Linking Multiple User Identities of Multiple Services From Massive Mobility Traces

HUANDONG WANG, Tsinghua University, China
YONG LI, Tsinghua University, China
GANG WANG, Virginia Tech, USA
DEPENG JIN, Tsinghua University, China

Understanding the linkability of online user identifiers (IDs) is critical to both service providers (for business intelligence) and individual users (for assessing privacy risks). Existing methods are designed to match IDs across *two services*, but face key challenges of matching multiple services in practice, particularly when users have multiple IDs per service. In this paper, we propose a novel system to link IDs across multiple services by exploring the spatial-temporal features of user activities, of which the core idea is that the same user's online IDs are more likely to repeatedly appear at the same location. Specifically, we first utilize a *contact graph* to capture the "co-location" of all IDs across multiple services. Based on this graph, we propose a set-wise matching algorithm to discover candidate ID sets, and use Bayesian inference to generate confidence scores for candidate ranking, which is proved to be optimal. We evaluate our system using two real-world ground-truth datasets from an Internet service provider (4 services, 815K IDs) and Twitter-Foursquare (2 services, 770 IDs). Extensive results show that our system significantly outperforms the state-of-the-art algorithms in accuracy (AUC is higher by 0.1-0.2), and it is highly robust against data quality, matching order and number of services.

Additional Key Words and Phrases: Identity linkage, spatio-temporal trajectory, online services, set-wise ID matching

1 INTRODUCTION

Online services are playing critical roles in almost all aspects of users' life. It is common for a user to have multiple online identifiers (IDs) in different services such as Instant messaging (IM), online social networks (OSN), e-commerce (EC), online review (OR), etc. Users may even have *multiple* IDs in *a single service*, where different IDs are used for different purposes [38].

Service providers have strong motivations to massively mining user data for monetization and optimizing user experience [22]. To capture a more comprehensive understanding of user behavior, it is increasingly intriguing to link user IDs across multiple services to fuse the separated data [25, 53, 56]. However, from the user perspective, linking IDs across services may have privacy implications since more information are exposed [17, 35].

To these ends, understanding the "linkability" of online IDs is critical to both service providers (for business intelligence) and users (for understanding privacy risks). Existing studies have explored various ways to link IDs by using service-specific information such as user attributes [14] and friendship graphs [21]. However, these approaches depend on whether these services have the same data type. For example, e-commerce services often do not have friendship graphs to match with an online social network. Moreover, users may fill in fake information (*e.g.*, name, gender) in their profiles, which makes the linkage even harder.

In this paper, we explore a more generic approach to link user IDs by leveraging the spatial-temporal features of user activities. The key intuition is that no matter what online services a user accesses, we can bind them to the user's *physical presence*, which is characterized by time and location. This becomes possible because most online services today have a mobile version with locations as parts of the service (*e.g.*, Uber, Yelp, Twiter). In addition, with some tolerance

Authors' addresses: Huandong Wang, Tsinghua University, China, whd14@tsinghua.org.cn; Yong Li, Tsinghua University, China, liyong07@tsinghua.edu.cn; Gang Wang, Virginia Tech, Blacksburg, VA, USA, gangwang@vt.edu; Depeng Jin, Tsinghua University, China, jindp@tsinghua.edu.cn.

on granularity, even network accessing related information can be translated into location [20]. Our goal is to link multiple online IDs that belong to the same users across different services. This requires solving three key challenges that have not be addressed in existing work [21, 41, 42]:

- *Service multiplicity*: existing methods mainly focus on linking IDs of two services [28, 41]. In practice, however, the number of services can easily go over two. We find that adapting existing methods by matching services one by one produces unreliable results (see §7), which is significantly influenced by the number of services and the matching orders.
- *ID multiplicity*: users may register more than one ID in a service [38]. ID matching between two services should be "set-wise", *i.e.*, linking a set of IDs. However, existing methods are designed for pair-wise ID matching, which fail to capture users' multiple IDs in a single service [41, 42].
- *Heterogeneous data quality*: since user mobility behavior is extremely heterogeneous [28], the quantity and resolution of location data are drastically different across services. Early works simply filter out a large number of IDs with low-quality data [21, 41], which significantly reduces the data coverage and usability.

To solve these challenges, we propose a *contact graph* model for multi-service ID linking. Instead of matching IDs of just two services in a bipartite graph, we directly map multiple services and all their IDs into one big graph. In this graph, each node is an ID (regardless of service), and an edge represents that the connected IDs visited the same physical location, which is weighted by the number of co-locations of them. The high-level intuition is that the users' daily movements are fairly predictable with repeated patterns [49]. If multiple IDs belong to the same user, they are more likely to be "co-located" to build edges and form distinguishable subgraph structures over time.

Based on the contact graph, we propose a Bayesian-based optimal ID matching algorithm, which identifies the most probable ID sets that belong to the same physical user with the *target ID*. The high-level intuition is to extract possible candidate sets from neighboring IDs, and *rank* candidate sets based on their joint probability to match the target ID. Then, we propose a Bayesian inference method to obtain the joint probability for ranking candidate sets, which is proved to be optimal.

Our system allows a service provider to match its own IDs (as target IDs) with multiple other services simultaneously. The matching is based on "sets", capturing users who have multiple IDs in the same service. In order to solve heterogeneous data quality, the contact graph is constructed by including all IDs without arbitrarily pre-filtering any data. In addition, multiple user behavior models are utilized to produce a confidence probability for the candidate sets, which have different requirement for data quality of IDs to be linked. This allows applications to make use of the available data based on specific contexts.

We evaluate our system based on two real-world *ground-truth* datasets. One is from a large Internet service provider (ISP) that contains 412,455 users (815,117 online IDs) and 31 million access records to 4 online services: instant message (QQ), social network (Weibo), e-commerce (Taobao) and online review (Dianping). The second dataset contains 24K check-ins from Twitter and Foursquare from 385 users (770 online IDs) [58]. We use the state-of-the-art pair-wise ID matching algorithms POIS [41] and WYCI [42] as baselines. The results show that our algorithm significantly outperforms baselines (by 0.1 in AUC), particularly on users with multiple IDs per service (by 0.2 in AUC). We have three novel contributions summarized as follows:

- First, we propose a generic and optimal ID linking algorithm utilizing the spatial-temporal locality of user activities. Our *contact graph* model achieves set-wise ID matching for multiple services. The model effectively captures users with multiple IDs per service, and mitigates the ordering effect of multi-service matching.

- Second, we propose a novel Bayesian-based method to produce confidence probability for ID matching with proof of optimality. This addresses the challenge of uneven data-quality across services: instead of arbitrarily pre-filtering low-quality data, our method keeps all IDs in the matching for maximum data utilization.
- Third, we evaluate our system based on two real-world datasets with ground-truth. The results show that our system significantly outperforms the start-of-the-art in accuracy, and it is robust against data quality, number of services, and matching order.

As explained above, users' spatial-temporal locality, which can be inferred from the location data as well as users' network accessing information, is a more generic information than service-specific information, *i.e.*, friendship graphs [21], or user attributes [14]. On the other hand, we addressed a series of practical challenges in ID linking based on users' spatial-temporal locality, involving *service multiplicity*, *ID multiplicity*, and *heterogeneous data quality*. We believe it paves the way toward solving ID linkage problems in practice.

A conference version of this paper was published in [52]. Compared with the conference version, we further introduce two different user behavior models, and two different parameter estimation strategies in this new version. In addition, a modified matching score function is proposed. Extensive experiments show that different user behavior models and parameter estimation strategies perform well in different situations. In addition, the proposed matching score function helps to improve the performance in terms of F1 score by over 9.3%, indicating its effectiveness.

2 RELATED WORK

Works related to our paper can be summarized by the following four topics: applications of ID linking, ID linking methods, linking IDs using location data, and privacy protection mechanisms.

Applications of ID Linking: Based on linking IDs across services, a number of applications can be improved, including friend recommendations [53, 56], video recommendations [54], opinion analysis [9], social network design [22], indicating the large benefit we can obtain from the identity linkage.

ID Linking Methods: A large number of methods have been proposed to link IDs by utilizing different information. For example, approaches [18, 19, 21, 45, 59, 61] focus on utilizing social graphs. Approaches [11, 14, 24, 32, 55] focus on utilizing user attributes and user profile. Other approaches focus on utilizing user generated contents, e.g., similarity of movie ratings [35], writing style of posts [13]. Zhong et al. [60] proposed an unsupervised co-training algorithm to link IDs belonging to the same users, which manipulates two independent models, the attribute-based model and the relationship-based model. Nie et al. [36] jointly considered users' social network structure and users' article content to link their IDs across social platforms, which employed the latent Dirichlet allocation (LDA) model to capture the users' intersection core circle of friends. Li et al. [23] proposed a weakly-supervised identity linkage algorithm based on adversarial learning, which utilized users' features on different social platforms. All these approaches rely on service-specific features (*e.g.*, user names), which are depended on whether two services have overlapped features. In this work, we explore the mobility features of users for ID linking, which utilizes more generic information from services.

Linking IDs using Location Data: Riederer et al. [41] linked two trajectory datasets with maximum weight matching. Rossi et al. [42] proposed a trajectory-based linking method based on their defined spatio-temporal distance. Mulder et al. [8] linked trajectory datasets by measuring similarity between their transition kernels. Wang et al. [50, 51] proposed ID linking algorithms based on Gaussian distribution and Gaussian mixture model which considers spatio-temporal mismatches between different datasets. Recently, a number of deep learning based identity linkage algorithms are proposed. Jie et al. [10] proposed a deep learning based algorithm to link user IDs,

which used a co-attention mechanism to overcome the mismatches between different datasets. Luo et al. [30] proposed a CNN-based identity linkage algorithm to incorporate Point of Interest (POI) embeddings. However, the large number of parameters in the deep learning based models require more training datasets, which further increases the challenge of overcoming the heterogeneous data quality. What's more, these approaches can only perform pair-wise ID matching between two services, and face key challenges in the multiplicity of IDs and services.

Privacy Protection Mechanisms: On the other hand, researchers have focused on preserving users' privacy against the identity linkage. A number of studies focus on developing privacy criterias, e.g., k-anonymity [46], l-diversity [31], and t-closeness [26]. Other approaches implement the techniques of generalization, perturbation, and suppression on datasets to reduce the risk of users to be linked as well as other privacy leakage [1–4, 15, 37]

3 CHALLENGES OF ID LINKING

In this section, we discuss the high-level challenges in ID linking in practice, and introduce our core idea for solutions.

Problem I: Service multiplicity. Today's users often have multiple IDs in different online services. A recent survey by Pew Research Center in 2016 [38] shows that 56% of online adults use more than one of the five social media platforms measured. Consequently, when implementing ID linking, the number of services can easily go beyond two. However, existing methods are heavily optimized for linking IDs of two services. When working on more than 3 services, the linking accuracy is significantly influenced by the number services and matching orders (validated by experiments later).

Problem II: Identity multiplicity. Many web services do not or even cannot forbid an individual from having multiple IDs, and creating multiple accounts is quite normal in some services. According a poll with over 1200 votes in 2008, a majority of Twitter users (53%) have more than one Twitter accounts [47]. Existing ID linking methods mainly focus on pair-wise linkage, which cannot capture IDs that should be linked within the same service.

Solution Overview. We propose a contact graph to include all IDs from multiple services. In this graph, each node is an ID (regardless of service), and the edge represents two IDs once visited the same location. In this way, a group of conjoint nodes in the graph becomes possible candidates of same users' IDs. Furthermore, we develop a set-wise ID matching algorithm on this graph and match IDs from different services simultaneously, to solve problem I and II.

Problem III: Data quality heterogeneity. ID linking faces a key challenge of heterogeneous data quality. According to [28], online user profiles are largely incomplete: 80% of users missed 2+ out of 6 popular attributes. Similar for location data, the number of data records belonging to different IDs is highly skewed [41]. For users (IDs) that have incomplete data, existing methods often choose to filter them out, which significantly reduces data coverage and usability.

Solution Overview. Our goal is to keep all the IDs in the matching process. Instead of filtering out them, we build a Bayesian inference method to calculate the "confidence" probability for the matched IDs. We can identify high-quality matching results, and also keep ones to be utilizable.

Application Scenario. Based on these ideas, we build a system to solve a practical ID linking problem across multiple services. Supposing one service provider obtained multiple datasets from other web services, our goal is to link its own IDs with IDs of other services. For a given target ID in the current service, our system will identify a list of candidate sets of IDs that are likely to belong to the same physical user.

The workflow of our proposed ID linking system is shown in Figure 1. First, based on user trajectory data from multiple services, we construct a big contact graph between IDs to extract the subgraph structures formed by IDs belonging to the same user. At the same time, we propose

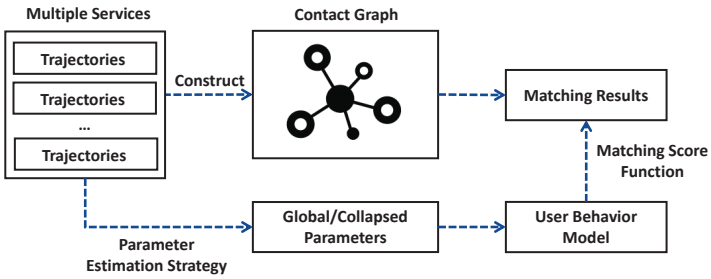


Fig. 1. The workflow of our proposed system.

two probabilistic user behavior models as well as two parameter estimation strategies to extract key mobility features of users. Then, based on the probabilistic user behavior model and subgraph structure in the contact graph, the likelihood of different combination of IDs belonging to same users are obtained. Finally, this system filters out unreliable matching results based on our proposed matching score function, and obtain the final results as the output.

4 PROBLEM FORMULATION

In this section, based on the introduced mathematical model, we formally define the set-wise ID matching problem. Then, we introduce two important concepts, *i.e.*, *contact graph* and *partition*, and present a probability model of users' behavior to solve the set-wise ID matching problem. For readability, we summarize the major notations used throughout the paper in Table 1.

4.1 Mathematical Model

We define a user identifier (ID) is a sequence of characters that identifies a unique user account of an online service. In addition, the service type has also been encoded in the user ID. Define $\boldsymbol{A}$ as the set of all IDs, and define $\boldsymbol{S}$ as the set of types (services) of IDs. For an arbitrary ID $v \in \boldsymbol{A}$, we denote $s(v)$ as its type (service). Then, $\forall s \in \boldsymbol{S}$, $\boldsymbol{A}^s$ is used to denote the set of all IDs of type s.

Since the timestamps have different resolutions in different services, we divide the time span of location traces into fixed sized time bins. On the other hand, recorded locations are also mapped into geographical regions, *e.g.*, the fixed-length grid or administrative regions. In addition, if the location data of different services are recorded in the same format with the same spatial resolution, such as access point (AP) or point of interest (POI), they can be used without change in ID linking. We define $\boldsymbol{T}$ and $\boldsymbol{L}$ as the set of all time bins and the set of all regions, respectively.

For a given online ID $u \in \boldsymbol{A}$, its mobility records are defined as $R(u) = \{(l_1^u, t_1^u), (l_2^u, t_2^u), \ldots, (l_{N_u}^u, t_{N_u}^u)\}$, where (l_i^u, t_i^u) represents a mobility record in location l_i^u at time t_i^u, and N_u is the number of records for ID u. Without loss of generality, we assume $t_1^u \leq t_2^u \leq \ldots \leq t_{N_u}^u$. For mobility records within the same time bin, they will be sorted based on their timestamp without discretized into time bins. For an ID set ξ, its mobility records are defined as $\mathcal{R}(\xi) = \{R(w) | w \in \xi\}$. Further, if all IDs in ξ belong to the same user, mobility records of different IDs will be merged into one trajectory $R(\xi) = \{(l_1^\xi, t_1^\xi), (l_2^\xi, t_2^\xi), \ldots, (l_{N_\xi}^\xi, t_{N_\xi}^\xi)\}$ by removing repeated records with the same space-time coordinates. Then, for each pair of online IDs $u, v \in \boldsymbol{A}$, let binary variable $X(u, v)$ indicate whether

Table 1. Notations and Descriptions.

Notation	***Description***
$\boldsymbol{A}$	The set of IDs.
$\boldsymbol{S}$	The set of types (services) for IDs.
$\boldsymbol{L}$	The set of locations.
$\boldsymbol{T}$	The set of time bins.
$\boldsymbol{A}^s$	The set of all IDs of type $s \in \boldsymbol{S}$.
$\boldsymbol{G}$	Contact graph of online IDs in physical world.
$\mathcal{N}(v)$	The neighbor of $v \in \boldsymbol{A}$.
$\mathcal{R}(V)$	The login records for a set of online IDs V.
$R(\xi)$	The login records for a set of online IDs $\xi \subseteq \boldsymbol{A}$ belonging to one user.
$X(\xi, v)$	Binary variable indicating whether IDs in ξ and v belong to the same users.
$s(v)$	The service type of online ID $v \in \boldsymbol{A}$.
$\mathcal{P}(\boldsymbol{A})$	The set of all partitions of $\boldsymbol{A}$.
$\mathcal{P}(\boldsymbol{A}, \xi)$	The set of all partitions in which all IDs in ξ are divided into one set.

they belong to the same user. That is,

$$X(u, v) = \begin{cases} 1, & \text{if } u, v \text{ belong to the same user,} \\ 0, & \text{otherwise.} \end{cases}$$

More generally, for a set of online IDs $\xi \subseteq \boldsymbol{A}$, we use $X(\xi)$ to indicate whether they belong to the same users. Thus, we have $X(\xi) = \prod_{u,v \in \xi} X(u, v)$. Similarly, for the target online IDs v and a set of IDs ξ, we let $X(\xi, v)$ indicate whether they belong to the same users, which can be computed by $X(\xi, v) = \prod_{u \in \xi} X(u, v)$. Then, based on these notations, we formally define it as the follows:

Set-wise Identity Matching Problem (SIMP)

Given: The target ID v, a list of candidate sets of IDs $\xi_1, ..., \xi_N \subseteq \boldsymbol{A}$, and their mobility records $R(v)$ and $R(\xi_i)$ for $i = 1, ..., N$.

Problem: Find a ranking function $\phi : \{\xi_1, ..., \xi_N\} \rightarrow \{1, ..., N\}$, such that the IDs belonging to the same user with v are ranked as high as possible, which can be expressed as:

$$\min_{\phi} \sum_{i=1}^{N} X(\xi_i, v)\phi(\xi_i). \tag{1}$$

4.2 Contact Graph and Partition

It has been found that users' daily mobility is fairly predictable with repeated patterns [49]. If multiple IDs belong to the same user, they are more likely to be "co-located". Thus, we can define the *contact graph* of IDs to extract the subgraph structures formed by IDs belonging to the same user.

Definition 1 (Contact Graph) Contact graph of IDs is defined as a graph $\boldsymbol{G} = (\boldsymbol{A}, \boldsymbol{E})$. For a pair of online IDs $u,\ v \in \boldsymbol{A}$, we say there exists an edge between u and v in $\boldsymbol{E}$, if u and v have mobility records at the same locations, *i.e.*, $\exists l \in \boldsymbol{L}$, such that $(l, t_1) \in R(u)$ and $(l, t_2) \in R(v)$ hold for some $t_1,\ t_2 \in \boldsymbol{T}$.

Note that the *contact graph* is defined only based on spatial information of mobility records of IDs, regardless of the temporal information. The intuition is to capture the candidate IDs belonging to the same user through the connectivity of the *contact graph* as much as possible. Actually, if two nodes have no common location in their trajectories, there is very low probability for them to

belong to the same user. Thus, for each ID v, we can limit the candidate online IDs belonging to the same user with it as a subset of its neighbor $\mathcal{N}(v) = \{b|b \in \boldsymbol{A}, (b, v) \in \boldsymbol{E}\}$. We further define the neighbor of v with a certain type $s \in \boldsymbol{S}$ as $\mathcal{N}^s(v) = \mathcal{N}(v) \cap A^s$. As for the temporal information, we will model it in the weight of edge. Specifically, for an edge $(u, v) \in \boldsymbol{E}$, we define its weight $w(u, v)$ as the probability that they belong to the same user, *i.e.*, $w(u, v) = P(X(u, v) = 1)$. For example, nodes with more frequent "co-locations" will have larger weight between them, which be introduced in detail in Section 6.

On the other hand, in order to describe the subgraph structures of IDs belonging to different users simultaneously, we define "partition" of IDs as follows:

Definition 2 (Partition) Given a node set $\boldsymbol{V}$, $p = \{\xi_1, \xi_2, ..., \xi_n\}$ is a set of non-empty subset of $\boldsymbol{V}$. That is, $\forall k \in \{1, ..., n\}$, we have $\xi_k \subseteq \boldsymbol{V}$. Then, p is a partition of $\boldsymbol{V}$ if the following three conditions hold: **(1)** $\emptyset \notin p$, **(2)** $\cup_{\xi \in p} \xi = \boldsymbol{A}$, **(3)** if $\xi_1, \xi_2 \in p$ and $\xi_1 \neq \xi_2$, we have $\xi_1 \cap \xi_2 = \emptyset$.

There is an inherent partition of $\boldsymbol{A}$ composed of the true set of online IDs belonging to each user. Specifically, we assume, the number of each user's IDs of type $s \in \boldsymbol{S}$ follows independent geometric distribution with parameter θ_s. That is, the prior probability of a set of IDs ξ belonging to the same user is $P(\xi) = \prod_{s \in \boldsymbol{S}} \theta_s^{n_s(\xi)}(1 - \theta_s)$, where $n_s(\xi)$ is the number of IDs of service s in ξ. Then, for a partition p, its prior $P(p)$ depends on the online IDs of each user, which can be expressed as follows:

$$P(p) = \prod_{\xi \in p} P(\xi). \tag{2}$$

In addition, we use $\mathcal{P}(\boldsymbol{V})$ to denote the set of all partitions of $\boldsymbol{V}$. Then, for a subset $\xi \subseteq \boldsymbol{V}$, we define $\mathcal{P}(\boldsymbol{V}, \xi)$ as the set of all partitions in which all IDs in ξ are divided into one set, *i.e.*, $\mathcal{P}(\boldsymbol{V}, \xi) = \{p|p \in \mathcal{P}(\boldsymbol{V}),\ \exists \zeta \in p, s.t.\ \xi \subseteq \zeta\}$.

Based on the concept of partitions, we can further express the probability of the observed records $\mathcal{R}(V)$ of an ID set V conditioned on one of its partition p as follows:

$$P(\mathcal{R}(V)|p) = \prod_{\xi \in p} P(R(\xi)|X(\xi) = 1), \tag{3}$$

which is an important indicator of the goodness of partition p compared with the true partition. In addition, we can demonstrate that ranking results of ID sets based on this probability is optimal (see §6.2). Next, we will further introduce how to calculate this probability.

5 PROBABILITY MODEL OF USER BEHAVIOR

In order to formally analyze our problem, it is necessary to build a mobility model which describes how users move and produce location records. Human mobility modelling has been studied in a number of works [5, 7]. Brockmann et al. [5] model human mobility as a Lévy flight, where the length of spatial displacement of individuals follows power-law distribution. Cho et al. [7] model human movements as periodic movement between their home and work place by Gaussian mixture model. However, these approaches are mainly designed for modelling continuous space-time coordinates of human trajectories, which do not consider the heterogeneous mobility data across services. In our work, we model human movement on the discrete time bins and geographical regions by two widely used probability models including multinomial model [34, 42] and Markov model [8, 43]. What's more, in order to estimate parameters used in the probability models, we present two strategies, including using global parameters and collapsed parameters. In the conference version of this paper [52], we use the Markov model and global parameters. In this journal version, we will compare advantages and disadvantages of the two probability model and two parameter estimation strategies, which will be introduced in detail in this section.

5.1 Probability Model

In this section, we will introduce two representative probability models given their parameters, including multinomial model [34, 42] and Markov model [8, 43].

Multinomial Model: In this model, users' movement is modelled to follow multinomial distribution, which is a generalized Bernoulli distribution on $|\boldsymbol{L}|$ locations. Specifically, for an ID $u \in \boldsymbol{A}$, for each location l_i^u and time bin t_i^u, the probability of visiting l_i^u at t_i^u by u is modelled by a $|\boldsymbol{L}|$-sized vector $H = (h_1, h_2, ..., h_{|\boldsymbol{L}|})$, which can be represented as follows,

$$p((l_{i+1}^u, t_{i+1}^u)|H) = h_{l_{i+1}^u}. \tag{4}$$

Note that H can be either global parameters independent with u or personalized parameters dependent with u. Then, the probability of all the mobility records of user u, *i.e.*, $R(u) = \{(l_1^u, t_1^u), (l_2^u, t_2^u), ..., (l_{N_u}^u, t_{N_u}^u)\}$ can be computed as:

$$P(R(u)|H) = \prod_{i=1}^{N_u} h_{l_i^u} = \prod_{l \in \boldsymbol{L}} h_l^{c_l^u}, \tag{5}$$

where c_l^u is the observed visit times of user u to location l. It can be calculated by

$$c_l^u = \sum_{i=1}^{N_u} I(l_i^u = l). \tag{6}$$

$I(\cdot)$ is the indicator function of the logical expression. Specifically, we have $I(true) = 1$ and $I(false) = 0$. For an ID set ξ belonging to the same user, the probability of all the mobility records in $R(\xi)$ can be also calculated based on (5) by replacing u with ξ.

Markov Model: Markov model is widely used to model human movement [27, 29]. Specifically, it models the movements of human as transitions among definite and countable states, and each state corresponds to a location. Users' movement is characterized by a transition matrix T of size $|\boldsymbol{L}| \times |\boldsymbol{L}|$, where $|\boldsymbol{L}|$ is the total number of locations. Similarly with H, T can be either global parameters independent with u or personalized parameters dependent with u. Then, each location record is modelled to be only dependent on the last location record. Denote $T_{l,k}$ as the probability that user u moves from location l to location k in adjacent records. Then, the probability of a location record can be calculated as follows,

$$P((l_{i+1}, t_{i+1})|(l_i, t_i), T) = T_{l_i, l_{i+1}}. \tag{7}$$

Similarly, the probability of all the mobility records of user u can be computed as:

$$P(R(u)|T) = \prod_{i=1}^{N_u - 1} T_{l_i^u, l_{i+1}^u} = \prod_{l,k \in \boldsymbol{L}} T_{l,k}^{c_{lk}^u}, \tag{8}$$

where c_{lk}^u is the observed transition counts between location l and k of user u. It can be calculated by

$$c_{lk}^u = \sum_{i=1}^{N_u - 1} I(l_i^u = l) \cdot I(l_{i+1}^u = k). \tag{9}$$

For an ID set ξ belonging to the same user, the probability of all the mobility records in $R(\xi)$ can be also calculated based on (8) by replacing u with ξ.

Compared with multinomial model, Markov model further consider correlation between time-adjacent records, which models users' mobility patterns more accurately. However, there are more parameters used in Markov model, which may lead to overfitting problem.

5.2 Parameter Estimation Strategy

Given the probability model, there is still a problems unsolved. That is, what is the value of parameters used in the model? An easy way to think about is to estimate parameters from trajectories of IDs. However, when matching multiple IDs, whose parameters we should use? In addition, for IDs

with sparse trajectories, their estimated parameters are easy to be wrong or biased. By elaborately design the parameter estimation strategies, we present two solutions including global parameters and collapsed parameters as follows.

Global Parameters: In this case, we do not distinguish individuals, and assume all users follow the same mobility model with equal parameters H or T. Thus, they are estimated from the records of all users with the following equations:

$$h_l = \frac{\sum_{u \in \boldsymbol{A}} c_l^u}{\sum_{l \in \boldsymbol{L}} \sum_{u \in \boldsymbol{A}} c_l^u}. \tag{10}$$

$$T_{lk} = \frac{\sum_{u \in \boldsymbol{A}} c_{lk}^u}{\sum_{k \in \boldsymbol{L}} \sum_{u \in \boldsymbol{A}} c_{lk}^u}. \tag{11}$$

Then, based on the estimated global parameter, probability of observing mobility records $R(u)$ can be calculated based on (5) and (8).

The information utilized to solve the ID matching problem by using multinomial model mainly comes from "co-locations" of IDs, *i.e.*, visiting the same location at the same time bin. If a set of IDs ξ are frequently "co-located" in their trajectories. When we divide them into the same cluster corresponding to one user, a number of repeated records will be removed in their merged trajectories, and the probability of the corresponding mobility records will be calculated only by one time in probability of the merged trajectories $P(R(\xi))$, leading to the decrement of likelihood of all records conditioned on the corresponding partition $P(\mathcal{R}(\boldsymbol{A})|p)$. On the other hand, the information utilized to solve the ID matching problem by using Markov model mainly comes from transitions between time-adjacent records. If there are more frequent transitions based on T estimated from all users in one merged trajectory, its probability will be larger. In contract, if there are more rare transitions in one merged trajectories, its probability will be smaller. Thus, both models with global parameters require target trajectories to have overlapped time span to construct co-locations and transitions in their merged trajectories. Overall, in each model, we can find correct ID matching results by maximizing the probability of observed records conditioned on the corresponding partition, which describes the subgraph structures of IDs belonging to different users.

On the whole, estimating parameters based on the aggregated statistics of all IDs can overcome sparsity issue. For example, when location records of some IDs are insufficient, the global parameters can help us to better model their movements. On the other hand, personalized mobility patterns (e.g., frequently-visited locations) are ignored in this way. The corresponding side effect is that when trajectories of two IDs have no overlapped time window, the probability of their records will be independent with partition p. Thus, in this case, these models are ineffective in solving the ID matching problem.

Collapsed Parameters: Different with using global parameters, which are estimated from the records of all users, another solution is to integrate with respect to the parameters, and obtain an unconditional version of (5) and (8) from the perspective of expectations.

In order to achieve this goal, we adopt Bayesian probability methods, and regard both H and T as random variables. Further, by denoting $T_{l,:}$ as the $|\boldsymbol{L}|$-sized vector $[T_{l,1}, T_{l,2}, ..., T_{l,|\boldsymbol{L}|}]^T$, we use the common conjugate prior distribution to model them [40, 48], which can be expressed as follow:

$$p(H) = Dirichlet(H|\gamma_1), \tag{12}$$

$$p(T) = \prod_{l \in \boldsymbol{L}} Dirichlet(T_{l,:}|\gamma_2). \tag{13}$$

where $Dirichlet(\cdot|\gamma)$ is Dirichlet distribution with hyper-parameter γ [33]. Specifically, γ is used to describe how much we believe this prior.

Finally, the unconditional probability of observing mobility records $R(u)$ is computed by integrating H and T. We denote the unconditional probability under multinomial model and Markov

ID	Trajectories									
u_1	l_A	l_B	l_A	l_A	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$
u_2	l_A	$\emptyset$	$\emptyset$	$\emptyset$	l_B	l_B	l_B	$\emptyset$	$\emptyset$	$\emptyset$
u_3	$\emptyset$	l_B	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	l_A	l_A	l_A

(a) Trajectories of IDs

Locations l	Probability H_l
l_A	0.3
l_B	0.7

(b) Global Parameters

Partition p	MN-G	MN-C
$\{u_1, u_2\}, \{u_3\}$	-9.0	-14.4
$\{u_1, u_3\}, \{u_2\}$	-9.8	-12.2
$\{u_1\}, \{u_2\}, \{u_3\}$	-10.2	-12.8

(c) $\log P(\mathcal{R}(V)|p)$

Fig. 2. Three example IDs to be matched.

model as $P_{\text{MN}}(\cdot)$ and $P_{\text{MK}}(\cdot)$ respectively, and they can be computed as follows:

$$P_{\text{MN}}(R(u)) = \int P(R(u)|H)p(H)dH = \lambda_1(\gamma_1)B(b^u), \tag{14}$$

$$P_{\text{MK}}(R(u)) = \int P(R(u)|T)p(T)dT = \lambda_2(\gamma_2)\prod_{l\in L} B(b_l^u), \tag{15}$$

where $B(\cdot)$ is the multivariate Beta function, and $\lambda_1(\gamma)$ and $\lambda_2(\gamma)$ are functions of γ and independent with $R(u)$. b^u and b_l^u are both $|\boldsymbol{L}|$–sized vectors with $b^u(k) = c_k^u + \gamma_1$ and $b_l^u(k) = c_{lk}^u + \gamma_2$ for each $k \in \boldsymbol{L}$, respectively. Similarly, for a cluster of IDs ξ belonging to the same user, the probability of all the mobility records in $R(\xi)$ can be also calculated based on (14) and (15) by replacing u with ξ.

Different with global parameter, personalized mobility patterns are also considered by using collapsed parameters. Let us consider a simple example of matching two IDs $\xi = \{u, v\}$ by multinomial model. In addition, trajectories of u and v have no over-lapped time window in their trajectories. Thus, we have $b^\xi(l) = c_l^u + c_l^v + \gamma_1$. In addition, the trajectory of u is significantly denser than v, *i.e.*, for each l, we have $c_l^u \gg c_l^v$. From (14), we can obtain:

$$P_{\text{MN}}(R(\xi)) = \lambda(\gamma_1)B(b^u) \cdot \frac{B(b^\xi)}{B(b^u)}. \tag{16}$$

Based on the property of beta function, since $b^u(l) \gg |b^\xi(l) - b^u(l)| = c_l^v \ \ \forall l$, we have:

$$P_{\text{MN}}(R(\xi)) \approx \lambda(\gamma_1)B(b^u) \cdot \prod_{l\in L}\left[\frac{b^u(l)}{\sum_{l\in L} b^u(l)}\right]^{c_l^v}. \tag{17}$$

Then, for a fix u, if ID v has more records at u's frequently visited locations, the probability of their merged trajectories is larger, and we are more likely to divided them into one cluster. In reverse, if v has more record at u's rare visited locations, the probability is smaller, and they are more likely to be divided into different clusters in the partitions. Similarly, Markov model with collapsed parameter also considers the similarity of transition patterns of different IDs. Overall, in each model with collapsed parameters, we can also find correct ID matching results by maximizing the probability of observed records conditioned on the corresponding partition.

For simplicity, we refer the combinations of two probability models of multinomial model and Markov model and two parameter estimation strategies of global parameters and collapsed parameters as MN-G, MN-C, MK-G, and MK-C, respectively.

To illustrate the definitions in the user behavior model, let us consider three example IDs with mobility records shown in Figure 2. For simplicity, there are only two locations l_A and l_B in this example. Their trajectories are shown in Figure 2(a), where each column represents a time bin, and the global parameters are shown in Figure 2(b). Finally, the probability of observed records conditioned on three different partitions based on two different user behavior models is shown in Figure 2(c). Note that comparison between probability of different models (different columns) is meaningless. Based on MN-G model, since u_1 and u_2 are "co-located" at a rarely visited location l_A compared with "co-location" at l_B between u_1 and u_3 ($H_{l_A} < H_{l_B}$), u_1 and u_2 should be matched with higher probability $P(\mathcal{R}(V)|p)$. However, visit times to different locations between u_1 and u_3 are more similar compared with u_1 and u_2. Thus, u_1 and u_3 have higher probability under the MN-C model and should be linked based on this model. Note that the information of "co-location" is also considered in MN-C model, but in this example frequently visited locations for different IDs are a more dominant factor to link IDs. Overall, two user behavior models capture different aspects of mobility patterns of IDs to match them.

Overall, we propose two proposed parameter estimation strategies to characterize mobility features of trajectories, which help us overcome the challenge of adopting the probabilistic mobility models in the ID linking problem. Compared with the asymmetric parameter estimation strategies used in existing approaches [8, 34, 42, 43], the proposed two parameter estimation strategies are all symmetric, which eliminate the influence of the order of ID pairs to the performance of ID linking. The parameter estimation strategies combined with probabilistic mobility models pave the way for linking IDs belonging to the same users.

6 ID MATCHING PROBLEM

The goal of the set-wise ID matching problem is to find a ranking function $\phi(\xi_k)$ for each candidate ID set ξ_k. Specifically, we rank candidate ID sets ξ_k based on the joint posterior probability of IDs in ξ_k belonging to the same user conditioned on the observed mobility records, *i.e.*, $P(X(\xi_k, v) = 1|\mathcal{R}(\boldsymbol{V}))$, where $\boldsymbol{V}$ is the set of candidate online IDs. As introduced in Section 4.2, we limit $\boldsymbol{V}$ to the neighbor of v in the constructed *contact graph*, *i.e.*, $\mathcal{N}(v)$. We now introduce how to calculate $P(X(\xi, v) = 1|\mathcal{R}(\boldsymbol{V}))$.

6.1 Pair-wise Matching Problem

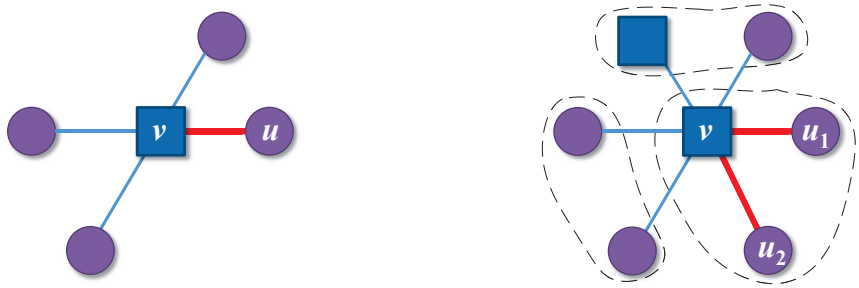
Let us first consider the case of pair-wise matching, *i.e.*, linking the pairs of IDs of two services that belong to the same users. In this case, each user is assumed to have at most one ID of each service. We denote the service of target ID as s_0, and the other service as s_1. Then, $\boldsymbol{V}$ can be further limited to $\boldsymbol{V} = \mathcal{N}^{s_1}(v) \cup v$, of which an example is shown in Figure 3(a).

For a target ID $v \in \boldsymbol{A}^{s_0}$ and an arbitrary ID $u \in \mathcal{N}^{s_1}(v)$, v can only belong to the same user with at most one ID in $\mathcal{N}^{s_1}(v)$. Thus, we have $\sum_{w \in \mathcal{N}^{s_1}(v)} X(w, v) \leq 1$. Considering it is possible that v does not belong to the owners of any IDs in $\mathcal{N}^{s_1}(v)$, we have:

$$\sum_{w \in \mathcal{N}^{s_1}(v)} P(X(w, v) = 1|\mathcal{R}(\boldsymbol{V})) + \beta(v) = 1, \tag{18}$$

where $\beta(v)$ is the probability that v does not belong to the same user with any IDs in $\mathcal{N}^{s_1}(v)$. On the other hand, through Bayes' theorem, we have:

$$P(X(w, v) = 1|\mathcal{R}(\boldsymbol{V})) = \frac{P(X(w, v) = 1)P(\mathcal{R}(\boldsymbol{V})|X(w, v) = 1)}{P(\mathcal{R}(\boldsymbol{V}))}. \tag{19}$$



(a) v and $u \in \mathcal{N}^{s_1}(v)$.

(b) A partition for $\mathcal{N}(v)$.

Fig. 3. Diagrams for relationship between candidate nodes or sets in $\mathcal{N}(v)$ and the target node v.

According to (2), for all $w \in \mathcal{N}^{s_1}(v)$, $X(w, v) = 1$ corresponds to a partition of $\boldsymbol{V}$ with the same prior probability, due to their similar component structure, *i.e.*, one 2-size set $\{w, v\}$ and other $|\boldsymbol{V}| - 2$ 1-size sets. We denote their prior probability as $P(X(w, v) = 1) = P(p_1)$. Further, we define $Q(w, v)$ as the joint probability of the observation $\mathcal{R}(\boldsymbol{V})$ and $X(w, v) = 1$, which can be expressed as follows:

$$\begin{aligned} Q(w, v) &\triangleq P(X(w, v) = 1) \cdot P(\mathcal{R}(\boldsymbol{V}) | X(w, v) = 1), \\ &= P(p_1) \cdot P(R(w, v) | X(w, v) = 1) \prod_{o \in \boldsymbol{V} \setminus \{w, v\}} P(R(o)). \end{aligned}$$

Then, (19) can be simplified as follows:

$$P(X(w, v) = 1 | \mathcal{R}(\boldsymbol{V})) = \frac{Q(w, v)}{P(\mathcal{R}(V))}. \tag{20}$$

Similarly, for $\beta(v)$, its corresponding partition is $p_0 = \{\{w\} | w \in \boldsymbol{V}\}$. Thus, we have:

$$\beta(v) = \frac{P(p_0) \cdot \prod_{w \in \boldsymbol{V}} P(R(w))}{\mathcal{R}(\boldsymbol{V})}.$$

By defining $\beta'(v)$ as the numerator of $\beta(v)$ and combining (18) and (20), we have:

$$P(X(u, v) = 1 | \mathcal{R}(\boldsymbol{V})) = \frac{Q(u, v)}{\sum\limits_{w \in \mathcal{N}^{s_1}(v)} Q(w, v) + \beta'(v)},$$

So far, we have obtained the probability that u belongs to the same user with v, which solves the pair-wise matching problem.

6.2 Multiplicity of IDs and Services

Based on the pair-wise matching problem, we further investigate the problem of multiple IDs and services. For each online ID $u \in \mathcal{N}^{s_1}(v)$, we have obtained the probability that u belongs to the same user with v. However, in the general SIMP, multiple IDs in $\mathcal{N}^{s_1}(v)$ can belong to the same user with v simultaneously. IDs in $\mathcal{N}^{s_0}(v)$ can also belong to the same user with v. In addition, IDs of multiple services should also be considered. Thus, we solve the problem in $\boldsymbol{V} = \mathcal{N}(v)$, of which an example is shown in Figure 3(b).

When considering multiple online IDs, *e.g.*, $u_1, u_2 \in \mathcal{N}(v)$, random variables $X(u_1, v)$ and $X(u_2, v)$ are not independent with each other. Thus, we cannot obtain the joint probability by simply

using the product of the probability for each ID. For a set of IDs $\xi \subseteq \boldsymbol{V}$, in order to calculate $P(X(\xi, v) = 1|\mathcal{R}(\boldsymbol{V}))$, we first apply condition probability formula, which obtains:

$$\begin{aligned} P(X(\xi, v) = 1|\mathcal{R}(\boldsymbol{V})) &= P(\mathcal{R}(\boldsymbol{V}), X(\xi, v) = 1)/P(\mathcal{R}(\boldsymbol{V})), \\ &\propto P(\mathcal{R}(\boldsymbol{V}), X(\xi, v) = 1). \end{aligned}$$

Then, we utilize *partition* to further simplify this equation. Specifically, by applying the formula of total probability with respect to all possible partitions of $\boldsymbol{V}$, we have:

$$P(\mathcal{R}(\boldsymbol{V}), X(\xi, v) = 1) = \sum_{p \in \mathcal{P}(\boldsymbol{V})} P(\mathcal{R}(\boldsymbol{V}), X(\xi, v) = 1|p)P(p),$$

where $P(p)$ is the prior probability of the partition p. Specifically, for an arbitrary partition p, if ξ and v are divided into one set in p, we have $P(X(\xi, v) = 1|p) = 1$. Otherwise, it equals to 0. We use $\mathcal{P}(\boldsymbol{A}, \xi \cup v)$ to denote the set of all partitions in which all IDs in $\xi \cup v$ are divided into one set. Then, the right hand can be limited to $\mathcal{P}(\boldsymbol{A}, \xi \cup v)$. Combining relation of $P(\mathcal{R}(\boldsymbol{V}), X(\xi, v) = 1)$ and $P(X(\xi, v) = 1|\mathcal{R}(\boldsymbol{V}))$ based on Bayes' theorem, we have:

$$P(X(\xi, v) = 1|\mathcal{R}(\boldsymbol{V})) \propto \sum_{p \in \mathcal{P}(\boldsymbol{V}, \xi \cup v)} P(\mathcal{R}(\boldsymbol{V})|p)P(p). \tag{21}$$

In addition, for an arbitrary partition $p \in \mathcal{P}(\boldsymbol{V})$, we use $D(p)$ to represent the likelihood $P(\mathcal{R}(\boldsymbol{V})|p)P(p)$, which can be calculated as follows,

$$D(p) = P(\mathcal{R}(\boldsymbol{V})|p)P(p) = P(p) \prod_{\eta \in p} P(R(\eta)|X(\eta) = 1).$$

Putting it into (21), we obtain:

$$P(X(\xi, v) = 1|\mathcal{R}(\boldsymbol{V})) \propto \sum_{p \in \mathcal{P}(\boldsymbol{V}, \xi \cup v)} D(p). \tag{22}$$

ALGORITHM 1: RS(v, $\mathcal{N}(v)$, ξ)

Input: The target ID v, its neighborhood $\mathcal{N}(v)$ and a set of IDs $\xi \subseteq \mathcal{N}(v)$.
Output: The probability-based ranking score $\mathrm{RS}(v, \mathcal{N}(v), \xi) = P(X(\xi, v) = 1|\mathcal{R}(\boldsymbol{V}))$.
Initialize:
  $D_{sum} \leftarrow 0$; $D_{target} \leftarrow 0$; $f \leftarrow 0$;
  if $|\mathcal{N}(v)| > N_{max}$ **then**
    $f = 1$;
if $f = 0$ **then**
  for $p \in \mathcal{P}(\mathcal{N}(v))$ **do**
    $D_{sum} = D_{sum} + D(p)$;
    if $\exists U \in p \ \ s.t. \ \ \xi \cup v \subseteq U$ **then**
      $D_{target} = D_{target} + D(p)$;
  $\mathrm{CP}(\xi, v) = D_{target}/D_{sum}$.
else
  $p = \{\xi \cup v\} \bigcup \{\{w\} | w \in \boldsymbol{V} \backslash \{v \cup \xi\}\}$;
  $\mathrm{CP}(\xi, v) = D(p)$.

On the contrary, $X(\xi, v) \neq 1$ corresponds to partitions in which all IDs in $\xi \cup v$ are not divided into one set. Thus, we also have:

$$P(X(\xi, v) \neq 1|\mathcal{R}(\boldsymbol{V})) \propto \sum_{p \in \mathcal{P}(\boldsymbol{V}) \setminus \mathcal{P}(\boldsymbol{V}, \xi \cup v)} D(p). \tag{23}$$

By combining (22) and (23), we have:

$$P(X(\xi, v) = 1|\mathcal{R}(\boldsymbol{V})) = \sum_{p \in \mathcal{P}(\boldsymbol{V}, \xi \cup v)} D(p) / \sum_{p \in \mathcal{P}(\boldsymbol{V})} D(p). \tag{24}$$

So far, we have obtained the probability that all IDs in ξ belongs to the same user with v, which solves the set-wise matching problem. This ranking function based on the posterior probability is optimal, which is proved by the following theorem:

Theorem 1 The ranking function based on the posterior probability $P(X(\xi, v) = 1|\mathcal{R}(\boldsymbol{V}))$ is the optimal solution of *Set-wise Identity Matching Problem.*

Proof: Let ϕ_0 denote the ranking function based on the posterior probability. Since the target function (1) is a random variable, we consider its expectation conditioned on the observations, and have:

$$E(\sum_{i=1}^{N} X(\xi_i, v)\phi(\xi_i)|\mathcal{R}(\boldsymbol{V})) = \sum_{i=1}^{N} P(X(\xi_i, v) = 1|\mathcal{R}(\boldsymbol{V}))\phi(\xi_i).$$

Without loss of generality, we assume $P(X(\xi_1, v)=1|\mathcal{R}(\boldsymbol{V})) \geq P(X(\xi_2, v)=1|\mathcal{R}(\boldsymbol{V})) \geq \ldots \geq P(X(\xi_N, v)= 1|\mathcal{R}(\boldsymbol{V}))$. On the other hand, according to our proposed method, we have $\phi_0(\xi_i) = i$. Thus, we have $\phi_0(\xi_1) \leq \phi_0(\xi_2) \leq \ldots \leq \phi_0(\xi_N)$. Combining these two inequalities and applying the rearrangement inequality, for an arbitrary ranking function ϕ_1, we have:

$$\sum_{i=1}^{N} P(X(\xi_i, v)=1|\mathcal{R}(\boldsymbol{V}))\phi_1(\xi_i) \geq \sum_{i=1}^{N} P(X(\xi_i, v)=1|\mathcal{R}(\boldsymbol{V}))\phi_0(\xi_i).$$

In another form, it can be expressed as:

$$E(\sum_{i=1}^{N} X(\xi_i, v)\phi_1(\xi_i)|\mathcal{R}(\boldsymbol{V})) \geq E(\sum_{i=1}^{N} X(\xi_i, v)\phi_0(\xi_i)|\mathcal{R}(\boldsymbol{V})).$$

Thus, the ranking function ϕ_0 generated by the algorithm minimizes the target function $\sum_{i=1}^{N} X(\xi_i, v)\phi(\xi_i)$ of the set-wise identity matching problem, which proves its optimality.■

6.3 Approximation Algorithm

The ranking function based on posterior probability is optimal. However, calculating it based on (24) suffers from high computational complexity growing exponentially with the size of $\boldsymbol{V}$. In order to solve this problem, we apply the three approximation methods as follows:

- *Ignoring non-adjacent IDs:* It is unreasonable to link two IDs of which the trajectories have no co-location. Thus, we limits the problem to the neighbor of v, *i.e.*, $\mathcal{N}(v)$, which significantly reduces the size of $\boldsymbol{V}$. For these IDs with a large number of neighbors, we implement further approximation methods to them.
- *Ignoring the denominator:* From (24), we can observe that the denominator is independent with ξ. Thus, we alternatively rank the candidate ID sets only based on the numerator of (24). By this way, only partitions in $\mathcal{P}(\boldsymbol{V}, \xi \cup v)$ need to be considered.
- *Reducing feasible partitions:* We further reduce feasible partitions in $\mathcal{P}(\boldsymbol{V}, \xi \cup v)$ to reduce computational complexity. Specifically, we use the constant parameter N_{max} to represent the maximum accepted $|\boldsymbol{V}|$. If $|\boldsymbol{V}| \geq N_{max}$, all IDs in $\boldsymbol{V} \setminus \{v \cup \xi\}$ are considered to belong to different users.

ALGORITHM 2: SIMP Algorithm

Input: The target ID v and its neighbor $\mathcal{N}(v)$, a list of candidate sets of IDs $\xi_1, ..., \xi_N \subseteq \mathcal{N}(v)$.
Output: The ranking function ϕ.
Initialize:
  $C \leftarrow \{1, ..., N\}$; $r \leftarrow 1$;
while $C \neq \emptyset$ **do**
  $k = \text{argmax}_{i \in C} \text{RS}(v, \mathcal{N}(v), \xi_i)$;
  $\phi(\xi_k) = r$;
  $C = C \backslash k$; $r = r + 1$;

As for the first approximation method of ignoring non-adjacent IDs, let us consider an example of using multinomial model with global parameters and given two non-adjacent IDs $V = \{u_1, u_2\}$. Two possible partitions of them are $p_1 = \{\{u_1\}, \{u_2\}\}$ and $p_2 = \{\{u_1, u_2\}\}$. Since u_1 and u_2 are non-adjacent, they do not have "co-locations". Thus, likelihood of their observed records conditioned on the two partitions, *i.e.*, $P(\mathcal{R}(V)|p_1)$ and $P(\mathcal{R}(V)|p_2)$, are completely equal, indicating that there is no correlation between trajectories of u_1 and u_2. Thus, it is reasonable to ignore u_2 when matching possible IDs belonging to the same user with u_1. As for the second approximation method of ignoring the denominator, this method does not influence ranking results of ID sets, but the left numerator does not range from 0 to 1 anymore and is only in proportion to the confidence probability. Overall, if we only care about ranking results, this approximation method can be adopted without any influence. If we also care about the confidence of top candidate sets, this method affects the utility in some way. The third approximation method is a compromise between calculating accurate probability of ID sets and the computational complexity.

Based on above approximation methods, the computational complexity is reduced from $O(2^{|V|})$ to $O(|V|)$. Based on them and (24), we propose an algorithm to calculate the confidence score, which is described in Algorithm 1. Given the target ID v and its neighborhood $\mathcal{N}(v)$, if $|\mathcal{N}(v)| \leq N_{max}$, the confidence score is computed by traversing all partitions in $\mathcal{P}(V)$ according to (24). Otherwise, the two proposed approximation methods are adopted to reduce the computational complexity. We further propose SIMP algorithm shown in Algorithm 2 to solve the general set-wise identity matching problem, in which confidence scores of candidate ID sets are calculated by invoking Algorithm 1, and ID sets with higher confidence probability are given higher rankings.

6.4 Matching Score Function

In certain application scenarios, a relatively small number of more reliable ID matching results are required rather than the whole ranking of all candidate sets of IDs. A naive solution to this problem is adding a threshold θ to filter out matched IDs with likelihood (24) less than θ. However, (24) cannot well distinguish the reliability of matching results when the likelihood is close to 1. Let us consider a simple example with two "isolated" IDs (w, v) to be matched. Here, "isolated" means they do not have other neighbor nodes. Then, the logarithmic likelihood can be represented as $\log Q(w, v) - \log(Q(w, v) + \beta)$. When we have $Q(w, v) \gg \beta$, due to the limitation of computational accuracy, the obtained logarithmic likelihood is always equal to 0. Thus, it is hard to characterize the reliability of ID pair (w, v). In order to solve this problem, we propose a modified matching score function based on the likelihood (24) as follows:

$$\begin{aligned} Score(\xi, v) &= \alpha_1 \log P(X(\xi, v)) - \alpha_2 \log(1 - P(X(\xi, v)), \\ &= (\alpha_2 - \alpha_1) \log \sum_{p \in \mathcal{P}(V)} D(p) + \alpha_1 \log \sum_{p \in \mathcal{P}(V, \xi \cup v)} D(p) - \alpha_2 \log \sum_{p \in \mathcal{P}(V) \backslash \mathcal{P}(V, \xi \cup v)} D(p), \end{aligned} \tag{25}$$

where α_1 and α_2 are hyper-parameters to adjust the influence of two parts. As we can observe, compared with the logarithmic likelihood, which ranges from negative infinity to 0, this modified matching score function ranges from negative infinity to positive infinity. Let us consider the same example of the two "isolated" IDs (w, v), and let $\alpha_1 = 0$. Then, we have $Score(w, v) \propto \log(Q(w, v) + \beta) - \log\beta$, which increases with $Q(w, v)$ even for $Q(w, v) \gg \beta$. Thus, we can better characterize the reliability of matching results, especially when the likelihood is close to 1. Then, we can use the threshold θ to filter out unreliable matched IDs with matching score (25) less than θ, and obtain the reliable matching results.

7 PERFORMANCE EVALUATION

We evaluate our system against two state-of-the-art pair-wise matching algorithms using two ground-truth datasets. Now we introduce the utilized datasets, evaluation metrics and experiment results.

7.1 Datasets

We have two *ground-truth* datasets for performance evaluation, including a new ISP dataset (4 services) and an existing dataset from a prior work (2 services). We summarize the basic statistics of utilized datasets in Table 4.

7.1.1 ISP Dataset. This dataset is collected by a large ISP in China, which covers 412,455 ground-truth users and their 815,117 IDs in 4 representative online services including instant messenger (QQ), online social network (Weibo), e-commerce (Taobao) and online review site (Dianping) during the full month of November 2015. All of them are the leading and most popular online services among the corresponding categories in China. In addition, they have covered the most commonly-used online service categories with different usage patterns. Thus, by conducting experiments based on them, performance of different ID linking algorithms can be comprehensively evaluated. We show details of these online services in Table 5. It records users' accessing activities via broadband network, which are associated to a physical locations, *e.g.*, a WiFi access point or a broadband interface. For simplicity, we refer them as access points (AP). Each record represents the user's login action in a given service, characterized by "service name", "ID", "AP name", and "timestamp". There are 31 million total records (on average 38.2 records per ID). The ground-truth is also provided by the same ISP, which collect users' online IDs via the cellular networks that are associated with the same device with unique cellular identifier.

We have taken a number of steps to protect the privacy of users involved in the utilized datasets in this paper. First, user identity information of all trajectories is removed by the providers of the datasets. Instead, they replace each user ID with an encrypted bit string. Second, the datasets are stored in separated servers belonging to each provider of the datasets, which are all well protected by their own firewalls. All data processing is finished on the services, which is overlooked by the collaborators.

7.1.2 Twitter-Foursquare. On Foursquare, users may display their Twitter account information, which makes it possible to obtain the ground-truth mapping between Twitter IDs and Foursquare IDs. This dataset is collected by Zhang et al. [58]. In total, it contains 385 users with location check-ins on both sides (770 online IDs), and totally 24,556 location check-ins collected from both Twitter and Foursquare. Compared with the ISP dataset, users' location contents on Twitter and

Table 2. Dataset summary.

Dataset	# Records	# Ground-Truth Users	Time	# Locations
ISP	31,137,469	412,455	Nov. 2015	3,362,800
Twitter-Foursquare	24,556	385	2008-2012	6,531

Table 3. Four services in the ISP dataset.

Service	Type	# of IDs
QQ	Instant messaging (IM)	725,621
Taobao	E-commerce (EC)	7,545
Weibo	Online social network (OSN)	2,545
Dianping	Online review (OR)	79,403

Foursquare are actively uploaded by them. Therefore, they are less sensitive compared with the ISP datasets. Even so, user IDs are all replaced by an encrypted bit string to protect users' privacy.

7.2 Baseline Algorithms

POIS: Riederer et al. [41] propose a probabilistic model to link IDs, which is also based on user behavior models. Specifically, they model the visit times of a user to a location by Poisson distribution. In addition, the occurrence of a login record of each service is modeled to follow independent Bernoulli distribution. Then, ID pairs with large matching scores obtained based on the probabilistic model are linked.

WYCI: Rossi et al. [42] model the probability of users' visit to different locations as a multinomial distribution. Specifically, they use one trajectory from the target ID pair to be linked to estimate the parameters of the model, and then use the left one to compute the matching score as follows,

$$S_{WYCI}(u,v) = \prod_{(l,t)\in R(u)} P(l|R(v)) = \prod_{(l,t)\in R(u)} \frac{c_l^v + \alpha}{\sum_{l\in L} c_l^v + \alpha|\boldsymbol{L}|}, \tag{26}$$

where c_l^v is the number of records of user v at location l. In addition, $|\boldsymbol{L}|$ is the number of distinctive locations, and α is the smoothing parameter, which can be also regarded as the parameter of the Dirichlet prior for the multinomial distribution.

UIDwST: Gao et al. [12] link users across social networks by measuring the similarity of users' check-ins in terms of spatio-temporal distribution. They further consider the importance of different check-in records by applying a weighting scheme based on the term frequency-inverse document frequency (TF-IDF). Finally, they also consider the conflictive check-in records with little time gap but large geographical distance. Then, they add a penalty to the above similarity which is inversely proportional to the number of conflictive check-in records between the target ID pairs. Finally, ID pairs with similarity larger than a predefined threshold are linked.

STUL: Chen et al. [6] also link users by measuring the similarity of their spatial and temporal behavior. In terms of spatial behavior, they measure the similarity between users' stay regions, which is also weighted by the importance of different regions by considering their popularity. In terms of temporal behavior, they measure the similarity of users' time distribution in each stay region.

Based on our analysis in Section 5.2, in the case of pair-wise ID matching for IDs with no overlapped time window and significant heterogeneous trajectory density ($c_l^u \gg c_l^v \ \ \forall l$), our proposed SIMP algorithm based on multinomial model and collapsed parameters is equivalent to WYCI algorithm.

These algorithms are all designed to match IDs between two services, and we apply them by matching multiple services one by one. For a given ID in one service, they produce a ranked list of the matched IDs (with ranking scores).

7.3 Evaluation Metrics

We use standard metrics including precision, recall and AUC to evaluate the algorithm performance with some adjustments. More specifically, for a target ID v, POIS or WYCI produces *a ranked list of IDs*: $[u_1, u_2, ..., u_k]$, where u_i is the i_{th} ID and k is the number of matched IDs. Our system SIMP produces *a ranked list of sets*: $[U_1, U_2, ..., U_k]$, where each item U_i is a set of IDs ($U_i = \{u_{i1}, ..., u_{is}\}$). For comparison, we need to either convert a set-list to an ID-list, or the other way around.

7.3.1 ID List Evaluation. First, we convert the set-list generated by our algorithm to an ID-list, by setting the set size as 1 ($|U_i| = 1$), and then compute the metrics using standard precision, recall, and AUC.

Precision & Recall: Given the list of matched IDs, precision is the fraction of IDs that truly belong to the same user within IDs that are matched by algorithms. Recall is the fraction of IDs that are correctly matched within IDs that actually belong to the same user. Then, F1 score is their harmonic mean [39]. Formally, we denote $\mathcal{D}$ as the set of IDs matched by the algorithm to be evaluated, and denote C as the set of all IDs belonging to the target users. Note that the target IDs are excluded in both C and $\mathcal{D}$. Then, these metrics can be expressed as follows:

$$\begin{cases} \text{Precision} &= \frac{|\mathcal{D} \cup C|}{|\mathcal{D}|}, \\ \text{Recall} &= \frac{|\mathcal{D} \cup C|}{|C|}, \\ \text{F1} &= \frac{2\text{Precision} \cdot \text{Recall}}{\text{Precision}+\text{Recall}}. \end{cases}$$

In these metrics, only the ID with the highest probability in the ID-list is evaluated. In addition, we can further trade off between precision and recall by adding constraints to filter out some poorly matched IDs. By this way, recall is decreased, while the precision is improved. Specifically, we add a threshold θ to filter out matched IDs with likelihood less than θ in our proposed SIMP algorithm. In addition, POIS algorithm adopts another threshold ϵ to the gap of matching scores of the best and second-best IDs. Thus, for baseline algorithms, we try both thresholds and select a better one.

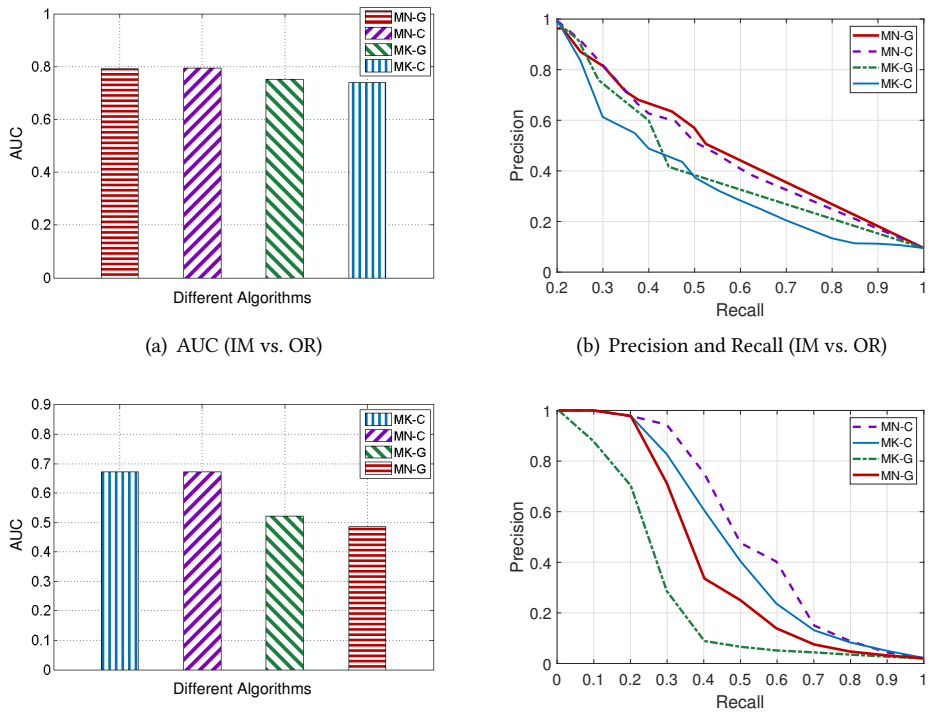
AUC pair-wise: By plotting the true positive rate (TPR) of the obtained results as a function of false positive rate (FPR), we obtain the ROC curve. Then, AUC is equal to the area under ROC curve [16, 44]. It is essentially evaluating the quality of a ranking, which is formulated as follows,

$$AUC = \frac{\sum_{i=1}^{n_0}(n_0 + n_1 - r_i) - n_0(n_0 + 1)/2}{n_0 n_1},$$

where r_i is the rank of i_{th} positive instance, and n_0 and n_1 are the number of positive and negative instances, respectively. Here, "positive" means the matched ID is correct based on ground-truth. We can observe that if a certain algorithm ranks the positive instance higher, the corresponding AUC is larger. Specifically, we set $k = 10$ IDs in the list, and thus $n_0 + n_1 = k = 10$. We refer this AUC (for pair-wise matching algorithms) as *AUC pair-wise*.

7.3.2 Set List Evaluation. Clearly, converting the set-list to an ID-list diminishes the key benefits of our system. Our system may correctly match all the IDs in the top candidate set, but had to shrink the set size to 1 for the comparison. Thus, we introduce a method to convert the ID list into set list (applied to results of POIS and WYCI), and use a AUC to evaluate the ranking quality of the set list. The basic idea is to group IDs into sets (of pre-defined size), and then we rank these sets based on the highest ranked ID in each set. For example, for a given ID list $[u_1, u_2, , ..., u_3]$, we can convert them to a set list (with set size = 2) as $[\{u_1, u_2\}, \{u_1, u_3\}, \{u_2, u_3\}]$.

AUC set-wise: For a given list of sets, we also use AUC to evaluate the ranking quality. We follow the same definition of AUC to calculate the probability of ranking a randomly chosen "positive set"



(a) AUC (IM vs. OR)

(b) Precision and Recall (IM vs. OR)

(c) AUC (Twitter vs. Foursquare)

(d) Precision and Recall (Twitter vs. Foursquare)

Fig. 4. Performance of different user behavior models on IDs of IM vs. OR and Twitter vs. Foursquare.

higher than a randomly chosen "negative set", where "positive set" means that *all* the IDs in this set belong to the same user as the target ID v. If any ID is incorrect, the set is a negative one.

7.4 Experiment and Results

We evaluate our system by experiments in different cases with IDs of one-to-one relation (one ID per service) and of one-to-many relation (multiple IDs per service and multiple services). For better understanding, we summary the default values of key parameters in our experiments in Table 4. Other parameters of baseline algorithms all follow the recommended settings [6, 12, 41, 42].

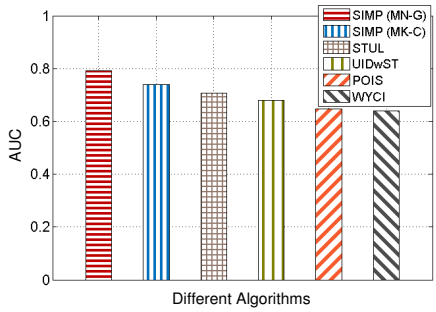
7.4.1 One-to-One Relation. We first select users who only have one IM identity and one OR identity of the ISP dataset as the ground-truth, and evaluate the performance of different user

Table 4. Default parameter settings.

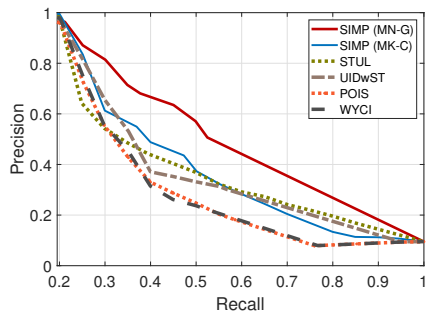
Parameters	Description	Value
-	Spatial Resolution (ISP dataset)	Access points (AP)
-	Spatial Resolution (Twitter-Foursquare dataset)	$1km \times 1km$ grids
-	Temporal Resolution	1 hour
k	#Instances considered in AUC	10
β	Default value of $\beta(v)$	e^{-10}
α_2/α_1	Parameters of matching score function (25)	10^5
γ_1	Prior Parameters for multinomial model	0.1
γ_2	Prior Parameters for Markov model	0.1

Table 5. Four services in the ISP dataset.

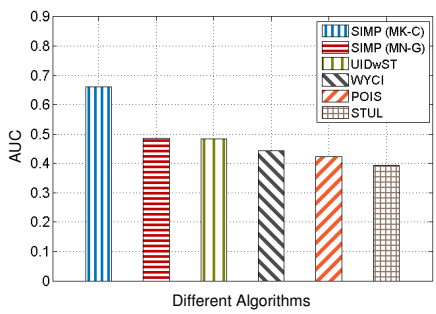
Service	Type	# of IDs
QQ	Instant messaging (IM)	725,621
Taobao	E-commerce (EC)	7,545
Weibo	Online social network (OSN)	2,545
Dianping	Online review (OR)	79,403



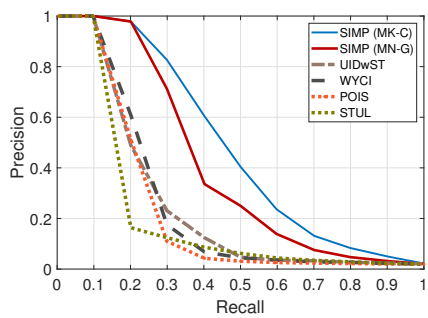
(a) AUC (IM vs. OR)



(b) Precision and Recall (IM vs. OR)



(c) AUC (Twitter vs. Foursquare)

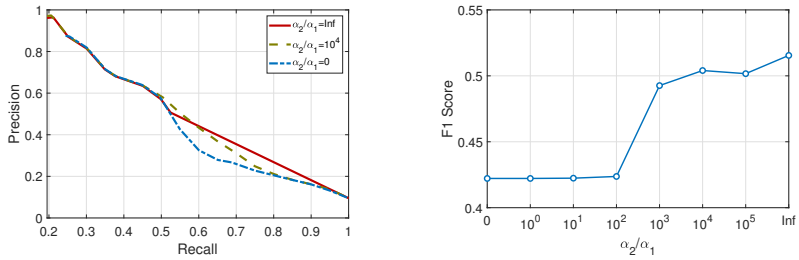


(d) Precision and Recall (Twitter vs. Foursquare)

Fig. 5. Performance of different algorithms on IDs of IM vs. OR and Twitter vs. Foursquare.

behavior models and different algorithms for IDs with one-to-one relation on them as well as the Twitter-Foursquare dataset. Note than in this case, the set size shrinks to 1, which is to our disadvantage.

We first examine the impact of different user behavior models to the performance of our proposed SIMP algorithm in Figure 4. Specifically, Figure 4(a) and (b) show the experimental results on the ISP dataset, where Figure 4(a) shows the mean AUC of SIMP algorithm with different user behavior models and Figure 4(b) shows the precision-recall curves by adjusting parameter θ or ϵ in our proposed algorithms and baselines. From Figure 4(b), we can observe that the proposed SIMP algorithms of both user behavior models with global parameters perform better. The reason lies in that global parameters utilize trajectories of all users to estimate parameters in the behavior models. Thus, they are more accurate than algorithms with collapsed parameters which only utilize trajectories of the two target trajectories.



(a) Precision-recall curves of different parameters

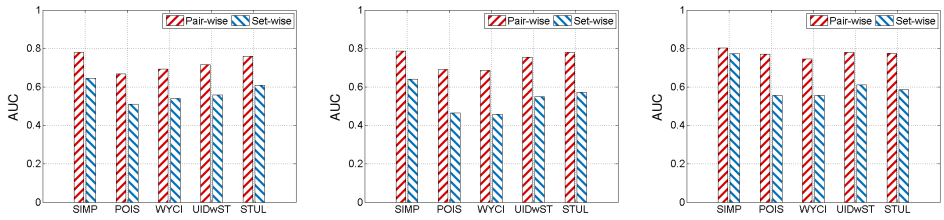
(b) F1 score vs. α_2/α_1

Fig. 6. Impact of parameters in the matching score function.

In Figure 4(c) and (d) conduct the same experiments using the Twitter-Foursquare dataset as in Figure 4(a) and (b), which show a different trend with results on the ISP dataset. That is, as shown in Figure 4(c) and (d), two user behavior models with collapsed parameters have the better overall performance. As we introduced in Sec. 5, both models with global parameters require target trajectories to have overlapped time span to construct co-locations and transitions in their merged trajectories. However, the time of location records in the Twitter-Foursquare dataset ranges from 2008 to 2012, while records in the ISP dataset concentrate on one month and are much denser. Thus, the constructed co-locations and transitions in the Twitter-Foursquare dataset are much less. For example, according to our statistics, only 7.5% records in the Twitter-Foursquare dataset have "co-located" records of another services within the period of one day, while for the ISP dataset the percentage is larger than 35.8%, which leads to the under-performance of the proposed algorithms with global parameters. Different with global parameters, personalized mobility patterns are also considered by using collapsed parameters. Thus, they have better performance of the Twitter-Foursquare dataset.

In addition, we can observe that MN-G and MN-C achieve the best performance on two datasets, respectively. However, it is hard to conclude that multinomial model is better than Markov model. On both datasets, MK-C algorithm has very close performance to the best one. For example, the performance gap between MK-C algorithm and the best MN-C algorithm in terms of AUC is less than 0.02 on Twitter-Foursquare dataset. Theoretically, Markov model has larger performance potential than multinomial model, since it further characterize the transition patterns of user trajectories compared with multinomial model. The major factor that limits it performance in our experiments is the sparsity of trajectories. Thus, the relative larger number of parameters in Markov model are not well estimated based on the insufficient trajectory points, leading to its under-performance compared with multinomial model.

In Figure 5, we conduct the same experiments as in Figure 4 to compare the performance of our proposed algorithms compared with baselines. As for our proposed algorithms, we only show their performance with two representative user behavior models, *i.e.*, multinomial model with global parameters (MN-C) and Markov model with collapsed parameters (MK-C). Results show that our algorithms have better performance on the most situations. Specifically, our algorithms outperform the baselines by 0.12 in terms of AUC on average on the ISP dataset, and outperform the baselines by 0.23 in terms of AUC on average on the Twitter-Foursquare dataset. In addition, we can observe that STUL algorithm outperforms other baselines on the ISP dataset. However, it suffers from under-performance on the Twitter-Foursquare dataset. The main reason is that this



(a) One OR to many IM (b) One IM to many OR (c) One OSN to many IM

Fig. 7. Performance on IDs with one-to-many relation, with different combinations of services.

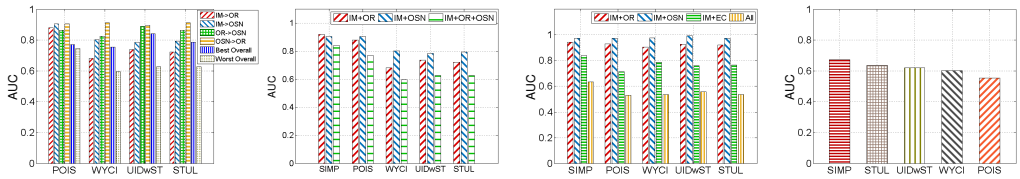
algorithm utilizes the aggregated features of user trajectories to characterize the importance of each location record, which is similar to the utilized global parameters in our proposed algorithms. However, trajectories of Twitter-Foursquare are insufficient compared with their large spatial coverage. Thus, the obtained features by aggregating user trajectories are not accurate, leading to its under-performance. Overall, experimental results indicate the effectiveness of our proposed algorithms.

Then, we evaluate the impact of parameter α_1 and α_2 in our proposed matching score function in Figure 6. First, we show the precision-recall curves of the proposed SIMP algorithm with three representative parameters in Figure 6(a), where we use the multinomial user mobility model as the default setting. Specifically, the matching score function with $\alpha_2/\alpha_1 = 0$ is equivalent to the original logarithmic likelihood. We can observe that a larger α_2/α_1 helps to improve the performance in terms of precision. Further, we fix the recall to be 0.6, and show the F1 score of the proposed SIMP algorithm as the function of α_2/α_1 in Figure 6(b). We can observe that by increasing α_2/α_1, we can obtain a performance gain of about 9.3% in terms of F1 score, indicating the effectiveness of our proposed matching score function.

7.4.2 One-to-Many Relation. Then, we select users with IDs of one-to-many relation, *e.g.*, users with one IM identity and multiple OR identities, and evaluate the performance of our system using both pair-wise and set-wise AUC. As for our proposed algorithms, we use the SIMP algorithm with multinomial model and global parameters (MN-C) as the default setting. The results are shown in Figure 7. From the results, we can have three key observations. First, our algorithm significantly outperforms the baselines over different combinations of services and AUC metrics. Second, the performance of our algorithm is more *consistent* over different services. The baselines, in contrary, have a larger variance in the AUC. Third, the gaps between our algorithm and baselines are larger for set-wise AUC than pair-wise AUC. This indicates the advantage of our algorithm in finding multiple IDs in one service simultaneously. The highest gain over baselines is 0.2 in terms of AUC (pair-wise), when matching IM to OSN.

7.4.3 Multiple Services. Next, we examine the performance of our algorithm in linking IDs over multiple services by selecting users who have three IDs (*i.e.*, one IM, one OR and one OSN) as the ground-truth. We examine the impact of *number of services* and *matching order*. In this particular experiment, we use IM identity as the target ID, and find the other two IDs belonging to the same user. There are three possible matching sequences: (IM-OR, IM-OSN), (IM-OR, OR-OSN), (IM-OSN, OSN-OR). We perform each sequence and obtain the set-wise AUC shown in Figure 8.

From Figure 8(a) and (b), we find that *number of services* has a significant influence on the two baseline algorithms. Both POIS and WYCI have a clear performance degradation from 2-service



(a) Different matching orders. (b) AUC for three services. (c) AUC for four services. (d) Overall Performance.

Fig. 8. Performance on IDs of multiple services and overall performance.

matching to 3-service matching, where the average AUC difference is 0.16. In addition, the *matching order* also matters, particularly for WYCI. In this case, the AUC difference between the best and worse sequences can be as large as 0.2. This confirms our design intuition, that the pair-wise matching has fundamental limitation to scale-up to multiple services. On the other hand, as we can observe from Figure 8(b), performance degradation of our proposed algorithm is only 0.05 in terms of AUC from 2-service matching to 3-service matching, while it is 0.16 on average for two baseline algorithms. It indicates that our algorithm is much less sensitive to the matching order nor the number of services.

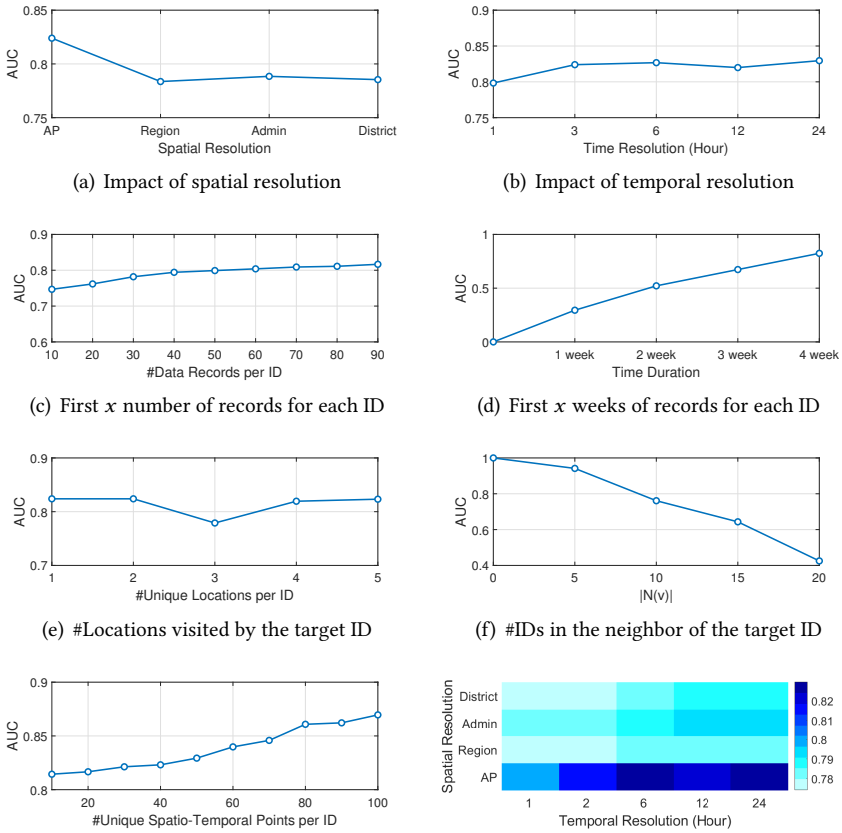
Figure 8(c) shows the result by extending the experiment scope to users who have four IDs (*i.e.*, one ID for each service). We can observe that our algorithm consistently outperforms other algorithms under these settings. The advantage is more obvious for multi-service matching. Finally, we discard all constraints and show the overall performance of different algorithms in Figure 8(d). We can observe that our algorithm outperforms other algorithms with performance gap of over 0.1 in terms of AUC.

7.4.4 Different Data Quality. Finally, we evaluate the sensitivity of our system in different data quality. We consider a number of key factors including the spatial/temporal resolution, number of records per ID, time range of the dataset, number of visited locations per ID, number of IDs in the neighbor, and number of spatio-temporal points per ID. For this group of experiments, we follow the same settings as Figure 5.

First, we evaluate the impact of temporal and spatial resolution. In terms of spatial resolution, we divide the city (where the ISP data is collected) into 7527 regions, 188 administration areas and 16 districts. We reduce the spatial resolution from AP-level all the way to the district-level. The results are shown in Figure 9(a). On the other hand, we also reduce the temporal resolution from 1 hour to 24 hour, and show results in Figure 9(b). We find our algorithm is not very sensitive to temporal/spatial resolution changes, *i.e.*, the variance is less than 5% in terms of AUC. It indicates the robustness of the model. The result is also consistent with a prior finding that human mobility has a high level of uniqueness even under a coarse-grained temporal/spatial resolution [57].

Then, we perform ID matching by using the first x number of data records for each ID. The obtained results are shown in Figure 9(c). We can observe that a larger number of data records help to improve matching accuracy. When there are 10 or more records, the AUC of the proposed algorithms reaches a stable state. We then change the time duration of the ISP dataset from 1 to 4 weeks, and show the results in Figure 9(d). It shows a similar trend: with more data records or a longer observation period, the matching accuracy increases.

We also evaluate the impact of the number of unique locations visited by the target ID. Specifically, we show the performance of algorithms on user groups with different number of visited locations in Figure 9(e). As we can observe, the performance is not much influenced by the number of visited



(a) Impact of spatial resolution

(b) Impact of temporal resolution

(c) First x number of records for each ID

(d) First x weeks of records for each ID

(e) #Locations visited by the target ID

(f) #IDs in the neighbor of the target ID

(g) #Unique spatio-temporal points per ID

(h) Impact of spato-temporal resolution

Fig. 9. Performance evaluation with different data quality metrics.

locations, indicating the time bins of mobility records has provided sufficient information for linking IDs. Then, we evaluate the impact of the number of IDs in the neighbor of target ID v, *i.e.*, $|\mathcal{N}(v)|$. Intuitively, as the number of IDs in the neighbor increases, the collision of similar spatial-temporal traces become larger. As we can observe from Figure 9(f), a larger $|\mathcal{N}(v)|$ leads to the decrement of matching accuracy, which coincides with our intuition.

Next, we also evaluate the impact of the number of unique spatio-temporal points of the target ID. We show the performance of algorithms on user groups with different number of unique spatio-temporal points in Figure 9(g). We can observe that performance for target ID with denser trajectories, *i.e.*, larger number of unique spatio-temporal points, is better than that of sparse trajectories. We further investigate the performance under different combination of spatial and temporal resolutions in Figure 9(h). We can observe that performance of our proposed algorithm is better under relatively larger temporal resolutions for most spatial resolutions. The main reason is the sparsity and noise of trajectory datasets. Overall, our algorithm is not very sensitive to temporal/spatial resolution changes, indicating its effectiveness.

Summary: The evaluation results show that our proposed system outperforms the baseline algorithms in different aspects, particularly for IDs of many-to-many relation across multiple

services. Its AUC beats baselines by 0.1 in overall performance, and by 0.2 in many-to-many ID matching, In addition, results show that our solution is robust to the changes in spatial-temporal resolution, demonstrating the effectiveness of our proposed system.

8 CONCLUSIONS

In this work, we propose an ID-linking algorithm across multiple services by modeling the spatial-temporal locality of user activities. We propose a novel contact graph and an optimal Bayesian-based inference method to link IDs across services. Our system solves a number of open problems in multi-service ID linking, including service and identity multiplicity and heterogeneous data quality. Experimental results on large scale and real-world datasets demonstrate the effectiveness of our system.

REFERENCES

[1] Osman Abul, Francesco Bonchi, and Mirco Nanni. 2008. Never walk alone: Uncertainty for anonymity in moving objects databases. In *Proceedings of the IEEE International Conference on Data Engineering (ICDE).*

[2] Osman Abul, Francesco Bonchi, and Mirco Nanni. 2010. Anonymization of moving objects databases by clustering and perturbation. *Information Systems* 35, 8 (2010), 884–910.

[3] Gergely Acs and Claude Castelluccia. 2014. A case study: Privacy preserving release of spatio-temporal density in paris. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD).*

[4] Miguel E Andrés, Nicolás E Bordenabe, Konstantinos Chatzikokolakis, and Catuscia Palamidessi. 2013. Geo-indistinguishability: Differential privacy for location-based systems. In *Proceedings of the ACM SIGSAC conference on Computer & communications security (CCS).*

[5] Dirk Brockmann, Lars Hufnagel, and Theo Geisel. 2006. The scaling laws of human travel. *Nature* 439, 7075 (2006), 462.

[6] Wei Chen, Hongzhi Yin, Weiqing Wang, Lei Zhao, Wen Hua, and Xiaofang Zhou. 2017. Exploiting spatio-temporal user behaviors for user linkage. In *Proc. ACM CIKM.* 517–526.

[7] Eunjoon Cho, Seth A Myers, and Jure Leskovec. 2011. Friendship and mobility: user movement in location-based social networks. In *Proc. KDD.*

[8] Yoni De Mulder, George Danezis, Lejla Batina, and Bart Preneel. 2008. Identification via location-profiling in GSM networks. In *Proc. ACM WPES.*

[9] Fatima Zohra Ennaji, Lobna Azaza, Zakaria Maamar, Abdelaziz El Fazziki, Marinette Savonnet, Mohamed Sadgal, Eric Leclercq, Idir Amine Amarouche, and Djamal Benslimane. 2018. Impact of Credibility on Opinion Analysis in Social Media. *Fundamenta Informaticae* 162, 4 (2018), 259–281.

[10] Jie Feng, Mingyang Zhang, Huandong Wang, Zeyu Yang, Chao Zhang, Yong Li, and Depeng Jin. 2019. DPLink: User Identity Linkage via Deep Neural Network From Heterogeneous Mobility Data. In *Proc. WWW.* 459–469.

[11] Ming Gao, Ee-Peng Lim, David Lo, Feida Zhu, Philips Kokoh Prasetyo, and Aoying Zhou. 2015. CNL: collective network linkage across heterogeneous social platforms. In *Proceedings of the IEEE International Conference on Data Mining (ICDM).*

[12] Xing Gao, Wenli Ji, Yongjun Li, Yao Deng, and Wei Dong. 2018. User Identification with Spatio-Temporal Awareness across Social Networks. In *Proc. ACM CIKM.*

[13] Oana Goga, Howard Lei, Sree Hari Krishnan Parthasarathi, Gerald Friedland, Robin Sommer, and Renata Teixeira. 2013. Exploiting innocuous activity for correlating users across sites. In *Proceedings of the 22nd international conference on World Wide Web (WWW).*

[14] Oana Goga, Patrick Loiseau, Robin Sommer, Renata Teixeira, and Krishna P Gummadi. 2015. On the reliability of profile matching across large online social networks. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD).*

[15] Marco Gramaglia and Marco Fiore. 2015. Hiding mobile traffic fingerprints with GLOVE. In *Proceedings of the 11th ACM Conference on Emerging Networking Experiments and Technologies (CoNext).*

[16] David J Hand and Robert J Till. 2001. A simple generalisation of the area under the ROC curve for multiple class classification problems. *Machine learning* 45, 2 (2001), 171–186.

[17] Danesh Irani, Steve Webb, Calton Pu, and Kang Li. 2011. Modeling unintended personal-information leakage from multiple online social networks. *Internet Computing* 15, 3 (2011), 13–19.

[18] Shouling Ji, Weiqing Li, Neil Zhenqiang Gong, Prateek Mittal, and Raheem A Beyah. 2015. On Your Social Network De-anonymizablity: Quantification and Large Scale Evaluation with Seed Knowledge.. In *Proceedings of the Network and Distributed System Security Symposium (NDSS).*

[19] Shouling Ji, Weiqing Li, Mudhakar Srivatsa, and Raheem Beyah. 2014. Structural data de-anonymization: Quantification, practice, and implications. In *Proceedings of the ACM Conference on Computer and Communications Security (CCS)*.

[20] Ethan Katz-Bassett, John P John, Arvind Krishnamurthy, David Wetherall, Thomas Anderson, and Yatin Chawathe. 2006. Towards IP geolocation using delay and topology measurements. In *Proceedings of the ACM SIGCOMM conference on Internet Measurement (IMC)*.

[21] Nitish Korula and Silvio Lattanzi. 2014. An efficient reconciliation algorithm for social networks. *Proceedings of the VLDB Endowment (PVLDB)* 7, 5 (2014), 377–388.

[22] Shamanth Kumar, Reza Zafarani, and Huan Liu. 2011. Understanding User Migration Patterns in Social Media. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.

[23] Chaozhuo Li, Senzhang Wang, Yukun Wang, Philip Yu, Yanbo Liang, Yun Liu, and Zhoujun Li. 2019. Adversarial learning for weakly-supervised social network alignment. In *Proc. AAAI*.

[24] Chaozhuo Li, Senzhang Wang, Philip S Yu, Lei Zheng, Xiaoming Zhang, Zhoujun Li, and Yanbo Liang. 2018. Distribution Distance Minimization for Unsupervised User Identity Linkage. In *Proc. ACM CIKM*.

[25] Chung-Yi Li and Shou-De Lin. 2014. Matching users and items across domains to improve the recommendation quality. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*.

[26] Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. 2007. t-closeness: Privacy beyond k-anonymity and l-diversity. In *Proceedings of the 23rd IEEE International Conference on Data Engineering (ICDE)*.

[27] George Liu and Gerald Maguire Jr. 1996. A class of mobile motion prediction algorithms for wireless mobile computing and communications. *Mobile Networks and Applications* 1, 2 (1996), 113–121.

[28] Siyuan Liu, Shuhui Wang, Feida Zhu, Jinbo Zhang, and Ramayya Krishnan. 2014. Hydra: Large-scale social identity linkage via heterogeneous behavior modeling. In *Proceedings of the ACM Conference on Management of Data (SIGMOD)*.

[29] Xin Lu, Erik Wetter, Nita Bharti, Andrew J Tatem, and Linus Bengtsson. 2013. Approaching the limit of predictability in human mobility. *Scientific reports* 3 (2013).

[30] Xucheng Luo, Shengyang Li, and Yuxiang Peng. 2020. CNNTOP: a CNN-based Trajectory Owner Prediction Method. *arXiv preprint arXiv:2001.01185* (2020).

[31] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. 2007. l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 1, 1 (2007), 3.

[32] Xin Mu, Feida Zhu, Ee Peng Lim, Jing Xiao, Jianzong Wang, and Zhi Hua Zhou. 2016. User Identity Linkage by Latent User Space Modelling. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*.

[33] Kevin P Murphy. 2012. Machine Learning: A Probabilistic Perspective. Adaptive Computation and Machine Learning.

[34] Farid M Naini, Jayakrishnan Unnikrishnan, Patrick Thiran, and Martin Vetterli. 2016. Where you are is who you are: User identification by matching statistics. *IEEE Transactions on Information Forensics and Security* 11, 2 (2016), 358–372.

[35] Arvind Narayanan and Vitaly Shmatikov. 2008. Robust de-anonymization of large sparse datasets. In *Proceedings of the IEEE Symposium on Security and Privacy (SP)*.

[36] Yuanping Nie, Yan Jia, Shudong Li, Xiang Zhu, Aiping Li, and Bin Zhou. 2016. Identifying users across social networks based on dynamic core interests. *Neurocomputing* 210 (2016), 107–115.

[37] Simon Oya, Carmela Troncoso, and Fernando Pérez-González. 2017. Back to the drawing board: Revisiting the design of optimal location privacy-preserving mechanisms. In *Proceedings of the ACM Conference on Computer and Communications Security (CCS)*.

[38] Pewinternet. [n. d.]. *Social Media Update 2016*. http://www.pewinternet.org/2016/11/11/social-media-update-2016/.

[39] David Martin Powers. 2011. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. (2011).

[40] Jan-Hendrik Prinz, Hao Wu, Marco Sarich, Bettina Keller, Martin Senne, Martin Held, John D Chodera, Christof Schütte, and Frank Noé. 2011. Markov models of molecular kinetics: Generation and validation. *The Journal of chemical physics* 134, 17 (2011), 174105.

[41] Christopher Riederer, Yunsung Kim, Augustin Chaintreau, Nitish Korula, and Silvio Lattanzi. 2016. Linking Users Across Domains with Location Data: Theory and Validation. In *Proceedings of the 25th International Conference on World Wide Web (WWW)*.

[42] Luca Rossi and Mirco Musolesi. 2014. It's the way you check-in: identifying users in location-based social networks. In *Proceedings of the 2nd ACM SIGCOMM Workshop on Online Social Networks (WOSN)*.

[43] Reza Shokri, George Theodorakopoulos, Jean-Yves Le Boudec, and Jean-Pierre Hubaux. 2011. Quantifying location privacy. In *Proceedings of the IEEE Symposium on Security and Privacy (SP)*.

[44] Kai Shu, Suhang Wang, Jiliang Tang, Reza Zafarani, and Huan Liu. 2017. User Identity Linkage across Online Social Networks: A Review. *ACM SIGKDD Explorations Newsletter* 18, 2 (2017), 5–17.

[45] Mudhakar Srivatsa and Mike Hicks. 2012. Deanonymizing mobility traces: Using social network as a side-channel. In *Proceedings of the ACM conference on Computer and communications security (CCS)*.

[46] Latanya Sweeney. 2002. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10, 05 (2002), 557–570.

[47] Techcrunch. [n. d.]. *How Many Twitter Accounts Do You Have?* https://techcrunch.com/2008/01/09/how-many-twitter-accounts-do-you-have/.

[48] Benjamin Trendelkamp-Schroer and Frank Noé. 2013. Efficient Bayesian estimation of Markov model transition matrices with given stationary distribution. *The Journal of chemical physics* 138, 16 (2013), 04B612.

[49] Ionut Trestian, Supranamaya Ranjan, Aleksandar Kuzmanovic, and Antonio Nucci. 2009. Measuring serendipity: connecting people, locations and interests in a mobile 3G network. In *Proceedings of the ACM SIGCOMM conference on Internet Measurement (IMC)*.

[50] Huandong Wang, Chen Gao, Yong Li, Zhi-Li Zhang, and Depeng Jin. 2017. From Fingerprint to Footprint: Revealing Physical World Privacy Leakage by Cyberspace Cookie Logs. In *Proc. ACM CIKM*. 1209–1218.

[51] Huandong Wang, Yong Li, Chen Gao, Gang Wang, Xiaoming Tao, and Depeng Jin. 2019. Anonymization and De-anonymization of Mobility Trajectories: Dissecting the Gaps between Theory and Practice. *IEEE Transactions on Mobile Computing* (2019).

[52] Huandong Wang, Yong Li, Gang Wang, and Depeng Jin. 2018. You Are How You Move: Linking Multiple User Identities From Massive Mobility Traces. In *Proc. SDM*.

[53] Ming Yan, Jitao Sang, Tao Mei, and Changsheng Xu. 2013. Friend transfer: cold-start friend recommendation with cross-platform transfer learning of social knowledge. In *Proceedings of the International Conference on Multimedia and Expo (ICME)*.

[54] Chunfeng Yang, Huan Yan, Donghan Yu, Yong Li, and Dah Ming Chiu. 2017. Multi-site User Behavior Modeling and Its Application in Video Recommendation. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*.

[55] Reza Zafarani and Huan Liu. 2013. Connecting users across social media sites: a behavioral-modeling approach. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*.

[56] Reza Zafarani and Huan Liu. 2014. Finding Friends on a New Site Using Minimum Information. In *Proceedings of the SIAM International Conference on Data Mining (SDM)*.

[57] Hui Zang and Jean Bolot. 2011. Anonymization of location data does not work: A large-scale measurement study. In *Proceedings of the 17th annual international conference on Mobile computing and networking (MobiCom)*.

[58] Jiawei Zhang, Xiangnan Kong, and Philip S Yu. 2014. Transferring heterogeneous links across location-based social networks. In *Proceedings of the ACM International Conference on Web Search and Data Mining (WSDM)*.

[59] Jiawei Zhang and S Yu Philip. 2015. Multiple anonymized social networks alignment. In *Proceedings of the IEEE International Conference on Data Mining (ICDM)*.

[60] Zexuan Zhong, Yong Cao, Mu Guo, and Zaiqing Nie. 2018. CoLink: An Unsupervised Framework for User Identity Linkage.. In *Proc. AAAI*.

[61] Xiaoping Zhou, Xun Liang, Xiaoyong Du, and Jichao Zhao. 2018. Structure based user identification across social networks. *IEEE Transactions on Knowledge and Data Engineering* 30, 6 (2018), 1178–1191.