Estimating Latent People Flow without Tracking Individuals

Yusuke Tanaka[1], **Tomoharu Iwata**[2], **Takeshi Kurashima**[1], **Hiroyuki Toda**[1], **Naonori Ueda**[2,3]
[1] NTT Service Evolution Laboratories, Kanagawa 239-0847, Japan
[2] NTT Communication Science Laboratories, Kyoto 619-0237, Japan
[3] RIKEN Center for AIP, Tokyo 103-0027, Japan
{tanaka.y, iwata.tomoharu, kurashima.takeshi, toda.hiroyuki, ueda.naonori}@lab.ntt.co.jp

Abstract

Analyzing people flows is important for better navigation and location-based advertising. Since the location information of people is often aggregated for protecting privacy, it is not straightforward to estimate transition populations between locations from aggregated data. Here, aggregated data are incoming and outgoing people counts at each location; they do not contain tracking information of individuals. This paper proposes a probabilistic model for estimating unobserved transition populations between locations from only aggregated data. With the proposed model, temporal dynamics of people flows are assumed to be probabilistic diffusion processes over a network, where nodes are locations and edges are paths between locations. By maximizing the likelihood with flow conservation constraints that incorporate travel duration distributions between locations, our model can robustly estimate transition populations between locations. The statistically significant improvement of our model is demonstrated using real-world datasets of pedestrian data in exhibition halls, bike trip data and taxi trip data in New York City.

1 Introduction

With recent advances in wireless and mobile networks, location information of people can be recorded in a variety of spaces such as exhibition halls, shopping malls, amusement parks, and urban cities. It is important to understand people's mobility patterns in these spaces, because it provides useful knowledge for optimizing navigation systems [Huang and Gartner, 2010], travel route recommendation [Kurashima *et al.*, 2010], location-based mobile advertising [Dhar and Varshney, 2011], urban planning [Yuan *et al.*, 2012], and disaster management [Song *et al.*, 2014]. For example, analyzing the routes of people helps to guide them to the locations of interest, and determine which advertisement to serve to the visitors to suit their current locations.

Many statistical methods that analyze trajectory data have been proposed to understand people's mobility patterns [Giannotti *et al.*, 2007; Monreale *et al.*, 2009; Zhuang *et al.*, 2017]. Trajectory data is a set of location points with time stamps for each individual (e.g., pedestrian, car, bike). However, since privacy concerns are increasing, location information of people is often aggregated, making the tracking of individuals impossible [Chow and Mokbel, 2011]. Aggregated data can also be obtained directly from people counting systems that use various sensors such as video cameras [Chan *et al.*, 2008] and inductive-loop traffic detectors [Klein, 2001]. These systems have been recently developed for low-cost and large-scale people count data collection. Giving the preference for data aggregation, existing methods that analyze trajectory data are inapplicable. Therefore, estimating latent people flows, i.e., transition populations between locations, from only aggregated data is of critical importance in understanding people's mobility patterns while protecting privacy.

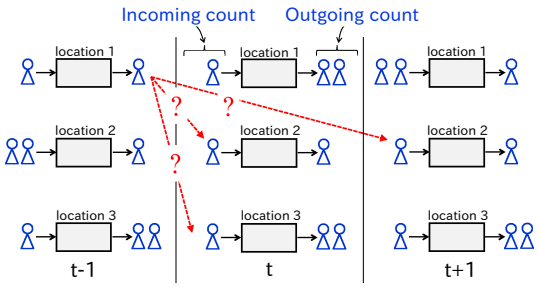


Figure 1: Aggregated data targeted. Observes only incoming and outgoing counts at each location and each time step (blue). The next location of each individual is unobserved (red).

The aggregated data considered consist of incoming and outgoing counts at each location and each time step as shown in Figure 1. For example, in an exhibition hall, we may be able to obtain only counts of pedestrians entering or leaving each event booth, and we cannot know the history of locations visited by each pedestrian. Other examples include pedestrian data for each attraction in an amusement park [Du *et al.*, 2014], each store in a shopping mall [Senior *et al.*, 2007], and traffic data for intersections [Kumar *et al.*, 2013].

The collective flow diffusion model (CFDM) was proposed to estimate transition populations between locations from ag-

gregated data [Kumar *et al.*, 2013]. CFDM is learned by maximizing a likelihood subject to constraints that represent people flow conservation. The constraint states that all people who leave a location at time step t always arrive at another location at the next time step, $t+1$. However, this assumption is too restrictive for many real-world applications, mainly because we do not always have a large enough number of sensor devices to cover large-scale spaces: People might be moving through passages where no sensor devices are placed at an observation time step. In that case, the people are not observed in any location, and CFDM fails to estimate transition populations accurately as it assumes that everyone is in one or other of the observed locations at the next time step.

The purpose of this paper is to robustly estimate unobserved transition populations between locations in practical situations, that is, the observation range is limited and some people are not observed in any location in some time periods. In order to achieve our purpose, we propose a new probabilistic model that incorporates people's travel duration between locations. This is, however, quite challenging to implement, because travel duration between locations is not present in the aggregated data, and, moreover, people are heterogeneous so travel duration depends on the individual. For example, in an exhibition hall, some people might directly visit their next location, but others might take a rest before arriving at another location; in an urban city, moving speeds might depend on the means of transport (e.g., walking, bike, car).

With the proposed model, the temporal dynamics of people flows are assumed to be probabilistic diffusion processes over a network, where the nodes are locations (e.g., booths and shops) and the edges are paths between these locations (e.g., passages and roads). People diffuse from node to node according to transition probabilities that depend on their locations. Since the transition populations between locations are not observed, we treat them as latent variables. Our key idea is to treat travel duration as random variable that follows a probability distribution. This enables us to capture the heterogeneity in travel duration among individuals. The distributions of travel duration are incorporated into the constraints that represent flow conservation; they aim to model that people who left one location in one time step and arrived at another location after some delay. We develop an efficient inference algorithm to estimate the transition probabilities, the transition populations, and the travel duration distributions.

We show the effectiveness of our model on real-world datasets, pedestrian location logs from large-scale exhibition halls, and bike trip data and taxi trip data in New York City. First, we show that our model achieves high estimation performance for the transition populations between locations. Second, we show that our model precisely captures the travel duration distributions between locations. Third, by using the estimated transition populations, we show that our model can analyze the routes of pedestrians in exhibition halls that are likely to be chosen by visitors without tracking individuals.

The major contributions of this paper include:

- We propose a probabilistic model that incorporates travel duration distributions for robustly estimating unobserved transition populations between locations from only aggregated data.
- We develop an efficient inference algorithm to learn the transition probabilities, the transition populations, and the travel duration distributions.
- Experiments on multiple real-world datasets show that our proposed model achieves statistically significant improvement over baseline methods.

2 Related Work

A number of works have been published recently that use the latest trends (e.g., deep learning) for analyzing aggregated data such as population or inflow and outflow at each location [Hoang *et al.*, 2016; Zhang *et al.*, 2017; Yao *et al.*, 2018]. These works can predict the population or inflow and outflow in the future with consideration of some external factors (e.g., weather). They cannot, however, estimate the transition populations between locations from inflow and outflow, which is the task considered in this paper. The task of estimating the transition populations between locations is important for better understanding people's mobility patterns; for example, it is useful for finding the popular routes of pedestrians in exhibition halls. While the method of [Xu *et al.*, 2017] can recover user trajectories given aggregated data and transition probabilities. The transition probabilities must be manually set by using other information such as distances between locations. In other words, the method cannot estimate the transition populations between locations from only aggregated data. Different from that method, our model can estimate not only the transition populations but also transition probabilities from only aggregated data. What is more, our model also enables us to estimate the travel duration distributions, a function not considered in the previous method.

Collective graphical models (CGMs) [Sheldon and Dietterich, 2011] have been recently developed as a general framework for analyzing aggregated data. The models have been applied for modeling contingency tables [Sheldon and Dietterich, 2011] and bird migration [Sheldon *et al.*, 2008]. Prior works provide efficient inference techniques for CGMs based on maximum a posteriori [Sheldon *et al.*, 2013], MCMC sampler [Sheldon and Dietterich, 2011], message passing [Sun *et al.*, 2015], and variational Bayesian inference [Iwata *et al.*, 2017]. The collective flow diffusion model (CFDM) [Kumar *et al.*, 2013] is the first application of efficient inference techniques developed for CGMs to the transportation domain. By maximizing a likelihood subject to constraints that represent people flow conservation, this model can estimate transition populations between locations from only aggregated data [Kumar *et al.*, 2013; Du *et al.*, 2014]. However, CFDM does not consider people's travel duration between locations, which is important for modeling the temporal dynamics of people flows. The proposed model is an extension of CFDM, and incorporates travel duration distributions into the constraints that represent flow conservation. Our model can estimate the transition populations between locations with consideration of the travel duration. Therefore, our model more precisely estimates transition populations between locations than previous models.

Information diffusion models are related to our work, because their main goal is to estimate latent flows of a piece of

Symbol	Description
$\mathbf{V}$	set of locations
i	location, $i \in \mathbf{V}$
$\mathbf{E}_i$	set of locations that are accessible from location i
T	#observation time steps
$N^{\text{out}}_{t,i}$	#outgoing people from location i at time step t, $N^{\text{out}}_{t,i} \geq 0$
$N^{\text{in}}_{t,i}$	#incoming people to location i at time step t, $N^{\text{in}}_{t,i} \geq 0$
M_{tij}	transition population who leave location i at time step t and whose next location is j, $M_{tij} \geq 0$
θ_{ij}	transition probability that people leave location i and move to location j, $\theta_{ij} \geq 0, \sum_{j \in \mathbf{E}_i} \theta_{ij} = 1$
α_{ji}	parameter of travel duration distribution from location j to i, $\alpha_{ji} > 0$
λ	hyperparameter for controlling the penalty, $\lambda \geq 0$

Table 1: Notation

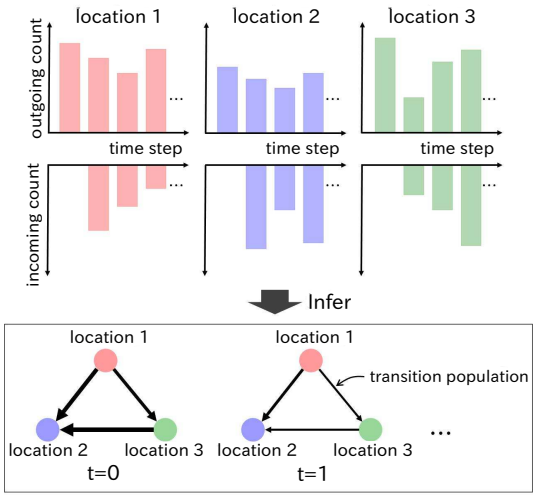


Figure 2: Illustration of the studied problem. (Upper figures) The aggregated data at each location. (Lower figures) Visualization of the estimated transition populations between locations at each time step, where arrow width is proportional to the transition population.

information over a network. These models have been used for analyzing the spread of information [Rodriguez *et al.*, 2011; Kurashima *et al.*, 2014] and user influence [Iwata *et al.*, 2013; Tanaka *et al.*, 2016] over social networks. These models incorporate a continuous time distribution to model the temporal dynamics of information flows, with the aim of describing the spread of information from node to node with delays over time. However, in analyzing information flows, it is not necessary to consider flow conservation, a critical factor in modeling the temporal dynamics of people flows based on aggregated data. Our model can infer latent people flows from only aggregated data by considering the flow conservation constraints that incorporate travel duration distributions.

3 Problem Formulation

In this section, we detail the aggregated data considered here, and define our problem of estimating latent people flows. The notations used in this paper are listed in Table 1.

Aggregated data: We illustrate an example of aggregated data for the three locations in the upper part of Figure 2. The data consist of incoming and outgoing counts at each location and each time step. Let $N^{\text{out}}_{t,i}$ be the count of outgoing people from location i at time step t, and $N^{\text{in}}_{t,i}$ be the count of incoming people to location i at time step t. Suppose that we are given a set of outgoing counts $\mathbf{N}^{\text{out}} = \{N^{\text{out}}_{t,i} | t = 0, \ldots, T-1, i \in \mathbf{V}\}$ and a set of incoming counts $\mathbf{N}^{\text{in}} = \{N^{\text{in}}_{t,i} | t = 1, \ldots, T, i \in \mathbf{V}\}$, where T is the number of observation time steps and $\mathbf{V}$ is the set of locations. The aggregated data of individuals (e.g., pedestrians, cars, bikes) can be gathered using various methods such as Wi-Fi [Musa and Eriksson, 2012], Bluetooth [Kotanen *et al.*, 2003], and video-based people counting systems [Chan *et al.*, 2008].

Problem: Our problem of estimating latent people flows, i.e., transition populations between locations, is formulated as follows. Figure 2 illustrates the studied problem for the case of three locations. Suppose that we have a set of incoming and outgoing counts at each location and each time step. We would like to estimate the transition population between locations at each time step, M_{tij}, which is the number of people who left location i at time step t and whose next location was j. Our goal is to infer a set of transition populations $\mathbf{M} = \{M_{tij} | t = 0, \ldots, T-1, i \in \mathbf{V}, j \in \mathbf{E}_i\}$, where $\mathbf{E}_i$ is the set of locations that are accessible from location i; people in location i can move only to the linked locations $j \in \mathbf{E}_i$.

4 Model

We propose a probabilistic model that incorporates people's travel duration between locations for estimating unobserved transition populations between locations from only aggregated data. Modeling the travel duration between locations is expected to estimate the transition populations more robustly in practical situations, that is, the observation range is limited and some people are not observed in any location in some time periods.

In the proposed model, the temporal dynamics of people flows are assumed to be probabilistic diffusion processes over a network, where the nodes are locations and the edges are paths between these locations. If the network structure such as road networks or a set of neighbor information is unknown, our model works by assuming a complete graph among locations. We assume that people diffuse from node to node in accordance with location-dependent transition probabilities. Let $\theta_{ij} \geq 0$ be the transition probability that people leave location i and move to location j, where $\sum_{j \in \mathbf{E}_i} \theta_{ij} = 1$. Given the outgoing count $N^{\text{out}}_{t,i}$ and the transition probabilities $\boldsymbol{\theta}_i = \{\theta_{ij} | j \in \mathbf{E}_i\}$, transition population $\mathbf{M}_{ti} = \{M_{tij} | j \in \mathbf{E}_i\}$ is given by the following multinomial distribution,

$$P(\mathbf{M}_{ti} | N^{\text{out}}_{t,i}, \boldsymbol{\theta}_i) = \frac{N^{\text{out}}_{t,i}!}{\prod_{j \in \mathbf{E}_i} M_{tij}!} \prod_{j \in \mathbf{E}_i} \theta_{ij}^{M_{tij}}. \quad (1)$$

Since we cannot obtain the tracking information of individuals in our aggregated data setting, $\mathbf{M}_{ti}$ is unobserved. Therefore, we treat it as a latent variable. Transition population

$\mathbf{M}_{ti}$ satisfies the following two relations that represent outgoing and incoming flow conservation:

$$N^{\text{out}}_{t,i} = \sum_{j \in \mathbf{E}_i} M_{tij}, \tag{2}$$

$$N^{\text{in}}_{t,i} = \sum_{j \in \mathbf{E}_i} \sum_{\tau=0}^{t-1} F(\Delta_{t,\tau}; \alpha_{ji}) M_{\tau ji}. \tag{3}$$

The outgoing flow conservation (2) indicates that the sum of people leaving location i at time step t equals the observed outgoing count at the same time step. The incoming flow conservation (3) indicates that the weighted sum of people leaving location j before time step t equals the observed incoming count at time step t. Weight $F(\Delta_{t,\tau}; \alpha_{ji})$ is travel duration probability, which is the probability that people who left location j at time step τ arrive at location i at time step t. Here, $\Delta_{t,\tau} = t - \tau$ is travel duration and α_{ji} is a parameter of the probability distribution. Details of $F(\Delta_{t,\tau}; \alpha_{ji})$ are given in the following paragraphs. The travel duration is treated as a random variable that follows the probability distribution; it is not a point estimate. This enables us to capture the heterogeneity in travel duration among individuals.

To derive the travel duration probability $F(\Delta_{t,\tau}; \alpha_{ji})$, we first introduce continuous travel duration distribution $f(\Delta; \alpha_{ji})$ as the probability density function of travel duration Δ from location j to i given parameter α_{ji}, where $\Delta > 0$ is a continuous random variable. Note that our model does not depend on the particular choice of the travel duration distribution. The travel duration probability $F(\Delta_{t,\tau}; \alpha_{ji})$ is calculated by the following integral of $f(\Delta; \alpha_{ji})$ over the interval from $\Delta_{t,\tau} - 1$ to $\Delta_{t,\tau}$:

$$F(\Delta_{t,\tau}; \alpha_{ji}) = \int_{\Delta_{t,\tau}-1}^{\Delta_{t,\tau}} f(\Delta; \alpha_{ji}) d\Delta. \tag{4}$$

Though our model can use any distribution as $f(\Delta; \alpha_{ji})$, for simplicity we consider the well-known one-parameter distribution, i.e., Rayleigh distribution, which is widely used for assessing the duration time in various fields such as user modeling [Wang *et al.*, 2008] and information diffusion modeling [Rodriguez *et al.*, 2011]. Figure 3 illustrates an example when using the Rayleigh distribution. The travel duration distribution is as follows:

$$f(\Delta; \alpha_{ji}) = \frac{\Delta}{\alpha^2_{ji}} \exp\Big(-\frac{\Delta^2}{2\alpha^2_{ji}}\Big), \tag{5}$$

the red line in Figure 3. Here, $\alpha_{ji} > 0$. The travel duration probability is given by

$$F(\Delta_{t,\tau}; \alpha_{ji}) = \exp\Big(-\frac{(\Delta_{t,\tau}-1)^2}{2\alpha^2_{ji}}\Big) - \exp\Big(-\frac{(\Delta_{t,\tau})^2}{2\alpha^2_{ji}}\Big), \tag{6}$$

the red area in Figure 3: The probability that people who left location j at time step τ arrive at location i at time step t.

5 Inference

Given outgoing counts $\mathbf{N}^{\text{out}}$ and incoming counts $\mathbf{N}^{\text{in}}$, we estimate the latent transition populations $\mathbf{M}$, the transition probabilities $\mathbf{\Theta} = \{\boldsymbol{\theta}_i | i \in \mathbf{V}\}$, and the parameters of travel duration distributions $\mathbf{A} = \{\alpha_{ji} | j \in \mathbf{V}, i \in \mathbf{E}_j\}$ by maximizing the likelihood with the flow conservation constraints.

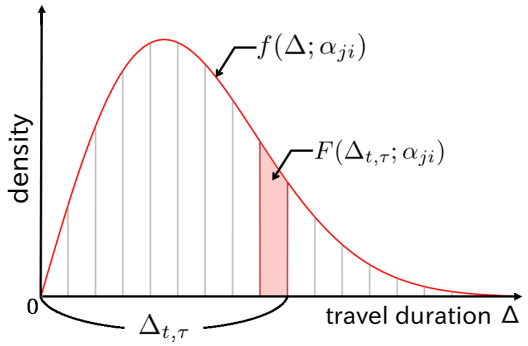


Figure 3: Travel duration probability from location j to i, where $\Delta_{t,\tau}$ is the time difference between outgoing time τ and incoming time t. The gray vertical lines are spaced by unit time steps.

The log-likelihood of $\mathbf{M}$ is given by

$$\begin{aligned}
&\log P(\mathbf{M} | \mathbf{N}^{\text{out}}, \mathbf{\Theta}) \\
&\propto \sum_{t=0}^{T-1} \sum_{i \in \mathbf{V}} \sum_{j \in \mathbf{E}_i} \Big(-\log M_{tij}! + M_{tij} \log \theta_{ij} \Big) \\
&\approx \sum_{t=0}^{T-1} \sum_{i \in \mathbf{V}} \sum_{j \in \mathbf{E}_i} \Big((1 + \log \theta_{ij}) M_{tij} - M_{tij} \log M_{tij} \Big),
\end{aligned} \tag{7}$$

where $\log N_{ti}!$ is omitted since it does not depend on $\mathbf{M}$, and we employ Stirling's approximation, $\log n! \approx n \log n - n$, in order to calculate $\log M_{tij}!$ efficiently [Sheldon *et al.*, 2013]. The parameters are estimated by maximizing the approximate log-likelihood (7) subject to constraints (2) and (3). The constraints might not strictly hold in real-world data, because the observations are noisy. To handle noisy observations, we treat the constraints as soft constraints, where the squared difference between the left- and right-hand sides in each of (2) and (3) are minimized. Then, the objective function to be maximized is described as follows:

$$\begin{aligned}
&\mathcal{J}(\mathbf{M}, \mathbf{\Theta}, \mathbf{A}) \\
&= \sum_{t=0}^{T-1} \sum_{i \in \mathbf{V}} \Bigg[\sum_{j \in \mathbf{E}_i} \Big((1 + \log \theta_{ij}) M_{tij} - M_{tij} \log M_{tij} \Big) \\
&- \frac{\lambda}{2} \| N^{\text{out}}_{t,i} - \sum_{j \in \mathbf{E}_i} M_{tij} \|^2 \\
&- \frac{\lambda}{2} \| N^{\text{in}}_{t+1,i} - \sum_{j \in \mathbf{E}_i} \sum_{\tau=0}^{t} F(\Delta_{t+1,\tau}; \alpha_{ji}) M_{\tau ji} \|^2 \Bigg].
\end{aligned} \tag{8}$$

Here, $\lambda \geq 0$ is a hyperparameter that controls the penalty for violating the constraints and $\| \cdot \|$ is the Euclidean norm. We update each parameter $\mathbf{M}$, $\mathbf{\Theta}$ and $\mathbf{A}$ alternately as described in the following paragraphs.

Update of $\mathbf{M}$: Given current estimates $\hat{\mathbf{\Theta}}$ and $\hat{\mathbf{A}}$, we arrive

Algorithm 1: Inference procedure for our model.

Input : $\mathbf{N}^{\text{in}}$, $\mathbf{N}^{\text{out}}$, $\mathbf{V}$, $\{\mathbf{E}_i\}_{i\in\mathbf{V}}$, T, λ
Output: $\mathbf{M}$, $\boldsymbol{\Theta}$, $\mathbf{A}$

```
Initialize M, Θ, A
repeat
    Update M by solving (9)
    for i ∈ V do
        for j ∈ E_i do
            Update θ_ij by (10)
        end for
    end for
    Update A by solving (11)
until Convergence
```

at the following optimization problem:

$$\begin{aligned} \underset{\mathbf{M}}{\text{maximize}} \quad & \mathcal{J}(\mathbf{M}, \hat{\boldsymbol{\Theta}}, \hat{\mathbf{A}}) \\ \text{subject to} \quad & M_{tij} \geq 0, \ \ t = 0, \ldots, T-1, \ i \in \mathbf{V}, \ j \in \mathbf{E}_i. \end{aligned} \tag{9}$$

We solve the optimization problem by using the L-BFGS-B method [Byrd *et al.*, 1995].

Update of $\boldsymbol{\Theta}$: The estimates of $\boldsymbol{\Theta}$ are obtained in closed form by maximizing (8) given current estimates $\hat{\mathbf{M}}$ as follows:

$$\theta_{ij} = \frac{\sum_{t=0}^{T-1} \hat{M}_{tij}}{\sum_{t=0}^{T-1} \sum_{j\in\mathbf{E}_i} \hat{M}_{tij}}. \tag{10}$$

Update of $\mathbf{A}$: Given current estimates $\hat{\mathbf{M}}$ and $\hat{\boldsymbol{\Theta}}$, we solve the following optimization problem:

$$\begin{aligned} \underset{\mathbf{A}}{\text{maximize}} \quad & \mathcal{J}(\hat{\mathbf{M}}, \hat{\boldsymbol{\Theta}}, \mathbf{A}) \\ \text{subject to} \quad & \alpha_{ji} > 0, \ \ j \in \mathbf{V}, i \in \mathbf{E}_j, \end{aligned} \tag{11}$$

by using the L-BFGS-B method [Byrd *et al.*, 1995]. We iteratively update each parameter until the value of (8) converges. Our inference procedure is shown in Algorithm 1, which always converges and guarantees that the local optima of the respective parameters are obtained by setting λ to a fixed value.

Validation of λ: Hyperparameter λ can be determined by the following validation procedure: (a) learn the model parameters by Algorithm 1 with each candidate value of λ; (b) predict the incoming counts at the next time step; (c) choose the value based on prediction performance in the training data. Given the preceding outgoing counts, the predicted incoming counts in location i at the next time step are given by

$$\hat{N}^{\text{in}}_{t+1,i} = \sum_{j\in\mathbf{E}_i} \sum_{\tau=0}^{t} F(\Delta_{t,\tau}; \hat{\alpha}_{ji}) \hat{\theta}_{ji} N^{\text{out}}_{\tau,j}. \tag{12}$$

6 Experiments

6.1 Data Description

We evaluated the proposed model using real-world pedestrian location logs collected in exhibition halls. The data was collected at an event that attracted large crowds, Niconico Chokaigi 2016 [1], held at Makuhari Messe located near Tokyo, Japan from 10:00 to 18:00 on April 29th. The event site was composed of four exhibition halls, Hall 1, Hall 2, Hall 3 and Hall 4 with sizes of 186.3m × 124.8m, 183.2m × 124.8m, 127.6m × 124.8m, and 190.9m × 108.3m, respectively. The number of event booths, $|\boldsymbol{V}|$, in Hall 1, Hall 2, Hall 3 and Hall 4 were 38, 27, 10, and 9, respectively. We gathered pedestrian location logs using Bluetooth beacons which were placed at each booth. The technology enables us to observe the times at which each user entered or left the observation range (at most 10 – 15 meters). The data consist of 3,727 mobile users who agreed to provide detailed information of location over time. The original data contains time stamps of arrival and departure at booths for each user; users can be tracked over time. Note that the tracking information of users was used only for evaluating the estimation performance for transition populations and travel duration probabilities; we did not use the tracking information in the inference process. In this experiment, we created aggregated incoming and outgoing pedestrian counts at each booth, where the time interval was 3 minutes. The number of observation time steps was 160. $\mathbf{E}_i$ was a set of edges assuming a complete graph for all datasets. The total sum of incoming and outgoing pedestrians at all booths in each hall are shown in Table 2.

Data	Pedestrian data				Bike trip data		Taxi trip data	
Area/Date	Hall 1	Hall 2	Hall 3	Hall 4	Mar. 1	Jun. 1	Mar. 1	Jun. 1
#outgoing	30,606	34,750	15,448	10,990	14,366	18,961	201,440	196,101
#incoming	30,606	34,750	15,448	10,990	14,640	19,312	204,476	199,232

Table 2: The total sum of incoming and outgoing counts

To additionally validate the performance of our model, we used two public datasets, bike trip data [2] and taxi trip data [3] in New York City. The data is a set of trip records consisting of: trip id, pickup location, dropoff location, pickup date and time, and dropoff date and time. Note that the data consist of location information only when people started and finished the trips; the trajectories during the trips are not recorded. We used the data from 8:00 to 24:00 on March 1st and June 1st, and created gridded incoming and outgoing count data, where the time interval in both datasets was 10 minutes. The number of observation time steps was 96. The grid sizes in the bike trip data and taxi trip data were 2km × 2km (12 × 12 grid cells) and 3km × 3km (18 × 18 grid cells), respectively. Here, we omitted grid cells if their incoming and outgoing counts were lower than a threshold. Then, bike trip data and taxi trip data held 11 and 14 grid cells, respectively. $\mathbf{E}_i$ was a set of edges assuming a complete graph for all datasets. The total sum of incoming and outgoing bikes/taxis at overall grid cells on each date are shown in Table 2.

6.2 Quantitative Evaluation

Transition Population Estimation

We evaluated our model in terms of its performance in estimating the transition populations $\mathbf{M}$. The evaluation metric

[1] http://www.chokaigi.jp/2016/en/

[2] https://www.citibikenyc.com/system-data

[3] http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml

Data	Pedestrian data				Bike trip data		Taxi trip data	
Area/Date	Hall 1	Hall 2	Hall 3	Hall 4	Mar. 1	Jun. 1	Mar. 1	Jun. 1
Proposed	**1.237 ± 0.024**	**1.028 ± 0.014**	**0.784 ± 0.034**	**0.675 ± 0.030**	**0.621 ± 0.015**	**0.704 ± 0.021**	**0.443 ± 0.006**	**0.454 ± 0.006**
CFDM	1.370 ± 0.030	1.140 ± 0.013	0.828 ± 0.038	0.709 ± 0.030	0.648 ± 0.015	0.740 ± 0.021	0.463 ± 0.006	0.469 ± 0.006
Popularity	1.751 ± 0.012	1.594 ± 0.013	1.180 ± 0.019	0.848 ± 0.025	0.724 ± 0.016	0.790 ± 0.021	0.510 ± 0.009	0.520 ± 0.008
Uniform	1.835 ± 0.011	1.676 ± 0.012	1.210 ± 0.019	1.221 ± 0.024	1.031 ± 0.013	1.081 ± 0.018	0.983 ± 0.010	0.987 ± 0.009

Table 3: NMAE L_1 and standard errors for the estimation of transition populations in the real-world datasets.

is the following normalized mean absolute error (NMAE) in transition populations:

$$L_1 = \frac{1}{T}\sum_{t=0}^{T-1}\frac{\sum_{i\in\mathbf{V}}\sum_{j\in\mathbf{E}_i}|M^{\star}_{tij} - \hat{M}_{tij}|}{\sum_{i\in\mathbf{V}}\sum_{j\in\mathbf{E}_i}M^{\star}_{tij}}, \quad (13)$$

where $M^{\star}_{tij}$ is the true transition population leaving location i at time step t and whose next location is j; $\hat{M}_{tij}$ is its estimate. We compared our proposed model with the collective flow diffusion model (CFDM) [Kumar *et al.*, 2013]. Different from our model, CFDM does not consider travel duration between locations. In addition, we compared our proposed model with the following two baseline methods: Popularity and Uniform methods. Popularity assumes that people move to other locations in proportion to location popularity regardless of current locations; the estimated transition population is given by

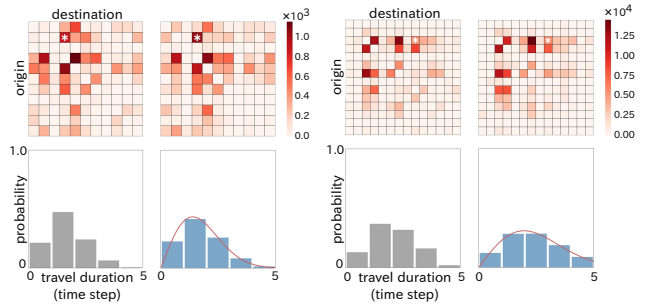
$$\hat{M}_{tij} = N^{\text{out}}_{t,i} \times \frac{\sum_{t=0}^{T-1} N^{\text{in}}_{t+1,j}}{\sum_{t=0}^{T-1}\sum_{j\in\mathbf{E}_i} N^{\text{in}}_{t+1,j}}. \quad (14)$$

Uniform estimates transition populations using a discrete uniform distribution; it assumes that people move to neighbor locations with equal probability $1/|\mathbf{E}_i|$. For our model, hyperparameter λ was set to the best value based on the validation procedure described in Section 5; λ was chosen from 0.1, 0.2, 0.5, 1, 2, and 5. We used a Rayleigh distribution for modeling travel duration distribution as shown in Section 4.

Table 3 shows NMAE L_1 of the proposed model, CFDM, Popularity and Uniform. For all datasets, the proposed model performed better than the other methods, and the differences between our model and CFDM were significant (t-test, $p < 0.01$). We found similar results using other evaluation metrics (e.g., MAE, RMSE). The upper parts of Figure 4(a) and 4(b) provide a heatmap visualization of the transition matrix for the bike trip data (Jun.1) and taxi trip data (Jun.1), respectively. The transition matrix is the total number of bikes/taxis that moved between each pair of origin and destination locations; its elements were calculated as follows: $M_{ij} = \sum_{t=0}^{T-1} M_{tij}$. The true matrix is shown on the left and the estimate of our model is shown on the right. As shown, our model accurately estimated the transition populations between locations. Qualitative comparisons of the estimated transition populations in pedestrian data are described at length in Section 6.3.

Travel Duration Probability Estimation

We evaluated the performance of estimating travel duration probabilities. We used the mean Kullback-Leibler (KL) divergence between the true and estimated travel duration probabilities over all pairs of locations as the evaluation metric:



(a) Bike trip data (Jun. 1) (b) Taxi trip data (Jun. 1)

Figure 4: (Upper figures) Heatmap visualization of transition matrix. The true matrix is shown on the left and the estimate by our model is shown on the right. (Lower figures) Travel duration probability specified by white asterisk ($*$) in the corresponding upper figure. Red line represents the estimated travel duration distribution.

$$L_2 = \frac{1}{|\mathbf{V}|}\sum_{j\in\mathbf{V}}\frac{1}{|\mathbf{E}_j|}\sum_{i\in\mathbf{E}_j}\sum_{\Delta=1}^{T} P^{\star}(\Delta)\log\frac{P^{\star}(\Delta)}{F(\Delta;\hat{\alpha}_{ji})}, \quad (15)$$

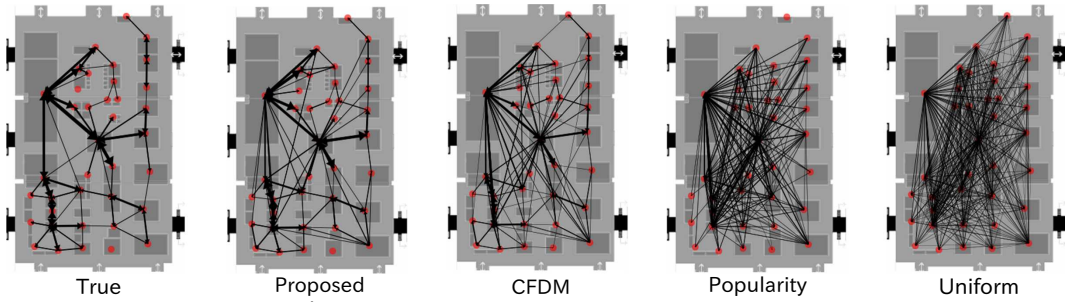
where $P^{\star}(\Delta)$ and $F(\Delta;\hat{\alpha}_{ji})$ are the true travel duration probability and the estimate of our model, respectively.

Table 4 shows the mean KL divergence L_2 of the proposed model and CFDM. In CFDM, the travel duration probability equals 1 if $\Delta = 1$ and 0 otherwise. As shown in Table 4, the proposed model achieved lower mean KL divergence values for all datasets. The results show that our model precisely estimated the travel duration probabilities. The lower parts of Figure 4(a) and 4(b) show visualization examples of travel duration probability, $F(\Delta;\alpha_{ji})$, specified by white asterisk ($*$) in the corresponding upper heatmap. Red line represents the travel duration distribution, $f(\Delta;\alpha_{ji})$, estimated by our model. As shown, our model well estimated the travel duration distributions and so allows us to capture the heterogeneity in travel duration among individuals.

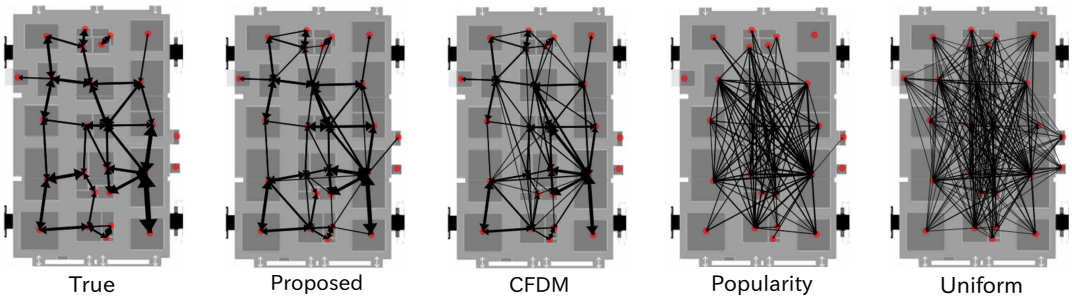
As shown in Table 3 and 4, our model achieved better performance in estimating both transition populations and travel duration probabilities. These results indicate that incorporating the people's travel duration into the model is important in estimating latent people flows.

6.3 Qualitative Evaluation

Figure 5 visualizes the estimated pedestrian flows in Hall 1 and Hall 2 from the pedestrian data. We illustrate the true pedestrian flow on the left in Figure 5, and the estimates of the proposed model, CFDM, Popularity and Uniform on the



(a) Hall 1



(b) Hall 2

Figure 5: Comparison of pedestrian flow visualizations in Hall 1 and Hall 2. The true pedestrian flow is shown on the left, and the estimates of the proposed model, CFDM, Popularity and Uniform on the right. Red dots represent the locations of booths, and the directed edges represent the pedestrian flows. The edge widths are proportional to the total numbers of pedestrians that moved between each pair of booths.

Data	Pedestrian data				Bike trip data		Taxi trip data	
Area/Date	Hall 1	Hall 2	Hall 3	Hall 4	Mar. 1	Jun. 1	Mar. 1	Jun. 1
Proposed	**1.629**	**2.243**	**1.492**	**1.841**	**2.656**	**2.560**	**2.109**	**2.033**
CFDM	1.802	2.748	2.217	2.961	3.370	3.325	3.352	3.256

Table 4: Mean KL divergence L_2 for the estimations of travel duration probabilities in the real-world datasets.

right in Figure 5. Here, the red dots represent the locations of booths, and the directed edges represent the pedestrian flows between locations. The edge widths are proportional to the total numbers of pedestrians that moved between each pair of booths. Note that we omitted those edges whose transition populations were lower than a threshold, and bidirectional edge widths are proportional to the average of the transition populations between the pair of locations. As shown in Figure 5, the proposed model better discerned the pedestrian flows than the other methods. CFDM tends to output some false flows. This is reasonable because, as was explained earlier, CFDM is based on the unreal assumption that all pedestrians who left a location at one time step should arrive at another location at the next time step. In other words, CFDM fails to estimate the pedestrian flows between locations when travel duration is significant. Our model, on the other hand, could more precisely estimate the pedestrian flows because it considers travel duration. The visualization results are useful for optimizing navigation systems and strategies of location-based advertising. For example, discovering popular routes of pedestrians yields better route recommendations.

7 Conclusion

This paper proposed a probabilistic model that incorporates people's travel duration between locations for estimating latent people flows, i.e., transition populations between locations, from only aggregated data. Incorporating the travel duration into the model enables us to robustly estimate transition populations in practical situations, that is, the observation range is limited and some people are not observed in any location in some time periods. Since travel duration is treated as a random variable that follows a probability distribution, our model can capture the heterogeneity in travel duration among individuals. The inference algorithm presented herein allows us to infer transition probabilities, transition populations, and travel duration distributions. We used three real-world datasets, pedestrian data, bike trip data, and taxi trip data, to confirm that our model can precisely estimate the transition populations between locations. Our future work is to conduct extended experiments considering temporal factors (e.g., time-of-day) and external factors (e.g., weather).

References

[Byrd *et al.*, 1995] Richard H. Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. A limited memory algorithm for bound constrained optimization. *SIAM J. Sci. Comput.*, 16:1190–1208, 1995.

[Chan *et al.*, 2008] Antoni B. Chan, Zhang-Sheng J. Liang, and Nuno Vasconcelos. Privacy preserving crowd monitoring: Counting people without people models or tracking. In *CVPR'08*, pages 1–7, 2008.

[Chow and Mokbel, 2011] Chi-Yin Chow and Mohamed F. Mokbel. Trajectory privacy in location-based services and data publication. *ACM SIGKDD Explorations Newsletter*, 13(1):19–29, 2011.

[Dhar and Varshney, 2011] Subhankar Dhar and Upkar Varshney. Challenges and business models for mobile location-based services and advertising. *Communications of the ACM*, 54(5):121–129, 2011.

[Du *et al.*, 2014] Jiali Du, Akshat Kumar, and Pradeep Varakantham. On understanding diffusion dynamics of patrons at a theme park. In *AAMAS'14*, pages 1501–1502, 2014.

[Giannotti *et al.*, 2007] Fosca Giannotti, Mirco Nanni, Fabio Pinelli, and Dino Pedreschi. Trajectory pattern mining. In *KDD'07*, pages 330–339, 2007.

[Hoang *et al.*, 2016] Minh Hoang, Yu Zheng, and Ambuj Singh. FCCF: Forecasting citywide crowd flows based on big data. In *SIGSPATIAL'16*, pages 1–10, 2016.

[Huang and Gartner, 2010] Haosheng Huang and Georg Gartner. A survey of mobile indoor navigation systems. *Cartography in Central and Eastern Europe*, pages 305–319, 2010.

[Iwata *et al.*, 2013] Tomoharu Iwata, Amar Shah, and Zoubin Ghahramani. Discovering latent influence in online social activities via shared cascade poisson processes. In *KDD'13*, pages 266–274, 2013.

[Iwata *et al.*, 2017] Tomoharu Iwata, Hitoshi Shimizu, Futoshi Naya, and Naonori Ueda. Estimating people flow from spatio-temporal population data via collective graphical mixture models. *ACM Transactions on Spatial Algorithms and Systems*, 3(1):1–18, 2017.

[Klein, 2001] Lawrence A. Klein. *Sensor Technologies and Data Requirements for ITS*. Artech House, 2001.

[Kotanen *et al.*, 2003] Antti Kotanen, Marko Hännikäinen, Helena Leppäkoski, and Timo Hämäläinen. Experiments on local positioning with bluetooth. In *ICIT'03*, pages 297–303, 2003.

[Kumar *et al.*, 2013] Akshat Kumar, Daniel Sheldon, and Biplav Srivastava. Collective diffusion over networks: Models and inference. In *UAI'13*, pages 351–360, 2013.

[Kurashima *et al.*, 2010] Takeshi Kurashima, Tomoharu Iwata, Go Irie, and Ko Fujimura. Travel route recommendation using geotags in photo sharing sites. In *CIKM'10*, pages 579–588, 2010.

[Kurashima *et al.*, 2014] Takeshi Kurashima, Tomoharu Iwata, Noriko Takaya, and Hiroshi Sawada. Probabilistic latent network visualization: Inferring and embedding diffusion networks. In *KDD'14*, pages 1236–1245, 2014.

[Monreale *et al.*, 2009] Anna Monreale, Fabio Pinelli, Roberto Trasarti, and Fosca Giannotti. WhereNext: A location predictor on trajectory pattern mining. In *KDD'09*, pages 637–646, 2009.

[Musa and Eriksson, 2012] A. B. M. Musa and Jakob Eriksson. Tracking unmodified smartphones using wi-fi monitors. In *SenSys'12*, pages 281–294, 2012.

[Rodriguez *et al.*, 2011] Manuel G. Rodriguez, David Balduzzi, and Bernhard Schölkopf. Uncovering the temporal dynamics of diffusion networks. In *ICML'11*, pages 561–568, 2011.

[Senior *et al.*, 2007] Andrew W. Senior, Lisa M. Brown, Arun Hampapur, Chiao-Fe Shu, Yun Zhai, Rogério Schmidt Feris, Ying li Tian, Sergio Borger, and Christopher R. Carlson. Video analytics for retail. In *AVSS'07*, pages 423–428, 2007.

[Sheldon and Dietterich, 2011] Daniel Sheldon and Thomas G. Dietterich. Collective graphical models. In *NIPS'11*, pages 1161–1169, 2011.

[Sheldon *et al.*, 2008] Daniel Sheldon, M. A. Saleh Elmohamed, and Dexter Kozen. Collective inference on markov models for modeling bird migration. In *NIPS'08*, pages 1321–1328, 2008.

[Sheldon *et al.*, 2013] Daniel Sheldon, Tao Sun, Akshat Kumar, and Thomas G. Dietterich. Approximate inference in collective graphical models. In *ICML'13*, pages 1004–1012, 2013.

[Song *et al.*, 2014] Xuan Song, Quanshi Zhang, Yoshihide Sekimoto, and Ryosuke Shibasaki. Prediction of human emergency behavior and their mobility following large-scale disaster. In *KDD'14*, pages 5–14, 2014.

[Sun *et al.*, 2015] Tao Sun, Daniel Sheldon, and Akshat Kumar. Message passing for collective graphical models. In *ICML'15*, pages 853–861, 2015.

[Tanaka *et al.*, 2016] Yusuke Tanaka, Takeshi Kurashima, Yasuhiro Fujiwara, Tomoharu Iwata, and Hiroshi Sawada. Inferring latent triggers of purchases with consideration of social effects and media advertisements. In *WSDM'16*, pages 543–552, 2016.

[Wang *et al.*, 2008] Longhao Wang, Yu Zheng, Xing Xie, and Wei-Ying Ma. A flexible spatio-temporal indexing scheme for large-scale GPS track retrieval. In *MDM'08*, pages 1–8, 2008.

[Xu *et al.*, 2017] Fengli Xu, Zhen Tu, Yong Li, Pengyu Zhang, Xiaoming Fu, and Depeng Jin. Trajectory recovery from ash: User privacy is not preserved in aggregated mobility data. In *WWW'17*, pages 1241–1250, 2017.

[Yao *et al.*, 2018] Huaxiu Yao, Fei Wu, Jintao Ke, Xianfeng Tang, Yitian Jia, Siyu Lu, Pinghua Gong, Jieping Ye, and Zhenhui Li. Deep multi-view spatial-temporal network for taxi demand prediction. In *AAAI'18*, 2018.

[Yuan *et al.*, 2012] Jing Yuan, Yu Zheng, and Xing Xie. Discovering regions of different functions in a city using human mobility and POIs. In *KDD'12*, pages 186–194, 2012.

[Zhang *et al.*, 2017] Junbo Zhang, Yu Zheng, and Dekang Qi. Deep spatio-temporal residual networks for citywide crowd flows prediction. In *AAAI'17*, 2017.

[Zhuang *et al.*, 2017] Chenyi Zhuang, Nicholas Jing Yuan, Ruihua Song, Xing Xie, and Qiang Ma. Understanding people lifestyles: Construction of urban movement knowledge graph from GPS trajectory. In *IJCAI'17*, pages 3616–3623, 2017.