Reconstructing Individual Mobility from Smart Card Transactions: A Space Alignment Approach

Nicholas Jing Yuan*, Yingzi Wang†*, Fuzheng Zhang†*, Xing Xie* and Guangzhong Sun†
* Microsoft Research
Email: {nicholas.yuan, xing.xie at microsoft.com}
† School of Computer Science and Technology, University of Science and Technology of China
Email: {yingzi, zhfzh} at mail.ustc.edu.cn, gzsun at ustc.edu.cn

***Abstract*—Smart card transactions capture rich information of human mobility and urban dynamics, therefore are of particular interest to urban planners and location-based service providers. However, since most transaction systems are only designated for billing purpose, typically, fine-grained location information, such as the exact boarding and alighting stops of a bus trip, is only partially or not available at all, which blocks deep exploitation of this rich and valuable data at individual level.**

This paper presents a "space alignment" framework to reconstruct individual mobility history from a large-scale smart card transaction dataset pertaining to a metropolitan city. Specifically, we show that by delicately aligning the monetary space and geospatial space with the temporal space, we are able to extrapolate a series of critical domain specific constraints. Later, these constraints are naturally incorporated into a semi-supervised conditional random field to infer the exact boarding and alighting stops of all transit routes with a surprisingly high accuracy, e.g., given only 10% trips with known alighting/boarding stops, we successfully inferred more than 78% alighting and boarding stops from all unlabeled trips. In addition, we demonstrated that the smart card data enriched by the proposed approach dramatically improved the performance of a conventional method for identifying users' home and work places (with 88% improvement on home detection and 35% improvement on work place detection).

The proposed method offers the possibility to mine individual mobility from common public transit transactions, and showcases how uncertain data can be leveraged with domain knowledge and constraints, to support cross-application data mining tasks.

I. Introduction

Many data mining tasks benefit from cross-application datasets. Such cases often follow a simple paradigm as illustrated in Figure 1: The source application generates enormous data, which intend to serve its own needs, but might also be significantly valuable to another target application, where data are limited or not easy to obtain. To name a few, taxi trajectories collected for security management can be leveraged to probe traffic flows [1]. Yet users' search queries can be employed to accurately detect pandemic influenza trends [2].

Mining smart card transactions gives another example that fall in this scope. Smart cards (such as credit cards, fuel cards[1], campus cards, and public transit cards) facilitate millions of people for digital payment and public transport ticketing in many metropolises. Examples include London's Oyster Card[2], San Fransisco's Clipper Card[3], and Beijing's BMAC Card[4]. Overwhelming amounts of transaction data are cumulated in the fare systems every day. Such data are of particular interest to urban planners and location-based service providers, since it reveals urban dynamics and human mobility patterns. Several attempts have been made in mining smart card transactions, and show promising prospects in various applications such as mobility modeling [3] and personalized recommendations [4, 5].

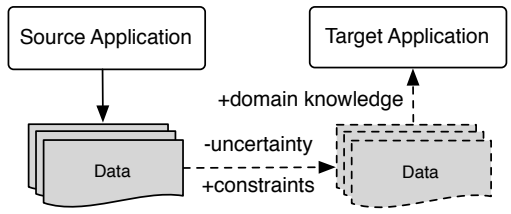


Fig. 1. The paradigm of mining cross-application data

However, most existing approaches in mining transactions of public transit smart cards suffer from the data uncertainty and incompleteness problems. This is also a challenge to a broad range of compelling applications dealing with cross-application datasets: Data generated from the source application often lacks information that is necessary to the target application, e.g., the public transit transactions sometimes do not include the information of trip destinations (the fare does not depend on the destinations thus there is no intention to record such data [5]), but knowing both the origin and destination of a trip is crucial for mining mobility patterns. As a consequence, a considerable amount of work either excludes these uncertain bus trips [4] or focuses on mining aggregated level instead of individual level patterns for uncertain bus trips [3]. A few existing methods have been proposed to recover public transit trips [6, 7, 8], but most of them assume that at least the origin or the destination is given for each trip, which is sometimes not the case.

To address this challenge, this paper provides a systematic solution to reconstruct fine-grained mobility history at individual level from common smart card transactions, which exemplifies how the data coming from the source application can be enriched in terms of granularity and availability to facilitate the target application. Typically, the data coming from

[1] https://en.wikipedia.org/wiki/Smart_card
[2] https://oyster.tfl.gov.uk/oyster/entry.do
[3] https://www.clippercard.com
[4] http://www.bmac.com.cn

TABLE I. EXPENSE RECORDS AND CHARGING RECORDS

(a) Expense Records

CardID	Bus	Boarding	Alighting	Time	Expense	Balance
1	$N2$	–	–	2013-03-14 09:02	0.8	12.3
2	$L3$	31	19	2013-03-14 17:45	0.4	32.2
3	$N1$	–	–	2013-03-15 08:45	0.4	10.6
3	$L1$	04	22	2013-03-16 18:20	0.8	49.8

(b) Charging Records

CardID	Time	Amount	Balance
3	2013-03-15 18:05	50.0	50.6
4	2013-03-15 18:05	20.0	21.6
5	2013-03-15 18:07	20.0	20.4
6	2013-03-15 18:08	30.0	40.8

the source application is uncertain yet constrained—-take as an example the smart card transactions–such constraints are not only related to the *geospatial* space (e.g., the distance that a passenger walks when transferring bus lines), and may also be implied in the *monetary* space (e.g., the balance of a card), or the *temporal* space (e.g., the time interval between two transits). As shown in Figure 1, the proposed solution reduces the uncertainty of the data generated from the source application (here, smart card transactions), by incorporating constraints implied in the source application as well as domain knowledge in the target application (mobility mining).

To the best of our knowledge, this is the first solution that can recover individual bus trips from course-grained smart card data, where the information of both the boarding and alighting stops may be unavailable. In summary, this paper mainly offers the following contributions:

- We have derived a space alignment framework that coalesces the monetary, temporal, and geospatial spaces, to segment all the trips and extract domain specific constraints, which significantly reduce the number of candidate bus stops, even without the information of the boarding or alighting stops.
- We have constructed a conditional random field based sequential model to infer the actual alighting and boarding stops for each trip, where the extracted constraints are naturally incorporated, and the known bus stops for some trips are leveraged for training the model in a semi-supervised way.
- We conducted extensive experiments to validate the proposed method with a large scale human labeled dataset as ground truth. The experimental results as well as the demonstrations validate the effectiveness of our method.

II. DATA

This section explicitly describes the smart card transaction dataset we used in this study, which consists of two tables: the expense records and the charging records, as illustrated in Table I(a) and Table I(b) respectively. The dataset covers a population of 701,250 card holders. We note that the smart card is not limited to the payment for bus transit, but can also be used for other types of payments in this city, such as taxis, subways, and shopping. However, the dataset we obtained only covers bus related expense records (but the charging records are fully available).

A. The Expense Records

This table contains in total 22.03 million bus-trip records, during the period from Aug. 2012 to May. 2013. Each trip is shown as a row in Table I(a), containing the following columns:

- **CardID**: the ID of a smart card, where each card has a unique ID, and typically an individual has only one smart card. Note that in this work, all CardIDs are **anonymized** and are not associated with any personally identifiable information or profiles, for protecting users' privacy.
- **Bus**: the line number (encoded by us from the original names) of a bus, where there are two types of bus lines as shown in Table I(a): **non-ladder-fare** lines (beginning with "N") such as $N2$ and $N1$, and **ladder-fare** lines (beginning with "L") such as $L1$ and $L3$. If you take a non-ladder-fare bus, the fare is identical for the whole line regardless of where you get on or get off the bus, thus you are only required to swipe the smart card once you get on the bus. Yet for ladder-fare bus lines, since the fare is calculated according to the distance between the boarding and alighting stops, you have to swipe the card twice: one swipe at the boarding stop, and the other at the alighting stop.
- **Boarding** and **Alighting**: the codes of the boarding and alighting stops. This information is only available for ladder-fare lines, since the fare of the non-ladder-fare lines is fixed as mentioned above, thus the public transport authority does not record the boarding nor alighting stops in the billing system. We note that even for ladder-fare lines, the recorded information is only a code of the bus stop, which identifies how long (in kilometers) the bus stop is apart from the bus stop with the code 0, where the 0-coded stop is unknown to us. That means, the direction of a bus line is not observable, since either the departure stop or the terminal stop of a bus line can be coded with 0.
- **Time**: the exact time that the fee of a bus trip is deducted from the smart card, which also depends on whether it is a ladder-fare line: For non-ladder-fare lines, the recorded time is the moment that you swipe the smart card when you get on the bus at the boarding stop, while for ladder-fare lines, it is the moment before you get off the bus at the alighting stop.
- **Expense**: the expense of a trip. For non-ladder-fare lines, it is a fixed amount, but for ladder-fare lines, it varies according to the distance between the boarding stop and the alighting stop, which can be calculated directly from the boarding column and the alighting column in the table, e.g.,

$$\mathrm{e} = a + b \cdot \max(|\text{boarding} - \text{alighting}| - c, 0), \qquad (1)$$

where e is the expense, and a, b, c are system parameters varied for different bus lines. It follows that if the distance between boarding and alighting stops is less than or equal to c kilometers, you should pay a, otherwise, you should pay additional b for every extra kilometer. In such a way, the whole fare system considering all possible boarding/alighting stops looks like a "ladder" (that's the reason it is called ladder fare).
- **Balance**: the remaining balance of the smart card after a trip.

B. The Charging Records

To maintain the usage of a smart card, people typically recharge it when necessary. The charging record, as exemplified in Table I(b), has a simple schema and is easy to understand. Our dataset contains 5.93 million charging records, each of which includes the columns of **CardID**, **Time** (the time you charge your smart card), **Amount** (how much you charge this time), and **Balance** (how much you have in the smart card after this charging).

C. Road Network

The road network G is a directed graph $G = (V, E)$, where V is a set of nodes, representing the terminal points of road segments, and E is a set of road segments. Each road segment $e \in E$ contains the information of limit (maximum) driving speed. The road network we used contains 148,110 nodes and 196,307 road segments.

D. Data Denoising and Data Labeling

Through a public map API[5], we can search for the geo-coordinates of the bus stops of each bus line in this city, as well as its pricing information (i.e., we obtain the parameters in Eq. (1) for the bus line), by providing the original bus line names in the expense record. Nevertheless, there are still some bus lines which we failed to find pricing information (95 lines) or the bus stop geo-coordinates (488 lines), since sometimes the name is not complete or ambiguous. Fortunately, these bus lines only account for a small part of all the expense records (about 20%) in the transaction data, as shown in the statistics of Table II. Therefore, we removed the records associated with these "unknown" bus lines in our study.

In addition, we conducted a data labeling program recruiting 102 selected participants to label the specified most frequent ladder-fare lines. In this 4-months program (from Dec. 2012 to Mar. 2013), each participant was provided with a free smart card, and her expense of daily bus transit was reimbursed for labeling the data. Specifically, after signing a consent form regarding the privacy and legal issues, each participant was required to manually record her **every trip** paid by the smart card during the 4 months, and label the corresponding expense records (as shown in Table I(a)) in the transaction data, indicating the names of the boarding and alighting stops, as well as the boarding and alighting time.

TABLE II. STATISTICS OF BUS LINES AND EXPENSE RECORDS

line type	#lines	ratio of records
lines without coordinates	95	4.16%
lines without price info	488	16.84%
non-ladder-fare	270	36.62%
labeled ladder-fare	124	26.54%
unlabeled ladder-fare	288	15.85%

As described in Section II-A, knowing the direction of a ladder-fare line is equivalent to knowing the mapping between codes and the real names (thus the locations) of the stops. We term these ladder-fare lines as **labeled lines**. If a trip recorded in the expense record belongs to a labeled line, the alighting/boarding stops are called **labeled stops** of this **labeled trip**. As a result, we find out the directions of 124 most frequent ladder-fare bus lines, as shown in Table II. The labeled lines cover more than 26% of all trips recorded in the expense records, and accounted for more than 62% ladder-fare trips (recall that alighting and boarding stops for non-ladder-fare lines are not even recorded in the raw transaction data). Figure 2 plots all the labeled bus lines (colored blue), where the underlying road networks (colored gray) delineate the urban area of this city. Clearly, the labeled lines cover a majority of the urban area. However, our main challenge is how to leverage these partially labeled trips to recover the rest (and much more) unlabeled trips.

[5] http://api.amap.com

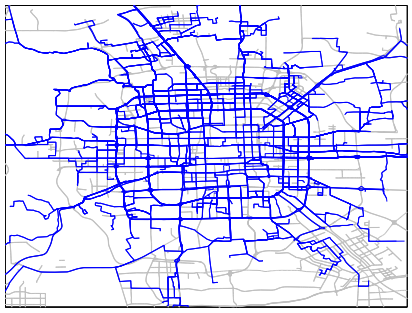


Fig. 2. Labeled ladder-fare bus lines

III. METHODOLOGY

There are three parallel spaces in the transaction data: the **monetary space** $\mathcal{M}$, the **temporal space** $\mathcal{T}$, and the **geospatial space** $\mathcal{S}$.

The balance, charging amount, and expense of a trip are associated with the monetary space, for a given smart card. As shown in the above line of Figure 3, the balance of a user's smart card rises after the user charges the card, and declines after a trip, where the timestamps of expense and charging are points in the temporal space (shown in the middle line of Figure 3). As described in Section II-A, for non-ladder-fare trips, the timestamps reflect the boarding time while for ladder-fare trips, they represent the alighting time.

For a certain trip (the timeslot of each trip is denoted as a colored solid line in the middle of Figure 3), each intermediate point in the temporal space is aligned with a spatial point located in the geospatial space, restricted by the bus line of this trip. In particular, the boarding and alighting timestamps can be mapped to the boarding and alighting stops respectively. Note that the doted lines in the temporal space denote that the time flows but no bus trips are recorded, e.g., when the users stay at home during night or work at office during daytime.

By superimposing the three spaces, the goal of recovering individual bus trips can be generally described as: *To identify the mapping from the temporal space $\mathcal{T}$ to the geospatial space $\mathcal{G}$ for each trip recorded in the expense records $\mathbb{E}$, given the charging records $\mathbb{C}$ in the monetary space $\mathcal{M}$ and a specified CardID.*

A. Preliminary

If the smart cards can only be used for bus trips, then we can continuously track a user's balance without any dis-

Algorithm 1: Segmentation

Input: CardId d, expense records $\mathbb{E}$, and charging records $\mathbb{C}$
Output: Segments $\mathbf{S}$

$\mathbf{I} \leftarrow \{1\}$; /* $\mathbf{I} = \{I_i\}_{i=1}^{|\mathbf{I}|}$ is the index of split points */
$\mathbf{E} \leftarrow$ select * from $\mathbb{E}$ where CardID=d order by time; /* $\mathbf{E} = \{E_i\}_{i=1}^{|\mathbf{E}|}$ */
$\mathbf{C} \leftarrow$ select * from $\mathbb{C}$ where CardID=d order by time; /* $\mathbf{C} = \{C_j\}_{j=1}^{|\mathbf{C}|}$ */
$c_i \leftarrow 0, i = 1, 2, \ldots, |\mathbf{E}|$;
$i \leftarrow 1, j \leftarrow 1$;
while $i \leq |\mathbf{E}| - 1$ **do**
  if $j \leq |\mathbf{C}|$ **and** $t_1 < \xi_j < t_i$ **then**
    $c_i \leftarrow c_i + c(\xi_j)$;
    /* $c(\xi_j)$ can be directly read from C_j */
    $j \leftarrow j + 1$;
  else
    $i \leftarrow i + 1$;
    if $b_i + e_i \neq b_{i-1} + c_{i-1}$ **then** /* b_i and e_i can be directly read from E_i */
      $\mathbf{I} \leftarrow \mathbf{I}.add(i)$;
return $\mathbf{S} = \{S_k\}_{k=1}^{|\mathbf{S}|}$, where $S_k = \{E_{i_k}\}_{i_k=I_k}^{I_{k+1}-1}$

continuous points. However, it is not the case for our data. In fact, the card holders can use their smart cards for other payments, such as taking a taxi or a private car, or even shopping. Though we have the full information of the charging records, the expense records only include bus related expense. Meanwhile, the "dirty" bus lines (without information of price or coordinates) are removed from our data as described in Section II-D. In order to make sure that the consecutive records are really two consecutive bus trips with known price information and coordinates (see Section II-D), we partition an individual's expense records into segments (which is essential for our modeling later), defined as below.

Definition 1 (Segment): Let $R = \{r_1, r_2, \ldots, r_n\}$ denote the expense records for a given smart card, with timestamps $t_1, t_2, \ldots, t_n \in \mathcal{T}$ sorted in the chronological order, and expense $e_1, e_2, \ldots, e_n \in \mathcal{M}$ (as recorded in the expense records). Let $\xi_1, \xi_2 \ldots, \xi_m$ be the timestamps when the smart card is charged with amount $c(\xi_1), c(\xi_2), \ldots, c(\xi_m) \in \mathcal{M}$. Let $b(t)$ be the balance of the smart card at time $t \in \mathcal{T}$. R is called a **segment** if the following condition holds for $1 \leq i \leq n-1$:

$$b_{i+1} + e_{i+1} = b_i + c_i, \tag{2}$$

where b_i is the balance of the smart card right after the ith trip[6], i.e., $b_i = \lim_{t \to t_i^+} b(t)$, and

$$c_i = \sum_{\substack{t_i < \xi_j < t_{i+1} \\ j=1,2,\ldots,m}} c(\xi_j), \tag{3}$$

which denotes the total charges between the ith and the $(i+1)$th trip. ■

Intuitively, a segment is a sequence of expense records where the balance of the card can be continuously tracked without any missing expense records. Note that we do not explicitly segment the records to days as done in many existing approaches [6, 7], instead, a segment can contain many days and nights as long as no missing points are found. Based on Definition 1, we propose an algorithm to perform the segmentation for a certain smart card with CardID d, as presented in Algorithm 1. This algorithm incrementally calculates c_i defined in Eq. (3) for each record of d, and checks whether condition (2) holds in place. Given the expense records $\mathbf{E}$ and charging records $\mathbf{C}$ of a certain card in the chronological order, the segmentation is obtained in $O(|\mathbf{E}| + |\mathbf{C}|)$ time.

[6]We take the right limit here since $b(t)$ is a step function as depicted on the top part of Figure 3.

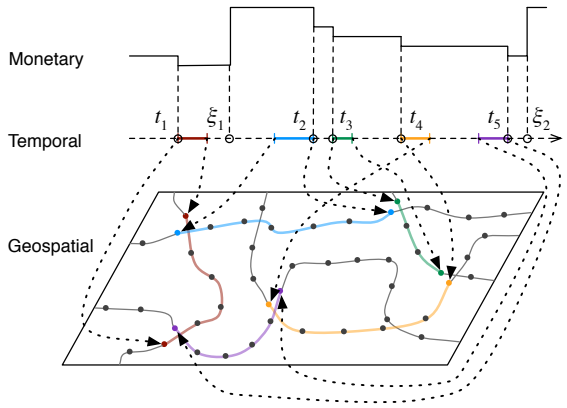


Fig. 3. Space alignment

B. Constraints for Transitions

Recall that in our data, there are non-ladder-fare trips and ladder-fare trips, where the directions are known only to part of the ladder-fare trips (by labeling, as described in Section II-D). Let $S = \{l_1, l_2, \ldots, l_m\}$ denote a segment with m trips in chronological order, where $l_i = (o_i, d_i)$ is the origin (i.e., boarding stop) and destination (i.e., alighting stop) of the ith trip. Assume all the trips in S are not labeled (without any information of the directions), and l_i has n_i bus stops. Assuming the alighting stop is different from the boarding stop for a trip, it follows that in the worst case, there are in total $n_i(n_i - 1)$ possible trips (pairs of boarding-alighting stops) for l_i. Thus we have

$$\prod_{i=1}^{m} n_i(n_i - 1) \tag{4}$$

candidates for S.

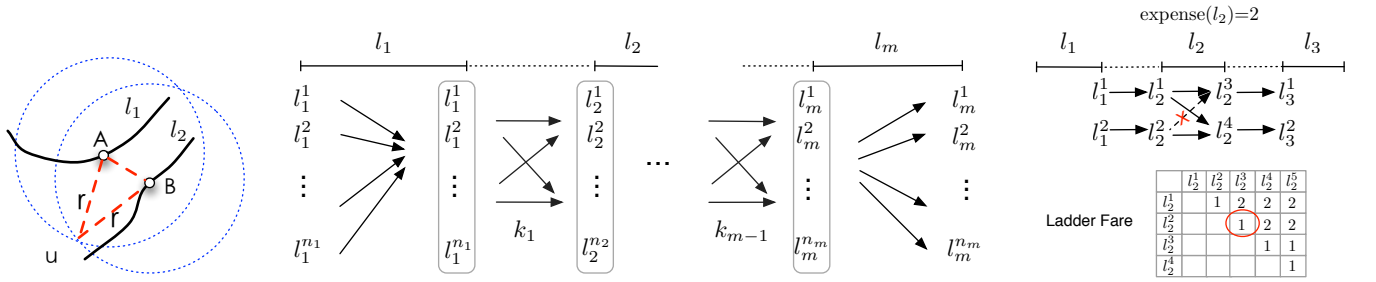
However, by considering several constraints in the monetary space, temporal space, and geospatial space, we can exert several constraints to dramatically reduce the number of candidate trips, even if all the trips are not labeled. In fact, there are two types of transitions (displacements in the geospatial space) in a segment, defined as follows:

Definition 2 (Inner-Transition and Outer-Transition): Given a segment $S = \{l_1, l_2, \ldots, l_m\}$ where l_i is a bus trip from boarding stop o_i to alighting stop d_i, we call each transition $o_i \rightarrow d_i$ an inner-transition, where the movement of a user is strictly restricted by the bus (along the bus line). We call each transition between two consecutive trips, i.e., $d_i \rightarrow o_{i+1}$, an outer-transition. ■

We introduce the constraints for both of the two transitions.

- **Proximity Constraints [for Outer-Transitions].**

Given the limits of walking speed and walking duration, as well as the highly developed transportation systems in metropolises, a citizen's walking scope is usually limited. This is also well supported by results in existing literatures, for example, Bassett Jr et al. [9] reported that nowadays American



(a) proximity constraints (b) reduction of outer-transitions candidates using proximity constraints (c) reduction of inner-transitions using fare constraints

Fig. 4. Proximity constraints and fare constraints

adults walk on average 5119 steps (about 2.5 miles) per day. Another recent study showed that more than 97.6% walking trips of Sydney citizens are less than 2km [10], using 3 years data of Sydney Household Travel Survey with 24,806 respondents.

As shown in Figure 4 (a), l_1 and l_2 are two *consecutive* bus trips of user u. Since u's total walking distance and walking duration are often limited in a day. If u *only walk on feet* during the time between l_1 and l_2, we can assume that a user u's walking scope is bounded in a circle of radius r centered on the alighting bus stop A, and also within a circle of radius r centered on the next boarding stop B. This is because people typically chooses to alight at the closest bus stop to her real destination, and board at the closest bus stop to her origin (of walking). Note that during the time between A and B, u may either stay at some places (e.g., staying at home during night or working in the office), or walk around (e.g, walking along a shopping street), but still, the walking scope is bounded and the above assumption holds.

By the triangle inequality, it is straightforward to show that the distance between A and B is less than $2r$, as illustrated in Figure 4 (a). In other words, if a user only travel on feet during an outer-transition, the distance between the alighting stop of the first trip and the boarding stop of the next trip should be less than $2r$, where the number of such possible outer-transition pairs is denoted as k_1. Using the labeled data, we found that the distance of outer-transitions are less than 3.2km for all the participants. In our experiments, we set $r = 2$km.

Note that k_1 is possible to be 0 for certain consecutive trips in a segment (obtained through Algorithm 1), which indicates that the user takes other vehicles such as private cars or taxis, instead of walking during the intermediate time between l_1 and l_2 (although the segmentation procedure eliminates non-bus outer-transitions that users pay by smart cards, people still can pay for the taxis by cash). We term such outer-transition as a *drifting point*. In this case, we snip the segment at the drifting point to two pieces, and consider each of them as a segment.

As illustrated in Figure 4 (b), without loss of generality, we assume the number of outer-transitions of $l_1 \rightarrow l_2, l_2 \rightarrow l_3, \ldots, l_{m-1} \rightarrow l_m$ to be $k_1, k_2, \ldots, k_{m-1}$ and l_i^j is the jth bus stop of bus line l_i, thus the total number of candidates for S should be at most

$$(n_1 - 1)(n_m - 1) \prod_{i=1}^{m-1} k_i. \tag{5}$$

Note that for any two bus lines l_i and l_{i+1}, k_i is typically much smaller than $n_i(n_i - 1)$. Even if l_i and $l_i + 1$ are the same bus line (with different directions), k_i should be less than $n_i r / L_i$, where L_i is the total length of the bus line that trip l_i belongs to, given the bus stops are uniformly distributed along the bus line. In most cases, k_i equals to the number of "transfer stops" between l_i and l_{i+1} for $i = 1, 2, \ldots, m-1$, which makes the candidates significantly less than Eq. (4).

In practice, we indexed the bus lines using an R-tree, and calculated the k_i pairs for the outer-transition between l_i and l_{i+1} using a range query. After each calculation, we stored the results in a hash table with key (l_i, l_{i+1}), thus next time we can directly retrieve the result from the hash table if we encounter this outer-transition later in the expense records.

• Fare Constraints [for Inner-Transitions].

Since we have crawled the fare system (price information) of all the trips that we deal with, we can further reduce the number of inner-transitions by considering the actual expense deducted from the smart card for each trip, after exerting the proximity transition. For example, if the expense calculated by Eq. (1) (using parameters obtained by the method described at Section II-D) for a candidate inner-transition (o_i, d_i) is larger than the actual expense recorded, we prune this candidate. Here, the distance between two bus stops is calculated in advance using the road network distance for all the bus lines (actually, for each stop, we only need to calculate the distance to the first stop of this bus line, and the other mutual distances are easy to obtain) to avoid replicated computation.

Note that the fare constraints not only work for the first and last inner-transition of a segment as shown in Figure 4 (b), but also help reduce the inner-transitions in between of two outer-transitions. For example, as illustrated in Figure 4 (c), suppose (l_1^1, l_2^1), (l_1^2, l_2^2) are 2 candidate outer-transitions from trip l_1 to trip l_2, and (l_2^3, l_3^1), (l_2^4, l_3^2) are 2 candidate outer-transitions from trip l_2 to trip l_3. Before exerting the fare constraints, there are 4 possible inner transitions in l_2: $l_2^1 \rightarrow l_2^3$, $l_2^1 \rightarrow l_2^4$, $l_2^2 \rightarrow l_2^3$, and $l_2^2 \rightarrow l_2^4$, however, by checking the price information shown in the ladder fare table, we find that the fare from l_2 to l_3 is less than the actual expense recorded. Therefore, $l_2^2 \rightarrow l_2^3$ is not a possible inner-transition for l_2, thus the total number of candidate transitions from the alighting stop of l_1 to the boarding stop of l_3 is decreased from 4 to 3.

• Temporal Constraints [for Inner and Outer Transitions].

Although we have only one timestamp for a trip, we can leverage it to further weed out unreasonable candidates. Let

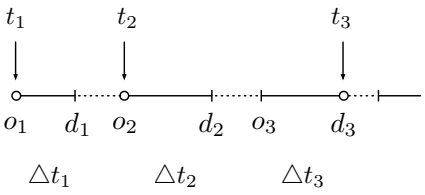


Fig. 5. Temporal constraints

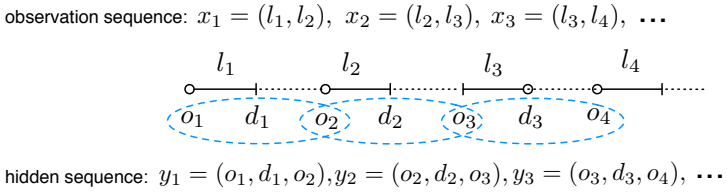


Fig. 6. Constructed observation sequence and hidden sequence in the linear-chain CRF

t_1, t_2, t_3 be the timestamps of three trips $l_i = (o_i, d_i)$, for $i = 1, 2, 3$. Recall that for ladder-fare trips, the timestamps are the alighting time, and for non-ladder-fare trips, the timestamps are boarding time. As shown in Figure 5, l_1 and l_2 are non-ladder-fare trips, and l_3 is a ladder-fare trip. The minimum travel time between o_i and d_i, denoted as $\triangle t_i$ can be calculated using the road network, where the travel speed is substituted with the limit driving speed (refer to Section II-C). Consequently, the following conditions should hold for this example:

$$\begin{cases} \triangle t_1 \leq t_2 - t_1, & (6) \\ \triangle t_2 + \triangle t_3 \leq t_3 - t_2. & (7) \\ \ldots \end{cases}$$

Thus the candidates which violate the above conditions are removed.

Similar to the fare constraints, temporal constraints take effect with both the beginning (ending) inner-transitions and the inner-transitions in between of two outer-transitions. To avoid duplicate calculations, we also pre-compute all the minimum travel time between a given bus stop to the 0-coded bus stop of each bus line (thus we have the minimum travel time between any pair of bus stops), and store the results using a hash table.

Note that the temporal constraint in Eq. (6) differs from that of Eq. (7) and all previous types of constraints in terms of the number of transitions (either inner-transition or outer-transition) involved. Actually, the proximity constraints are associated with 1-step outer-transition; the fare constraints are associated with 1-step inner-transition, which is also true for temporal constraints with the form of Eq. (6). The constraints with the form as Eq. (7), however, are described using 2-step inner-transitions and 1-step outer-transition. Next, we introduce a unified model to deal with all the above constraints, and incorporate the labeled trips (note that until now we do not rely on any labeled data) to infer the most possible candidate for a segment.

C. Semi-supervised CRF with Constraints

1) Model: Conditional Random Fields (CRFs) [11] have been successfully applied to many sequential labeling applications in data mining and machine learning, where the most widely used one is the linear-chain CRF. A linear-chain CRF is an undirected graphical model, which defines a conditional probability over a hidden label sequence $\mathbf{y} = \{y_1, y_2 \ldots, y_m\}$ conditioned on an observation sequence $\mathbf{x} = \{x_1, x_2, \ldots, x_m\}$, with the form

$$p_\lambda(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\lambda)} \exp\left(\sum_{i=1}^{m-1}\sum_{k=1}^{K} \lambda_k f_k(y_i, y_{i+1}, \mathbf{x})\right), \quad (8)$$

where $\{f_k(\cdot)\}_{k=1}^K$ are real valued feature functions (which are typically binary functions), $\{\lambda_k\}_{k=1}^K$ are parameters, and $Z(\lambda)$ is a normalization function (also called *partition* function)

$$Z(\lambda) = \sum_{\mathbf{y}} \exp\left(\sum_{i=1}^{m-1}\sum_{k=1}^{K} \lambda_k f_k(y_i, y_{i+1}, \mathbf{x})\right). \quad (9)$$

Actually, given the candidates generated for a segment, our problem can be formulated as a sequential labeling problem. Specifically, we construct a linear chain CRF as follows. Given a segment $S = \{l_1, l_2, \ldots, l_m\}$, let $\mathbf{x} = \{x_1, x_2, \ldots, x_{m-1}\}$ be the observation sequence, where $x_i = (l_i, l_{i+1})$ for $i = 1, 2, \ldots, m-1$. That is, the outer-transition between consecutive lines is regarded as a node in the CRF chain (note that here each node is a pair of trips, as shown in Figure 6). Later, let $y_i = (y_i^1, y_i^2, y_i^3)$ denote the triple (o_i, d_i, o_{i+1}) for $i = 1, 2, \ldots, m-1$, which is an inner-transition coalesced with an outer-transition (we will crystallize the reason for this later). The sequence $\mathbf{y} = \{y_1, y_2, \ldots, y_{m-1}\}$ is thus the label sequence.

With fully labeled sequences, CRF is typically trained by maximizing the penalized conditional log-likelihood on the training sequences $\mathcal{D}$ with length N

$$L(\lambda, \mathcal{D}) = \sum_{i=1}^{N} \log p(\mathbf{y}^{(i)}|\mathbf{x}^{(i)}) - \frac{\sum_k \lambda_k^2}{2\sigma^2}, \quad (10)$$

which can be optimized using the gradient-based method or Expectation Maximization (EM) [12]. Here, to avoid overfitting, we include a Gaussian prior with zero-mean and variance σ^2=10.

However, in our dataset, a significant number of segments are partially labeled (for certain bus lines), a semi-supervised training approach that can take the full use of available labels is more preferred. More importantly, rich prior knowledge and constraints we derived are not thoroughly leveraged. Therefore, we employ the Generalized Expectation Criterion [13] as an objective function, which enables semi-supervised CRF training with constraints as side information. Given a real-valued constraint function $G(\mathbf{y}, \mathbf{x})$ and unlabeled data $\mathcal{U}$, the Generalized Expectation Criterion is given by

$$O(\lambda, \mathcal{D}, \mathcal{U}) = L(\lambda, \mathcal{D}) - S(E_{\tilde{p}(\mathbf{x})}\left[E_{p_\lambda(\mathbf{y}|\mathbf{x})}\left[G(\mathbf{y}, \mathbf{x})\right]\right]), \quad (11)$$

where $\tilde{p}(\mathbf{x})$ is the empirical distribution over unlabeled data $\mathcal{U}$, $E[\cdot]$ stands for the expectation, and S is a score function[7] expressing the distance between the model expectation and

[7]We employed the square distance as the score function in our implementation, following [14].

a targeted expectation. The optimization of Eq. (11) can be performed using gradient-based methods [14].

2) Features: Compared with other sequential models such as Hidden Markov Model (HMM), CRF is more flexible in incorporating features, e.g., the transition from one label to another can also depend on the whole observation sequence. As such, designing features is the most critical part for applying CRF to various applications.

We use both uni-gram features and bi-gram features. For each uni-gram label y_i, the features we used for training include:

- the indicator function of l_i and x_i;
- timestamp t_i, which is discretized to the following time slots: 12am-6am, 6am-9am, 9am-12pm, 12pm-5pm, 5pm-8pm, 8pm-12am (considering typical commute patterns in the studied city);
- the trip type (non-ladder or ladder);
- time interval $\triangle t = t_{i+1} - t_i$, which is discretized to hours;
- expense of trip l_i, which is integer multiples of the unit price.

The bi-gram features include:

- the indicator functions of (y_i, y_{i+1}), (x_i, y_{i+1}), and (x_i, y_i, y_{i+1});
- time interval $\triangle t_{i+2} - t_i$, which is discretized to hours;
- whether $y_i^3 = y_{i+1}^1$;
- whether l_i and l_{i+2} are the same bus line.

3) Constraints: Our constraints are categorized into two types: one-label constraints and two-label constraints. one-label constraints, which restrict the candidates of y_i, are associated with both inner-transitions and outer-transitions, such as proximity constraints, fare constraints and temporal constraints with the form of Eq. 6. One-label constraints eliminate certain states given a specified observation, e.g., outer-transitions with a distance larger than $2r$. Two-label constraints include the temporal constraint with the form of Eq. (7), and

$$y_i^3 = y_{i+1}^1, \forall i = 1, 2 \ldots, m-1, \tag{12}$$

where $y_i = (y_i^1, y_i^2, y_i^3)$. This is to ensure the chain is connected as shown in Figure 6. For one-label constraints, we assign a high probability to the labels which satisfy the one-label constraints, following the method described in [15]; and for two-label constraints, a probability transition matrix [14] is built for calculating the target expectation, based on the Kirchoff matrix (refer to [14] for details) and whether the two-label constraints are satisfied.

As a result, all the constraints we derived before are incorporated into this framework succinctly and consistently, which is the reason that we model a triple as a node in a linear chain CRF. Regarding each boarding or alighting stop of a bus trip as a hidden state could yet be an alternative way to model a segment, which forms a high-order CRF, however, additional computation cost is exponential to the order of the CRF [16] (in our scenario, the order should be 3 due to the distinct properties of inner and outer transitions). On the contrary, by connecting the inner and outer transitions with a triple, we can naturally restrict the candidates to a relatively small set and thus considerably accelerate the inference. In addition, our model does not need to separately tackle different higher-order constraints and features with various forms, which might clutter the model, yet some intrinsic prior knowledge and strict conditions, such as fare constraints between inner-transitions, are naturally leveraged to pre-exclude irrelevant labels and redundant features.

IV. Evaluation

A. Settings

The dataset we used is described in Section II. Here we introduce 1) which baseline methods were compared with, and 2) how we evaluated these methods.

1) Baselines: We compared our method (semi-supervised CRF with constraints generated using space alignment, shortened as "CRF+C"), with the following baselines.

- CRF without constraints ("CRF" for short). This algorithm uses the same setting as CRF+C, except that it does not incorporate constraints. This is for evaluating whether the constraints are useful to detect the bus stops in a trip.
- Trip-Chaining with maximum frequency ("TC+MF" for short). The Trip-Chaining (TC) algorithm is adopted by most existing approaches [6, 8, 17, 18] for inferring origin-destination pairs. TC is based on several explicit assumptions such as the proximity between consecutive trips, and "the first trip of a day starts from the alighting station of last night" [6]. Since TC requires at least one stop is known for all trips, in case TC fails to find the stop of some trip in a segment, we assigned it with the most frequent label in the labeled test data.
- Trip-Chaining with maximum similarity ("TC+MS" for short), which is a state-of-the-art variation of the trip-chaining method [7]. A major difference between TC+MS and TC is that TC+MS assigns similar destinations (origins) to trips that have similar origins (destinations) when other rules in TC fail.

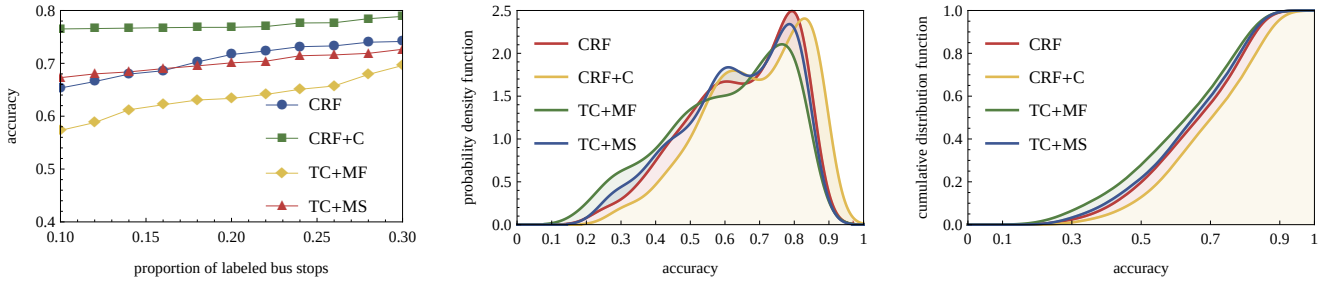
2) Criteria: We measured the performance of each algorithm using accuracy, calculated by

$$\text{Accuracy} = \frac{\text{correctedly identified unlabeled bus stops}}{\text{unlabeled bus stops}}. \tag{13}$$

For each individual, we calculate the accuracy after running a test for each method. The overall accuracy is calculated by an average of 10-fold cross-validation.

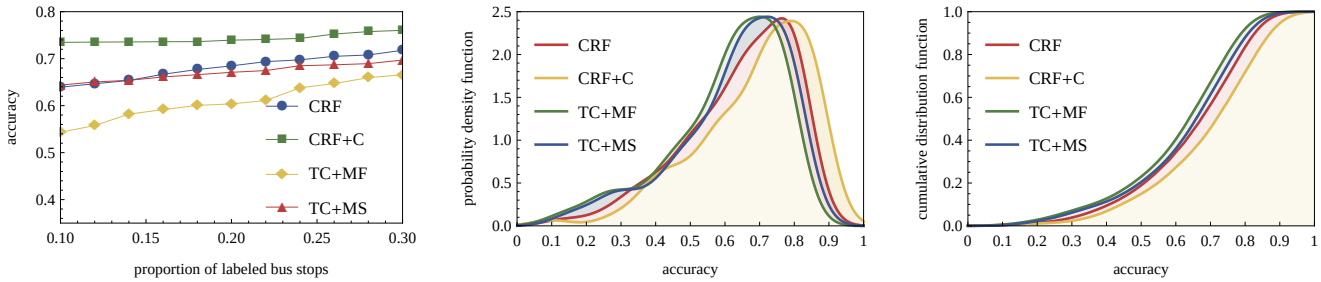
B. Evaluation on All Card Holders' Data

We first evaluated our method on all card holders' data using labeled trips as the testing set (otherwise, the ground truth is not known). Specifically, we first selected fully-labeled segments (of which we removed 8.5% segments with length less than 3) after performing Algorithm 1. In order to reveal the performance of these methods on both labeled and unlabeled trips, we randomly removed labels for 70%–90% trips, which resulted in the remaining 10%–30% labeled trips to fit the same scale of labels as the whole dataset (note that in the entire dataset 26.54% trips are labeled). Next, we further randomly removed a bus stop (either boarding or alighting) for each trip, so as to compare our methods against TC-based approaches (where they require at least one bus stop is known for each trip). Then, we conducted 10-fold cross-validation to calculate the accuracy of each method.



(a) overall accuracy (b) pdf (c) cdf

Fig. 7. Overall accuracy, probability density function, and cumulative distribution function of all users' results



(a) overall accuracy (b) pdf (c) cdf

Fig. 8. Overall accuracy, probability density function, and cumulative distribution function of participants' results

Figure 7(a) plots the overall accuracy, where the x-axis is the proportion of labeled bus stops. It is clear that our method significantly outperforms the competitors. For example, even with only 10% labeled bus stops, CRF+C achieves a high accuracy at 0.78, while the performance of other methods, especially the TC+MF method, rely much more on the labeled data. Note that the result of TC+MS is in good agreement with the reported (66%) performance mentioned in [7].

Next, we investigated the distribution of accuracy among all users. Figure 7(b) and Figure 7(c) respectively show the probability distribution function (pdf) (fitted from the histogram) and the cumulative distribution function (cdf) of the accuracy among all users. The results validate the advantage of our method, e.g., Figure 7(b) shows that the accuracy of our method still has a high probability density within the interval $[0.8, 0.9]$.

C. Evaluation on Completely Labeled Participants' Data

As mentioned in Section II-D, the 102 participants manually labeled all their bus trips, including non-ladder-fare and ladder-fare trips. Basic demographic information of their age and gender is presented in Table III. Thus we have both non-ladder-fare trips and ladder-fare trips in the testing set, which is exactly the situation in the real data. Similarly, as the above experiment, we constantly fed 10%-30% (randomly chosen) trips with the boarding or alighting stops as labels, then compare all methods using 10-fold cross validation.

As shown in Figure 8, the accuracy, pdf and cdf exhibit consistent trend with the previous results in Figure 7. We found that the accuracy for all the methods are actually a little lower than the previous experiment, however, our method still shows clear advantage compared with other methods. In particular, the overall accuracy of our method still exceeds 0.75 when we have 25% labeled stops.

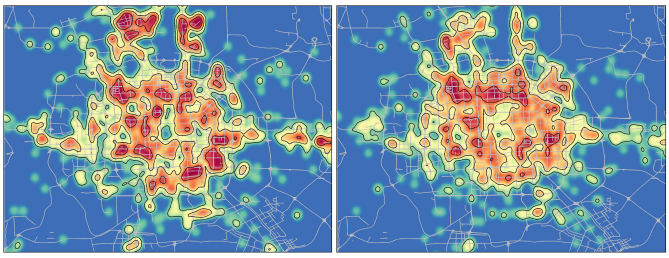
TABLE III. DEMOGRAPHICS OF THE PARTICIPANTS

gender		age			
male	female	19-24	25-30	31-36	37-47
57.6%	42.4%	39.4%	45.5%	10.6%	4.5%

D. Detection of Important Places

As a demonstration, we show how the enriched smart card transactions can be utilized to mine important locations such as home and work places of users. As shown in Figure 7, the overall accuracy of our method is higher than 0.78 given the 25% labeled stops. Hence, we applied the proposed method to the entire dataset, and reconstructed mobility history for each individual. We employed the clustering-based method proposed in [19] to identify home and work places. Later, we performed a 2D Kernel Density Estimation (KDE) given the detected home and work places, as shown in Figure 9(a) and Figure 9(b) respectively. The identified hot spots for both home and work places coincide well with the local household surveys. For the 102 participants, we compared the identified home and work places with the real ones provided by themselves, where we successfully identified 96 home places and 92 work places[8]. However, if we directly use the smart card data without applying our space alignment approach, only 51 home places and 68 work places are successfully identified, i.e., the space alignment approach increases the performance by 88% for home identification and 35% for work place

[8]According to our privacy agreement with the participants, we cannot show the density distribution of their home and work places (as Figure 9) here.



(a) home (b) work

Fig. 9. Identified distribution of home and working places of all card users using 2D Kernel Density Estimation

identification. The reason behind is that many bus stops that are close to the real home or work places, are actually derived from the non-ladder-fare trips, which are implicitly learned by the proposed method.

V. Related Work

A. Human Mobility Analytics

Human mobility is attracting more researchers' attentions thanks to the increasing availability of mobility data. Recent studies converge to suggest that human mobility is highly regular, predictable, and unique. Ground-breaking work on studying human mobility patterns from large scale mobile phone traces is established in [20]. In a series of work, they reported that human mobility patterns show high degree of spatial and temporal regularity, and each individual has a significant probability to return to a few highly frequented locations. Based on a 3-month mobile records captured from 50,000 individuals, Song et al. [21] suggested that human mobility has a predictability of 93%. More recently, de Montjoye et al. [22] mathematically formulated the uniqueness of mobility, and showed that four spatial-temporal points are enough to uniquely identify 95% of the individuals, based on an investigation of a 15 months mobility data covering 1.5 million individuals.

Mining human mobility data has also enabled a variety of emerging applications. For example, Yuan et al. [1] introduced a driving direction system with the intelligence mined from local taxi drivers. Ge et al. [23] presented a recommendation system in order to maximize the profit of a taxi driver, based on taxi trajectories. Hoh et al. [24] designed a time-to-confusion metric and a cloaking algorithm to help users avoid privacy risks based on vehicle GPS trajectories. Meanwhile, mobility data have been utilized for studying several research topics in social science such as friendships and social ties [25]. For example, Cranshaw et al. [26] showed that human mobility patterns have strong connections with the structure of their underlying social network.

In this paper, motived by the above work, we restrict ourselves to the problem of recovering human mobility data from transaction data associated with bus trips. Compared with other kinds of human mobility data such as mobile phone traces or check-ins, public transit data characteristically reveal individual's daily transits between important locations such as home and work places, which may complement existing approaches and findings founded on other types of mobility data. Additionally, the methods provided in this paper might help identify new opportunities in human mobility analytics dealing with cross-application data.

B. Mining Smart Card Transactions

Smart cards and integrated ticketing are supported by public transit operators in many cities, which provides convenience to both citizens and governments for public transit ticketing. The overwhelming usage of smart card makes the transaction data invaluable resources for understanding urban commute patterns and human dynamics.

In transportation research area, numerous studies have attempted to mine users' travel behaviors from smart card transactions [27], For example, Utsunomiya et al. [28] reported several findings on walking access distance, frequency and consistency of daily travel patterns, and variability of smart card customer behaviors by residential area, based on smart card transaction data in combine with card holders' personal information, and proposed to improve user trust in transit service, and adjust fare according to users' needs. Recently, [29] investigated the crowdedness of London Underground by mining the spatial-temporal patterns from the Oyster Card Data. Their results indicate that the crowdedness is highly regular and predicable, and suggest users slightly adjust their travel time to avoid congestion peak. Pelletier et al. [30] provided a comprehensive review of recent literatures on the usage of smart card data in public transit planning.

Existing work on mining smart card transactions, especially bus trip transactions, often encounters the problem of data incompleteness. This is because most Automatic Fare Collection (AFC) systems record the bus trip boarding location coarsely at the bus-route level (without the information of specific boarding and alighting stops). For example, as mentioned in [5] and [8], the London Oyster Card data only contain the information of origin and start time for bus trips, since the pricing of a bus trip does not depend on the destination (but for rail/tube trips, the destination is also recorded). Similarly, in the dataset used by [3] to mine collective mobility patterns, the bus trips only have information of boarding time and travel fare.

Several approaches have been proposed to infer the boarding stops [6] or alighting stops [7] of a bus trip. Nevertheless, these approaches still require that at least one location (either the boarding or alighting stop) is available. For example, Trépanier et al. [7] addressed the problem of inferring trip destinations with smart card transactions where the boarding stop is recorded. Most of these approaches employ the Trip-Chaining method or its variations[6, 7, 8, 17, 18], which is based on several assumptions, such as the users return to the first boarding station at the end of a day [7]. Cui [6] suggested that side information such as the Automatic Vehicle Location (AVL) data could be leveraged to infer the origin and destinations when location information is not available in the smart card transactions.

Our work is different from the above methods in the following aspects: First, we provided a space-alignment framework to coalesce the information in the monetary space (rarely considered before), temporal space (with a historical view instead of separated days adopted by many existing approaches), and geospatial space, which is flexible enough to be applied to

different datasets with various types of missing information, e.g., even for trips that neither alighting nor boarding stops are available, our approach can still infer the origins and destinations with a high accuracy (75%), which improves the state-of-the-art with 10% when labels are rare (less than 25%). Second, instead of using hard-coded inference rules and assumptions, we employed a probabilistic model, which naturally incorporates domain constraints, and inherits the advantage of statistical modeling to achieve a global optimization. Finally, due to the lack of large-scale ground truth data for testing the accuracy of a model, few existing approaches have evaluated the rate of correctly inferred bus stops. In contrast, we directly validated the proposed method using a large scale human-labeled data, where every trip that appeared in the transactions during the 4 months is labeled by the participant herself.

Nevertheless, our method is motivated by existing approaches, and we believe the proposed method as well as the reconstructed data would be beneficial for urban planners, transportation engineers, and researchers in related fields.

VI. Conclusion

We have provided a systematic way to recover individual mobility history from urban scale smart card transactions. By aligning data in different dimensions, we formulated several underlying constraints from the transaction data, and incorporated these constraints into a semi-supervised probabilistic model. Extensive experiments validated that the proposed method has a considerably high accuracy given very limited number of known alighting or boarding stops.

Although the work reported in this paper is based on a public transit transaction dataset, we believe the proposed space alignment framework can be easily adapted to other location-related transaction data, and may also provide implications to data miners who deal with cross-application datasets.

References

[1] J. Yuan, Y. Zheng, X. Xie, and G. Sun, "T-drive: Enhancing driving directions with taxi drivers' intelligence," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 1, pp. 220–232, 2013.

[2] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant, "Detecting influenza epidemics using search engine query data," *Nature*, vol. 457, no. 7232, pp. 1012–1014, 2008.

[3] L. Liu, A. Hou, A. Biderman, C. Ratti, and J. Chen, "Understanding individual and collective mobility patterns from smart card records: A case study in shenzhen," in *Intelligent Transportation Systems, 2009. ITSC'09*. IEEE, 2009, pp. 1–6.

[4] N. Lathia, J. Froehlich, and L. Capra, "Mining public transport usage for personalised intelligent transport systems," in *Data Mining (ICDM), 2010 IEEE 10th International Conference on*. IEEE, 2010, pp. 887–892.

[5] N. Lathia and L. Capra, "Mining mobility data to minimise travellers' spending on public transport," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2011, pp. 1181–1189.

[6] A. Cui, "Bus passenger origin-destination matrix estimation using automated data collection systems," Master's thesis, Massachusetts Institute of Technology, 2006.

[7] M. Trépanier, N. Tranchant, and R. Chapleau, "Individual trip destination estimation in a transit smart card automated fare collection system," *Journal of Intelligent Transportation Systems*, vol. 11, no. 1, pp. 1–14, 2007.

[8] W. Wang, J. P. Attanucci, and N. H. Wilson, "Bus passenger origin-destination estimation and related analyses," *Journal of Public Transportation*, 2011.

[9] D. R. Bassett Jr, H. R. Wyatt, H. Thompson, J. C. Peters, and J. O. Hill, "Pedometer-measured physical activity and health behaviors in united states adults," *Medicine and science in sports and exercise*, vol. 42, no. 10, p. 1819, 2010.

[10] R. Daniels and C. Mulley, "Explaining walking distance to public transport: the dominance of public transport supply," *World*, vol. 28, p. 30, 2011.

[11] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proceedings of the Eighteenth International Conference on Machine Learning*, ser. ICML '01, 2001, pp. 282–289.

[12] C. Sutton and A. McCallum, "An introduction to conditional random fields for relational learning," *Introduction to statistical relational learning*, vol. 93, pp. 142–146, 2007.

[13] G. S. Mann and A. McCallum, "Generalized expectation criteria for semi-supervised learning with weakly labeled data," *The Journal of Machine Learning Research*, vol. 11, pp. 955–984, 2010.

[14] G. Druck, G. Mann, and A. McCallum, "Semi-supervised learning of dependency parsers using generalized expectation criteria," in *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*, 2009, pp. 360–368.

[15] G. Mann and A. McCallum, "Generalized expectation criteria for semi-supervised learning of conditional random fields," in *Proc. ACL*, 2008, pp. 870–878.

[16] S. Sarawagi and W. W. Cohen, "Semi-markov conditional random fields for information extraction," *Advances in Neural Information Processing Systems*, vol. 17, pp. 1185–1192, 2004.

[17] J. Zhao, A. Rahbee, and N. H. Wilson, "Estimating a rail passenger trip origin-destination matrix using automatic data collection systems," *Computer-Aided Civil and Infrastructure Engineering*, vol. 22, no. 5, pp. 376–387, 2007.

[18] J. J. Barry, R. Freimer, and H. Slavin, "Use of entry-only automatic fare collection data to estimate linked transit trips in new york city," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2112, no. 1, pp. 53–61, 2009.

[19] S. Isaacman, R. Becker, R. Cáceres, S. Kobourov, M. Martonosi, J. Rowland, and A. Varshavsky, "Identifying important places in peoples lives from cellular network data," in *Pervasive Computing*, 2011, pp. 133–151.

[20] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi, "Understanding individual human mobility patterns," *Nature*, vol. 453, no. 7196, pp. 779–782, 2008.

[21] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási, "Limits of predictability in human mobility," *Science*, vol. 327, no. 5968, pp. 1018–1021, 2010.

[22] Y.-A. de Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel, "Unique in the crowd: The privacy bounds of human mobility," *Scientific reports*, vol. 3, 2013.

[23] Y. Ge, H. Xiong, A. Tuzhilin, K. Xiao, M. Gruteser, and M. Pazzani, "An energy-efficient mobile recommender system," in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2010, pp. 899–908.

[24] B. Hoh, M. Gruteser, H. Xiong, and A. Alrabady, "Achieving guaranteed anonymity in gps traces via uncertainty-aware path cloaking," *IEEE Transactions on Mobile Computing*, vol. 9, no. 8, pp. 1089–1107, 2010.

[25] D. Wang, D. Pedreschi, C. Song, F. Giannotti, and A.-L. Barabasi, "Human mobility, social ties, and link prediction," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2011, pp. 1100–1108.

[26] J. Cranshaw, E. Toch, J. Hong, A. Kittur, and N. Sadeh, "Bridging the gap between physical location and online social networks," in *Ubicomp*, 2010, pp. 119–128.

[27] B. Agard, C. Morency, and M. Trépanier, "Mining public transport user behaviour from smart card data," in *12th IFAC Symposium on Information Control Problems in Manufacturing-INCOM*, 2006, pp. 17–19.

[28] M. Utsunomiya, J. Attanucci, and N. Wilson, "Potential uses of transit smart card registration and transaction data to improve transit planning," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1971, no. 1, pp. 119–126, 2006.

[29] I. Ceapa, C. Smith, and L. Capra, "Avoiding the crowds: understanding tube station congestion patterns from trip data," in *Proceedings of the ACM SIGKDD International Workshop on Urban Computing*, 2012, pp. 134–141.

[30] M.-P. Pelletier, M. Trépanier, and C. Morency, "Smart card data use in public transit: A literature review," *Transportation Research Part C: Emerging Technologies*, vol. 19, no. 4, pp. 557–568, 2011.