decomposition and reconstruction with cD. And then, the user identities are linked according to the similarity between their representations. An example is shown in Fig. 2, in which we extracted 166 key points from 754 original data points to be the user identity representation. We can find that the extracted key points are very representative. Meanwhile, the data size has been greatly reduced, which is beneficial for improving the efficiency of user identity linkage.

C. The Algorithm

a) Boundary box filter: The spatiotemporal data records of a user identity are within a certain latitude and longitude range. We call this range the boundary box, composed of $min\_latitude$, $max\_latitude$, $min\_longitude$, $max\_longitude$, as shown in Fig. 3. If the boundary boxes of two user identities do not intersect, the probability that they belong to the same natural person is minimal. In that case, we do not consider these two identities as anchor links (anchor links refer to multiple user identities from different social networks belonging to the same natural person) and do not calculate their similarity. This method can not only improve the accuracy of identity linkage but also improve efficiency.

b) Key point: We first decompose the original spatiotemporal data records by DWT with $db1$ wavelet function to obtain the DWT coefficients cA and cD. Then only use the detail coefficient cD for DWT reconstruction (different from tradi-

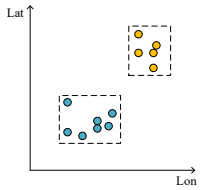

Fig. 3. Boundary box of spatiotemporal data records of two user identities

tional cA) with results denoted by sig. Finally, We consider the data points whose absolute value after reconstruction is not greater than the parameter $noise\_lower$ as key points because points near zero in data reconstructed by our method indicate representative points in the original signal (high-frequency noise removed). If there are no key points, the original data will be considered key points. We extract up to n_r key points from the spatiotemporal data records as the user identity representation, as shown in Algorithm 1. This user identity representation method can significantly improve the efficiency of user identity linkage while preserving the original data information as much as possible. For details, see the later ablation study.

Algorithm 1: Extract key points from spatiotemporal data records via DWT

Input: Spatiotemporal data records $[r_1, r_2, ..., r_{n_r}]$, $noise\_lower$

Output: Up to n_r key points $[p_1, p_2, ...]$

(cA, cD) = DWT($[r_1, r_2, ..., r_{n_r}]$, 'db1');
sig = IDWT($None$, cD, 'db1');
$\mathbb{K}$ = []; `// key points`
for i *in range(len(sig))* **do**
  if $abs(sig[i]) \leq noise\_lower$ *and* $i < n_r$ **then**
    $\mathbb{K}$.append(r_i);
  end
end
if $\mathbb{K}$ *is empty* **then**
  return $[r_1, r_2, ..., r_{n_r}]$
end
return $\mathbb{K}$;

The specific steps are as follows:

(1) Calculate the boundary box of each user identity.

$$boundary\_box = \{min\_lat, max\_lat, min\_lon, max\_lon\}$$

(2) According to Algorithm 1, a maximum of n_r key points are extracted from the spatiotemporal data records of each user identity as the user identity representation.

$$a_i = \{p_{i1}, p_{i2}, ...\}, b_j = \{p^{'}_{j1}, p^{'}_{j2}, ...\}$$

Each user identity is represented by several key points, where a_i is a user identity in social network A (OSN A), b_j is a user identity in social network B (OSN B), p is a key point.

(3) According to the boundary box filter algorithm, filter out several user identity pairs.

(4) Calculate the similarities of other user identity pairs (a_i, b_j)

$$user\_sim(a_i, b_j) = \frac{|a_i \cap b_j|}{\sqrt{|a_i| \times |b_j|}}$$

The similarity calculation adopts the form of the Ochiai coefficient. However, because the probability of two longitude and latitude positions completely coincident in the actual data is very low, the key points whose distance is not greater than a specific parameter (*coincidence_radius*) are regarded as coincidence when we are calculating $|a_i \cap b_j|$. When the following condition is met:

$$dist(p_1, p_2) \leq coincidence\_radius$$

p_1 and p_2 are regarded as a point in $a_i \cap b_j$, where $dist(p_1, p_2)$ represents the distance of two key points p_1 and p_2, $coincidence\_radius$ represents the distance threshold of considering two points as one point in $a_i \cap b_j$. We call this metric **fuzzy Ochiai coefficient**, whose pseudo-code is shown in Algorithm 2.

We have also tried other similarity measures, such as Mean Distance, Jaccard Coefficent, and DTW Distance. Their performance in the effectiveness of user identity linkage is less than the fuzzy Ochiai coefficient we proposed. For details, see the later ablation study. The reason is that, on the one hand, the Ochiai coefficient can more accurately measure the contribution of coincidence points to similarity. On the other hand, the coincidence radius parameter added in the fuzzy Ochiai coefficient is more suitable for sparse datasets.

(5) User identity linkage

Take b_j which is most similar to a_i, if satisfied:

$$user\_sim(a_i, b_j) \geq sim\_lower$$

then take b_j as user identity linkage result of a_i, where sim_lower represents the lower threshold of similarity.

The time complexity of the WTLink algorithm is $O(n^2)$, where n represents the number of user identities in the spatiotemporal dataset. Because we use the boundary box filter algorithm, which can filter out most of the user identity pairs, the actual operating efficiency of the WTLink algorithm is much higher than $O(n^2)$.

D. Parameters

In order to obtain the best performance, we set some parameters for the WTLink algorithm.

- ***noise_lower.*** *noise_lower* represents the threshold for selecting key points after DWT reconstruction. We take the reconstructed data points whose absolute value is not greater than *noise_lower* as the key points.
- ***coincidence_radius.*** *coincidence_radius* refers to the coincidence radius. When calculating the similarity of user identity pairs, if the distance between two points is not

Algorithm 2: Calculate the similarity of user identity pair by fuzzy Ochiai coefficient

Input: Two sets of key points $ts1$ and $ts2$, $coincidence\_radius$
Output: The fuzzy Ochiai coefficient of $ts1$ and $ts2$

Ensure that $ts1$ is greater than $ts2$;
$\mathbb{N}$ = 0; // intersected count
$\mathbb{L}$ = []; // intersected ts2 index list
for i *in range(len($ts1$))* **do**
  $min\_dist = -1$;
  $\mathbb{I} = -1$; // intersected ts2 index
  for j *in range(len($ts2$))* **do**
    if j *not in* $\mathbb{L}$ **then**
      $dist$ = get_distance($ts1[i]$, $ts2[j]$);
      if $min\_dist < 0$ *or* $dist < min\_dist$ **then**
        $min\_dist = dist$;
        $\mathbb{I} = j$;
      end
    end
  end
  if $min\_dist \leq coincidence\_radius$ **then**
    $\mathbb{N} += 1$;
    $\mathbb{L}$.append($\mathbb{I}$);
  end
end
$similarity = \mathbb{N}/\sqrt{len(ts1) * len(ts2)}$;
return $similarity$;

greater than this value, they are considered to be the same point and included in the $|a_i \cap b_j|$ part of the Ochiai coefficient.

- ***sim_lower*.** *sim_lower* represents the lower threshold of the similarity of user identity pairs. If the similarity is not less than the threshold, it is considered valid before entering the user identity linkage stage. Otherwise, the user identity pair will be discarded.

IV. Experiment

The experiments were conducted on a 4-core (3.4GHz) desktop computer with 8GB memory and 1TB hard disk. It runs Windows 10 operating system and python 3.8.

A. Datasets

Foursquare and Twitter are two widely used social networks in which users can post statuses related to location information. Instagram is another popular photo-sharing application and service where users can share photos and videos with location information through mobile phones, desktops, laptops and tablets.

We use two datasets, FS-TW and IG-TW, which are crawled from social networking sites [1], [5], [10], where FS, TW and IG represent Foursquare, Twitter and Instagram, respectively. The two datasets each contain two domains. The records in both datasets are in the form of (userid, latitude, longitude, and timestamp). We anonymized the two datasets, i.e. the userid was replaced with a meaningless number. Similar to Literature [1], [5], [10], we select users with records in both domains. After processing, the statistical information of the datasets is shown in Table I.

B. Compared Methods

We compare our method, WTLink, with several state-of-the-art user identity linkage methods based on spatiotemporal data.

DG. This method is proposed by Chen et al. [5], in which the density-based clustering method was used to extract the user's stay area, and the Gaussian Mixture Model (GMM) based method was used to model the user's time behaviour. Then, the similarity between user identities is measured based on these features.

EPOCH. This method is proposed by Seglem et al. [13], in which the input data is temporarily divided into equal size time intervals. Then kNN classification algorithm is used for user identity linkage.

GS. This method is proposed by Cao et al. [14], which use $(g_1, o_1), \ldots, (g_m, o_m)$ to indicate that two user identities appear at the same time are observed, where o_i is the corresponding frequency, and then measure the similarity of user identities with the weight and frequency of g_i.

GKR. This method is proposed by Chen et al. [1]. The spatial domain is divided into grids. Only the $k \times k$ neighbours of the target grid is considered in the calculation. Then the grid weight is calculated based on Renyi entropy. Finally, the user identity is linked according to the grid-based similarity.

DPLink. This method is proposed by Feng et al. [2]. First, the potential representation of the motion trajectory was obtained using the deep learning tool. Then the similarity between the two vectors was calculated as the similarity of the trajectory.

WTCoef. This is the naive method we proposed. First, the boundary box of each user identity is calculated, and the approximate coefficient cA after DWT decomposition is used as the user identity representation. Then, the similarity of each user identity pair is calculated, and finally, the user identity linking is performed according to the similarity.

WTLink. This is our proposed method. First, we calculate the boundary box of each user identity and use DWT decomposition and reconstruction to extract the key points of each user identity as their representations. Then calculate the similarity of each user identity pair, and finally, link the user identities according to the similarity.

We use $precision$, $recall$ and $F1$, which are widely used [1], [11], [15], as evaluation metrics of user identity linkage results to measure its quality.

C. Results

(1) Effectiveness Evaluation

We compare our method WTLink with baseline methods on FS-TW and IG-TW datasets, and the results are shown in

TABLE I
DATASETS STATISTICS INFORMATION

Domain	Users count	Records count	Records count after de-duplication	Records count per capita after de-duplication	Time span
FS in FS-TW	862	13177	13164	15	2008-10-20 $\sim$ 2012-11-22
TW in FS-TW	862	174618	171146	199	1994-08-30 $\sim$ 2012-11-24
IG in IG-TW	1717	337934	334288	195	2010-10-16 $\sim$ 2013-09-08
TW in IG-TW	1717	447366	437665	255	2010-09-13 $\sim$ 2015-04-02

Table II. In this experiment, we choose the optimal parameters for WTLink on the FS-TW dataset, *noise_lower* is 1, *coincidence_radius* is 0.06, *sim_lower* is 0.01; The optimal parameters are selected on IG-TW dataset, *noise_lower* is 1, *coincidence_radius* is 0.1, *sim_lower* is 0.02. We report the best performance of baseline methods DG, EPOCH, GS, GKR and DPLink in the datasets FS-TW and IG-TW.

It can be found that WTLink outperforms all baseline methods in *F1* on these two datasets and improves the F1 score by 2.6% on the FS-TW dataset and 12.9% on IG-TW dataset compared to the best result of baseline methods. Because the IG-TW dataset is larger and denser, all methods perform better in this dataset.

WTLink does not use trajectories to process spatiotemporal data compared to other methods. Instead, key points are extracted through DWT decomposition and reconstruction to represent user identity, which reduces the sparsity, heterogeneity, imbalance and noise of spatiotemporal data, and avoids information distortion caused by gridding, thus improving the effectiveness of user identity linkage.

(2) Efficiency Evaluation

Average running time refers to the average running time per user, which is another essential factor to consider when linking user identities. We show the average running time of DG, EPOCH, GS, GKR, DPLink, WTCoef, and WTLink on FS-TW and IG-TW in Table III.

DG needs much time to extract the dwelling area and calculate the weight. It is time-consuming for EPOCH to use kNN to classify trajectories. GS needs much time to find the public grids of two user identities and then calculate the user similarity based on these grids. GKR needs time to calculate the user's representation using KDE. DPLink takes a long time to train the deep learning model. On both datasets, the average running time of WTCoef was the shortest, followed by WTLink. The average running time of WTLink is longer than that of WTCoef, because WTCoef only uses DWT decomposition, while WTLink uses DWT decomposition and reconstruction. Compared with other methods, WTLink uses several key points to represent each user identity instead of the record set or grid representation, which significantly reduces the amount of calculation in the user identity linking stage and effectively improves the efficiency and scalability of the algorithm.

Only the experimental results of the serial version of WTLink are reported here. All stages of WTLink support parallelism.

(3) Impact of Parameters

a) Varying $noise\_lower$*:* We conducted experiments when $noise\_lower$ is taken as [0, 0.0001, 0.001, 0.01, 0.1, 1], and the results are shown in Table IV. In this experiment, we selected the optimal parameters for the FS-TW dataset, $coincidance\_radius$ is 0.06, and sim_lower is 0.01. The optimal parameters were selected for the IG-TW dataset, $coincidence\_radius$ is 0.1, and sim_lower is 0.02.

We can find that in these two datasets, with the increase of $noise\_lower$, $precision$ fluctuates, and $recall$ and $F1$ gradually increase. Because the larger the $noise\_lower$ is, the more key points extracted by WTLink, and the less information is lost, but the amount of calculation will also increase. Overall, when $noise\_lower$ is 1, $precision$ and $recall$ have a good balance, and $F1$ is maximum. We also conducted experiments on larger $noise\_lower$ (such as 10, 100), and the results showed that $F1$ was not significantly improved, but the amount of calculation was greatly increased. Therefore, these experimental results were not shown.

b) Varying $coincidence\_radius$ *and* sim_lower*:* We carried out the experiment when $coincidence\_radius$ is 0 $\sim$ 0.2 with an interval of 0.02, sim_lower is 0 $\sim$ 0.1 with an interval of 0.01, and the results are shown in Fig. 4, in which different colours represent different *F1* values. In order to shorten the running time, $noise\_lower$ in this experiment is taken as 0.

We can find that *F1* value changes according to the combination of $coincidence\_radius$ and sim_lower, when $coincidence\_radius$ is 0, it is minimum, and when sim_lower is 0, it is also tiny. This phenomenon indicates that the setting of these two parameters is necessary. According to subgraph (a), we choose the optimal parameters for the FS-TW dataset (when *F1* reaches the maximum value), i.e. $coincidence\_radius$ is 0.06, and sim_lower is 0.01. According to subgraph (b), we choose the optimal parameters for the IG-TW dataset (when *F1* reaches the maximum value), i.e. $coincidence\_radius$ is 0.1, and sim_lower is 0.02. It should be noted that we also carried out experiments with larger $coincidence\_radius$ (e.g. 0.5, 1) and sim_lower (e.g. 0.1 $\sim$ 1), but can not get greater *F1*, so these experimental results are not shown.

In practice, if the ground truth is unknown, we recommend

TABLE II
PERFORMANCE OF DIFFERENT METHODS ON FS-TW AND IG-TW

		DG	EPOCH	GS	GKR	DPLink	WTCoef	WTLink
FS-TW	Precision	0.152	0.173	0.123	0.411	0.366	0.112	**0.469**
	Recall	0.150	0.177	0.119	**0.401**	0.352	0.083	0.399
	F1	0.152	0.175	0.121	0.405	0.359	0.095	**0.431**
IG-TW	Precision	0.601	0.625	0.397	0.701	0.601	0.132	**0.873**
	Recall	0.607	0.623	0.399	0.699	0.572	0.103	**0.787**
	F1	0.603	0.625	0.397	0.699	0.586	0.116	**0.828**

TABLE III
AVERAGE RUNNING TIME (SECONDS) OF DIFFERENT METHODS

	DG	EPOCH	GS	GKR	DPLink	WTCoef	WTLink
FS-TW	3.084	2.697	2.245	0.751	2.832	**0.619**	0.736
IG-TW	4.336	3.827	3.41	0.835	3.018	**0.752**	0.952

TABLE IV
PERFORMANCE OF WTLINK W.R.T. VARIED $noise\_lower$

Dataset	Noise_lower	Precision	Recall	F1
FS-TW	0	0.385	0.147	0.213
	0.0001	0.376	0.157	0.221
	0.001	0.364	0.19	0.25
	0.01	0.4	0.284	0.332
	0.1	0.436	0.362	0.396
	1	0.469	0.399	**0.431**
IG-TW	0	0.838	0.556	0.668
	0.0001	0.85	0.606	0.708
	0.001	0.831	0.646	0.727
	0.01	0.821	0.689	0.749
	0.1	0.825	0.738	0.779
	1	0.873	0.787	**0.828**

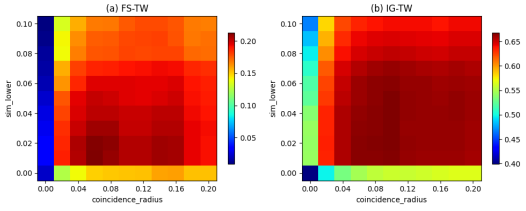

Fig. 4. Performance of WTLink w.r.t. varied $coincidence\_radius$ and sim_lower

to set $noise\_lower$ as 1, $coincidence\_radius$ as 0.06, and sim_lower as 0.02. Another way is to manually label part of the data as the ground truth.

(4) Ablation Study

In order to verify the effectiveness of various components proposed in WTLink, we compared WTLink and its six variants on the FS-TW and IG-TW datasets, as shown in Table V (WTLink is in bold). It can be found that when boundary box filtering or DWT denoising is not used, the F1 value slightly increases, but the running time dramatically increases. In particular, the running time on the more extensive dataset IG-TW can be as high as 3∼10 times that of WTLink, which significantly reduces the scalability of user identity linkage. The other variants replace the user identity metric with DTW, mean distance and Jaccard coefficient, which all cause the F1 value to drop to different degrees. The results show that all components of WTLink are indispensable.

V. CONCLUSION

The existing user identity linkage methods across social networks based on spatiotemporal data have some problems, such as trajectory processing is not suitable for sparse data, and grid processing leads to information loss and abnormality. Because of the above problems, we propose a method WTLink based on frequency domain analysis. According to the sparsity, heterogeneity and imbalance of spatiotemporal data in social networks, this method represents user identity through several key points and can effectively link user identity. We compare this method with several of the most advanced user identity linkage methods based on spatiotemporal data on real datasets. The results show that this method exceeds the baseline methods in terms of effectiveness and efficiency. When dealing with large datasets, to improve the efficiency of user identity linkage, further studies can be done on screening user identity pairs in the future, such as using possible conflicts between spatiotemporal data from different individuals for screening.

REFERENCES

[1] W. Chen, H. Yin, W. Wang, L. Zhao, and X. Zhou, "Effective and Efficient User Account Linkage across Location Based Social Networks," in *2018 IEEE 34th International Conference on Data Engineering (ICDE)*. IEEE, Apr 2018, pp. 1085–1096. [Online]. Available: https://ieeexplore.ieee.org/document/8509322/

[2] J. Feng, Y. Li, Z. Yang, M. Zhang, H. Wang, H. Cao, and D. Jin, "User Identity Linkage via Co-Attentive Neural Network From Heterogeneous Mobility Data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 2, pp. 954–968, Feb 2022. [Online]. Available: https://doi.org/10.1109/TKDE.2020.2989732 https://ieeexplore.ieee.org/document/9076832/

[3] Y. Yu, H. Tang, F. Wang, L. Wu, T. Qian, T. Sun, and Y. Xu, "TULSN: Siamese Network for Trajectory-user Linking," in *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, Jul 2020, pp. 1–8. [Online]. Available: https://ieeexplore.ieee.org/document/9206609/

[4] C. Miao, J. Wang, H. Yu, W. Zhang, and Y. Qi, "Trajectory-User Linking with Attentive Recurrent Network," in *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, ser. AAMAS '20. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems, 2020, pp. 878–886.

TABLE V
ABLATION RESULTS ON FS-TW DATASET AND IG-TW DATASET

Dataset	Boundary Box Filter	Preprocessing / Representation	Metric	F1	Average Running Time (s)
FS-TW	✓	**DWT Denoising**	**Fuzzy Ochiai Coefficient**	**0.431**	**0.736**
		DWT Denoising	Fuzzy Ochiai Coefficient	0.432	1.092
	✓	Original Data	Fuzzy Ochiai Coefficient	0.449	1.115
	✓	Original Data	DTW	0.099	3.375
	✓	DWT Denoising	DTW	0.103	1.419
	✓	DWT Denoising	Mean Distance	0.147	0.011
	✓	DWT Denoising	Jaccard Coefficient	0.334	0.751
IG-TW	✓	**DWT Denoising**	**Fuzzy Ochiai Coefficient**	**0.828**	**0.952**
		DWT Denoising	Fuzzy Ochiai Coefficient	0.831	3.443
	✓	Original Data	Fuzzy Ochiai Coefficient	0.842	9.549
	✓	Original Data	DTW	0.357	6.342
	✓	DWT Denoising	DTW	0.369	3.261
	✓	DWT Denoising	Mean Distance	0.279	0.007
	✓	DWT Denoising	Jaccard Coefficient	0.764	0.905

[5] W. Chen, H. Yin, W. Wang, L. Zhao, W. Hua, and X. Zhou, "Exploiting Spatio-Temporal User Behaviors for User Linkage," in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management - CIKM '17*, vol. 1. New York, New York, USA: ACM Press, 2017, pp. 517–526. [Online]. Available: http://dl.acm.org/citation.cfm?doid=3132847.3132898

[6] E. Brusa, C. Delprete, S. Gargiuli, and L. Giorio, "Screening of Discrete Wavelet Transform Parameters for the Denoising of Rolling Bearing Signals in Presence of Localised Defects," *Sensors*, vol. 23, no. 1, p. 8, Dec 2022. [Online]. Available: https://doi.org/10.3390/s23010008 https://www.mdpi.com/1424-8220/23/1/8

[7] J. Dogani, F. Khunjush, and M. Seydali, "Host load prediction in cloud computing with Discrete Wavelet Transformation (DWT) and Bidirectional Gated Recurrent Unit (BiGRU) network," *Computer Communications*, vol. 198, pp. 157–174, Jan 2023. [Online]. Available: https://doi.org/10.1016/j.comcom.2022.11.018 https://linkinghub.elsevier.com/retrieve/pii/S0140366422004479

[8] A. Gon and A. Mukherjee, "FPGA-Based Low-Cost Architecture for R-Peak Detection and Heart-Rate Calculation Using Lifting-Based Discrete Wavelet Transform," *Circuits, Systems, and Signal Processing*, vol. 42, no. 1, pp. 580–600, Jan 2023. [Online]. Available: https://doi.org/10.1007/s00034-022-02148-7 https://link.springer.com/10.1007/s00034-022-02148-7

[9] X. Han, L. Wang, L. Xu, and S. Zhang, "Social Media Account Linkage Using User-generated Geo-location Data," in *2016 IEEE Conference on Intelligence and Security Informatics (ISI)*. IEEE, Sep 2016, pp. 157–162. [Online]. Available: http://ieeexplore.ieee.org/document/7745460/

[10] C. Riederer, Y. Kim, A. Chaintreau, N. Korula, and S. Lattanzi, "Linking Users Across Domains with Location Data:Theory and Validation," in *Proceedings of the 25th International Conference on World Wide Web - WWW '16*. New York, New York, USA: ACM Press, 2016, pp. 707–719. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2872427.2883002

[11] H. Wang, Y. Li, G. Wang, and D. Jin, "You Are How You Move: Linking Multiple User Identities From Massive Mobility Traces," in *Proceedings of the 2018 SIAM International Conference on Data Mining*. Philadelphia, PA: Society for Industrial and Applied Mathematics, May 2018, pp. 189–197. [Online]. Available: https://epubs.siam.org/doi/10.1137/1.9781611975321.22

[12] W. Chen, W. Wang, H. Yin, J. Fang, and L. Zhao, "User Account Linkage Across Multiple Platforms with Location Data," *Journal of Computer Science and Technology*, vol. 35, no. 4, pp. 751–768, Jul 2020. [Online]. Available: https://doi.org/10.1007/s11390-020-0250-7 http://link.springer.com/10.1007/s11390-020-0250-7

[13] E. Seglem, A. Züfle, J. Stutzki, F. Borutta, E. Faerman, and M. Schubert, "On Privacy in Spatio-Temporal Data: User Identification Using Microblog Data," *Advances in Spatial and Temporal Databases - 15th International Symposium, SSTD 2017, Arlington, VA, USA, August 21-23, 2017, Proceedings*, pp. 43–61, 2017. [Online]. Available: http://link.springer.com/10.1007/978-3-319-64367-0_3

[14] W. Cao, Z. Wu, D. Wang, J. Li, and H. Wu, "Automatic User Identification Method across Heterogeneous Mobility Data Sources," in *2016 IEEE 32nd International Conference on Data Engineering (ICDE)*. IEEE, May 2016, pp. 978–989. [Online]. Available: http://ieeexplore.ieee.org/document/7498306/

[15] Z. Zhong, Y. Cao, M. Guo, and Z. Nie, "CoLink: An Unsupervised Framework for User Identity Linkage," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, AAAI-18*, S. A. McIlraith and K. Q. Weinberger, Eds. AAAI Press, 2018, pp. 5714–5721. [Online]. Available: https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17287