2013 IEEE International Conference on Green Computing and Communications and IEEE Internet of Things and IEEE Cyber, Physical and Social Computing

Privacy-Preserving Point-to-Point Transportation Traffic Measurement through Bit Array Masking in Intelligent Cyber-Physical Road Systems

Yian Zhou*, Qingjun Xiao*, Zhen Mo*, Shigang Chen* and Yafeng Yin†
*Department of Computer & Information Science & Engineering
University of Florida, Gainesville, FL 32611, USA
†Department of Civil and Coastal Engineering
University of Florida, Gainesville, FL 32611, USA

Abstract—**Traffic measurement is a critical function in transportation engineering. We consider privacy-preserving point-to-point traffic measurement in this paper. We measure the number of vehicles traveling from one geographical location to another by taking advantage of capabilities provided by the intelligent cyber-physical road systems that enable automatic collection of traffic data. The challenge is to allow the collection of aggregate point-to-point data while preserving the privacy of individual vehicles. We propose a novel measurement scheme which utilizes bit arrays to collect "masked" data and adopts maximum likelihood estimation (MLE) to obtain the measurement result. Both mathematical proof and simulation demonstrate the practicality and scalability of our scheme.**

Keywords-**Transportation traffic measurement, privacy, cyber-physical systems, maximum likelihood estimation**

I. INTRODUCTION

Traffic measurement is a critical function in transportation engineering [1]. There are two categories of traffic statistics, *"point"* statistics and *"point-to-point"* statistics. Point statistics tell the number of vehicles traversing a specific *point* (location). Various prediction models have been proposed to estimate them [2], [3], [4]. Point-to-point statistics describe the number of vehicles traveling between two *points* (locations). They are essential inputs to a variety of studies including estimation of traffic link flow distribution as part of investment plan and calculation of road exposure rates as part of safety analysis. Though some point-to-point statistics may be inferred from point data [5], the practicality is limited by either high computation overhead or degraded measurement accuracy. As for direct measurement of "point-to-point" traffic, little work has been done especially when drivers' location privacy is concerned.

This paper considers the important problem of privacy-preserving *point-to-point* transportation traffic measurement. The set of vehicles traveling from one geographical location to another is modeled as a traffic flow, and the flow size is the number of vehicles in the set. To enable automatic collection of traffic flow data, we take advantage of intelligent cyber-physical road systems (CPRS), which integrate the latest technologies in wireless communications and on-board computer processing into transportation systems [6] [7]. In particular, IntelliDrive [8] from USDOT [9] envisions a nationwide system where vehicles communicate with roadside equipments (RSE) in real time via dedicated short range communications. In CPRS, vehicles may report their IDs to RSEs when they pass by, and that information can be used by the authority to measure traffic flows. However, if a vehicle keeps transmitting its unique identifier to RSEs, that information will enable others to track its entire moving history. As more and more people concern about their location privacy, the degree of privacy that a scheme preserves will directly affect its applicability.

To address the concerns of privacy, there are many issues that we need to consider: First of all, we need a criteria to tell what is good privacy and what is not. In this paper, we capture the essence of privacy in traffic flow measurement, and quantify it as a probability that a potential tracker cannot identify any trace of any vehicle. Secondly, given that criteria, how can we preserve the optimal privacy? Apparently, the better the privacy, the more applicable the measurement scheme. Furthermore, to protect the privacy of vehicles, only randomized and de-identified information is collected. How can we achieve sound measurement accuracy based on information that looks totally random?

In this paper, we propose a novel scheme for privacy-preserving traffic flow measurement. It utilizes bit arrays to encode "masked" data sent from vehicles to RSEs, and adopts maximum likelihood estimation (MLE) to obtain measurement results. We analyze its performance through both mathematical proof and simulations, which demonstrate the applicability of our scheme.

The rest of the paper is organized as follows: Section II gives the system and threat model, problem statement, and the performance metrics. Section III discusses some straightforward solutions and their limitations. Section IV presents our novel solution and its performance analysis. Section V shows simulation results. Section VI summarizes the related work. Finally, Section VII draws the conclusion.

II. PRELIMINARIES

A. System Model

We consider an intelligent cyber-physical road system involving three groups of entities: vehicles, roadside

978-0-7695-5046-6/13 $26.00 © 2013 IEEE
DOI 10.1109/GreenCom-iThings-CPSCom.2013.146

equipments (RSE), and a central server. Each vehicle has a unique ID, e.g., its VIN. Each RSE also has its unique ID. Both vehicles and RSEs are equipped with computing and communication capabilities, e.g., on-board computer chips and communication modules. Vehicles communicate with RSEs in real time via dedicated short range communications (DSRC) [9]. RSEs are connected to the central server through wired or wireless means. They collect information from vehicles and transfer it to the central server on a periodical basis.

B. Problem Statement

We define a traffic flow between one RSE-equipped location and another RSE-equipped location as the set of vehicles traveling between the two locations during a measurement period. The size of the traffic flow is the number of vehicles in this set. Our problem is to measure the sizes of traffic flows in a road system between all pairs of locations where RSEs are installed while protecting vehicles' privacy. To achieve the privacy-preserving end, we need a solution in which a vehicle never transmits any unique identifier. Ideally, the information transmitted by the vehicles to the RSEs looks totally random, out of which neither the identity nor the trajectory of any vehicle can be pried with high probability.

We also assume that a special MAC protocol is applied to support privacy preservation such that the MAC address of a vehicle is not fixed. Vehicles may pick an MAC address randomly from a large space for one-time use when needed.

C. Threat Model

We assume a semi-honest model for the RSEs. On the one hand, all RSEs are from trustworthy authorities, which can be enforced by authentication based on PKI. The vehicles can use the public-key certificate broadcasted by RSEs, which they obtained from the trusted third parties, to verify the RSEs. On the other hand, the authorities may exploit the information collected by RSEs to track individual vehicles when they need to do so. For instance, if a vehicle transmits any unique identifier upon each query, that identifier can be used for tracking purpose.

Note that there are also other ways to track a vehicle, for example, tailgating the vehicle, or setting cameras near RSEs to take photos and using image processing to recognize it. These methods are beyond the scope of this paper. In this paper, we focus on preventing automatical tracking caused by the traffic flow measurement scheme itself.

D. Performance Metrics

In this paper, we consider three performance metrics to evaluate a traffic flow measurement scheme: preserved privacy, measurement accuracy, and computation overhead. They are defined in the following.

1) Preserved Privacy: We capture the essence of privacy preservation in point-to-point transportation traffic measurement, which is allowing the tracker only a limited chance to identify any partial or full trajectory of any vehicle. Accordingly, we quantify the privacy of a scheme through a parameter p which satisfies the following requirement: the probability for any "trace" of any vehicle not to be identified must be at least p, where a trace of a vehicle is a pair of RSEs it has passed by. A larger value of p means better privacy. Intuitively, a scheme with $p = 0.5$ is better than one with $p = 0.1$ in terms of privacy because the latter gives the tracker a better chance to link traces of a vehicle to obtain its trajectory since it allows the traces to be identified with a higher probability, i.e., $1 - p$.

2) Measurement Accuracy: Let n_c be the true size of a traffic flow between a pair of locations and $\hat{n}_c$ be the corresponding measured result. We define the measurement accuracy through the absolute difference between $\hat{n}_c$ and n_c, namely $|\hat{n}_c - n_c|$. Clearly, the smaller the difference, the more accurate the measurement.

3) Computation Overhead: We consider the computation overhead for vehicles, RSEs, and the central server. For vehicles, we measure the computation overhead for each vehicle per RSE en route. For RSEs, we measure the computation overhead for each RSE per passing vehicle. For the central server, we measure the computation overhead for it to measure the traffic flow size for a pair of RSEs.

III. Straightforward Approaches and Their Limitations

To measure the traffic flow sizes between all pairs of RSEs in the road system, a straightforward approach is making vehicles report their IDs to all RSEs that they pass by. RSEs collect the IDs from the passing vehicles. At the end of each measurement period, all RSEs send their collected ID sets to the central server, which then measures the traffic flow size between each pair of RSEs by simply comparing the two corresponding ID sets: if a vehicle ID appears in both ID sets, then the vehicle must have passed both RSEs. Thus, the number of IDs that appear in both ID sets equals the true traffic flow size between the two corresponding RSEs. However, this simple approach leads to serious privacy breaching as it reveals vehicles' identities along the way.

A natural follow-up thinking is making vehicles report their encrypted IDs (EIDs) to the RSEs en route. The central server will compute traffic flow sizes based on the EID sets collected by RSEs. To prevent the tracker from using fixed EIDs to identify vehicles, each vehicle has many EIDs encrypted by different keys. However, the EIDs of a vehicle must satisfy the following property: they will produce the same result after a certain procedure of computations, allowing the central server to find out they represent the same vehicle. In this scheme, although vehicles' true identities are hidden, traces of each vehicle are still revealed and can be linked to obtain its full trajectory.

An alternative approach is having the RSEs broadcast their IDs (RIDs). Each vehicle will record the RIDs of all RSEs it has passed by, and transmit them to every RSE that it passes in the future. RSEs collect those RIDs from passing vehicles, and send them to the central server at the end of each measurement period. To compute the size of a traffic flow between two RSEs, R_x and R_y, the central server simply goes through the RID set collected by R_y (R_x), and count the number of times that R_x (R_y) appears in this set. This is the size of the directional traffic flow from R_x (R_y) to R_y (R_x). The undirectional traffic flow between R_x and R_y is the sum of both directional flow sizes. Clearly, this approach also reveals a vehicle's trajectory in the form of a list of RIDs sent to each RSE that it passes.[1]

IV. Privacy Preserving Point-to-Point Transportation Traffic Measurement

In this section, we present our novel scheme for privacy preserving point-to-point transportation traffic measurement. There are two phases for each measurement period: online coding and offline decoding. Online coding is an interaction between vehicles and RSEs, where necessary information for traffic flow measurement are securely collected. Later in the offline decoding phase, the central server will use those information to compute traffic flow sizes. In the following, we first illustrate the two measurement phases, and then mathematically analyze the performance of our scheme.

A. *Online Coding Phase*

In our scheme, each RSE R_x maintains a counter n_x, which keeps track of the total number of vehicles passing by during the current measurement period. R_x also maintains a bit array B_x with a fixed length m ($m > 1$) to mask vehicle identities. At the beginning of each measurement period, n_x and all the bits in B_x are set to zeros. In addition, each vehicle v has a logical bit array LB_v, which consists of s ($1 < s < m$) bits randomly selected from B_x. The indices of these bits in B_x are $H(v \oplus K_v \oplus X[0])$,..., $H(v \oplus K_v \oplus X[s-1])$, where $\oplus$ is the bitwise XOR, $H(...)$ is a hash function whose range is $[0, m)$, X is an integer array of randomly chosen constants whose purpose is to arbitrarily alter the hash result, and K_v is the private key of v whose purpose is to protect the privacy of its logical bit array.

The online coding phase is quite simple. RSEs broadcast queries in pre-set intervals (e.g., once a second), ensuring that each passing vehicle receives at least one query and meanwhile giving enough time for the vehicle to reply. Collisions can be resolved through well-established CSMA or TDMA protocols, which are not the focus of this paper. Every query that an RSE sends out includes the RSE's RID and its public-key certificate. Suppose a vehicle, whose ID is v, receives a query from an RSE, whose ID is R_x. The vehicle first verifies the certificate, and then uses the RSE's public key to authenticate the RSE. After verifying that R_x is from the trustworthy authority, the vehicle v will randomly select a bit from its logical bit array LB_v by computing an index $b = H(v \oplus K_v \oplus X[H(R_x) \, mod \, s])$. The vehicle v then sends the resulting index b to the RSE R_x. Upon receiving the index b, R_x will first increase its counter n_x by 1, and then set the bth bit in B_x to 1:

$$B_x[H(v \oplus K_v \oplus X[H(R_x) \, mod \, s])] = 1. \tag{1}$$

B. *Offline Decoding Phase*

At the end of each measurement period, all RSEs will send their counters and bit arrays to the central server, which then performs the offline measurement. We employs the maximum likelihood estimation (MLE) [10] to measure the sizes of traffic flows based on the counters and bit arrays.

Suppose the set of vehicles that pass RSE R_x (R_y) is denoted as S_x (S_y) with cardinality $|S_x| = n_x$ ($|S_y| = n_y$). Clearly, the set of vehicles that pass both RSE R_x and R_y is $S_x \cap S_y$. Denote its cardinality as n_c, which is the value that we want to measure. Furthermore, denote by S the subset of vehicles in $S_x \cap S_y$ that happen to set the same bit in B_x and B_y, where B_x and B_y are the bit arrays at R_x and R_y, respectively. Let n_o be the cardinality of S, i.e., $n_o = |S|$. Clearly, $S \subseteq S_x \cap S_y$ and $0 \le n_o \le n_c$. For any vehicle, it has the same probability $\frac{1}{s}$ to set any bit in its s-bit logical bit array. As a result, the probability for an arbitrary vehicle v from $S_x \cap S_y$ to select the same bit in both B_x and B_y is $s \times \frac{1}{s} \times \frac{1}{s} = \frac{1}{s}$. Therefore, the number of such vehicles, n_o, is binomially distributed according to $B(n_c, \frac{1}{s})$. Accordingly, the probability for $n_o = z (0 \le z \le n_c)$ is

$$P(n_o = z) = \binom{n_c}{z} (\frac{1}{s})^z (1 - \frac{1}{s})^{n_c - z}. \tag{2}$$

Given the counters n_x and n_y, and bit arrays B_x and B_y, we measure n_c as follows: First, take a bitwise AND of B_x and B_y, and denote the resulting bit array as B_c. Namely,

$$B_c[i] = B_x[i] \wedge B_y[i], \; \forall i \in [0, m-1]. \tag{3}$$

We can easily find out the number of 0's in B_c. Suppose it is denoted by U_c. In the following, we will analyze the probability for an arbitrary bit in B_c to remain '0' after the online coding phase, and use it to establish the likelihood function for us to observe U_c '0' bits in B_c. Maximizing that likelihood function with respect to n_c will give the MLE estimate of n_c.

Clearly, the event for an arbitrary bit b in B_c to remain '0' after online coding is equivalent to the combination of

[1] The identity of a vehicle may be revealed by a photograph triggered by the vehicle rushing a red light or by a police car stopping the vehicle. When the identity is combined with the trajectory transmitted by the vehicle, the entire traveling path of the driver will be revealed.

the following two events: (1) *Event 1: None of the vehicles in S has chosen b at R_x and R_y*. If a vehicle $v \in S$ chooses b, then bit b in B_x and B_y are both set to '1' by v (hence bit b in B_c is also '1'). Since each vehicle has probability $\frac{1}{m}$ to set bit b to '1', the probability for the vehicle not to choose bit b is $1-\frac{1}{m}$. There are n_o vehicles in S. Therefore, the probability for the first event to happen is

$$q_1 = (1 - \frac{1}{m})^{n_o}. \tag{4}$$

(2) *Event 2: Either none of the vehicles in $S_x - S$ has chosen b at R_x or none of the vehicles in $S_y - S$ has chosen b at R_y*. Otherwise, bit b in both B_x and B_y will be '1' (hence bit b in B_c is '1'). The probability for bit b not chosen by any vehicle in $S_x - S$ is $(1-\frac{1}{m})^{n_x - n_o}$, and the probability for bit b not chosen by any vehicle in $S_y - S$ is $(1-\frac{1}{m})^{n_y - n_o}$. Therefore, the probability for the second event to happen is

$$\begin{aligned} q_2 &= 1 - (1 - (1 - \frac{1}{m})^{n_x - n_o}) \times (1 - (1 - \frac{1}{m})^{n_y - n_o}) \\ &= (1 - \frac{1}{m})^{n_x - n_o} + (1 - \frac{1}{m})^{n_y - n_o} \\ &\quad -(1 - \frac{1}{m})^{n_x + n_y - 2 \times n_o}. \end{aligned} \tag{5}$$

Combining above analysis, the conditional probability for bit b in B_c to remain '0' given $n_o = z$ is $q_1 \times q_2$, namely,

$$\begin{aligned} q(n_c | n_o = z) &= (1 - \frac{1}{m})^{n_x} + (1 - \frac{1}{m})^{n_y} \\ &\quad -(1 - \frac{1}{m})^{n_x + n_y - z}. \end{aligned} \tag{6}$$

Given $q(n_c|n_o = z)$ and the distribution of n_o, the overall probability $q(n_c)$ for bit b in B_c to remain '0' is

$$\begin{aligned} q(n_c) &= \sum_{z=0}^{n_c} q(n_c|n_o = z) \times P(n_o = z) \\ &= (1 - \frac{1}{m})^{n_x} + (1 - \frac{1}{m})^{n_y} - (1 - \frac{1}{m})^{n_x + n_y} \\ &\quad \times \left(\frac{\frac{1}{s} + (1 - \frac{1}{s})(1 - \frac{1}{m})}{1 - \frac{1}{m}} \right)^{n_c}. \end{aligned} \tag{7}$$

Knowing that each bit in B_c has a probability $q(n_c)$ to remain '0', we can establish the likelihood function for us to observe U_c '0' bits in B_c (hence $m - U_c$ '1' bits in B_c):

$$L = (q(n_c))^{U_c} \times (1 - q(n_c))^{m - U_c}. \tag{8}$$

The MLE estimate of n_c is the optimal value of n_c that maximizes the likelihood function in (8):

$$\hat{n_c} = \arg\max_{n_c} \{L\} \tag{9}$$

To find $\hat{n_c}$, we take logarithm on both sides of (8):

$$\ln L = U_c \times \ln q(n_c) + (m - U_c) \times \ln(1 - q(n_c)). \tag{10}$$

Take the first order derivative of (10), we have:

$$\frac{d \ln L}{d n_c} = \left(\frac{U_c}{q(n_c)} - \frac{m - U_c}{1 - q(n_c)} \right) \times q'(n_c), \tag{11}$$

where $q'(n_c)$ can be computed from (7) as follows:

$$\begin{aligned} q'(n_c) &= \frac{dq(n_c)}{dn_c} \\ &= -(1 - \frac{1}{m})^{n_x + n_y} \times \left(\frac{\frac{1}{s} + (1 - \frac{1}{s})(1 - \frac{1}{m})}{1 - \frac{1}{m}} \right)^{n_c} \\ &\quad \times \ln \left(\frac{\frac{1}{s} + (1 - \frac{1}{s})(1 - \frac{1}{m})}{1 - \frac{1}{m}} \right). \end{aligned} \tag{12}$$

To compute $\hat{n_c}$, we set the right side of (11) to 0:

$$\left(\frac{U_c}{q(n_c)} - \frac{m - U_c}{1 - q(n_c)} \right) \times q'(n_c) = 0. \tag{13}$$

Observe from (12) that $q'(n_c)$ cannot be 0 when $m > 1$ and $s > 1$. Therefore, we have:

$$\frac{U_c}{q(n_c)} - \frac{m - U_c}{1 - q(n_c)} = 0. \tag{14}$$

Substituting (7) to (14), we obtain the MLE estimator $\hat{n_c}$ of the desired traffic flow size n_c as follows:

$$\begin{aligned} \hat{n_c} &= \frac{1}{\ln \left(\frac{\frac{1}{s} + (1 - \frac{1}{s})(1 - \frac{1}{m})}{1 - \frac{1}{m}} \right)} \Big[-(n_x + n_y) \ln(1 - \frac{1}{m}) \\ &\quad + \ln \left((1 - \frac{1}{m})^{n_x} + (1 - \frac{1}{m})^{n_y} - \frac{U_c}{m} \right) \Big]. \end{aligned} \tag{15}$$

C. Privacy Guarantee

The previous two subsections give a detailed description of the two measurement phases of our scheme. In this subsection, we evaluate the privacy that our scheme preserves. Note that in our scheme, the only information that a vehicle v ever transmits to an RSE en route is an index of a bit b randomly selected from its s-bit logical bit array, LB_v. From the tracker's point of view, it can only identify the trace of a vehicle passing by two RSEs R_x and R_y through the observation of the bits that are set to '1' in both B_x and B_y; these bits will be '1' in B_c. Therefore, the preserved privacy of our scheme is actually a conditional probability which tells to what degree an observed '1' in B_c

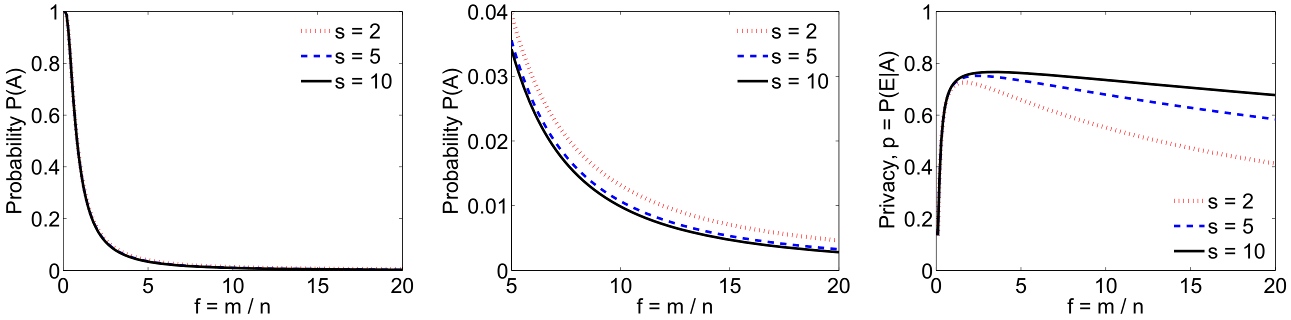

Figure 1. $n_x = n_y = n = 50,000$, $n_c = 5,000$; *First Plot*: probability $P(A)$ when m varies from $0.1n$ to $20n$, controlled by different $s = 2, 5, 10$; *Second Plot*: a zoom-in of the first plot when m varies from $5n$ to $20n$; *Third Plot*: probability $P(E|A)$ when m varies from $0.1n$ to $20n$, controlled by different $s = 2, 5, 10$.

does not represent a common vehicle passing by both R_x and R_y. We derive this conditional probability below.

Firstly, consider the probability for the tracker to observe an arbitrary bit, b, to be set to '1' in both B_x and B_y (event A), $P(A)$. Obviously, the probability $P(A)$ equals 1 minus $q(n_c)$ given our analysis in Section IV-B:

$$\begin{aligned} P(A) &= 1-(1-\frac{1}{m})^{n_x}-(1-\frac{1}{m})^{n_y}+(1-\frac{1}{m})^{n_x+n_y} \\ &\quad \times\left(\frac{\frac{1}{s}+(1-\frac{1}{s})(1-\frac{1}{m})}{1-\frac{1}{m}}\right)^{n_c} \qquad (16) \end{aligned}$$

Secondly, consider the conditional probability for such a bit, b, to not represent a common vehicle passing both R_x and R_y (event E), $P(E|A)$. This is the privacy p that we want to derive. Note that event E happens if and only if bit b in B_x is set only by vehicles passing only RSE R_x (i.e., in set $S_x - S_y$), and bit b in B_y is set only by vehicles passing only RSE R_y (i.e., in set $S_y - S_x$). Denote these two events as E_x and E_y, respectively. There are n_x (n_y) vehicles passing R_x (R_y), and n_c vehicles among them pass both R_x and R_y. Since each vehicle has a probability $\frac{1}{m}$ to set bit b to '1', the probability for E_x (E_y) to happen is:

$$P(E_x) = (1-(1-\frac{1}{m})^{n_x-n_c}) \times (1-\frac{1}{m})^{n_c}, \qquad (17)$$

$$P(E_y) = (1-(1-\frac{1}{m})^{n_y-n_c}) \times (1-\frac{1}{m})^{n_c}. \qquad (18)$$

Combining the above analysis, we have the formula for the preserved privacy of our scheme as follows:

$$\begin{aligned} p &= P(E|A) = \frac{P(E_x) \times P(E_y)}{P(A)} \\ &= \frac{1}{P(A)} \times ((1-\frac{1}{m})^{n_c} - (1-\frac{1}{m})^{n_x}) \\ &\quad \times ((1-\frac{1}{m})^{n_c} - (1-\frac{1}{m})^{n_y}), \qquad (19) \end{aligned}$$

where $P(A)$ is given in (16).

Observe that there are 2 parameters, s and m, that determine the value of $P(E|A)$. Among them, s only appears in the denominator $P(A)$, and it influences $P(E|A)$ through varying the value of $P(A)$. m influences both the denominator and the numerator. In the following, we consider the influence of s and m on $P(E|A)$ by first examining the influence of s on $P(A)$ (hence that on $P(E|A)$) under various values of m, and then analyzing how m determines the value of $P(E|A)$ given values for s.

1) Influence of s on $P(A)$: To examine how s effects $P(A)$, we take partial derivative of (16) with respect to s

$$\frac{\partial P(A)}{\partial s} = -(1-\frac{1}{m})^{n_x+n_y} \times \frac{n_c}{(m-1)s^2} C^{n_c-1}. \qquad (20)$$

where $C = \frac{\frac{1}{s}+(1-\frac{1}{s})(1-\frac{1}{m})}{1-\frac{1}{m}}$.

Clearly, $\frac{\partial P(A)}{\partial s} < 0$. Therefore, with the increment of s, the value of $P(A)$ decreases, and in turn, the value of $P(E|A)$ increases. In other words, the preserved privacy will be better with a larger value of s. The numerical results are shown in the first two plots of Figure 1 where $n_x = n_y = n = 50,000$, $n_c = 5,000$, and $s = 2, 5, 10$, corresponding to three curves in each plot. Clearly, as s increases, the probability $P(A)$ decreases.

Another observation from the numerical results gives that when $s > 5$, the difference in probability $P(A)$ under different s becomes quite small. For instance, when $m \in [5n, 20n]$, the difference in $P(A)$ when $s = 5$ and $s = 10$ is smaller than 0.0005 (see the two lower curves in the second plot of Figure 1). When $n > 10$, that difference becomes negligible. Therefore, when we analyze the effect of m on $P(E|A)$ in the following subsection, and set up the parameters for our simulations, we only consider the cases when $s = 2, 5, 10$, with established understanding that larger values of s will only make negligible differences.

2) Influence of m on $P(E|A)$: To examine the effects of m on $P(E|A)$, we take the partial derivative of (19) with respect to m and obtain the following:

$$\frac{\partial P(E|A)}{\partial m} = \frac{\frac{\partial P(E)}{\partial m} \times P(A) - \frac{\partial P(A)}{\partial m} \times P(E)}{P(A)^2} \tag{21}$$

where $P(E) = P(E_x) \times P(E_y)$. $P(E_x)$ and $P(E_y)$ are given in (17) and (18), respectively. Therefore, the partial derivative of $P(E)$ with respect to m is:

$$\begin{aligned} \frac{\partial P(E)}{\partial m} &= \frac{m-1}{m^3}\Big[(n_x+n_y)(1-\frac{1}{m})^{n_x+n_y} \\ &+ 2n_c(1-\frac{1}{m})^{2n_c} - (n_c+n_x)(1-\frac{1}{m})^{n_c+n_x} \\ &- (n_c+n_y)(1-\frac{1}{m})^{n_c+n_y}\Big]. \end{aligned} \tag{22}$$

In addition, from (16), we can compute the derivative of $P(A)$ with respect to m:

$$\begin{aligned} &\frac{\partial P(A)}{\partial m} = \frac{1}{m^2}\Big[- n_x(1-\frac{1}{m})^{n_x-1} - n_y(1-\frac{1}{m})^{n_y-1} + \\ &(1-\frac{1}{m})^{n_x+n_y-2} \cdot C^{n_c} \cdot \bigg((n_x+n_y)(1-\frac{1}{m}) - \frac{n_c}{s \cdot C}\bigg)\Big] \end{aligned} \tag{23}$$

Through analysis, we know that both $\frac{\partial P(E)}{\partial m}$ and $\frac{\partial P(A)}{\partial m}$ are negative when m exceeds a certain value, which means both $P(E)$ and $P(A)$ will decrease with the increment of m afterwards. Intuitively, increasing m gives each vehicle a smaller chance $\frac{1}{m}$ to set an arbitrary bit, b. Hence, $P(E)$ and $P(A)$ also drop. The effects that m has on $P(E|A)$ are twofold: on the one hand, the increment of m decreases the denominator $P(A)$, which pulls the privacy up; on the other hand, the increment of m decreases the numerator $P(E)$, which drags the privacy down. The combination of these two effects gives that the partial derivative of $P(E|A)$ with respect to m can be positive, negative, or 0, according to (21). Therefore, given s, we can choose an optimal m to achieve the best degree of privacy. The optimal m is obtained by setting the right side of (21) to 0.

The third plot of Figure 1 shows the numerical results for the preserved privacy under different m when $n_x = n_y = n = 50,000$, $n_c = 5,000$, and $s = 2, 5, 10$. Clearly, along each curve (controlled by s), there is an optimal value of m that gives the optimal privacy, p. For instance, $m = 3.8n$ gives the optimal privacy $p = 0.7661$ when $s = 10$. Another observation is, when s is large (5 or 10), there always exists a smooth interval of m near its extreme point that can achieve comparable privacy as the optimal. For example, when $s = 10$, the values of m within the interval $[3.8n, 13.2n]$ achieves privacy that is within 5% of the optimal privacy 0.7661. In practice, this smooth interval for privacy will allow us to adjust the value of m to achieve better measurement results while preserving comparable privacy.

Table I
VALUES FOR m TO ACHIEVE OPTIMAL p UNDER DIFFERENT s.

s	2	5	10
optimal m	$1.7n$	$2.7n$	$3.8n$
optimal p	0.7258	0.7513	0.7661

D. Computation Overhead

In our scheme, when a vehicle v passes an RSE R_x, the vehicle v only needs to compute two hashes to obtain an index of a random bit in its logical array LB_v, and the RSE R_x only needs to set 1 bit in its bit array B_x, as described in Section IV-A. Therefore, the computation overhead for each vehicle per RSE as well as that for each RSE per vehicle are both $O(1)$. As for the central server, in order to compute a traffic flow size between a pair of locations, it only needs to do a bitwise AND over two m-bit bit arrays, count the number of '0' in the resulting bit array, and use (15) to compute the MLE estimator. Therefore, the computation overhead for the central server is also $O(1)$.

V. Simulation

In this section, we evaluate the performance of our measurement scheme through simulations. The simulations are performed under five system parameters, n_x, n_y, n_c, s, and m. For a pair of RSEs, R_x and R_y, n_x (n_y) is the number of vehicles passing by R_x (R_y). There are n_c vehicles passing both R_x and R_y, which means the true traffic flow size is n_c. s is the number of bits that each vehicle chooses in its logical bit array, and m is the number of bits in the RSEs' bit array. In the simulation, we choose the five parameters as follows: $n_x = n_y = n = 50,000$, $100,000$, or $500,000$, and n_c varies from $1\%n$ to $50\%n$, with step size of $0.1\%n$; $s = 2, 5, 10$, and m is chosen to achieve the optimal privacy p, as determined in Section IV-C. Table I lists the values for the bit array size m to achieve the optimal privacy p under different values of s.

Figure 2, 3, and 4 show our simulation results when $n = 50,000$, $100,000$, and $500,000$, respectively. For each figure, there are three plots, corresponding to the results of three sets of simulations controlled by parameter s, where s = 2, 5, and 10. Each plot shows the measured traffic flow sizes $\hat{n_c}$ (y-axis) with respect to different true traffic flow sizes n_c (x-axis) under a given setting of n, s, and m, where m is chosen as described in Table I so that the optimal privacy is achieved. We also draw the equality line $y = x$ in each plot for reference. Clearly, the closer a point is to the equality line, the smaller difference between the measured traffic flow and the real traffic flow, and in turn, the more accurate the measurement result.

From the figures, one can see that our measurement scheme is quite accurate because most of the points in all plots of all figures lie closely to the equality line. In particular, given other parameters, our MLE estimator produces almost perfect results when $s = 2$ (the first plot in

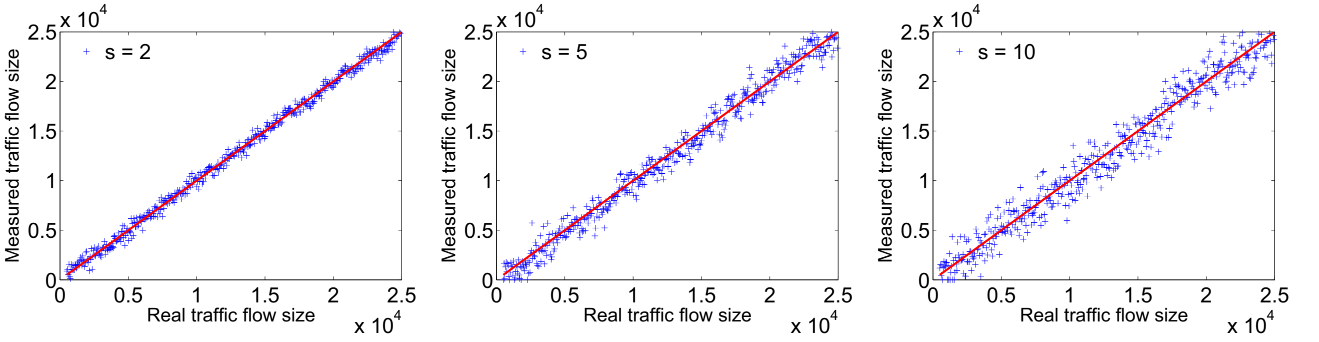

Figure 2. Measurement accuracy with optimal privacy, $n_x = n_y = n = 50,000$, $n_c = [0.01n, 0.5n]$. The x-axis shows true traffic flow sizes, and the y-axis shows the corresponding measured traffic flow sizes. The three plots are controlled by s. *First Plot*: $s = 2$; *Second Plot*: $s = 5$; *Third Plot*: $s = 10$.

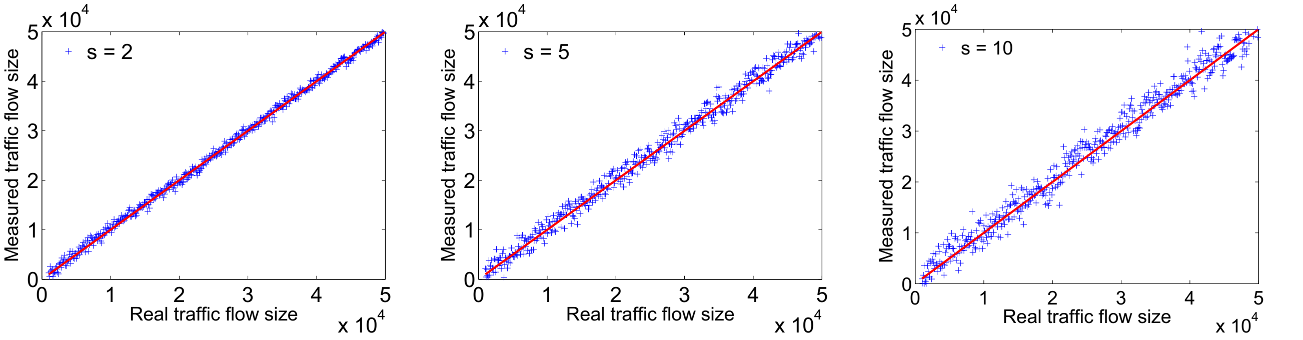

Figure 3. Measurement accuracy with optimal privacy, $n_x = n_y = n = 100,000$, $n_c = [0.01n, 0.5n]$. The x-axis shows true traffic flow sizes, and the y-axis shows the corresponding measured traffic flow sizes. The three plots are controlled by s. *First Plot*: $s = 2$; *Second Plot*: $s = 5$; *Third Plot*: $s = 10$.

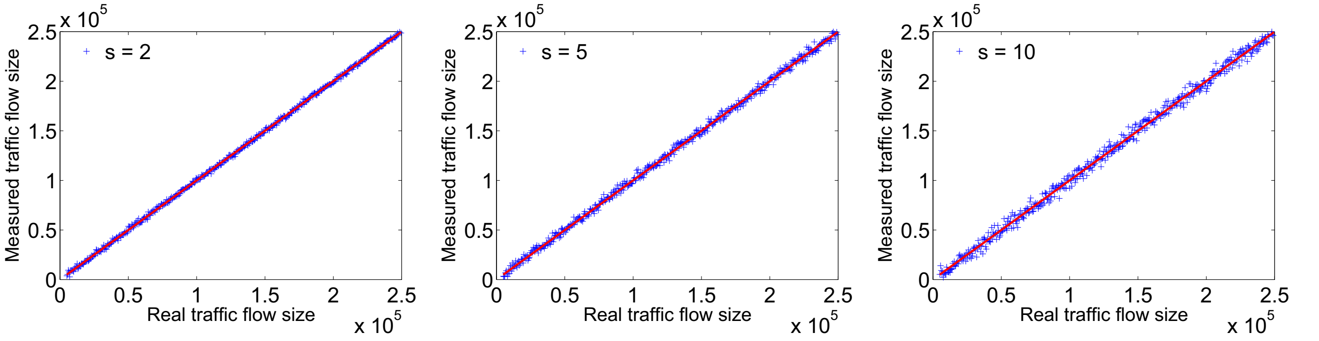

Figure 4. Measurement accuracy with optimal privacy, $n_x = n_y = n = 500,000$, $n_c = [0.01n, 0.5n]$. The x-axis shows true traffic flow sizes, and the y-axis shows the corresponding measured traffic flow sizes. The three plots are controlled by s. *First Plot*: $s = 2$; *Second Plot*: $s = 5$; *Third Plot*: $s = 10$.

Figure 2, 3, and 4). When s becomes larger, the variant for our estimator also becomes larger, producing relatively more points not close to the equality line (the third plot in Figure 2, 3, and 4), which indicates larger values of s yield less accurate measurement results.

Recall that a larger value of s brings better privacy (Table I). For example, the optimal privacy is 0.7661 when $s = 10$, better than the optimal privacy of 0.7258 when $s = 2$. This implies a tradeoff between the preserved privacy and the measurement accuracy. From Section IV-C, we know when s is large, there always exists a smooth interval of m near its extreme point that can achieve comparable privacy as the optimal. In reality, one can choose a relatively large value for s (e.g., 5 or 10), and adjust the value of m to achieve better measurement results while still preserving comparable privacy as the optimal.

Finally, the measurement results are more accurate with larger values of n. There are fewer points deviating from the equality line $\hat{n_c} = n_c$ in the three plots of Figure 4 than those in the corresponding plots of Figure 2. This is also a natural phenomenon given that the measurement result is obtained through an MLE estimator.

VI. Related Work

A. Transportation Traffic Measurement

In the area of transportation traffic measurement, various prediction models have been proposed to measure "point" traffic statistics, using data recorded by automatic traffic recorders (ATR) installed at road sections. For example, the multiple linear regression model in [2], artificial neural network in [3], and spatial statistical method in [4], etc. Those solutions, though elegant, are not appropriate for "point-to-point" transportation traffic measurement. The recent work in [5] tries to infer "point-to-point" statistics from "point" data, but the high computation overhead limits its practicability. We prefer a more accurate and efficient direct-measurement approach that should also address the privacy concern. Although Google recently announced to provide real-time traffic data service in Google maps [11], their approach cannot assure vehicle's privacy since it uses GPS and Wi-Fi in phones to track locations [12].

B. Network Traffic Measurement

Another branch of research that relates to (but is also significantly different from) ours is network traffic measurement, where researchers have proposed various methods for traffic flow measurement. Though having the same name, their problem is different from ours: to measure the network traffic between two network routers. The solutions can be summarized into two categories. One is indirect estimation based on link load and network routing, by employing statistical techniques [13] [14]. These methods cannot achieve high accuracy since their estimations are based on the unknown traffic volume. The other is direct measurement by different counting methods [15] [16]. In particular, Li et al. [16] develop a bitmap-based counting method for traffic flow measurement, which is most related to our work. However, all these solutions are not appropriate for our problem, because they measure traffic in network environment where the privacy of packets is not a concern, and counting can be done directly based on the packet IDs. In our problem, the privacy of vehicles is the major concern. Therefore, the solutions must incorporate randomization and de-identification techniques to protect vehicles' privacy, and do counting based on information that looks totally random.

VII. Conclusion

In this paper, we focus on the problem of privacy-preserving "point-to-point" transportation traffic monitoring in intelligent cyber-physical road systems. We formalize "point-to-point" traffic as traffic flows, and quantify privacy as a probability. We propose a novel scheme which allows the collection of aggregate traffic flow data while preserving the optimal privacy of individual vehicles. The proposed scheme utilizes bit arrays to collect "masked" data and adopts maximum likelihood estimation (MLE) to obtain the measurement result. Its feasibility and scalability are shown by both mathematical proofs and simulations.

VIII. Acknowledgment

This work is supported in part by the National Science Foundation under grant CPS 0931969 and by a grant from the Center for Multimodal Solutions for Congestion Mitigation sponsored by the US Department of Transportation.

References

[1] USDOT, "Traffic Monitoring Guide," *http://www.fhwa.dot.gov/ohim/tmguide/tmg3.htm*, 2001.

[2] D. Mohamad, K. C. Sinha, T. Kuczek, and C. F. Scholer, "Annual Average Daily Traffic Prediction Model for County Roads," *Journal of the Transportation Research Board*, vol. 1617/1998, pp. 69–77, 1998.

[3] W. Lam and J. Xu, "Estimation of AADT from Short Period Counts in Hong Kong – A Comparison Between Neural Network Method and Regression Analysis," *Journal of Advanced Transportation*, 2000.

[4] J. K. Eom, M. S. Park, T. Heo, and L. F. Huntsinger, "Improving the Prediction of Annual Average Daily Traffic for Nonfreeway Facilities by Applying a Spatial Statistical Method," *Journal of the Transportation Research Board*, vol. 1968/2006, pp. 20–29, 2006.

[5] Y. Lou and Y. Yin, "A Decomposition Scheme for Estimating Dynamic Origin-destination Flows on Actuation-controlled Signalized Arterials," *Transportation Research Part C*, vol. 18, 2010.

[6] J. Eriksson, H. Balakrishnan, and S. Madden, "Cabernet: Vehicular Content Delivery Using WiFi," *Proc. of MOBICOM*, 2008.

[7] U. Lee, J. Lee, J. Park, and M. Gerla, "FleaNet: A Virtual Market Place on Vehicular Networks," *IEEE Trans. on Vehicular Technology*, 2010.

[8] [Online]. Available: http://www.its.dot.gov/press/2010/vii2intellidrive

[9] [Online]. Available: http://www.dot.gov/

[10] G. Casella and R. L. Berger, "Statistical Inference," *2nd edition, Duxbury Press*, 2002.

[11] "Google map's time-in-traffic feature." [Online]. Available: http://mashable.com/2012/03/29/google-maps-traffic-data/

[12] T. Jeske, "Floating Car Data from Smartphones: What Google and Waze Know About You and How Hackers Can Control Traffic," *Proc. of the BlackHat Europe*, 2013.

[13] Y. Zhang, M. Roughan, C. Lund, and D. Donoho, "An information-theoretic approach to traffic matrix estimation," *Proc. of SIGCOMM*, 2003.

[14] Y. Zhang, M. Roughan, N. Duffield, and A. Greenberg, "Fast accurate computation of large-scale IP traffic matrices from link loads," *Proc. of SIGMETRICS*, 2003.

[15] J. Cao, A. Chen, and T. Bu, "A Quasi-Likelihood Approach for Accurate Traffic Matrix Estimation in a High Speed Network," *Proc. of INFOCOM*, 2008.

[16] T. Li, S. Chen, and Y. Qiao, "Origin-Destination Flow Measurement in High-Speed Networks," *Proc. of INFOCOM*, 2012.