Privacy-Preserving Cycle-Based Arrival Profile Estimation Based on Cross-Company Connected Vehicles

Tan, Chaopeng; Yao, Jiarong; Tang, Keshuang; Liang, Jinhao; Yin, Guodong

DOI
10.1109/TCE.2025.3525847

Publication date
2025

Document Version
Final published version

Published in
IEEE Transactions on Consumer Electronics

Privacy-Preserving Cycle-based Arrival Profile Estimation Based on Cross-company Connected Vehicles

Chaopeng Tan, Jiarong Yao*, Keshuang Tang, Jinhao Liang*, Guodong Yin

***Abstract*—Cycle-based arrival profiles can describe temporal demand distribution within a signal cycle for signalized intersections, which can be used to calculate indicators such as traffic volume, queue length, and facilitate fine-grained signal control. However, few studies address cycle-level arrival profile estimation based on connected vehicles (CVs). Besides, studies addressing privacy issues for cross-company collaboration in traffic management are still in their infancy. To fill these research gaps, this study proposes a data-driven method for privacy-preserving cycle-based arrival profile estimation using cross-company CV data. The cyclic arrival curve is discretized as an arrival rate vector whose elements are calculated using sampled CV trajectories, thus transforming the arrival profile estimation into a matrix completion problem. Considering cross-company collaboration, a privacy-preserving technique, secure multi-party computation, is used to encrypt initial arrival rate matrices of multiple companies. In particular, a perturbation approach is combined to enhance protection against inference attacks with prior knowledge of the matrix construction process. Then, matrix completion is realized through a singular value thresholding (SVT) algorithm, meanwhile achieving denoising. Empirical evaluation shows that the estimation accuracy of traffic volume and queue length derived from the proposed arrival profile estimation method can reach 87.6% and 78.4%, respectively, meanwhile protecting the privacy of multiple participating companies and outperforming existing methods. Simulation evaluation on a large-scale network further demonstrates the reliability of the proposed method considering ever-changing demand scenarios. A comprehensive sensitivity analyses exhibit its robustness to CV sample size, number of participating parties and data disparity, showing wide popularization and application prospects.**

***Index Terms*—Privacy preservation, arrival profile estimation, matrix completion, secure multi-party computation, singular value thresholding, connected vehicle**

I. Introduction

TRAFFIC volume plays a significant role in urban traffic management, which provides valuable insights into the traffic flow patterns and contributes to traffic signal evaluation and optimization at signalized intersections [1], [2] . In contrast to hourly traffic volume, which is usually an input of the design of fixed-time signal timing schemes and reflects the average level of traffic arrivals, cyclic traffic volume describes the fluctuation of traffic demand during a specific period and can serve for real-time traffic signal control. Recently, the study of Tan et al. [3] pointed out that the traffic signal performance can be further improved by considering the arrival distribution during the cycle, i.e., the cyclic arrival profile. In addition, in the context of fine-grained traffic management, the cyclic arrival profile depicts the operation of the traffic flow in more detail and can be further used to estimate more cycle-based parameters such as traffic volume, queue length, delay, and the utilization of green time. Therefore, accurate restoration of cycle-level arrival profiles at signalized intersections is crucial for the current refined urban traffic management.

With the advancement of connected technologies, connected vehicles (CVs) are becoming more and more popular, including floating cars, probe vehicles (PVs), autonomous vehicles (AVs), etc. Such connected vehicles can be utilized to infer traffic information while cruising around the roadway networks, thus bringing new opportunities for traffic flow parameter estimation, e.g., traffic arrival estimation [4]–[8], queue length estimation [9]–[14], and total delay estimation [15]–[17]. However, there are challenges in applying CV data, i.e., CVs only account for a portion of the total population, and we need to restore the operation of the traffic flow based on these sampled data [18]. Specific to the object of this study, i.e., traffic arrival estimation based on CVs, early studies focused on estimating aggregated volume over a relatively long period, ranging from 15 min to 60 min. For instance, Zheng and Liu [6] aggregated CV trajectories into a common cycle to obtain fixed vehicle arrival patterns, based on which a maximum likelihood estimator was formulated for the average arrival rate of a certain period. Later, to obtain the traffic variations during the period, studies have been devoted to estimating cycle-based traffic volume based on CV data, utilizing model-based [8], [19]–[21] or data-driven approaches [7]. For example, Yao et al. [21] developed a mode-based hybrid method integrating shock-wave theory and maximum likelihood estimation for queued volume and non-queued volume, respectively. Tan et al. [8] proposed a maximum a posteriori method to jointly

*Corresponding author: Jiarong Yao & Jinhao Liang

Manuscript received xxx; revised xxx. This research is supported by National Natural Science Foundation of China (NSFC) (52372319).

Chaopeng Tan is with the Department of Transport and Planning, Delft University of Technology, Delft, Netherlands 2628 CN (email: tantantan951122@gmail.com).

Jiarong Yao is with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798 (e-mail: jiarong.yao@ntu.edu.sg).

Keshaung Tang is with the Key Laboratory of Road and Traffic Engineering of the Ministry of Education, College of Transportation Engineering, Tongji University, Shanghai, China 201804 (e-mail: tang@tongji.edu.cn).

Jinhao Liang is with the Department of Civil and Environmental Engineering, National University of Singapore, Singapore, 119077, (e-mail: jh.liang@nus.edu.sg).

Guodong Yin is with the School of Mechanical Engineering, Southeast University, Nanjing, China 211189. (e-mail: ygd@seu.edu.cn).

estimate multi-stream traffic demands cycle by cycle. As a pioneer study using data-driven approaches, Tang et al. [7] formulated the cycle-based volume estimation into a tensor completion problem and solved it efficiently with Tucker decomposition. These studies have demonstrated promising estimation accuracy of traffic volume, especially when the penetration rate is above 20%.

To the best of our knowledge, the most detailed traffic arrivals estimated by existing studies based on CVs are cycle-based. While cycle-based arrivals can reflect traffic variations from cycle to cycle during the period, it does not reflect the vehicle arrival patterns within the cycle, i.e., the arrival profile during the cycle. Such arrival profiles can be used to further calculate parameters such as queue length and vehicle delay, which pave the way for refined evaluation and optimization of signalized intersections. For instance, Tan et al. [17] developed a CV-based cumulative flow diagram method to profile the period average traffic flow operations in the scale of a signal cycle, which is further extended to a cumulative flow diagram model for fixed-time traffic signal optimization [3]. Nevertheless, these studies focus on the period average arrival profile and do not achieve cycle-based arrival profile estimation.

On the other hand, while the overall penetration rate of CVs, which are mainly operated by mobility companies including ride-hailing companies such as Uber and Wejo, and navigation companies such as Amap and TomTom, in the current transport network has exceeded 20%, the penetration rate of CV data available to an individual company may not exceed 10% in practical applications. This is because CV data is precious to every company and contains commercial secrets, so companies have concerns about sharing CV data with others from a privacy-preservation point of view. Currently, although a few studies have attempted to protect personal-level privacy in the context of traffic systems to resolve potential legal and ethical considerations associated with CV data sharing [22]–[24], studies explicitly addressing the privacy issue that arise between the collaboration of different companies are in infancy, let along studies focusing on traffic state estimation based on CVs from multiple parties.

In summary, we conclude two research gaps among existing studies: (i) few studies have achieved detailed cycle-based arrival profile estimation based on CVs; (ii) few studies have considered company-level privacy protection in CV-based traffic state parameter estimation. To fill these research gaps, this study proposes a privacy-preserving matrix completion method for cycle-based arrival profile estimation at signalized intersections based on cross-company CV data. The proposed method transforms the cycle arrival profile estimation problem into a matrix completion problem and efficiently solves it with the Singular Value Thresholding (SVT) algorithm. Considering privacy-preserving, a secure multi-party computation (SMPC) technique and a perturbation approach are integrated to collaboratively generate the incomplete initial matrix without revealing the privately held CV data of the participating companies. Given such a detailed arrival profile, the cycle-based traffic state parameters, e.g., traffic volume and queue length, can be further estimated.

The major contribution of this study is three-fold.

- We transform the cycle-based arrival profile estimation problem into a purely data-driven matrix completion problem that is free from the assumption of arrival distribution/pattern or prior historical information. Such a data-driven approach is simple to implement and can be easily scaled to large-scale road networks.
- To the best of our knowledge, we achieve the most detailed arrival profile estimation sorely based on CV data, which can be further used to estimate traffic volumes or other traffic state parameters. The proposed method also shows its capability in oversaturated situation and real-time application, which is crucial for refined evaluation and optimization of signalized intersections.
- This study integrates SMPC and a perturbation method that resists inference attacks in third-party untrustworthy scenarios, which is among the pioneering work that considers privacy issues at the company level for traffic management based on CV data. The study can be effective in facilitating collaboration between mobility companies to better exploit all available sampled CV data from the road network for more accurate traffic state estimation.

II. Problem Statement

We consider the following scenarios where several mobility companies operate their own CVs, and the traffic signal operation center (TSOC) requires these companies to share their CV data in a certain form for traffic arrival estimation at signalized intersections. Motivated primarily by their commercial interests, all involved companies are considered honest but curious, which means that they are legitimate participants who will faithfully follow the communication protocol but may attempt to infer sensitive information about other companies from intermediate results exchanged during the collaboration. Second, TSOC is considered untrusted. This implies that the TSOC may be subject to attacks or attempts to access the private data of the participating companies.

The objective of this study is to estimate the cycle-based vehicle arrival profiles at the signalized intersection. As presented in Fig. 1, with signal timing parameters and CV trajectories as input, we can derive the arrival rate between queued CVs given their queuing position and expected arrival time. Nevertheless, vehicle arrivals after the last queued CVs are still unknown. Then our task is to utilize the incomplete arrival profile to estimate the complete arrival profile. Essentially, the arrival profiles in all cycles can be regarded as arrival rate vectors in terms of time in cycles with vacant elements. Therefore, for a specific analysis period, an incomplete arrival rate matrix can be modeled and our task is to complete this matrix. Note that, in the context of the scenarios with CVs operated by several mobility companies, a privacy-preserving mechanism is required for these companies to collaboratively generate the incomplete arrival rate matrix without revealing their private data. Then, upon receiving the incomplete arrival rate matrix, the TSOC assumes the responsibility for cycle-based arrival profile estimation by completing the matrix. In

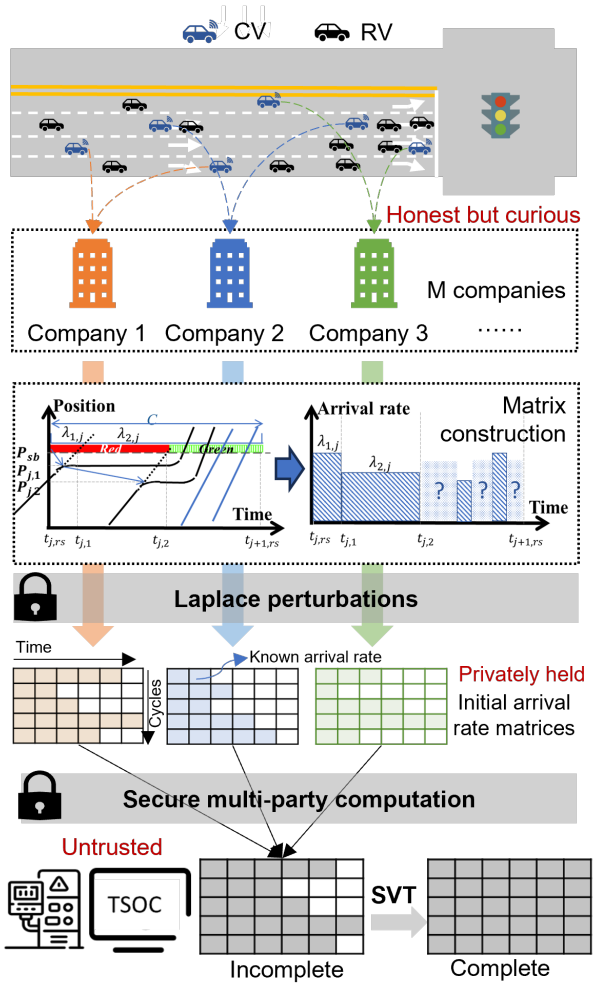

Fig. 1. Research scenario and problem statement.

particular, although the construction of the initial matrix is done by the companies, we assume that TSOC has a priori knowledge of the initial matrix construction process.

III. Methodology

The core of the proposed method is to construct an initial incomplete arrival rate matrix for all the cycles in the time-of-day (TOD) period based on cross-company CV data in a privacy-preserving manner, and then use the SVT matrix completion algorithm to obtain a complete arrival rate matrix. The advantage of the proposed data-driven method is that it eliminates the need for traffic flow assumptions required by model-based methods and is easy to apply by traffic engineers. Note that although a similar data-driven method by [7] implemented cycle-based volume estimation using tensor decomposition, the method cannot be further extended for more detailed cyclic arrival profile estimation in this study. In addition, privacy issues and oversaturated traffic conditions were not considered in [7].

Specifically, the proposed method comprises three major steps. In Section III-A, each company first exploits CVs to construct the initial incomplete arrival rate matrix, which is privately held by the companies. Then, in Section III-B,

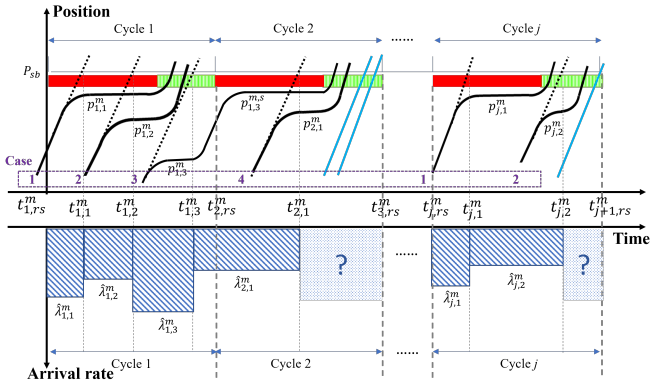

Fig. 2. Known arrival rate derived from stopped CVs.

all companies apply perturbation approaches to encrypt their private arrival rate matrices and use SMPC for aggregation to obtain an incomplete arrival rate matrix which is sent to TSOC. Finally, in Section III-C, TSOC applies the SVT algorithm to complete the incomplete arrival rate matrix aggregated in a privacy-preserving manner, thus achieving cycle-based arrival profile estimation.

A. Initial arrival rate matrix construction based on CVs

We assume that there are $M(M \geq 2)$ companies operating CVs, indexed by superscript m, involved in traffic arrival profile estimation. The given analysis period consists of $J(J \geq 2)$ cycles, indexed by subscript j. For the j-th cycle, K_j^m denotes the total number of queued CVs operated by company m, indexed by subscript k. Specifically, we use $\hat{\lambda}_{j,k}^m$ to denote the arrival rate of the interval between the expected arrival time of two consecutive CV k and CV k' operated by company m in cycle j. $p_{j,k}^m$ and $t_{j,k}^m$ denote the corresponding queuing position (measured from the upstream of the link) and the expected arrival time (i.e., arrival time assuming no delay will be experienced). Note that in the context of oversaturated intersections, CVs may stop multiple times. For ease of methodology introduction, we assume that the CV is queued up at most twice within the scale of the studied intersection. In particular, for a twice-queued CV, $p_{j,k}^m$ denotes the first-time queuing position while $p_{j,k}^{m,s}$ specifically denotes the second-time queuing position. Given the above critical parameters extracted from CVs, for each company m with privately operating CV trajectories, we can construct a privately held arrival rate matrix as follows.

As presented in Fig. 2, for each queued CV trajectory observed during the analysis period, we can classify it into four cases based on its own and previous CV trajectory types as below,

Case 1: The subject CV k is the first queued CV during the cycle. In this case, the subject CV has no previous queued CV, then the arrival rate before this CV is calculated as

$$\begin{cases} \hat{\lambda}_{j,k}^m \mid \hat{T}_{j,k}^m = \frac{p_{sb}-p_{j,k}^m}{t_{j,k}^m - t_{j,rs}} \frac{1}{L_s}, \\ \hat{T}_{j,k}^m = t_{j,k}^m - t_{j,rs}, \end{cases} \tag{1}$$

where $\hat{\lambda}^m_{j,k} \mid \hat{T}^m_{j,k}$ indicates that the arrival rate during the period $\hat{T}^m_{j,k}$ is $\hat{\lambda}^m_{j,k}$. $\hat{T}^m_{j,k}$ denotes the corresponding interval of the arrival rate, whose physical meaning is the time interval $[t_{j,rs}, t^m_{j,k}]$. L_s is the average spacing headway of queued vehicles. p_{sb} denotes the position of the stop bar. $t_{j,rs}$ denotes the start time of the red phase.

Case 2: The subject CV k is a once-queued CV following a once-queued CV k'. In this case, we can easily calculate the arrival rate between these two consecutive queued CVs as

$$\begin{cases} \hat{\lambda}^m_{j,k} \mid \hat{T}^m_{j,k} = \frac{p^m_{j,k'} - p^m_{j,k}}{t^m_{j,k} - t^m_{j,k'}} \frac{1}{L_s}, \\ \hat{T}^m_{j,k} = t^m_{j,k} - t^m_{j,k'}, \end{cases} \tag{2}$$

Case 3: The subject CV k is a twice-queued CV following a once/twice-queued CV k'. In this case, the calculation of the arrival rate is the same as Eq. (2), where only the first-time queuing positions of CV k and k' are used.

Case 4: The subject CV k is a once-queued CV following a twice-queued CV k'. In this case, the twice-queued CV k' is a residual vehicle from the last cycle, then its second-time queuing position is used for arrival rate calculation as below

$$\begin{cases} \hat{\lambda}^m_{j,k} \mid \hat{T}^m_{j,k} = \frac{p^{m,s}_{j,k'} - p^m_{j,k}}{t^m_{j,k} - t^m_{j-1,k'}} \frac{1}{L_s}, \\ \hat{T}^m_{j,k} = t^m_{j,k} - t^m_{j-1,k'}, \end{cases} \tag{3}$$

In particular, $\hat{T}^m_{j,k}$ overlaps two cycles in Case 4, as $t^m_{j-1,k'}$ in this case is typically smaller than $t_{j,rs}$ and $t^m_{j,k}$ is greater than $t_{j,rs}$. Thereby, for each cycle j, an arrival rate vector $\hat{\boldsymbol{\lambda}}^m_j$ can be obtained by queued CVs as below

$$\hat{\boldsymbol{\lambda}}^m_j = \left[\{ \hat{\lambda}^m_{j,k} \mid \hat{T}^m_{j,k} \}_{k=1,2,\cdots,K^m_j} \right], \tag{4}$$

Note that, the number of queued CVs of different companies in different cycles varies, resulting in a different size K^m_j of the arrival rate vectors. In order to construct the arrival rate matrix for later matrix completion algorithm and to pave the way for privacy-preserving aggregation of such arrival rate matrices of different companies, a standardization process is applied to normalize the size of $\hat{\boldsymbol{\lambda}}^m_j$, as presented in Fig. 3. Here we use Δ_j to denote a unit time period of cycle j (e.g., 5 s or 10 s), which makes $\lceil C_j/\Delta_j \rceil = I$. C_j denotes the cycle length of cycle j and I denotes the number of columns of the standard arrival rate vector (the same for all cycles and all participating companies). The normalized arrival rate vector $\boldsymbol{\lambda}^m_j$ is written as

$$\boldsymbol{\lambda}^m_j = \left[\{ \lambda^m_{j,i} \}_{i=1,2,\cdots,I} \right], \tag{5}$$

where $\lambda^m_{j,i}$ is the initial arrival rate of the i-th time period $T^m_{j,i}$ in cycle j of company m, which is calculated as

$$\lambda^m_{j,i} = \frac{\sum_{\delta \in T_{j,i}} \lambda^m_{j,\delta} \delta}{\Delta_j} \tag{6}$$

$$T_{j,i} = \tau_{j,i} - \tau_{j,i-1}, \tag{7}$$

$$\tau_{j,i} = t_{j,rs} + i\Delta_j. \tag{8}$$

where δ is the minimum time unit after discretization of time (e.g., 0.5 s or 1 s). $\tau_{j,i}$ is the right bound of the unit time period i in cycle j of company m. $\lambda^m_{j,\delta}$ is the corresponding arrival rate according to Eq. (4). Note that for those time periods that only partially overlap time units with known arrival rates, the weighted value of the arrival rates for these overlapping time units is used directly as the estimate. Besides, for those time periods located in the late cycle where no queued CVs are observed, whose $\tau^m_{j,i-1} > t^m_{j,K^m_j}$, the number of overlapping time units is 0, making $\lambda^m_{j,i}$ unknown.

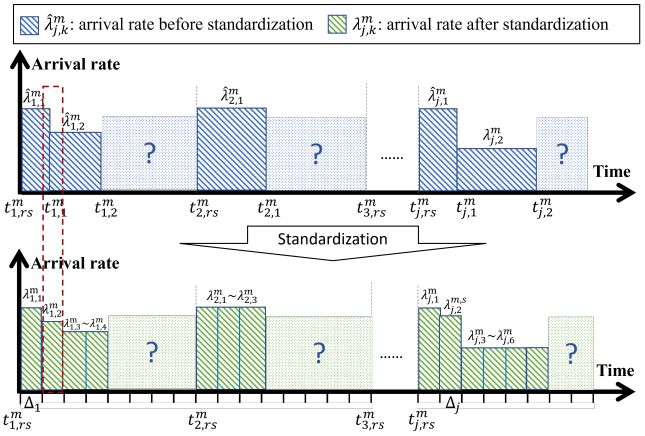

Fig. 3. Standardization of initial arrival rate matrix.

In this way, the arrival rate vectors of different cycles, whose cycle length may vary, can be normalized into the same size, which is however incomplete. Therefore, given these normalized arrival rate vectors of J cycles in the analysis period, an incomplete arrival rate matrix $\boldsymbol{\Lambda}^m$, privately held by company m, with size $J \times I$ can be obtained as below

$$\boldsymbol{\Lambda}^m = [\boldsymbol{\lambda}^m_1, \boldsymbol{\lambda}^m_2, \cdots, \boldsymbol{\lambda}^m_J]^T, \tag{9}$$

It is worth noting that, $\boldsymbol{\Lambda}^m$ is constructed using queued CVs only, while non-queued CVs also provide certain arrival information of the traffic flow. Here we further use non-queued CVs to estimate some unknown $\lambda^m_{j,i}$ in the arrival rate matrix $\boldsymbol{\Lambda}^m$.

For each cycle j, we can first estimate a prior penetration rate p^m_j as below,

$$p^m_j = \frac{K^m_j}{l \sum_{k=1}^{K^m_j} \hat{\lambda}^m_{j,k} \hat{T}^m_{j,k}} \tag{10}$$

where l is the number of lanes of the studied movement.

Then, for each time period $T^m_{j,i}$ of the arrival rate matrix whose value is unknown, we can count the number $N^m_{j,i}$ of non-queued CVs arriving within this period and obtain an estimate of the arrival rate of this period as below,

$$\lambda^m_{j,i} = \frac{N^m_{j,i}}{l p^m_j \Delta_j} \tag{11}$$

Note that for those steps whose $N^m_{j,i} = 0$, their corresponding arrival rate values in the arrival rate matrix are still unknown, thus the incomplete arrival rate matrix $\boldsymbol{\Lambda}^m$ privately held by company m is constructed using the information of all CV trajectories it has, no matter queued or non-queued.

B. Privacy-preserving aggregation of cross-company matrices

In the context of multiple mobility companies collaborating for arrival profile estimation, each company privately held an initial incomplete arrival rate matrix based on its operating CV data. We aim to aggregate these arrival rate matrices into one for arrival profile estimation by matrix completion methods, which makes fuller use of cross-company CV data than the matrix of a single company, without compromising confidential information of each company. This section presents a privacy-preserving aggregation method based on SMPC and perturbation techniques for cross-company arrival rate matrices.

1) Adding perturbations against inference attacks: In particular, from Fig. 2 we can find that, the moment of arrival of the CV trajectory corresponds exactly to the moment when the arrival rate changes. Although the subsequent standardization process presented in Fig. 3 can blur out this exact moment, we can still locate the unit time period at which the CV arrives through the relative magnitude relationship of the values in the initial matrix. Note that aggregating initial matrices from different companies in Eq. (16) does not change this phenomenon, even if additive secret sharing is used, because it does not change the aggregation result. Of course, when there are many CVs and the unit time period is large, almost every unit time period has a different value, which does not compromise privacy. But, when CVs are very limited and the unit time period is small, it is easy to infer which unit time periods have CVs passing through. For example, given a vector of arrival rates $[0.1, 0.05, 0.05, 0.05, ...]$ with 5 s for each time unit, adversaries with a priori knowledge of the initial matrix construction process can infer that there is a CV arriving around the moment of 5 s. In such a case, although we cannot know which company the CV belongs to so as not to compromise the privacy of companies, it still discloses personal privacy.

The key to resisting such an inference attack is to perturb consecutively identical arrival rate values (e.g., $[..., 0.05, 0.05, 0.05, ...]$ in the example), in this way the arrival rate varies from unit to unit and therefore the arrival time of CVs cannot be inferred. In this study, we propose a perturbation approach below.

Assuming that a sequence of Z arrival rates $[a_1, a_2, ..., a_Z]$ is extracted from a row of arrival rate matrix by company m and $a_z = a > 0$ ($z = 1, 2, \cdots, Z$). For brevity, the company index is omitted. The corresponding perturbed sequence is denoted as $[p_1, p_2, ..., p_Z]$. The proposed approach requires that

$$p_z = a_z + \eta_z, \quad z = 1, 2, \cdots, Z, \tag{12}$$

$$\eta_z \sim Lap(\varepsilon), \quad z = 1, 2, \cdots, Z. \tag{13}$$

where η_z is a randomly added perturbation following the Laplace distribution $Lap(\varepsilon)$, whose probability density function is $f(\eta) = \frac{1}{2\varepsilon} e^{-\frac{|\eta|}{\varepsilon}}$. Therefore, we have

$$\sum_{z=1}^{Z} p_z = Za + \sum_{z=1}^{Z} \eta_z \sim Lap(Za, Z\varepsilon) \tag{14}$$

$$\text{where} \quad \mathbb{E}[\sum_{z=1}^{Z} p_z] = Za \text{ and } \mathbb{V}[\sum_{z=1}^{Z} p_z] = 2Z^2\varepsilon^2. \tag{15}$$

This suggested that the expected summation of perturbed arrival rates is the same as the real one. This is desired because this sum indicates the number of vehicle arrivals between neighboring CVs and the value is reliable. As for the specific arrival rate for each time unit, the initial matrix itself is making an approximation with the average value, which is not always accurate. Here we add the perturbation which is actually equivalent to a randomized approximation method.

2) SMPC based on additive secret sharing: Since we have normalized the incomplete arrival rate matrix of each company into the same size, without considering privacy, the final arrival rate matrix should be aggregated by TSOC as follows,

$$\mathbf{\Lambda}^0 = \frac{\mathbf{\Lambda}^1 + \mathbf{\Lambda}^2 + \cdots + \mathbf{\Lambda}^M}{M} \tag{16}$$

Element-wise, we have

$$\lambda_{j,i}^0 = \frac{\sum_{m=1}^{M} \lambda_{j,i}^m}{M} \tag{17}$$

Note that, Eq. (17) only involves the summation of arrival rate values as the number of participating companies M is a known value to the TSOC. Therefore, we can apply an SMPC technique, additive secret sharing [23], [25], to achieve privacy-preserving summation of private data from participating companies. Note that additive secret sharing is a randomly-shared secure summation protocol. In practice, there are many similar protocals that can be used as alternatives to achieve the same effect, such as basic secure summation, encrypted secure summation, distributed secure summation, etc. A detailed review can be referred to [26]. The basic idea of additive secret sharing is shown in Fig. 4 and presented as follows. Recall that there are M companies in our cases and one company wants to share a secret value λ with the others. Instead of directly disclosing the value of λ, this agent creates N shares $s_1, s_2, ..., s_M$ out of λ such that

$$\lambda = \sum_{m=1}^{M} s_m \pmod{\rho} \tag{18}$$

where shares $s_1, s_2, ..., s_{M-1}$ are chosen independently and uniformly at random from the set $\{0, 1, 2, ..., \rho - 1\}$; ρ is a given large prime integer; s_M is determined by

$$s_M = \lambda - \sum_{m=1}^{M-1} s_m \pmod{\rho} \tag{19}$$

Then, each share s_m ($m = 1, 2, \cdots, M$) is sent to company m. To reconstruct λ, M companies, or the third party TSOC, can simply add all shares modulo ρ together since we have Eq. (18). Note that additive secret sharing can withstand any collusion of up to $M - 1$ agents since colluded agents only have random values that do not depend on λ. Besides, the additive secret sharing protocol maintains the accurate summation result. Note that, in our cases, private data λ, indicating the arrival rate, is smaller than 1. Thus we can let all companies and third parties agree to share the integer part of $10,000\lambda$, and the third party can divide the result by 10,000

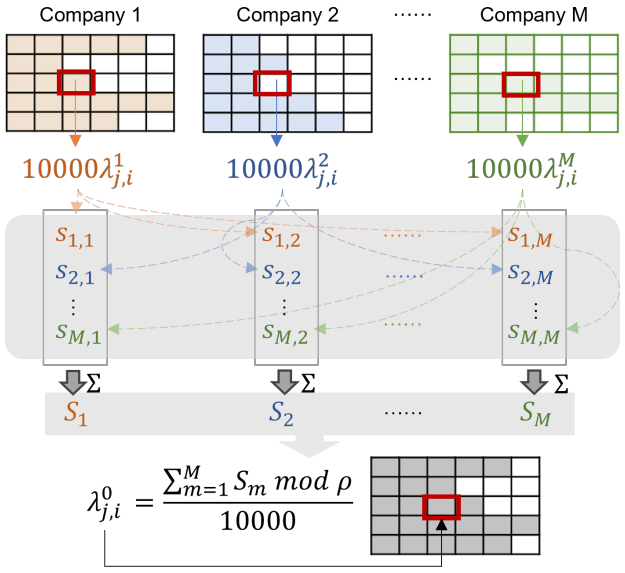

Fig. 4. Illustration of privacy-preserving aggregation of arrival rate matrices from companies.

after reconstructing it. Such a multiplier can be larger to obtain higher calculation accuracy.

In summary, we can apply Algorithm 1 to achieve Eq. (16) based on additive secret sharing protocol without compromising any private data of participating companies.

Algorithm 1 Privacy-preserving aggregation of arrival rate matrices by additive secret sharing

INPUT: Arrival rate matrices $\mathbf{\Lambda}^1, \mathbf{\Lambda}^2, \cdots, \mathbf{\Lambda}^M$ of M participating companies (multiplied by 10,000 already) with size $J \times I$; a big prime ρ;
For $i = 1, 2, \cdots, I$ **do**
  For $j = 1, 2, \cdots, J$ **do**
    For $m = 1, 2, \cdots, M$ **do**
      Company m uniformly and randomly samples s_{mn} from $\{1, 2, ..., \rho - 1\}$ and sends data s_{mn} to company n **for** $n = 1, 2, \cdots, M$ and $n \neq m$;
      Company m calculates $s_{mm} = \lambda_{j,i}^m - \sum_{n \neq m} s_{mn} \mod \rho$;
    end
    Company n calculates $S_n = \sum_{m=1,2,\cdots,M} s_{mn} \mod \rho$ and sends it to TSOC **for** $n = 1, 2, \cdots, M$;
    TSOC calculates $\lambda_{j,i}^0 = \frac{1}{10,000}(\sum_{m=1,2,\cdots,M} S_m \mod \rho)$
  end
end
RETURN: $\mathbf{\Lambda}^0$

C. Matrix completion for arrival profile estimation

In this step, TSOC has received the arrival rate matrix calculated from cross-company CV data, which is however incomplete and noisy. Therefore, estimating cycle-based arrival profiles is transformed into a noisy matrix completion problem. In this section, a popular numerical algorithm, SVT algorithm [27], is adopted for matrix completion, which can excavate the temporal correlation of intra-cycle arrival and inter-cycle difference of arrival fluctuation, finally obtaining a complete arrival rate matrix after iterative convergence of the nuclear norm of the matrix. Specifically, the SVT algorithm decomposes the matrix into its singular values and corresponding vectors, then isolates and thresholds these singular values. This approach filters out noise by shrinking the smaller singular values, which are more susceptible to noise, while preserving the larger, more significant singular values that represent the true underlying arrival patterns.

According to [27]–[29], low-rank matrix recovery problems such as matrix completion can be transformed into solving nuclear norm minimization, as given by Eq. (20) -(21),

$$\min_{\mathbf{\Lambda}} \quad ||\mathbf{\Lambda}||_* \tag{20}$$

$$s.t. \quad P_\Omega(\mathbf{\Lambda}) = P_\Omega(\mathbf{\Lambda}^0), \tag{21}$$

where $\mathbf{\Lambda}$ is the completed arrival rate matrix. $||\cdot||_*$ denotes the calculation of the nuclear norm, i.e., the summation of singular values of the matrix. Ω is the set of indices of known elements obtained by Section III-A. P_Ω is the operator selecting elements according to set Ω.

Then, starting with $\boldsymbol{X}^0 = \mathbf{0} \in \mathbb{R}^{J \times I}$, the SVT algorithm [27] gives the iterative equations below to solve the problem with high quality and efficiency,

$$\begin{cases} \mathbf{\Lambda}^a = \mathcal{D}_\gamma(\boldsymbol{X}^{a-1}), \\ \boldsymbol{X}^a = \boldsymbol{X}^{a-1} + \mu P_\Omega(\mathbf{\Lambda}^0 - \mathbf{\Lambda}^a), \end{cases} \tag{22}$$

where the superscript a denotes the number of iterations. γ is a constant comparison threshold value and $\gamma > 0$. μ is the step size at the iteration a. $\mathcal{D}_\gamma$ is the soft-thresholding operator, which is defined as

$$\mathcal{D}_\gamma(\boldsymbol{X}) = \boldsymbol{U}\mathcal{D}_\gamma(\mathbf{\Sigma})\boldsymbol{V}^T, \tag{23}$$

$$\mathcal{D}_\gamma(\mathbf{\Sigma}) = diag(\{(\sigma_r - \gamma)_+\}_{i=1,2,\cdots,R}), \tag{24}$$

where we use R to denote the rank of the matrix $\boldsymbol{X}$. $\{\sigma_r - \gamma\}_+$ is an operation that filters positive numbers, i.e., $\{\sigma_r - \gamma\}_+ = \max\{0, \sigma_r - \gamma\}$. According to singular value decomposition (SVD), we have

$$\boldsymbol{X} = \boldsymbol{U}\mathbf{\Sigma}\boldsymbol{V}^T, \tag{25}$$

$$\mathbf{\Sigma} = diag(\{\sigma_i\}_{i=1,2,\cdots,R}). \tag{26}$$

where $\boldsymbol{U} \in \mathbb{R}^{J \times R}$ and $\boldsymbol{V} \in \mathbb{R}^{I \times R}$ are matrices with orthonormal columns. σ_i $(i = 1, 2, \cdots, R)$ are positive singular values. That is to say, when we apply the soft-thresholding operator $\mathcal{D}_\gamma(\boldsymbol{X})$ to a matrix $\boldsymbol{X}$ as presented in Eq. (23), we actually only need to perform the soft-thresholding operator $\mathcal{D}_\gamma(\mathbf{\Sigma})$ on its SVD-decomposed diagonal matrix $\mathbf{\Sigma}$, and then restored the matrix by integrating $\boldsymbol{U}$ and $\boldsymbol{V}$. This property is what makes the SVT algorithm so efficient.

When the iteration reaches predefined stopping criteria, the final solution of Eq. (22) is the approximated solution of Eq. (20), namely, the completed matrix $\mathbf{\Lambda}$ can be obtained. The implementation process of the SVT algorithm is summarized in Algorithm 2.

In Algorithm 2, $||\cdot||_F$ indicates the Frobenius norm. $[\cdot]_\beta$ indicates the first β columns/elements for a matrix/vector. As

Algorithm 2 Matrix completion by SVT algorithm

INPUT: Incomplete arrival rate matrix $\boldsymbol{\Lambda}^0$ with size $J \times I$; the set of indices of known element Ω, comparison threshold value γ, step size μ, maximal iteration number a_m, stopping criteria ϵ.
Set the iteration number $a = 0$ and initialize $\boldsymbol{X}^0$.
Calculate the residual $res = \frac{||P_\Omega(\boldsymbol{X}^a - \boldsymbol{\Lambda}^0)||_F}{||P_\Omega(\boldsymbol{\Lambda}^0)||_F}$.
While ($res > \epsilon$ and $a < a_m$) **do**
  Apply singular value decomposition to $\boldsymbol{X}^a$ using Eq. (25) and obtain $\boldsymbol{U}, \boldsymbol{\Sigma}, \boldsymbol{V}$.
  Extract the diagonal elements of $\boldsymbol{\Sigma}$ as a vector σ.
  Set $\beta = \max\{j : \sigma_j > \gamma\}$.
  Apply singular value thresholding to $\boldsymbol{U}, \boldsymbol{\Sigma}, \boldsymbol{V}$ as below, i.e.,
    $\boldsymbol{U} \leftarrow [\boldsymbol{U}]_\beta$,
    $\boldsymbol{V} \leftarrow [\boldsymbol{V}]_\beta$,
    $\sigma \leftarrow [\sigma]_\beta - \gamma$,
    $\boldsymbol{\Sigma} \leftarrow diag(\sigma)$.
  Compute $\boldsymbol{\Lambda}^{a+1} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$, $\boldsymbol{X}^{a+1} = \boldsymbol{X}^a + \mu P_\Omega(\boldsymbol{\Lambda}^0 - \boldsymbol{\Lambda}^{a+1})$.
  $a \leftarrow a + 1$.
  Calculate $res = \frac{||P_\Omega(\boldsymbol{X}^{a+1} - \boldsymbol{\Lambda}^0)||_F}{||P_\Omega(\boldsymbol{\Lambda}^0)||_F}$.
end While
RETURN: $\boldsymbol{\Lambda} \leftarrow \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$

suggested by [27], μ can be determined based on the size of the matrix to be completed, shown as below

$$\mu = \frac{1.2 * J * I}{|\Omega|}, \tag{27}$$

which determines the convergence speed. γ is a parameter to threshold singular values, which shrinks the smaller singular values by Eq. (24) thus filtering out noises, i.e., Laplace perturbations in this study. According to [30], γ is suggested to be greater than $r\sqrt{J * I}$ to filter out noises and r is the root mean square error of the noisy data which can be calibrated using a data-driven method by [31] (or standard deviation of added noise if known). In our cases, we use γ suggested by [27]:

$$\gamma = 5\sqrt{J * I}, \tag{28}$$

which satisfies the threshold criterion by [30] considering the magnitude of the matrix values. As for ϵ and a_m, it is common knowledge that a smaller ϵ and a larger a_m can take longer time to reach convergence, although the accuracy can be improved. In this study, we let $\epsilon = 0.0001$ and $a_m = 1000$.

Further, $\boldsymbol{X}^0$ can be initialized as below to save work,

$$\boldsymbol{X}^0 = \lceil \frac{\gamma}{\mu ||P_\Omega(\boldsymbol{\Lambda}^0)||_2} \rceil * \mu P_\Omega(\boldsymbol{\Lambda}^0). \tag{29}$$

where $\lceil \cdot \rceil$ indicates rounding toward positive infinity.

With the estimated complete arrival rate matrix, each row vector of the completed $\boldsymbol{\Lambda}$ can be used to calculate the arrival profile as presented by Eq. (30).

$$q_{j,i} = l * \Delta_j * \lambda_{j,i} \tag{30}$$

where $q_{j,i}$ is the arrival flow of interval i in cycle j (veh).

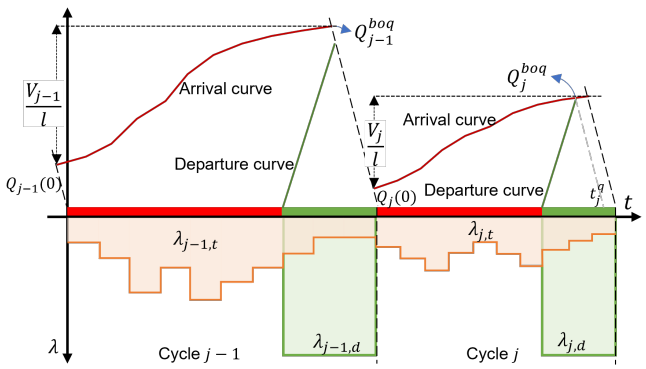

Fig. 5. Illustration of traffic state parameters estimation based on arrival profiles.

D. Traffic state parameter estimation based on arrival profiles

Combining signal timing information, we can further estimate the cycle-based traffic demand as well as the queue length based on these detailed arrival profiles, as illustrated in Fig. 5. Note that here traffic demand refers to vehicle arrivals during the cycle, which are different from throughput volume limited by green length. Traffic demand is independent of the signal timings and can describe traffic flow operations in both undersaturated and oversaturated scenarios. For cycle-based traffic demand, it is calculated as the sum of the products of the unit time period with the corresponding arrival rate based on the estimated arrival profile,

$$V_j = l * \Delta_j * \sum_{i=1}^{I} \lambda_{j,i}, \tag{31}$$

where V_j is the traffic demand of cycle j (veh).

Regarding the queue length, the real-time queue length can be estimated as the difference between the arrival curve, i.e., the integral of the arrival profile, and the dissipation curve that starts from the green start time with the saturated departure rate. For the sake of brevity, we represent the arrival profile as a continuum, $\lambda_{j,t}$ $(0 \le t \le C_j)$.

$$Q_j(t) = \max\{0, Q_j(0) + \int_0^t \lambda_{j,t}\, dt - \int_0^t \alpha_{j,t}\lambda_{j,d}\, dt\}, \tag{32}$$

$$\alpha_{j,t} = \begin{cases} 0, & \text{if } t \le g_j^s \\ 1, & \text{if } t > g_j^s \end{cases} \tag{33}$$

where g_j^s is the green start time of cycle j. $\lambda_{j,d}$ is the saturated departure rate of cycle j. $Q_j(0)$ is the initial queue of cycle j, which can be estimated by the arrival profile of the last cycle based on the flow conservation law, i.e., $Q_j(0) = \max\{0, Q_{j-1}(0) + \int_0^{C_{j-1}} \lambda_{j-1,t}\, dt - \int_0^{C_{j-1}} \alpha_{j-1,t}\lambda_{j-1,d}\, dt\}$.

Then, for undersaturated conditions where the queue can be cleared, the maximum back of queue (BOQ), namely the cumulative vehicles arriving before the moment that the physical queue is cleared, is estimated as

$$Q_j^{boq} = Q_j(0) + \int_0^{t_j^q} \lambda_{j,t}\, dt, \text{ if } Q_j(C_j) = 0 \tag{34}$$

where t_j^q is the minimum solution of solving $Q_j(t) = 0$. Note that, in the case that the cycle is oversaturated, i.e., $Q_j(t) > 0$

always holds, the queue length will continue to accumulate into the next cycle. At this point, all vehicles arriving during this cycle are stopped and accumulated in the queue, then we have

$$Q_j^{boq} = Q_j(0) + \frac{V_j}{l}, \text{ if } Q_j(C_j) > 0 \tag{35}$$

In particular, based on the detailed arrivals, we can further obtain other traffic state parameters such as total delay [17] and achieve fine-grained traffic signal timing optimization [3], which will not be discussed here.

IV. Evaluation

The proposed method is evaluated through one empirical case and one simulation case. The empirical case is used to evaluate the performance of arrival profile estimation by assessing the accuracy of traffic demands and queue lengths derived from arrival profiles. Then, the impact of the privacy-preserving mechanism on the arrival profile estimation and the capability of the proposed method in real-time estimation is also evaluated. In the simulation case, the proposed method is implemented on a large-scale network to show its applicability under various traffic conditions. Sensitive analysis in terms of the penetration rate of CV and the unit time period for matrix completion is also conducted to test the robustness of the proposed method.

A. Empirical Case

The study site of the empirical case is the intersection of Fuzhong Road and Huanggang Road in downtown Shenzhen, China, where the four through lanes in the southbound approach are chosen as the study object, as shown in Fig. 6. The analysis period started from 9:30 to 14:30 on April 13, 2017, comprising five signal schemes operated by SMOOTH adaptive signal control, as presented in Table I. Note that the actual signal timing varies from cycle to cycle and is averaged here. Video data were recorded to obtain the actual cycle-based traffic flow and back of queue during the analysis period, as shown in Fig. 6(b). The connected vehicle data, as shown in Fig. 6(c), of the straight-through flow were provided by DiDi Chuxing, a ride-sharing company. Preprocessing of the raw CV data is conducted through two steps. First, the position of the stop-line is determined using the most forward queuing position of the sample CVs, which is set to the spatial origin (coordinates increase in the direction of travel). Then, the initial and final positions of the trajectory were used to determine whether the trajectory was complete, i.e., passed the stop line, as illustrated in Fig. 6(d). If the initial position is greater than 0, then the trajectory starts at downstream of the stopline and is recognized as an anomaly. If the last position is less than 0, then the trajectory ends upstream of the stopline and is also recognized as an anomaly. Such anomalies or outliers are removed before the CV data are input to the proposed method. Some general statistics of the CV data after preprocessing are given in Table I, from which we can find that, although the average penetration rate of the study period was 8.57%, the average number of samples captured in each

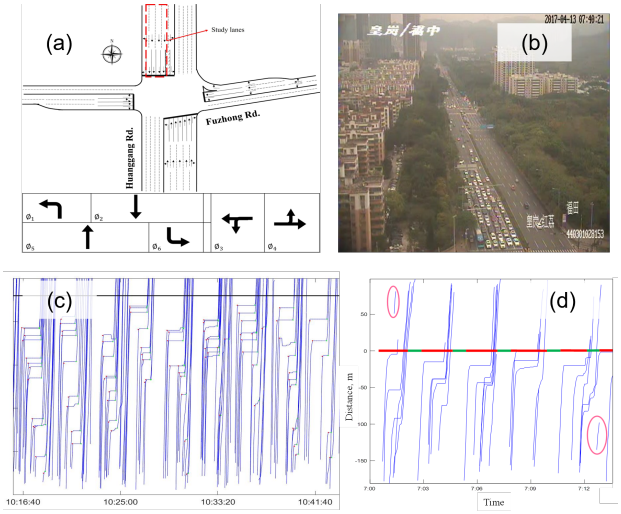

Fig. 6. Study site of the empirical case: (a) intersection layout, (b) screenshot of video data, (c) example CV trajectories, and (d) anomalies in raw CV data

TABLE I
Signal Timing Schemes and CV Data Information

TOD period	1	2	3	4	5
	9:30-10:30	10:30-11:30	11:30-12:30	12:30-13:30	13:30-14:30
Cycle length, C_m (s)	200	176	157	151	169
Green time (s)	105	83	80	79	84
Period volume (veh)	2677	2420	2148	1889	2376
Number of CVs	250	188	193	153	202
Penetration rate, p (%)	9.34	7.77	8.99	8.10	8.50
CVs per cycle (veh)	13.9	9.0	8.4	6.7	9.6

cycle was 9 veh as there are four lanes, which is a relatively good data quality from the perspective of absolute sample size.

Though the cycle-based arrival profile cannot be extracted and evaluated from the video data, its estimation accuracy can be evaluated laterally by traffic flow indicators derived from it, i.e., cycle-based traffic demand and queue length. Therefore, the mean absolute error (MAE) and mean average percentage error (MAPE) are used to evaluate the estimation accuracy of traffic demand and queue length derived from arrival profiles as below

$$MAE = \frac{1}{N}\sum_{n=1}^{N}|\hat{Y} - Y|, \tag{36}$$

$$MAPE = \frac{1}{N}\sum_{n=1}^{N}\frac{|\hat{Y} - Y|}{Y} \times 100\%, \tag{37}$$

where N is the number of estimates of traffic demand or queue length, $\hat{Y}$ (veh/cycle) denotes the estimates, and Y (veh/cycle) denotes the corresponding ground truth values.

1) Comparison with different existing methods:

Considering that the traffic demand level and pattern of vehicle arrivals vary over TOD periods, we implement the proposed matrix completion method separately for each TOD period for arrival profile estimation. For each TOD period, we let $\lceil I = C_m/\Delta_m \rceil$ for the standardization of the incomplete

TABLE II
OVERALL ESTIMATION ACCURACY OF CYCLE-BASED TRAFFIC DEMAND AND QUEUE LENGTH

Traffic state	MAE, veh (MAPE, %) TOD / Method	1	2	3	4	5	Overall
Traffic demand	SMPC-SVT	**9.9 (6.6)**	**12.8 (11.2)**	**11.4 (12.4)**	**6.1 (7.6)**	11.5 (10.8)	**10.3 (9.8)**
	Ban's [16]	42.8 (29.9)	44.6 (38.7)	34.2 (36.9)	32.4 (40.4)	45.5 (42.9)	38.5 (37.2)
	Zheng's [6]	39.7 (28.8)	23.4 (20.0)	22.2 (24.6)	32.0 (40.5)	30.0 (27.9)	29 (28.5)
	Hybrid [21]	31.2 (21.0)	22.5 (19.9)	17.9 (19.4)	18.3 (22.1)	**11.2 (10.3)**	19.8 (18.6)
Queue length	SMPC-SVT	**2.28 (11.30)**	**2.25 (11.18)**	**2.14 (13.42)**	2.59 (25.06)	**3.04 (11.97)**	**2.48 (14.87)**
	Yin's [14]	6.17 (29.96)	3.29 (16.69)	3.03 (18.95)	**1.83 (17.72)**	3.20 (12.26)	3.37 (18.68)
	Cheng's [10]	2.93 (14.99)	3.72 (19.47)	8.18 (60.93)	9.49 (91.60)	5.11 (19.42)	6.30 (47.49)

arrival matrix. In this case, we set $\Delta_m = 5$. For example, I of the first TOD period is calculated as $\lceil 200/5 = 40 \rceil$.

For *traffic demand* estimation, three representative methods were chosen for comparison.

- Ban's method [16]: A deterministic method using delay pattern of CVs. This method assumes that the vehicle arrival within a cycle obeys uniform arrival and volume can be deduced as a time-dependent function of the delay pattern of CVs based on the shockwave theory.
- Zheng's method [6]: A stochastic method based on maximum likelihood estimation. This method considers vehicle arrivals to be a time-dependent Poisson process and formulates the arrival rate estimation into a maximum likelihood estimation problem based on vehicle arrivals that are derived between consecutive queued CVs.
- Hybrid method [21]: A hybrid method integrating shockwave theory and probability theory. This method first uses shockwave theory to estimate the queued volume (i.e., queue length) and then uses maximum likelihood estimation for non-queued volume.

While for *queue length* estimation, two representative methods were chosen for comparison.

- Cheng's method [10]: A deterministic method using shockwave theory. This method first extracts critical points of CV trajectories and then estimates the queue length by fitting shockwaves.
- Yin's method [14]: A stochastic method using Kalman Filter to fit the queue accumulation process. This method considers the positioning errors of CVs and uses the Kalman Filter to handle such errors when fitting the queue shockwave.

The proposed method is referred to as the SMPC-SVT method, which can provide estimates of both traffic demand and queue length as the detailed arrival profile is estimated.

- SMPC-SVT method: the proposed privacy-preserving matrix completion method for arrival profile estimation. In the case that the privacy of mobility companies is considered, a perturbation approach and additive secret sharing are applied to protect the private CV data of each company for initial incomplete arrival matrix construction. Then, the SVT method is adopted to solve the matrix completion problem efficiently, thus estimating arrival profiles.

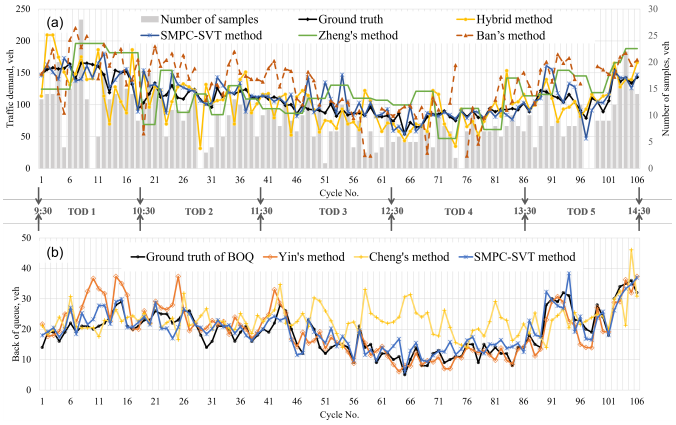

Fig. 7. Cycle-based traffic demand and queue length estimates by different methods: (a) traffic demand estimates and (b) queue length estimates

Overall performance: The comparative analysis of the proposed SMPC-SVT method and five existing methodologies on traffic volume and queue length estimation is summarized in Table II. Regarding traffic volume estimation, the SMPC-SVT method demonstrates superior performance compared to three other methods, achieving an MAE and MAPE of 10.3 veh/cycle and 9.8%, respectively, over the entire 5-hour analysis period. For specific TOD periods, the SMPC-SVT method excels in four out of five periods, exhibiting significantly smaller MAE values compared to alternative traffic volume estimation methods. Concerning queue length estimation, the proposed SMPC-SVT method attains the most favorable overall performance, yielding an MAE and MAPE of 2.48 veh and 14.87%, respectively, surpassing all other queue length estimation methods. These results underscore the efficacy of the SMPC-SVT approach in both traffic volume and queue length estimation tasks and provide a side note on the effectiveness of the method in estimating cycle-based arrival profiles.

Cycle-based performance: In order to have a more intuitive understanding of the ability of the proposed SMPC-SVT method to cope with dynamically varying traffic flows, Fig. 7 presents the cycle-based estimates of traffic volume and queue length by different methods. Note that, Zheng's method can only estimate the arrival rate for a longer time period. As is shown, for traffic demand estimation, the overall trends of different methods are consistent with the ground

TABLE III
TRAFFIC DEMAND ESTIMATION UNDER DIFFERENT CV SAMPLE SIZE

Reduction	0	20%	40%	60%
p	8.6	6.7	5.0	3.4
No. of CVs	9.0	7.2	5.4	3.6
Methods	MAE, veh (MAPE, %)			
SMPC-SVT	**10.3 (9.8)**	**11.5 (10.6)**	**14.7 (13.5)**	**24.5 (23.3)**
Ban's [16]	38.5 (37.2)	34.0 (34.1)	36.3 (36.7)	39.1 (39.7)
Zheng's [6]	29.0 (28.5)	39.9 (37.2)	54.9 (37.4)	58.6 (39.9)
Hybrid [21]	19.8 (18.6)	44.9 (39.4)	57.5 (40.1)	68.7 (46.7)

TABLE IV
TRAFFIC DEMAND ESTIMATION UNDER DIFFERENT Δ

Δ	Indicator	TOD					Overall
		1	2	3	4	5	
5s	MAE, veh	9.94	12.81	11.35	6.13	**11.52**	10.3
	MAPE, %	6.56	11.24	12.38	7.59	**10.79**	9.8
10s	MAE, veh	**8.17**	12.00	9.87	7.13	11.57	9.8
	MAPE, %	**5.47**	10.46	10.92	9.58	10.28	9.5
15s	MAE, veh	9.56	**10.33**	**5.83**	**3.65**	12.52	**8.2**
	MAPE, %	6.29	**9.33**	**6.61**	**4.66**	10.77	**7.5**

truth, but the fluctuation amplitudes of the four methods are different. Specifically, Ban's method has the greatest degree of volatility, followed by Zheng's method and the Hybrid method, and finally the proposed SMPC-SVT method that is most consistent with the ground truth. Regarding queue length estimation, the proposed SMPC-SVT method also shows the best fit to the ground truth, while Yin's method overestimates in TOD 1 and Cheng's method overestimates in both TOD 3 and 4, respectively.

2) Impact of the number of CV samples:

In order to give a more comprehensive insight into the reliability of the proposed method, sensitivity analysis is conducted in terms of the number of CV samples. Here the traffic demand estimation is used for sensitivity analysis. As we mentioned, four lanes exist for the straight-through flow of the empirical case, and despite a penetration rate of only 8.4%, its average number of trajectories still reaches 9.0 veh per cycle, which is relatively friendly to all methods. Therefore, here we test the effectiveness of the proposed SMPC-SVT method in the cases of less CV trajectories by further randomly reducing CV trajectories.

As shown in Table III, 20% – 60% reduction cases are further tested. Three parallel experiments are conducted to remove the influence of the stochastic sampling process. The MAE and MAPE values of each method here are the average values of three parallel experiments. The proposed method shows the best accuracy among the four methods, even when the penetration rate is about 3.4% (the 60%-reduction case), an estimation accuracy (i.e., $1-$MAPE) of 76.7% can be obtained. In terms of the growth rate of estimation error, the proposed method presents a slow increase when the sample reduction is no larger than 40% and an obvious increase while only 40% of the samples are used for estimation. It may be due to the fact that the increasing sparsity of the incomplete matrix actually lost some property of the original matrix, which limited the filling ability of the proposed method to some extent. Thus, the penetration of 5.0% can be regarded as the critical minimum penetration rate for the empirical case.

3) Impact of the unit time period:

Regarding the unit time period Δ used for matrix construction, two more cases, $\Delta = 10s$ and $\Delta = 15s$ were tested for the analysis period. With a greater Δ, the matrix size is reduced to a smaller number of matrix columns.

As shown in Table IV, the estimation errors, MAE and MAPE, gradually decrease as Δ increases from 5s to 15 s, although such decreases are slight. This may be due to the fact that the sparsity of the incomplete arrival rate matrix may be less severe with a greater Δ, as the matrix size is reduced with a smaller number of matrix columns. In addition, considering the randomization of CV observations, known values over longer Δ are also relatively more accurate. Nevertheless, the other side of the coin is that a larger Δ makes the estimated arrival profile smoother and provides less detailed time-varying arrival information during the cycle.

4) Impact of privacy-preserving mechanism:

Note that the privacy-preserving mechanism used in this study, SMPC, may have an impact on the estimation of arrival profiles by matrix completion method. This is because, in the scenario where company privacy is considered, each company holds only a portion of CV trajectories, constructs the initial arrival rate matrix separately, and builds the final matrix by privacy-preserving aggregation. Such a matrix is different from the directly generated matrix in the scenario where the companies share trajectories without considering privacy. Therefore, in this section, the impact of the privacy-preserving mechanism on arrival profile estimation accuracy is evaluated. For the initial arrival rate matrix provided by each party, a Laplacian perturbation with a coefficient of variation (COV) of 0.1 is added to the consecutively identical arrival rate values.

Number of parties: We first present the cycle-based traffic demand and queue length estimation results in Fig. 8 using the same total number of CV trajectories as the input of the SMPC-SVT method but considering the different number of participant companies, i.e., single party, two-party, and three-party. As is shown, for both traffic demand and queue length estimation, the results of all three different scenarios are practically consistent with trends of the ground truth, while the estimates of two-party or three-party may have larger fluctuations. Specifically, the gap of MAE values for both traffic demand and queue length estimation is not greater than 1 veh for each additional participant company while the total sample size remains unchanged. In particular, even though the multi-company collaboration scenarios degrade the accuracy of the traffic demand and queue length estimates slightly, the MAPE values, 12.4% for traffic demand and 23.4% for queue length, are still superior to existing methods selected for horizontal comparison, except for the state-of-the-art queue length estimation method by Yin et al., [14].

To dive deeper into the impact of privacy-preserving mechanism on arrival profile estimation, Fig. 9 shows the comparison of arrival profile estimation in TOD Period 5 under different

scenarios of multi-party collaboration. The consistency of the overall trend supports the conclusion that the estimation accuracy of arrival profiles can still be guaranteed through privacy protection, taking into account multi-party collaboration. The differences between them, in turn, are indicative of the impact that privacy-preserving methods can have on the performance of the proposed arrival rate matrix completion method. Noted that the later periods of the cycle, where the missing values of the incomplete arrival rate matrix are clustered, have a more significant difference in arrival rate estimation, which makes sense as the contribution of non-queued CVs to infer arrival information is relatively modest as compared with the queued CVs.

Different from the sensitivity analysis on CV sample size, here an analysis is further conducted to demonstrate the marginal benefit of the participation of more companies considering privacy preservation. The three-party case in Fig. 8, where each privately holds one of three pieces of CV data, is used here to compare with cases of removing one party and two parties respectively. Thus a total of three scenarios are considered, i.e., only one company performs the estimation (using 33.3% CV data), two companies collaborate on the estimation (using 66.6% CV data), and three companies collaborate on the estimation (using all CV data). As shown in Fig. 10 , the estimation accuracy increases with the participation of more parties, by about 2% in both traffic demand and queue length estimation. Specifically, the single-party scenario, despite not requiring privacy concerns, has estimates that significantly deviate from the ground truth value in many cycles due to the small number of CVs they can provide. As the number of participants increases, the number of significant deviations decreases, while our privacy-preserving mechanism, SMPC, protects the private data of each company from being revealed. These results demonstrate that, with an effective privacy-preserving mechanism, our method can allay concerns about privacy leakage across companies and facilitate more companies to participate in collaborations to increase the data size of available CV for more accurate and reliable traffic management.

Data disparity of parties: In practical application, it is common that data imparity may exist in the participating parties. The impact of the unevenness of data size on the effectiveness of the SMPC-SVT method is further explored using the three-party collaboration scenario where the participating parties are assumed to have a data size ratio of 1:2:3 to simulate the uneven data contribution of participants (hereinafter referred to as MPUC scenario). Comparison is conducted with the results of the multi-party collaboration scenario considering even contribution (hereinafter referred to as MPEC scenario) in Fig. 8 and Section IV-A (4). The dataset of each participant in MPEC scenario is 33.3% of the CV dataset, while the data sizes of each participant in MPUC scenario are 50%, 33.3%, and 16.7%, respectively.

As shown in Fig. 11 , the estimation accuracy of the SMPC-SVT method is still robust under MPUC scenario and slightly better than that under MPEC scenario. It makes sense that the estimation of Party 1 (50% of CV data) and Party 2 (33.3% of CV data) may slightly dominate in the aggregation process

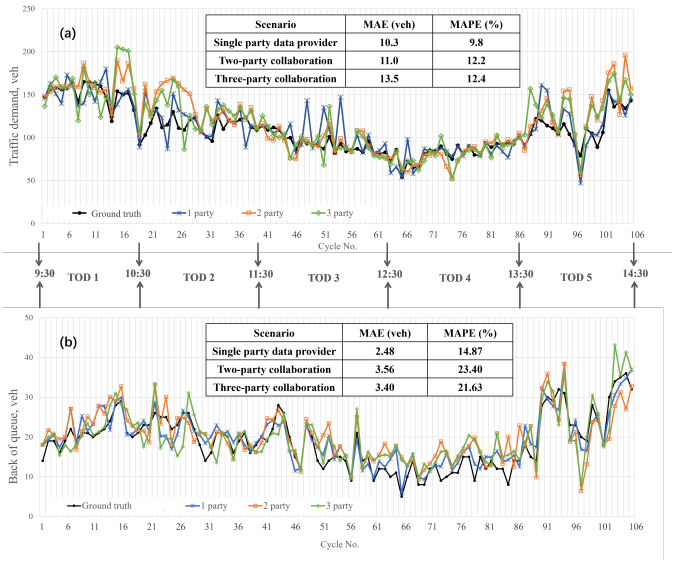

Fig. 8. Impact of SMPC on arrival rate matrix completion: (a) traffic demand estimation and (b) queue length estimation

conducted by TSOC. It can be further inferred that as long as the dataset of each party meets the minimum data size to realize a relatively accurate estimation, the unevenness in data size has little impact on the estimation accuracy.

Laplace perturbation: A sensitivity analysis is conducted here to test the robustness of the SMPC-SVT method to Laplace perturbations. A series of parallel experiments using different Coefficient of Variation (COV) values for Laplacian perturbation, ranging from 0 to 2, are done for the cyclic traffic demand estimation of both the two-party and three-party collaboration scenarios. The results of MAEs and MAPEs are shown in Fig. 12 with the previous test using a COV of 0.1. As shown in the figure, when the COV of Laplacian perturbation is not more than 1, the estimation error shows a slow wavelike increase, which is about 2 veh and 2% for MAE and MAPE, respectively. As the COV of perturbation increases from 1 to 2, the estimation error increases obviously. Besides, the growth of estimation error of the three-party scenario is a bit larger than that of the two-party scenario, which may be due to the accumulation of more perturbation with the increase of participating parties. As COV is actually the ratio of standard deviation to mean value, the results of sensitivity analysis imply that the proposed SVT-SMPC method is relatively robust to the added Laplacian perturbation for privacy preservation when the COV of the perturbed initial arrival rate is not more than 1.

5) Extension to real-time estimation:

Although the above validation is conducted offline, real-time application is also feasible as long as the sample CVs from different participating parties within a cycle are obtained in time to input into the method, then the cyclic arrival profile can be estimated when the next cycle just starts. The CV data in the empirical case study are used to explore the capacity of real-time estimation here. Instead of using the CV data of the whole TOD period as input and estimating the cyclic arrival profile of all the cycles at one time, in this test, the

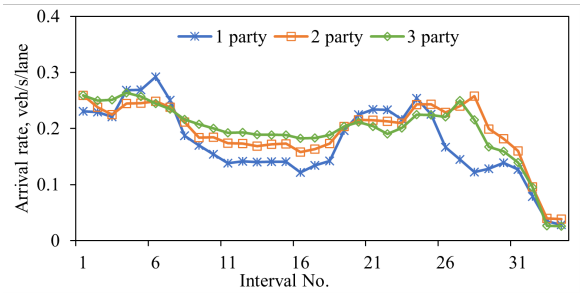

Fig. 9. An example of estimated arrival profile in scenarios with different numbers of parties (same sample size)

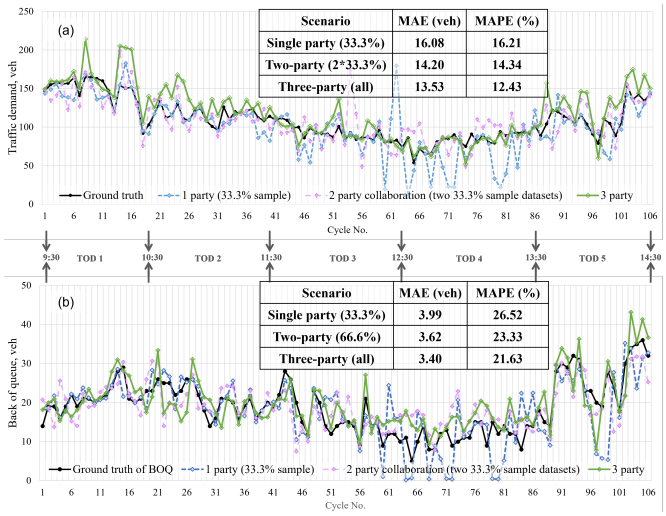

Fig. 10. Results for different numbers of participating companies (different amounts of CV data)

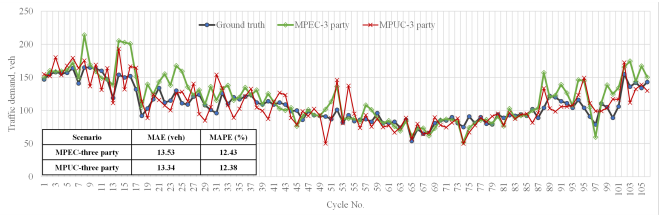

Fig. 11. Traffic demand estimation comparison of three-party collaboration between MPEC and MPUC

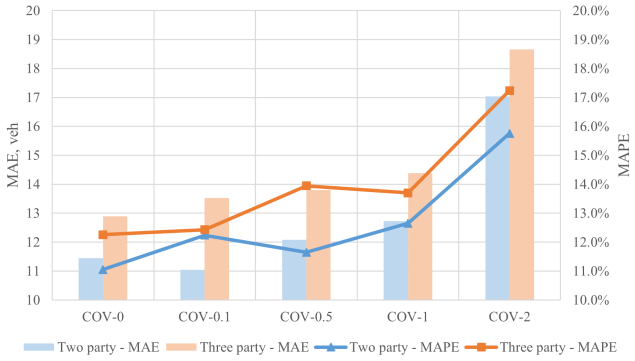

Fig. 12. Impact of different perturbance degrees on traffic demand estimation

TABLE V
TRAFFIC DEMAND ESTIMATION COMPARISON BETWEEN OFFLINE ESTIMATION AND REAL-TIME ESTIMATION SCENARIOS

Scenario	Offline estimation	Real-time estimation
Two-party collaboration	11.04 veh (12.24%)	13.69 veh (14.40%)
Three-party collaboration	13.53 veh (12.43%)	13.76 veh (14.42%)

proposed SMPC-SVT method is used once every cycle when the sampled CV data from multiple parties within the cycle are obtained when the cycle ends. Regarding the input data, the CV data of the past cycles counting from the start of the TOD period are used to construct the arrival rate matrix, thus the cyclic arrival profile estimation is carried out in an incremental pattern, as the matrix size will increase (in the row number) with every cycle operated, so will the number of sampled CVs.

The scenarios of two-party collaboration three-party are evaluated for real-time estimation, with a unit time step of $\Delta = 5\,s$. As shown in the results in Table V, a comparison is conducted between the real-time estimation and offline estimation. It is noted that both the MAE and MAPE are listed in the table, with the values of MAPE shown in the parenthesis. The estimation accuracy of real-time estimation decreases by about 2.16% and 1.99% for two-party and three-party collaboration scenario, respectively, which is still satisfactory for cyclic traffic demand estimation. The reason of the increase in estimation error when changing from offline estimation to real-time estimation can be explained by Fig. 13, showing the trend of cyclic demand estimation accuracy with the increasing number of cycles in TOD period 1. It is obvious that the MAPE decreases sharply with an increasing number of cycles, which means more available sample CVs and more information implying inner-cycle arrival pattern with the size of the constructed arrival rate matrix increasing. When the number of available cycles is larger than 4, the estimation accuracy becomes relatively stable. This also implies the minimum number of cycles required by the proposed matrix completion method, requiring more than four cycles to be estimated together. In practice, the real-time estimation of such first several cycles can be improved using historical data in the same period in the past days.

As for the computation efficiency of real-time estimation, the computation cost increases with the growth of the matrix size. Considering that the proposed method was encoded in Matlab and run using a desktop with a 2.50 GHz 8-core i7-11700 CPU and 16 GB RAM, an average cost of less than 0.1s can be obtained for the real-time estimation of the cyclic arrival profile. In short, the proposed SMPC-SVT method shows prospects in real-time estimation with relatively satisfactory performance and computation efficiency.

B. Simulation Case

As a data-driven method using connected vehicle data, one advantage of the proposed SMPC-SVT method is that the application scale can be enlarged as long as adequate data are available. Thus, to further explore the ability of the proposed method in large-scale applications, a simulation model is set up in a network scenario to evaluate its performance by estimating

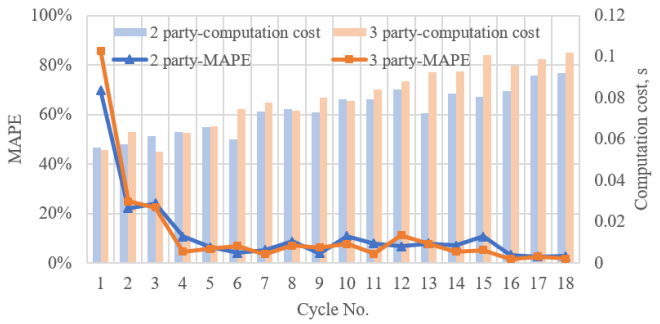

Fig. 13. Performance in real-time estimation.

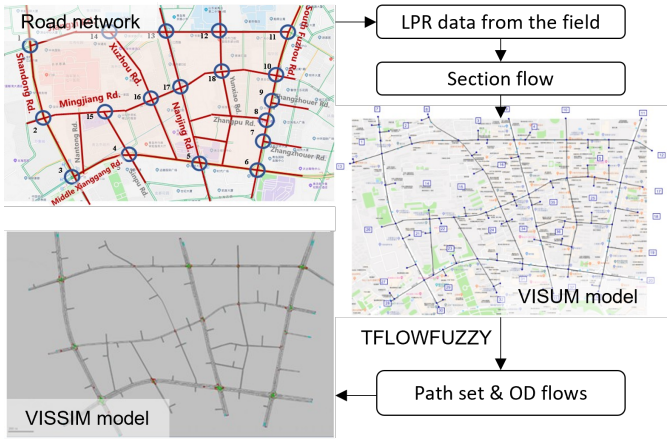

Fig. 14. Simulation of large scale road network

the total demand of all the controlled flows. The total demand here is calculated by summing up the cycle-based demand of all the cycles in a TOD period. To objectively assess the performance of the proposed method by removing the influence of the difference in the orders of magnitude, the weighted mean absolute percentage error (WMAPE) is further selected, as given by (38)

$$WMAPE = \frac{\sum_{z=1}^{Z} |\hat{Y}_z - Y_z|}{\sum_{z=1}^{Z} Y_z} \times 100 \tag{38}$$

Here, Z is the total number of controlled flows in the network. $\hat{Y}_z, Y_z$ (veh) are the estimated demand and ground truth demand, respectively.

The simulation case study site is a roadway network in Shinan District, Qingdao, Shandong Province, China. As shown in Fig. 14, there are 25 intersections in the study site, among them 18 intersections are signalized (denoted by blue circles), including 14 typical crossroads and 4 T-type intersections. In the network, 12 intersections are installed with license plate recognition (LPR) detectors in the field, and LPR data was collected from March 1 - March 12, 2019, from 7:00 - 9:00 p.m. to calibrate the simulation model. Two simulation models were built using the software VISSIM and VISUM. Based on the built-in TFLOWFUZZY model of the VISUM model, the path set and OD flow matrix are inversely estimated using the section passing flow obtained from the LPR detector data. With the calibrated OD flow matrix and route choice, the VISSIM model was loaded with traffic flow restoring the ground truth, and the signal timing data in March 2019 were also input into the model. Under the calibrated demand input, the mean absolute error between the actual network flow and the simulated network flow after a simulation run of the VISSIM model was about 5.64 veh/h, while the mean average percentage error was no more than 9%, which implies a satisfactory calibration performance to restore the actual network traffic flow operation.

For the total demand estimation, there are 88 controlled flows for all 18 intersections in the network to be estimated. The simulation period was set as 9000s in the VISSIM model, taking the first 1800s as the warm-up period while the remaining 7200s was the study period, namely the total demand of this 2-hour will be evaluated. After a simulation run, the vehicle trajectory data were extracted from the study period and sampled as input into the proposed method in order to simulate the connected vehicles in the mixed traffic flow in practical application. The penetration rates were set as 5%, 10%, 15%, 20%, 30%, 40% and 50% for parallel group experiments. Under each penetration rate scenario, three parallel experiments were done to guarantee the generality of the results.

1) Results of large-scale road network applications:

Based on the above-mentioned computation configuration, it took less than 10s for one parallel experiment to obtain the estimated volumes of all the 88 flows in the network. As shown in Table VI, the standard deviation of estimation errors across three parallel experiments decreases as the penetration rate increases. At low penetration rates, the proposed method exhibits instability. However, as the penetration rate increases from 5% to 10%, the estimation error decreases by more than half. With further increases in the penetration rate, the rate of error reduction slows down. When the penetration rate is 10%, the WMAPE is 4.21%, and the MAE is 46.68 vehicles over the 2-hour study period, equivalent to 23.34 vehicles per hour. Even at a penetration rate of 5% (with an average sample size of approximately 29 CVs per hour for each controlled flow), the WMAPE remains at 9.14%, indicating relatively satisfactory performance even in sparse scenarios. When the penetration rate is larger than 10%, the estimation is relatively robust, while the accuracy drops obviously when the penetration decreases to 5%. Thus, the penetration rate of 10% can be regarded as the critical minimum penetration rate for the simulation case.

Fig. 15 presents the detailed estimates of each controlled flow in terms of the ground truth values. The scatter distribution shows that the proposed method achieves accurate estimation for both flows with larger and smaller volumes, even for oversaturated flows, in the 10% penetration rate scenario.

C. Privacy analysis

Here we further clarify how additive secret sharing protects CV data in conjunction with Fig. 4. First, the initial matrix of each company is processed from its privately held CV data. Sharing such processed matrix rather than CV data somewhat

TABLE VI
PERFORMANCE OF THE PROPOSED METHOD IN A LARGE SCALE NETWORK

p, %	Indicator	Parallel experiments			Average
		1	2	3	
5	MAE, veh	142.52	104.52	92.76	113.27
	WMAPE, %	11.48	8.45	7.5	9.14
10	MAE, veh	37.68	71.30	47.34	52.11
	WMAPE, %	3.04	5.76	3.83	4.21
15	MAE, veh	48.00	46.08	45.96	46.68
	WMAPE, %	3.88	3.72	3.71	3.77
20	MAE, veh	47.17	24.72	37.97	36.62
	WMAPE, %	3.81	2.00	3.07	2.96
30	MAE, veh	37.33	31.73	29.76	32.93
	WMAPE, %	3.02	2.56	2.40	2.66
40	MAE, veh	38.19	25.90	29.51	31.20
	WMAPE, %	3.09	2.09	2.68	2.62
50	MAE, veh	34.52	29.35	35.68	33.18
	WMAPE, %	2.79	2.37	2.38	2.51

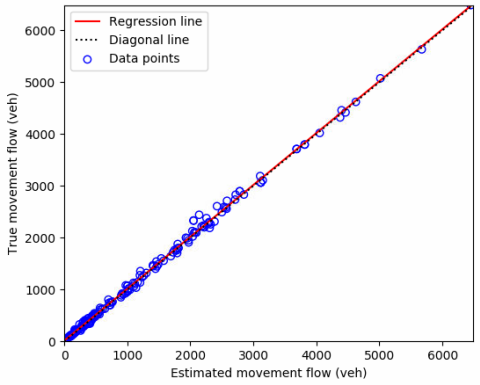

Fig. 15. Estimates versus ground truth values for all controlled flows

protects the CV data. However, this protection is vulnerable because it is possible for a third party to further infer some individual CV information through the unencrypted initial matrix in case of having prior knowledge of the matrix construction process. For example, if two neighboring elements of the same row of the initial matrix are different, it can be concluded that there exists a CV whose expected arrival time is near the corresponding time. Further, combined with the value of the elements, the stopping location of the CV is possible to be further inferred. So, adding perturbations and additive secret sharing provide the second layer of protection for private CV data. Adding perturbations can make all neighboring elements of the same row different thus against the inference attack mentioned in Section III-B. Additive secret sharing makes the values of the initial matrix independent and randomly generated while ensuring that the summation result remains unchanged. In such a case, the adversary is unable to infer the expected arrival time or stopping positions of the CVs based on the relative changes of the values in the matrices shared by each company.

In addition to secure multi-party computation, some studies also incorporate differential privacy (DP) to further enhance the protection of the output. The basic idea of DP is to add carefully calibrated noise to the results of queries or computations performed on the dataset, in such a way that the presence or absence of any individual's data has a negligible impact on the output. This ensures that the statistical properties of the data remain intact while protecting the privacy of individuals. In the case where the third party is not trustworthy, local differential privacy or distributed differential privacy methods can be used. However, in our problem, we protect the initial matrices that have been processed by each mobility company and aggregated by summation. It is also difficult to perform a difference attack on the result of this operation to infer information about any individual CV. Considering that DP will perturb the summation result and thus degrade the performance of the proposed model, the necessity of incorporating DP in the proposed method to enhance privacy protection can be one of the subjects of follow-up studies.

V. CONCLUSION

In this paper, a privacy-preserving matrix completion method that uses cross-company CV data is proposed for cycle-based arrival profile estimation at signalized intersections. Different from existing model-driven methods, the proposed method transforms cycle-based arrival profile estimation into a matrix completion problem and is free from arrival distribution assumption or parameter calibration using historical data; hence it is purely data-driven. Modeling traffic arrivals in a matrix form captures the temporal correlation of intra-cycle arrivals and inter-cycle fluctuations while providing a standardized format that enables crowdsourced CV data from different companies to work together to reconstruct actual traffic conditions. Combined with privacy-preserving techniques, i.e., additive secret sharing with a perturbation approach, such co-reconstruction process can effectively ensure that the privacy of each company will not be compromised, even in the face of inference attacks with prior knowledge of the model.

The proposed method is evaluated through an empirical case and a simulation case. In the empirical case, we evaluate the performance of the proposed method on cycle-based arrival profile estimation by verifying the accuracy of the derived estimates of traffic demand and queue length. The results show that (a) In the scenario without considering privacy issues, the proposed matrix completion method achieves the accuracies of 90.2% and 85% for cycle-based traffic demand and queue length estimation, respectively, given a penetration rate of 8.57%, which is superior to several existing methods. (b) The number of CV samples has significant impacts on the proposed matrix completion method as well as existing methods, while the proposed method still performs best even at a generation rate of 3.4%, where only 3.6 CVs are observed per cycle. (c) A larger unit time period improves the performance of the proposed matrix completion method but provides less detailed time-varying arrival information. (d) With the same data size, the collaborative estimation using data subsets considering perturbation of privacy-preserving mechanism and data disparity only slightly degrades the performance of the proposed matrix completion method. (e) Extending the proposed method to online estimation still achieves satisfactory estimation accuracy and efficiency.

Through further simulation tests on road networks, we demonstrate the reliability of the proposed matrix completion method when applied to large-scale road networks. (f) The method shows good estimation accuracy at both unsaturated and oversaturated flows. (g) The sensitivity analysis of the penetration rate shows that with more deployments of CVs, the estimation accuracy of the proposed method can be further improved, especially when the penetration rate is low.

Future research could work in the following directions. First, this study only discusses scenarios in which multiple companies hold homogeneous CV data, and future research could consider scenarios in which multiple companies hold heterogeneous data, such as CVs and detectors, to implement traffic state estimation with privacy preservation of all parties. Second, future research will continue to refine traffic signal optimization based on estimated time-varying cycle-based arrival rates.

Credit Authorship Contribution Statement

Chaopeng Tan: Conceptualization, Methodology, Writing - original draft, Writing - review & editing. **Jiarong Yao**: Conceptualization, Methodology, Validation, Writing - original draft, Writing - review & editing. **Keshuang Tang**: Funding acquisition, Methodology, Writing - review & editing. **Jinhao Liang**: Methodology, Writing - review & editing. **Guodong Yin**: Methodology, Writing - review & editing.

References

[1] M. Obayya, F. N. Al-Wesabi, R. Alabdan, M. Khalid, M. Assiri, M. I. Alsaid, A. E. Osman, and A. A. Alneil, "Artificial intelligence for traffic prediction and estimation in intelligent cyber-physical transportation systems," *IEEE Transactions on Consumer Electronics*, vol. 70, no. 1, pp. 1706–1715, 2024.

[2] C. Dai, S. Lu, C. Ma, S. Garg, and M. Alrashoud, "An adaptive rank-based tensor ring completion model for intelligent transportation systems," *IEEE Transactions on Consumer Electronics*, vol. 70, no. 1, pp. 2235–2243, 2024.

[3] C. Tan, Y. Cao, X. Ban, and K. Tang, "Connected vehicle data-driven fixed-time traffic signal control considering cyclic time-dependent vehicle arrivals based on cumulative flow diagram," *IEEE Transactions on Intelligent Transportation Systems*, 2024.

[4] I. U. M. S. A. S. M. A. K. X. B. L. Q. A., Ali, "A resource-aware multi-graph neural network for urban traffic flow prediction in multi-access edge computing systems," *IEEE Transactions on Consumer Electronics*, pp. 1–1, 2024, (Early Access).

[5] X. Zhan, Y. Zheng, X. Yi, and S. V. Ukkusuri, "Citywide traffic volume estimation using trajectory data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 2, pp. 272–285, 2016.

[6] J. Zheng and H. X. Liu, "Estimating traffic volumes for signalized intersections using connected vehicle data," *Transportation Research Part C: Emerging Technologies*, vol. 79, pp. 347–362, 2017.

[7] K. Tang, C. Tan, Y. Cao, J. Yao, and J. Sun, "A tensor decomposition method for cycle-based traffic volume estimation using sampled vehicle trajectories," *Transportation research part C: emerging technologies*, vol. 118, p. 102739, 2020.

[8] C. Tan, J. Yao, K. Tang *et al.*, "Joint estimation of multi-phase traffic demands at signalized intersections based on connected vehicle trajectories," *arXiv preprint arXiv:2210.10516*, 2022.

[9] G. Comert and M. Cetin, "Queue length estimation from probe vehicle location and the impacts of sample size," *European Journal of Operational Research*, vol. 197, no. 1, pp. 196–202, 2009.

[10] Y. Cheng, X. Qin, J. Jin, and B. Ran, "An exploratory shockwave approach to estimating queue length using probe trajectories," *Journal of intelligent transportation systems*, vol. 16, no. 1, pp. 12–23, 2012.

[11] G. Comert and M. Cetin, "Analytical evaluation of the error in queue length estimation at traffic signals from probe vehicle data," *IEEE Transactions on Intelligent Transportation Systems*, vol. 12, no. 2, pp. 563–573, 2011.

[12] C. Tan, J. Yao, K. Tang, and J. Sun, "Cycle-based queue length estimation for signalized intersections using sparse vehicle trajectory data," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 1, pp. 91–106, 2019.

[13] C. Tan, L. Liu, H. Wu, Y. Cao, and K. Tang, "Fuzing license plate recognition data and vehicle trajectory data for lane-based queue length estimation at signalized intersections," *Journal of Intelligent Transportation Systems*, vol. 24, no. 5, pp. 449–466, 2020.

[14] J. Yin, J. Sun, and K. Tang, "A kalman filter-based queue length estimation method with low-penetration mobile sensor data at signalized intersections," *Transportation Research Record*, vol. 2672, no. 45, pp. 253–264, 2018.

[15] W. Li, J. Wang, R. Fan, Y. Zhang, Q. Guo, C. Siddique, and X. J. Ban, "Short-term traffic state prediction from latent structures: Accuracy vs. efficiency," *Transportation Research Part C: Emerging Technologies*, vol. 111, pp. 72–90, 2020.

[16] X. Ban and M. Gruteser, "Mobile sensors as traffic probes: addressing transportation modeling and privacy protection in an integrated framework," in *Traffic and Transportation Studies 2010*, 2010, pp. 750–767.

[17] C. Tan, J. Yao, X. Ban, and K. Tang, "Cumulative flow diagram estimation and prediction based on sampled vehicle trajectories at signalized intersections," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 8, pp. 11 325–11 337, 2021.

[18] H. A. F. A.-M. K. M. O. A. A. S. A. M. A. H. G. P. M. M., Aljebreen, "Enhancing traffic flow prediction in intelligent cyber-physical systems: A novel bi-lstm-based approach with kalman filter integration," *IEEE Transactions on Consumer Electronics*, vol. 70, no. 1, pp. 1889–1902, 2024.

[19] A. Emami, M. Sarvi, and S. Asadi Bagloee, "Using kalman filter algorithm for short-term traffic flow prediction in a connected vehicle environment," *Journal of Modern Transportation*, vol. 27, pp. 222–232, 2019.

[20] Z. Zhang, S. Zhang, L. Mo, M. Guo, F. Liu, and X. Qi, "Traffic volume estimate based on low penetration connected vehicle data at signalized intersections: A bayesian deduction approach," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 8, pp. 10 596–10 609, 2021.

[21] J. Yao, F. Li, K. Tang, and S. Jian, "Sampled trajectory data-driven method of cycle-based volume estimation for signalized intersections by hybridizing shockwave theory and probability distribution," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 6, pp. 2615–2627, 2019.

[22] C.-M. Chen, Z. Li, S. Kumari, G. Srivastava, K. Lakshmanna, and T. R. Gadekallu, "A provably secure key transfer protocol for the fog-enabled social internet of vehicles based on a confidential computing environment," *Vehicular Communications*, vol. 39, p. 100567, 2023.

[23] C. Tan and K. Yang, "Privacy-preserving adaptive traffic signal control in a connected vehicle environment," *Transportation research part C: emerging technologies*, vol. 158, p. 104453, 2024.

[24] C.-M. Chen, Z. Li, A. K. Das, S. A. Chaudhry, and P. Lorenz, "Provably secure authentication scheme for fog computing-enabled intelligent social internet of vehicles," *IEEE Transactions on Vehicular Technology*, 2024.

[25] R. Talviste *et al.*, "Applying secure multi-party computation in practice," *Ph. D. dissertation*, 2016.

[26] T. Ranbaduge, D. Vatsalan, P. Christen *et al.*, "Secure multi-party summation protocols: Are they secure enough under collusion?" *Trans. Data Priv.*, vol. 13, no. 1, pp. 25–60, 2020.

[27] J.-F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM Journal on optimization*, vol. 20, no. 4, pp. 1956–1982, 2010.

[28] B. Recht, M. Fazel, and P. A. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," *SIAM review*, vol. 52, no. 3, pp. 471–501, 2010.

[29] E. Candes and B. Recht, "Exact matrix completion via convex optimization," *Communications of the ACM*, vol. 55, no. 6, pp. 111–119, 2012.

[30] B. P. Epps and A. H. Techet, "An error threshold criterion for singular value decomposition modes extracted from piv data," *Experiments in fluids*, vol. 48, no. 2, pp. 355–367, 2010.

[31] B. P. Epps and E. M. Krivitzky, "Singular value decomposition of noisy data: noise filtering," *Experiments in Fluids*, vol. 60, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:199131035

Chaopeng Tan received his B.S. and Ph.D. degrees in traffic engineering from Tongji University, Shanghai, China in 2017 and 2022, respectively. He was also a visiting Ph.D. student at the University of Washington (Seattle) from 2019 to 2021 and a Postdoctoral Research Fellow with the Department of Civil and Environment Engineering, National University of Singapore from 2022 to 2024. He is currently a Postdoctoral Research Fellow with the Department of Transport and Planning, Delft University of Technology, The Netherlands. His main research interests include intelligent transportation systems, traffic modeling and control with connected vehicles, and privacy-preserving traffic control.

Jiarong Yao received the B.S. degree in traffic engineering in 2016 and the Ph.D. degree in department of comprehensive traffic information and control engineering in 2021, both from Tongji University. Now she is a research fellow in School of Electrical and Electronic Engineering, Nanyang Technological University. Her main research interest is traffic control and Intelligent Transportation Systems.

Keshuang Tang received his doctor's degree in traffic engineering from Nagoya University in 2008. Afterward, he worked at The University of Tokyo as a postdoctoral research fellow, and then at Tohoku University as a Project Assistant Professor. He was also a visiting scholar of the University of California, Berkeley in 2010. He is currently a full professor at the College of Transportation Engineering, Tongji University, China. His main research interests include driver behavior, signal control, and Intelligent Transportation Systems.

Jinhao Liang received the B.S. degree from School of Mechanical Engineering, Nanjing University of Science and Technology, Nanjing, China, in 2017, and Ph.D. degree from School of Mechanical Engineering, Southeast University, Nanjing, China, in 2022. Now he is a Research Fellow with Department of Civil and Environmental Engineering, National University of Singapore. His research interests include vehicle dynamics and control, autonomous vehicles, and vehicle safety assistance system.

Guodong Yin received the Ph.D. degree in mechanical engineering from Southeast University, Nanjing, China, in 2007. From August 2011 to August 2012, he was a Visiting Scholar with the Department of Mechanical and Aerospace Engineering, The Ohio State University, Columbus, OH, USA. He is currently a Professor with the School of Mechanical Engineering, Southeast University. He was the recipient of the National Science Fund for Distinguished Young Scholars and is a China-SAE Fellow. His research interests include vehicle dynamics and control, automated vehicles, and connected vehicles.