Appendix A
The validation of mobility computation from trajectory data

The mobility metrics defined in Sec. 4.1 is theoretic in that the whole trajectory of each urban user should be observed in order to compute the exact value of each metric. This assumption is infeasible as the real-world trajectory measurement reports user's locations in discrete time intervals. Nevertheless, we show empirically that due to the characteristic of trajectory data set in this work, the mobility computation method in Sec. 4.2 can approximate the exact value of mobility metrics with statistically bounded errors. In details, these errors can happen in the computation of people/segment entropy of user trajectories (Part A) or in the aggregation of individual trajectory's entropy on each grid cell of the city (Part B). In addition, we demonstrate the semantics of the Vibrancy and Commutation metrics by correlating them with side information using real-world data capture (Part C).

A. Analysis on the Computation of Mobility Metrics

On each trajectory, the people entropy and segment entropy are defined in the paper as below.

$$H(P) = -\sum_{j=1}^{M} p_j \cdot \log p_j \quad where \quad p_j = \sum_{i=1}^{N} q_{ij} \cdot \frac{T_i}{T} \tag{1}$$

$$H(Q_i, P) = -\sum_{j=1}^{M} q_{ij} \cdot \log p_j \tag{2}$$

With the urban trajectory data set, these metrics are computed by

$$\widetilde{H}(\widetilde{P}) = -\sum_{j=1}^{M} \tilde{p}_j \cdot \log \tilde{p}_j \quad where \quad \tilde{p}_j = \frac{\sum_{i=1}^{n} \tilde{q}_{ij}}{n} \tag{3}$$

$$\widetilde{H}(\widetilde{Q}_i, \widetilde{P}) = -\sum_{j=1}^{M} \tilde{q}_{ij} \cdot \log \tilde{p}_j \tag{4}$$

In theoretic and empirical computations, the semantic distribution on each stay segment, i.e., $Q_i =< q_{i1}, \cdots, q_{iM} >$, are both estimated by feature extraction on each grid cell (stay location). The error on q_{ij} can then be omitted by $q_{ij} = \tilde{q}_{ij}$ on the same cell. Hence, we mainly consider the error induced on estimating p_j.

Examine the equation to compute $\tilde{p}_j$ in Eq. (3). We assume the record distribution within each stay segment is uniform (see Appendix B for details). Let the ith stay segment have a consecutive record interval of t_i. The number of records in the ith stay segment can be computed by $n_i = \lfloor \frac{T_i}{t_i} \rfloor \ or \ \lfloor \frac{T_i}{t_i} \rfloor + 1$. $\tilde{p}_j$ can be rewritten as

$$\tilde{p}_j = \frac{\sum_{i=1}^{N} n_i \cdot q_{ij}}{n} = \varepsilon_1 + \frac{\sum_{i=1}^{N} \frac{T_i}{t_i} \cdot q_{ij}}{n} \tag{5}$$

The measurement error of p_j can be decomposed into two parts by

$$\varepsilon(p_j) = \tilde{p}_j - p_j = \varepsilon_1 + \frac{\sum_{i=1}^{N} \frac{T_i}{t_i} \cdot q_{ij}}{n} - \sum_{i=1}^{N} q_{ij} \cdot \frac{T_i}{T} = \varepsilon_1 + \varepsilon_2 \tag{6}$$

where the two error terms are bounded by

$$|\varepsilon_1| = \frac{1}{n} \cdot \sum_{i=1}^{N} q_{ij} \cdot |\lfloor \frac{T_i}{t_i} \rfloor - \frac{T_i}{t_i}| < \frac{1}{n} \cdot \sum_{i=1}^{N} q_{ij} \approx p_j \cdot \frac{N}{n} \tag{7}$$

$$|\varepsilon_2| = \sum_{i=1}^{N} q_{ij} \cdot T_i \cdot |\frac{1}{n \cdot t_i} - \frac{1}{T}| = \sum_{i=1}^{N} q_{ij} \cdot \frac{T_i}{T} \cdot |\frac{T}{n} \cdot \frac{1}{t_i} - 1| \tag{8}$$

By a third-order Taylor expansion of the expectation of $\frac{1}{t_i}$:

$$\begin{aligned} E(\frac{1}{t_i}) \approx & \quad \tfrac{1}{E(t_i)} \cdot (1 + \tfrac{VAR(t_i) \cdot E(t_i)}{E(t_i^3)}) \\ = & \quad \tfrac{1}{E(t_i)} \cdot (1 + \tfrac{\rho^2(t_i)}{\gamma(t_i) \cdot \rho^3(t_i) + 3\rho^2(t_i) + 1}) = \tfrac{1 + \psi(t_i)}{E(t_i)} \end{aligned} \tag{9}$$

where $\gamma(t_i)$ and $\rho(t_i)$ are skewness and coefficient of variance (COV) of t_i. Consider the expectation of $n \cdot t_i$:

$$\begin{aligned} & E(n \cdot t_i) = E(\textstyle\sum_{k=1}^{N} n_k \cdot t_i) \\ \in & (\textstyle\sum_{k=1}^{N} T_k \cdot E(\frac{t_i}{t_k}) - N \cdot E(t_i), \ \ \textstyle\sum_{k=1}^{N} T_k \cdot E(\frac{t_i}{t_k}) + N \cdot E(t_i)) \end{aligned} \tag{10}$$

Because t_i has an independent identical distribution for any $1 \leq i \leq N$, then $E(\frac{t_i}{t_k}) = 1$ for any $k \neq i$, Eq. (10) becomes

$$E(n \cdot t_i) \in (T - N \cdot E(t_i), T + N \cdot E(t_i)) \tag{11}$$

Then we have $\frac{1}{E(t_i)} \in (\frac{n-N}{T}, \frac{n+N}{T})$. Substituting it into Eq. (8), the expectation of the second error term is bounded by:

$$\begin{aligned} E(|\varepsilon_2|) < & \quad \textstyle\sum_{i=1}^{N} q_{ij} \cdot \frac{T_i}{T} \cdot \psi(t_i) + \textstyle\sum_{i=1}^{N} q_{ij} \cdot \frac{T_i}{T} \cdot (1 + \psi(t_i)) \cdot \frac{N}{n} \\ < & \quad p_j \cdot (\psi(t_i) + (1 + \psi(t_i)) \cdot \tfrac{N}{n}) \end{aligned} \tag{12}$$

In our data set measured at Beijing (∼30M trajectories) , the average length of stay segments is 5183 seconds and the average record interval within a stay segment is 157 seconds, $\frac{N}{n} \sim \frac{157}{5183} = 0.03$. Also, we have $\psi(t_i) = 0.11$ by computing the skewness and COV of t_i. Finally, the error expectation on p_j is bounded by $E(|\varepsilon_1|) + E(|\varepsilon_2|) < p_j \cdot 0.173$. Empirically, the computation error for each average semantic distribution p_j is bounded by 17.3%.

From the error on the semantic distribution p_j, we further derive error rates in computing the final people (segment) entropy through numerical experiments. For Eq. (1) or Eq. (2), we randomly generate a distribution $P =< p_1, \cdots, p_M >$ or $Q_i =< q_{i1}, \cdots, q_{iM} >$ that sums to one. On each probability of the distribution (p_j or q_{ij}), a random noise with rates of +17.3% or -17.3% of the original probability is added. We compute the actual entropy using the original distribution and the erroneous entropy using the distribution with noise. The error rate of the entropy can be calculated. The same process is repeated 10^6 times and we obtain the expectation and 95% confidence interval (upper bound) of the error rate in computing entropies. As shown in Figure 1, the expectation of the error rate is below 3% consistently and the 95% CI drops as we have more semantic categories (a larger M). In our scenario with 10 POI types, the 95% CI of error rate is about 7% ($M = 10$). Consider that the entropy of many trajectories will be aggregated on a cell. Under an i.i.d assumption of the entropy distribution, the variance of error rates in computing the average entropy in each cell further drops by the square

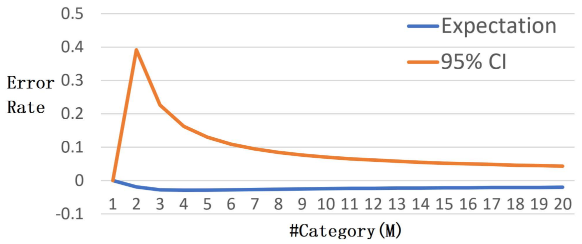


Fig. 1: The expectation and standard deviation of error rates in the entropy computation.

root of the number of records in that cell. In our data at Beijing, there are 53k cells occupying 99.5% of location records in total. The minimal number of records in each of these cells are 3000. The error rate in computing the average entropy of each cell will be bounded by $2.4\% + \frac{4.6\%}{\sqrt{3000}} = 2.5\%$ ($M = 10$), which is sufficiently small for actual usage.

B. Analysis on the Aggregation of Mobility Metrics in Grid Cells

In this paper, the basic aggregation method of entropy values on individual grid cells adopts the average operation. In theory, the aggregated entropy value on each cell should be the entropy value of individual records on this cell weighted by the length of the corresponding stay segments. We show that the simplified computation can approximate the actual entropy value with statistically bounded errors, according to the law of large numbers.

Consider a cell in the city grid with N records located on the cell in the entire data set. Without loss of generality, we assume that there are no two records belonging to the same stay segment of a trajectory, otherwise we will have removed one of them in the pre-processing for aggregation computation. Denote the entropy value of the ith record to be h_i and the length of the corresponding stay segment to be T_i. The exact aggregated entropy value of the cell is computed by

$$H_C = \frac{\sum_{i=1}^{N} h_i \cdot T_i}{\sum_{i=1}^{N} T_i} \tag{13}$$

In comparison, the actual computation is conducted by

$$\widetilde{H}_C = \frac{\sum_{i=1}^{N} h_i}{N} \tag{14}$$

It is shown that the expectation of the exact aggregated value in Eq. (13) equals our computation result in Eq. (14).

$$E(H_C) = \sum_{i=1}^{N} h_i \cdot E(T_i) \cdot E(\frac{1}{\sum_{i=1}^{N} T_i}) \tag{15}$$

By a Taylor expansion of the expectation of $\frac{1}{\sum_{i=1}^{N} T_i}$:

$$\begin{aligned} E(H_C) = & \quad \textstyle\sum_{i=1}^{N} h_i \cdot E(T_i) \cdot (\frac{1}{N \cdot E(T_i)} + \frac{N \cdot VAR(T_i)}{E^3(\sum_{i=1}^{N} T_i)}) \\ = & \quad \textstyle\sum_{i=1}^{N} h_i \cdot E(T_i) \cdot \frac{1}{N \cdot E(T_i)} = \frac{\sum_{i=1}^{N} h_i}{N} = \widetilde{H}_C \end{aligned} \tag{16}$$

The variance of the exact aggregated value in Eq. (13) can be estimated by

$$\begin{aligned} \sigma^2(H_C) = & \quad \frac{\sum_{i=1}^{N} h_i \cdot T_i}{\sum_{i=1}^{N} T_i} \\ = & \quad E(\textstyle\sum_{i=1}^{N} h_i T_i)^2 E(\frac{1}{(\sum_{i=1}^{N} T_i)^2}) - E^2(\sum_{i=1}^{N} h_i T_i) E^2(\frac{1}{\sum_{i=1}^{N} T_i}) \\ = & \quad \frac{\sum_{i=1}^{N} h_i^2 \cdot E(T_i^2) + \sum_{i=1}^{N} \sum_{j \neq i} h_i \cdot h_j \cdot E^2(T_i)}{N \cdot E(T_i^2) + (N^2 - N) E^2(T_i)} - \\ & \quad (\textstyle\sum_{i=1}^{N} h_i \cdot E(T_i))^2 \cdot (\frac{1}{N \cdot E(T_i)})^2 \\ = & \quad (\textstyle\sum_{i=1}^{N} h_i^2) \cdot (\frac{E(T_i^2)}{N^2 \cdot E^2(T_i) + N \cdot \sigma(T_i)} - \frac{1}{N^2}) + \\ & \quad (\textstyle\sum_{i=1}^{N} \sum_{j \neq i} h_i \cdot h_j) \cdot (\frac{E^2(T_i)}{N^2 \cdot E^2(T_i) + N \cdot \sigma(T_i)} - \frac{1}{N^2}) \end{aligned} \tag{17}$$

Because $\frac{E^2(T_i)}{N^2 \cdot E^2(T_i) + N \cdot \sigma(T_i)} - \frac{1}{N^2} < 0$, we have

$$\begin{aligned} \sigma^2(H_C) < & \quad (\textstyle\sum_{i=1}^{N} h_i^2) \cdot (\frac{E(T_i^2)}{N^2 \cdot E^2(T_i) + N \cdot \sigma(T_i)} - \frac{1}{N^2}) \\ < & \quad (\textstyle\sum_{i=1}^{N} h_i^2) \cdot (\frac{\rho^2(T_i) + 1}{N^2 + N \cdot \rho^2(T_i)} - \frac{1}{N^2}) \\ < & \quad (\textstyle\sum_{i=1}^{N} h_i^2) \cdot \frac{\rho^2(T_i)}{N^2} = \frac{E^2(h_i) \cdot (\rho^2(h_i) + 1) \cdot \rho^2(T_i)}{N} \end{aligned} \tag{18}$$

The standard deviation is then bounded by

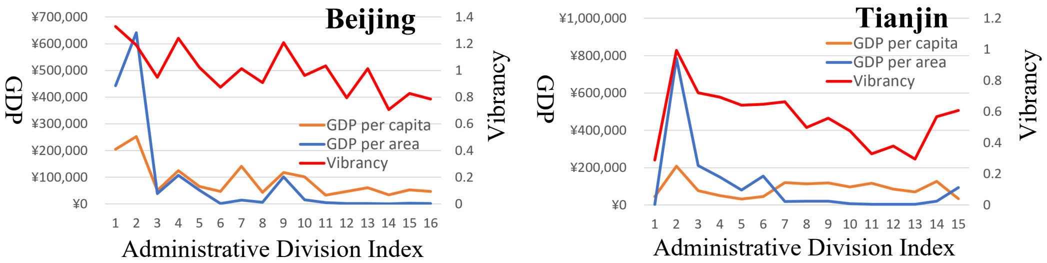
$$\sigma(H_C) < E(H_C) \cdot \frac{\rho(T_i) \cdot \sqrt{\rho^2(h_i) + 1}}{\sqrt{N}} \tag{19}$$

In our data set measured at Beijing, we estimate the above measures as $\rho(T_i) = 1.414$, $\rho(h_i) = 1.95$ (vibrancy), $\rho(h_i) = 1.18$ (commutation). For more than 99.5% location records in our data, the underlying grid cell has a minimal number of records larger than 3000 ($N \geq 3000$). Finally, we have $\sigma^{(}H_C) < 0.057 \cdot E(H_C)$. The 95% confidence interval of the error rate of $\widetilde{H}_C$ is 11.4% ($2 \cdot 0.057$).

C. Semantics of Vibrancy and Commutation metrics

Vibrancy: We compute the average vibrancy of location records in each administrative division (DIV) in Beijing and Tianjin using our urban trajectory data set. We also obtain the GDP, population, and area size information of these divisions from national survey data. It is hypothesized that the vibrancy metric can indicate people's living quality. The region with more high vibrancy people can have stronger economy in general. As shown in Figure 2, in both the city of Beijing and Tianjin, the vibrancy metric in all DIVs shows a correlation with GDP per capita and GDP per area in the same DIV. The Pearson correlation coefficients between vibrancy and these two GDP metrics are 0.77 and 0.66 in Beijing, and 0.37 and 0.79 in Tianjin. The observation on urban data supports our hypothesis.

Commutation: Similarly, we compute the commutation metric for each trajectory in our urban data set. We also compute the displacement (travel distance) of each trajectory for comparison. Because of the sparse nature of the data set, we can not detect most of travel segments in the whole trajectory. Instead, we sum together the displacement between any two consecutive stay segments, for which the detection rate is much higher. The Pearson correlation coefficient between the



(a) (b)

Fig. 2: The correlation among the vibrancy metric and GDP per capita/per area in the city of Beijing and Tianjin.

displacement per record and the commutation metric reaches 0.34, indicating a moderate correlation that supports the usage of the commutation metric. Note that we use displacement per record because the total travel distance measured in a trajectory is correlated with the number of location records on that trajectory data.

For diversity and fluidity, the semantics of metrics can be derived directly from their mathematical definitions.