Compressive Sensing based Monitoring with Vehicular Networks

Hongjian Wang†, Yanmin Zhu†, Qian Zhang‡
† Shanghai Jiao Tong University
‡ Hong Kong University of Science and Technology
† {thekingofkings,yzhu}@sjtu.edu.cn; ‡qianzh@cse.ust.hk

***Abstract*—Vehicles are becoming powerful mobile sensors, and vehicular networks provide a promising platform to support a wide range of existing large-scale monitoring applications such as road surface monitoring, and etc. In vehicular networks, inter-vehicle contacts are scarce resources for data delivery. This presents a major challenge for monitoring applications with vehicular networks. By analyzing a large dataset of taxi traces collected from around 2,600 taxis in Shanghai, China, we reveal that there is strong correlation with data readings on vehicles. Motivated by this important observation, we propose a compressive sensing based approach called `CSM` to monitor with vehicular networks. Two key issues must be addressed. *First*, there is an intrinsic tradeoff between communication cost and estimation accuracy. *Second*, guaranteed estimation accuracy should be provided over the highly dynamic network. To address the above issues, we first characterize the relationship between estimation error (ℓ_2 error) and sparsity property of a dataset. Then, we determine two critical parameters: the minimum number of seeds and the minimum transmission hop length for compressive measurements in the network. The selection of the two parameters can reduce the communication cost while guaranteeing the required estimation accuracy. Extensive simulations based on real vehicular GPS traces collected in Shanghai, China have been performed and results demonstrate that `CSM` achieves much higher estimation accuracy at the same communication cost compared with other alternative schemes.**

I. Introduction

With the rapid deployment of inter-vehicle radio technologies such as dedicated short range communication (DSRC), it has been practical to realize vehicle-to-vehicle and vehicle-to-infrastructure communications, making vehicular networks a reality. However, as vehicles may be distributed over a vast area and move at a high speed, it is often difficult to find a connected path between any two vehicles. We may have to leverage the opportunity that vehicles meet with each other or meet with Access Points (APs) deployed along the road for relaying information in an opportunistic manner. We also call such vehicular networks vehicular delay-tolerant networks.

Recently, more and more vehicles are equipped with various sensors, such as accelerometers, pollution sensors and Global Positioning System (GPS) receivers. Thus, vehicles are becoming powerful mobile sensors, and vehicular networks provide a promising platform supporting a wide range of existing large-scale monitoring applications, such as road surface monitoring [1], urban monitoring [2] and etc. In this paper we consider monitoring with vehicular networks. As a motivating example shown in Fig. 1, each vehicle detects road surface conditions and such readings are collected by a monitoring center responsible for detecting defects of urban roads in the city.

The main objective of a monitoring application is to achieve high accuracy. A straightforward approach to monitoring with vehicular networks is to have each vehicle to report its readings independently to one of the APs by using multi-hop routing algorithms [3] [4]. Although this approach may be able to collect all the readings, there are several main defects. First, in a vehicular network, a vehicle can only communicate with another vehicle when they encounter with each other and such encounter opportunities may be scarce and should be well utilized. This straightforward approach however introduces a high communication overhead. Second, a complete round of monitoring finishes only when each of the vehicle delivers its reading to the APs. It will take a very long time. Thus, such straightforward approach is inappropriate for monitoring with

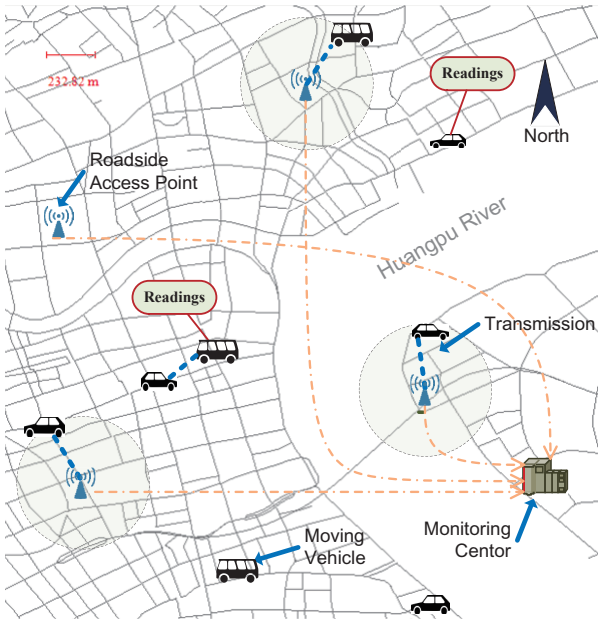


Fig. 1. Motivating example for monitoring with vehicular networks: vehicles running in the urban area of Shanghai, China. Each vehicle periodically detects the road surface condition and a monitoring center wants to collect road conditions collected by the vehicles. The vehicles can report their readings to the road side access points (APs) by multi-hop transmissions, which then relay the readings to the monitoring center.

vehicular networks.

To achieve better performance, we explore data readings on vehicles in the real world, which actually exhibit some interesting hidden structures. We mine a large dataset of taxi traces collected from around 2,600 taxis in Shanghai, China. In this dataset, we have recorded data readings from the taxis, including instant speed and position with a time granularity from 15 seconds to several minutes. The dataset spans a duration of nearly 2 years. Through analyzing data readings from these real datasets using entropy analysis, we demonstrate that there is strong correlation with the data readings.

Motivated by this important observation, we propose a compressive sensing based approach called `CSM` to monitor with vehicular networks. *Compressive sensing* is an effective technique for recovering data with sparsely sampled data [5]. With `CSM`, only a small number of vehicles are selected as seeds to start compressive measurements. As the measurement messages propagate across the network, they coalesce new readings when vehicles encounter with each other. Vehicles offload measurements at APs as they pass by. After receiving the measurement data from APs, the monitoring center can recover the readings by applying the compressive sensing technique on the set of measurement data.

Two key issues must be addressed for this compressive sensing based approach. *First*, there is an intrinsic tradeoff between communication cost and estimation accuracy. To save the precious communication opportunities in vehicular networks, it is highly desirable to use a smaller number of seeds and a short propagation hop length for compressive measurements. However, the performance of data estimation may also be degraded. *Second*, monitoring applications usually pose a rigid requirement on the estimation error. It is difficult to guarantee the estimation accuracy given the highly dynamic nature of vehicular networks.

In [6], Li et al. propose to recover average speeds of road segments with sparsely sampled data from vehicles with a compressive sensing based technique. However, in their approach the sensed data on vehicles are transmitted over the cellular network. The inter-vehicle communication is not considered. Some techniques based on compressive sensing have been proposed for collecting data from sensor networks. In [7], [8], [9] some compressive sensing-based methods have been proposed for data collection in sensor networks. In [10], the validity of the compressive sensing theorem for data collection in sensor networks is rigorously justified. These existing techniques for sensor networks cannot be applied in monitoring with vehicular networks. They usually rely on a fixed routing structure which is unavailable in vehicular networks that are highly mobile and link availability is scarce. Furthermore, a main focus of sensor networks is energy efficiency which is not a problem for vehicular networks.

To address the challenges mentioned above, we first characterize the relationship between estimation error (ℓ_2 error) and sparsity property of a dataset. Then, by modeling a vehicular network as a contact graph, next, we determine two critical parameters: the minimum number of seeds and the minimum transmission hop length for compressive measurements in the network. The selection of the two parameters can reduce the communication cost while guaranteeing the required estimation accuracy. Finally, extensive simulations based on real vehicular GPS traces collected in Shanghai, China have been performed and results demonstrate the efficacy of our approach.

We have made the following intellectual contributions:

- By analyzing a large dataset of real data readings on vehicles using entropy analysis, we reveal that there is strong correlation with data readings from different vehicles.
- We propose an approach called `CSM` based on compressive sensing for monitoring with mobile vehicles. `CSM` guarantees the data estimation accuracy while significantly reducing the communication cost.
- We analytically derive the expected delay and the communication cost for compressive sensing based data collection from a vehicular network.
- Trace-driven experiments have been conducted, which confirm that the required accuracy is guaranteed and significant reduction of communication cost is achieved.

The remainder of the paper will be organized as follows. Section II gives the preliminary on the network model, empirical study and compressive sensing. The detailed design of `CSM` is described in Section III. In Section IV we give theoretical analysis. Section V discusses evaluation results. In Section VI we review related work. Finally, we conclude the paper in Section VII.

II. Preliminary

In this section we first present the network model, then introduce our empirical study with real vehicular data readings, and finally introduce the preliminary of compressive sensing.

A. System Model

There are a set C of n vehicles which can communicate with each other as they encounter, i.e., within the communication range, $C = \{c_1, c_2, \cdots, c_n\}$. A set A of wireless access points (APs) are deployed in the road network. The APs are connected through a wired network, such as the Internet. A vehicle can also communicate with an AP when they are within the communication range.

Each vehicle periodically generates data readings about the vehicle itself or the surrounding environment. Let the reading of vehicle c_i be denoted by x_i. We refer to a *message* as a packet carrying data readings. A monitoring center connects to all the APs and wants to retrieve all data readings of the vehicles in the network.

B. Empirical Study with Real Traces

We next reveal the existence of correlation with data readings from vehicles by analyzing the dataset of real traces. The traces were collected from around 2,600 taxis in Shanghai, China. Equipped with the Global Positioning System (GPS)

receiver, each taxis periodically reads information such as speed, position, and occupancy. A record of the readings as follows is saved. The traces span a duration of over TWO

Vehicle ID	Speed	Longitude	Latitude	Occupancy	Timestamp

years, from January, 2006 to December, 2007. Thus, the traces provide a real dataset of data readings on vehicles.

We use entropy analysis on the real datasets of speeds and positions. Before presenting entropy analysis, we give some notations for the datasets. We denote the time instants by $T = \{t_1, t_2, \cdots, t_m\}$. Then, the data readings for a vehicle, c_i, at the j-th time slot t_j is $x_{i,j}$. For all vehicles, we have a matrix representing all readings

$$\mathbb{X} = \begin{bmatrix} x_{1,1}, & x_{1,2}, & \cdots, & x_{1,m} \\ x_{2,1}, & x_{2,2}, & \cdots, & x_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1}, & x_{n,2}, & \cdots, & x_{n,m} \end{bmatrix}, \tag{1}$$

where the i-th row is a sequence of data readings for c_i, and the j-th column refers to the set of readings at time t_j.

We next show the calculation of the marginal entropy of data readings of a given vehicle. For each vehicle c_i, we take $X_i = \{x_{i,1}, x_{i,2}, \cdots, x_{i,n}\}$ as a series of observations for its readings. Therefore, the entropy of each X_i can be calculated by

$$H(X_i) = \sum_{k=1}^{n} \Pr(x_{i,k}) \times \log_2 \frac{1}{\Pr(x_{i,k})}. \tag{2}$$

We can get the probability of $x_{i,k}$ in $\vec{x_i}$ by taking statistics. Suppose that $x_{i,k}$ appears σ_k times in set X_i, and X_i has a length n. The probability of $x_{i,k}$ is

$$\Pr(x_{i,k}) = \frac{\sigma_k}{n}. \tag{3}$$

After getting the marginal entropy, we show the calculation of the conditional entropy of $\mathbb{X}$. Given X_{i1}, the conditional entropy of X_i can be calculated as,

$$H(X_i|X_{i1}) = H(X_{i1}, X_i) - H(X_{i1}). \tag{4}$$

The joint entropy $H(X_{i1}, X_i)$ can be calculated as,

$$\begin{aligned} &H(X_{i1}, X_i) \\ &\quad = \sum_{k=1}^{n} \Pr((x_{i1,k}, x_{i,k})) \times \log_2 \frac{1}{\Pr((x_{i1,k}, x_{i,k}))}. \end{aligned} \tag{5}$$

We construct a joint set from two reading sets X_{i1} and X_i

$$X_i' = \begin{bmatrix} (x_{i1,1}, & x_{i,1}) \\ (x_{i1,2}, & x_{i,2}) \\ \vdots & \vdots \\ (x_{i1,n}, & x_{i,n}) \end{bmatrix}, \tag{6}$$

where X_{i1} is the readings of another vehicle c_{i1} different from c_i. We count the number of occurrences of $(x_{i1,k}, x_{i,k})$ as σ_k; the probability of $(x_{i1,k}, x_{i,k})$ can be calculated by Eq. (3)

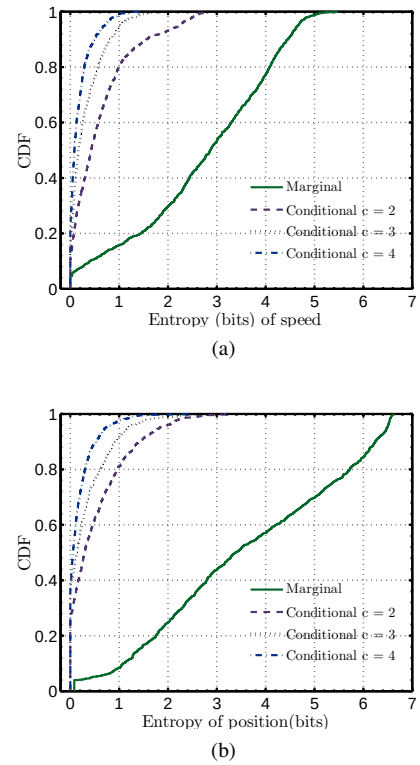


Fig. 2. CDFs of marginal entropy and conditional entropy: (a) Speed readings and (b) Position readings.

We can further generalize Eq. (4) as,

$$\begin{aligned} &H(X_i|X_{i1}, X_{i2}, \cdots) \\ &\quad = H(\cdots, X_{i2}, X_{i1}, X_i) - H(X_{i1}, X_{i2}, \cdots). \end{aligned} \tag{7}$$

In Fig. 2, we plot the CDF of marginal entropy and conditional entropy of speed readings and position readings, respectively. We can find that the conditional entropy is much smaller than the marginal entropy for both speeds and positions. This demonstrates that these data readings from vehicles are strongly correlated. In addition, we can see position readings have a larger marginal entropy than speed readings, while position readings have a smaller conditional entropy than speed readings. This indicates that position readings are more correlated than speed readings.

C. Compressive Sensing

We give some preliminary of compressive sensing. The compressive sensing can produce a good approximation of an original n-dimensional vector with only m samples, where $m < n$. More importantly, data loss tolerance is an innate ability of compressive sensing.

According to the compressive sensing theory [5], a K-sparse dataset (with K number of non-zero coefficients) can be fully recovered by solving a programming optimization problem with non-adaptive linear projections which preserve the structure of the sparse dataset. Suppose we have the original data readings, an n-by-1 vector $\vec{x}$. It has the K-sparse

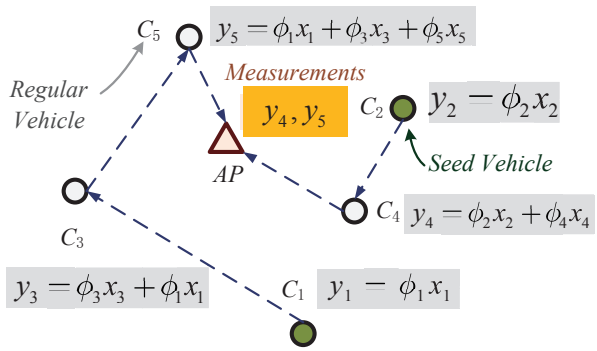


Fig. 3. An example for explaining the basic idea of the CSM approach.

representation $\vec{d}$ in some domain with representation basis Ψ,

$$\vec{d} = \Psi \vec{x} \tag{8}$$

Furthermore, if $K \ll n$, then $\vec{d}$ can be fully recovered by a small number of measurements using an m-by-n $(m < n)$ compressive sampling basis Φ, with

$$\vec{y} = \Phi \Psi^{-1} \vec{d}, \tag{9}$$

where $\vec{y}$ is acquired from $\vec{y} = \Phi \vec{x}$.

Eq. (9) illustrates the core of the compressive sensing technique. By solving the following ℓ_1-norm optimization problem,

$$(L_1) \quad \min_{d} ||\vec{d}||_{\ell_1} \text{ subject to } \vec{y} = \Theta \vec{d}, \tag{10}$$

where the ℓ_1-norm of $\vec{d}$, $||\vec{d}||_{\ell_1} = \sum_i^n |d_i|$, and $\Theta = \Phi \Psi^{-1}$. We can find out an n-by-1 vector $\vec{d'}$ and $\vec{x'} = \Psi^{-1} \vec{d'}$, which is a good approximation to the original $\vec{x}$.

III. Design of CSM

In this section we present the detailed design of CSM.

A. Basic Idea

In this section we give the basic idea of our approach using compressive sensing for data collection with a vehicular network. We illustrate the basic idea in Fig. 3. In the example, there are 5 vehicles $\{c_i | 1 \leq i \leq 5\}$ and one AP. And each vehicle c_i has one reading x_i. The AP wants to obtain all readings of the five vehicles.

With CSM, a small number of vehicles are selected as *seed vehicles*. The rest of the vehicles are *regular vehicles*. Vehicle c_1 and c_2 are such two seed vehicles, and other vehicles including $c_3, c_4,$ and c_5 are regular vehicles. Initially, a seed vehicle generates a message containing its own data reading multiplied by a random coefficient. Let's take c_1 for example. It generates message $y_1 = \phi_1 x_1$, where x_1 is the data reading of c_1 and ϕ_1 is a random coefficient.

Upon encountering with other regular vehicles, a seed vehicle forwards the message to the regular vehicle. Upon receiving a message, a regular vehicle combines its own reading with the received data by taking the weighted sum. For example, c_1 forwards its message containing y_1 to c_3. On receiving y_1 from c_1, c_3 combines x_3 and y_1 and generates y_3.

Eventually, the AP receives a set of combined data, or *measurements*. It then recovers the readings using the compressive sensing technique as introduced in Section II-C. For example, the AP receives two measurements y_4 and y_5 from c_4 and c_5, respectively. It then tries to recover the readings x_1-x_5 based on compressive sensing.

B. Key Issues

Given a fixed number of measurements, the estimation error of compressive sensing is strongly dependent on the sparsity of the dataset of readings. The estimation error is smaller if the sparsity of the dataset, K, is smaller. For a dataset in the real world, however, the sparsity is relatively large. To fully recover the dataset without error requires an extremely large number of measurements. This is impractical for vehicular networks where inter-vehicle contacts are scarce resources.

Many monitoring applications can tolerate a certain relatively small estimation error defined by ϵ_0. The objective of CSM is to provide a guaranteed estimation accuracy while minimizing the communication cost.

We have to address two key issues in the design of CSM as follows:

- **Key issue 1**: What is the minimum number of seeds to ensure the estimation accuracy?
- **Key issue 2**: What is the minimum number of transmission hops that a measurement should be forwarded in order to coalesce a sufficient number of data readings?

In this section we address the key issues. First, we characterize the relationship between the relative estimation error and the sparsity property of data readings. Then, we determine the minimum number of seeds and the minimum number of hops required for measurements.

C. Characterizing Estimation Error

To tolerate estimation error, we optimistically underestimate the real sparsity K of the dataset by ignoring the small tail of the dataset represented in another domain. Let us call it *virtual sparsity*, denoted by K'. We characterize the relationship between the estimation error and the virtual sparsity. *Essentially, a smaller virtual sparsity requires fewer measurements but results in a higher estimation error.*

We consider the ℓ_2 error of the estimation, $||\vec{x'} - \vec{x}||_{\ell_2}$, which is the ℓ_2 norm of $(\vec{x'} - \vec{x})$ given by $\left(\sum_i^n |x'_i - x_i|^2\right)^{1/2}$. Let $E(X)$ denote this estimation ℓ_2 error, and we have

$$E(X) = ||\vec{x'} - \vec{x}||_{\ell_2} \leq C_0 K'^{-\frac{1}{2}} \sum_i^n |d_i - d_{K'i}|, \tag{11}$$

where C_0 is a constant coefficient, and $\vec{d}$ is the representation of $\vec{x}$ in some transform domain. This upper bound is given in [11], and $\vec{d}_{K'}$ is the vector with the exact locations and amplitudes of the K'-largest entities of $\vec{d}$.

Usually, there are few real datasets that are strictly K-sparse so we adopt the following method to approximate a coefficients vector. We sort the coefficients in the descending order denoted as $|\theta_1| \geq |\theta_2| \geq \cdots \geq |\theta_n|$. Then, we keep the

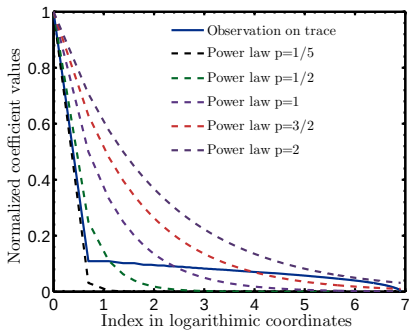


Fig. 4. Normalized coefficient values of a dataset of real taxi speeds in the traces. This shows that a dataset of real data readings approximately follows the power law.

largest K' coefficients and discard all other entities. We have shown that data readings are redundant with entropy analysis. Furthermore, the magnitude of its transform coefficients decays in power law, i.e., the ith largest coefficient satisfies,

$$|\theta_i| \leq Ri^{-1/p}, \tag{12}$$

where R is a constant and $0 < p < 2$.

To verify this, we project a bunch of real speed readings in the traces into the discrete Fourier domain. We sort all the Fourier coefficients in descending order, calculate the mean value in each position, and finally normalize all the coefficients. Notice that we plot on the logarithmic coordinates, thus depicting the fast decreasing property of the speed values. In Fig. 4, we show the normalized coefficients of a real taxi speed dataset. We can find that the dataset of vehicle speeds approximately follow the power law.

In fact, the power law is an empirical description of the sparsity of a dataset. Theoretically, we have

$$||\vec{\theta}||_p = \left(\sum |\theta_i|^p\right)^{1/p} \leq R, \tag{13}$$

to formulate the constraint on sparsity, which is generally obeyed on natural classes. ℓ_p norm with small p is the natural mathematical measure of sparsity, and as p decreases below 1, more and more sparsity is required. Also, from this viewpoint, an ℓ_p constraint based on $p = 2$ requires no sparsity at all.

With the K'-term approximation and constraints on sparsity we can reduce the ℓ_2 error in Eq. (11) to

$$||\vec{x'} - \vec{x}||_{\ell_2} \leq \alpha_p R K'^{1/2-1/p}. \tag{14}$$

Here p is the inherent property of a specific dataset, and α_p and R are constants determined by p.

The ℓ_2 error bound given in Eq. (14) claims that the estimation error can be determined by the selecting the K'-strongest term. The larger K' we use, the smaller the ℓ_2 error is. Meanwhile, a larger overhead is incurred.

We use the relative ℓ_2 error ϵ as our optimization objective, and

$$\epsilon = \frac{||\vec{x'} - \vec{x}||_{\ell_2}}{||\vec{x}||_{\ell_2}}. \tag{15}$$

The $||\vec{x}||_{\ell_2}$ is the property of a specific dataset, just as given in Eq. (13).

D. Determining m and h

With the previous characterization of the relationship between relative estimation error and virtual sparsity, we next determine two important parameters, i.e., m and h, representing the number of measurements and the number of nonzeros. These two parameters directly depict the properties of the compressive sensing sample matrix. Therefore, by analyzing the property of the sample matrix, we can determine m and h.

According to compressive sensing theory, Φ needs to preserve the Restricted Isometry Property (RIP), and the original dataset can then be recovered from m compressed measurements. We give the definition of the RIP [12]: let F be an m-by-n random sample matrix. If F has RIP constants δ_s for every $x \in \{x \in \mathbf{R}^n : ||x||_0 \leq s\}$ satisfying

$$\delta_s \triangleq \arg\min_{\delta}(1-\delta)||x||_2^2 \leq ||Fx||_2^2 \leq (1+\delta)||x||_2^2, \tag{16}$$

then F is said to satisfy the RIP.

The number δ_s measures how close the vectors in F are to behaving like an orthonormal system. The parameter m tells how many of rows Φ has, and h is the number of nonzeros in each row. They jointly determine whether Φ satisfies the RIP.

Number of measurements m. It is a well-known result in compressive sensing theory that a K-sparse dataset can be recovered from Φ with $m = O(K \log N)$ measurements. Specifically, m should satisfy the following condition [13]:

$$m \geq c \cdot \mu^2(\Phi, \Psi) \cdot K \cdot \log N, \tag{17}$$

where $\mu(\Phi, \Psi)$ is the coherence between sampling basis Φ and transform basis Ψ.

Number of nonzeros h. According to [14], we know that a sparse random sample matrix Φ is also sufficient to recover the data readings. A peak-to-total energy condition needs to be satisfied, which is $\frac{||\vec{d}||_\infty}{||\vec{d}||_2} \leq \omega$. This peak-to-total condition bounds the largest component of the data readings.

We define the sparsity of Φ as the number of nonzeros in each row, which is exactly h. The sparsity of Φ will produce an extra factor of $1/p\omega^2$ in the number of measurements, where $p = h/n$ refers to the probability of an entity in Φ to be non-zero. It turns out that as h becomes smaller, we need a larger m to compensate the information loss. For data that is K-sparse in Fourier transform, if $h = \log^2 n$, then $1/p\omega^2 = O(1)$, in which case there is no side effect on the number of measurements m. If $h = \log n$, there is an extra factor of $1/p\omega^2 = O(\log n)$ in m. In our approach we set the $h = \log^2 n$ to achieve a better performance.

E. Determining Number of Seeds

With the optimal number of measurements and the optimal number of nonzeros, we next determine the minimum number of seeds. Since the success of a message forwarding from a vehicle to the APs in the vehicular network is probabilistic, we have to determine what is the minimum number of seeds required to ensure the APs can collect m measurements with a time constraint.

To facilitate our analysis, we model the vehicular network as a contact graph $G = \{V, E\}$ $(V = C)$. An edge $e_{i,j} \in E$ between two vehicles c_i and c_j represents the contact process. As shown by previous studies [15] [16], the inter-contact time of two vehicles is exponentially distributed. Thus, the contact process is a Poisson process. Let $\lambda_{i,j}$ denote the contact rate between c_i and c_j.

The minimum number of seeds is given by the following theorem.

Theorem 1. *The minimum number of seeds for time constraint D is $\frac{1}{\sigma} \times m$, where $\sigma = \dfrac{e^{-\lambda_{t,t+D}D}(\lambda_{t,t+D}D)^h}{h!}$.*

Proof: Noticing that in the inter-contact time model, the encounter event model between vehicles is just the same as that between vehicle and AP. Thus, we use *node* to represent both vehicle and AP without lose of generality.

For a node tuple (n_p, n_q), the inter contact time $\tau(n_p, n_q)$ is exponentially distributed with rate $\lambda_{p,q}$. Thus, for a specific packet forwarding h hops within time D is a Poisson Process. We use $N(t+\Delta t)$ to denote the number of encounters during time interval Δt. We have

$$\Pr(N(t+D) - N(t) = h) = \frac{e^{-\lambda_{t,t+D}D}(\lambda_{t,t+D}D)^h}{h!}, \quad (18)$$

where $\lambda_{t,t+D} = \int_t^{t+D} \lambda(t)dt$ for the rate parameter may change over time.

Eq. (18) gives us the probability that one message can be finally collected at the APs, i.e., the delivery ratio. Let σ denote the delivery ratio. Then, we can derive the number of measurements needed at the very beginning is $\frac{1}{\sigma}m$. ■

F. Determining Number of Hops

It is easy to find that the number of nonzeros h in each row of the sample matrix Φ actually indicates the number of data readings that a measurement should include, i.e. the number of hops. Essentially, it is the number of vehicles that are involved in one measurement. Thus, the message from a seed vehicle initially takes a random walk. As the number of hops of the measurement message equals to h, it should be quickly moved to one of the APs. To this end, a number of routing algorithms designed to optimize the delivery delay can be used, such as Delegation Forwarding [4], and etc.

IV. Analysis

In this section we provide analytical results on the expected delay and the communication cost of our approach.

The overall delay is determined by the delay of the last measurement arriving at the APs. This delay is analytically derived, as shown in the following theorem.

Theorem 2. *Based on an inter-contact time model, our compressive sensing based approach has a total delay D,*

$$D = \sum_{i=0}^{h-1} \frac{\binom{n}{2}}{\binom{n}{2} - i} E(\tau), \quad (19)$$

where n is the number of vehicles and τ is the inter contact time.

Proof: Clearly, we have

$$D = h \times E(\kappa), \quad (20)$$

where κ is the single hop delay.

In our approach each message is only forwarded to a vehicle that has not involved in the measurement. All encounter events among vehicles are independent.

For a vehicle carrying a q-hop message, the next encounter may not result in the inclusion of a new reading in the measurement. Thus, for the message currently having q hops, we take x as the number of encounters needed to get one success, where $x \sim \text{Geometric}(p)$. Thus, for message m with q hops, $x^{(q)} \sim \text{Geometric}(\frac{\binom{n}{2}-q}{\binom{n}{2}})$. The expectation of the next-hop forwarding time is

$$\begin{aligned} E(m^{(q)} \to m^{(q+1)}) &= E(x^{(q)}) \times E(\tau) \\ &= \frac{\binom{n}{2}}{\binom{n}{2} - q} \times E(\tau). \end{aligned} \quad (21)$$

Therefore,

$$\begin{aligned} E(\kappa) &= \frac{1}{h} \sum_{i=0}^{h-1} E(m^{(i)} \to m^{(i+1)}) \\ &= \frac{1}{h} \sum_{i=0}^{h-1} \frac{\binom{n}{2}}{\binom{n}{2} - i} \times E(\tau). \end{aligned} \quad (22)$$

Replacing Eq. (22) in Eq. (20), we finish the proof. ■

We also analytically derive the communication cost of our approach, as shown in the following theorem.

Theorem 3. *The overall communication cost of* `CSM` *is bounded by $O(K \log^3 N)$.*

Proof: From the proof for Theorem 2, we can easily derive the forwarding overhead o_i for msg_i is

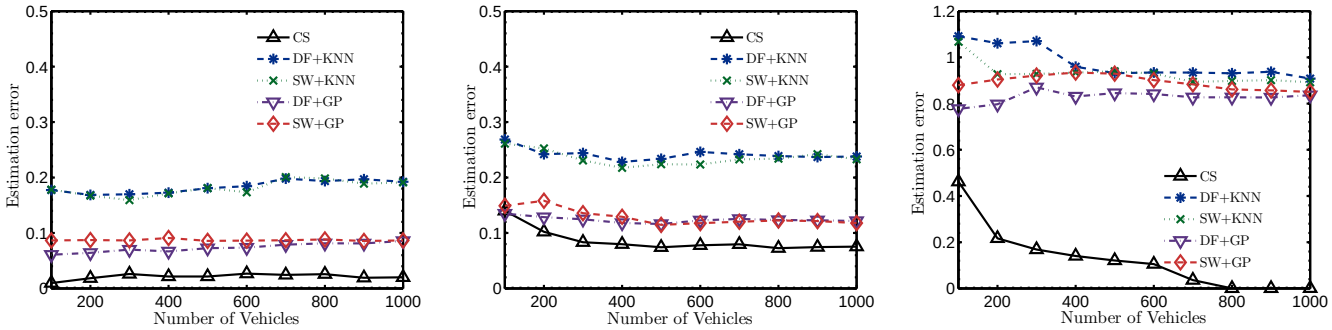
$$E(o_i) = \sum_{i=0}^{h-1} \frac{\binom{n}{2}}{\binom{n}{2} - i}.$$

We use a sparse sample matrix. Thus, $h \ll \binom{n}{2}$. Then, we have $E(o_i) = h$. Because h is small, we can treat it as a constant in Theorem 1. We end up with $m' = O(m)$. Finally, we have the overall communication cost $O(mh)$. Use the constraints for m and h, we can derive the result. ■

We should notice that by using a sparse sample matrix, we can have a small h in our approach. Our approach can achieve the overhead as low as $O(\log^3 N)$, while the required estimation error can be guaranteed.

V. Performance Evaluation

In this section, we first present the methodology and experimental setup, and then present evaluation results.



(a) IDEAL (b) POSITION (c) SPEED

Fig. 5. The estimation error against the number of vehicles for all the three datasets.

TABLE I
SUMMARY OF TRACES

Parameter	Value
Date	Feb. 19, 2007
Duration (hours)	24
Granularity (seconds)	15 - 60
Total number of contacts	685,000

A. Methodology and Experimental Setup

To evaluate the performance of our approach, we have performed simulations driven by real vehicular traces. In simulation, a vehicle moves on the road network in Shanghai, China, strictly following a trajectory recorded in the traces. The traces that we have used for simulations is summarized in Table I. The communication range is 150 meters.

We use three datasets of readings. Two are from the real readings included in the taxi traces, as introduced in Section II. One synthetic dataset of readings is generated, which has a sparsity of around 20. The three datasets of readings are called SPEED, POSITION and SYNTHETIC, respectively.

We adopt comparative study, comparing our approach with other alternative approaches. We divide the data collection task into two phases. In the first phase, readings are forwarded to the APs. In the second phase, data recovery is performed. Therefore, we use two different schemes in each phase. In total, there are four combinations, resulting in four approaches for comparison.

- **Delegation forwarding (*DF*)**. It forwards a packet only to a node with higher probability of delivering the packet to the destination. We implement it with statistical information of encounters between vehicles from the traces.
- **Spray and wait (*SW*)**. It forwards multiple copies of each packet. Initially, a source node has a copy budget of eight for each packet. Then, the relay node will spray half of the budget to each encountered vehicle.

We use two representative estimation schemes to recover readings of all vehicles.

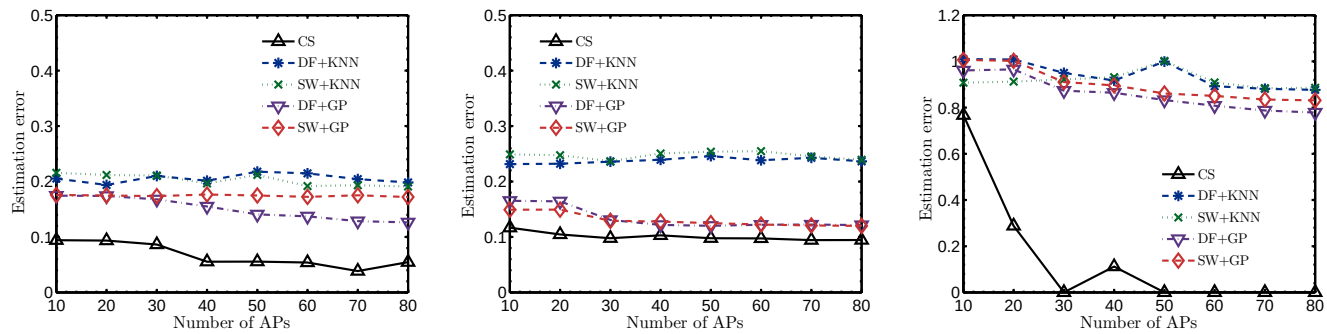
- **k-NN estimate (*KNN*)**. It computes the mean value of the k nearest neighbors of the vehicle as its sensor reading. Using the k-NN estimate, we can get a local approximation, which fits the real case that the nearer neighbors contribute to the estimates.
- **Gaussian process regression (*GP*)**. It is also known as Kriging, which is a geostatistical technique to interpolate the value of a random field at unobserved locations from observations of its value at nearby locations.

For fair comparison among different approaches, we first run our approach and record the communication cost. Other approaches can use the same communication cost to deliver readings towards the APs. Readings successfully received by the APs are then used for data recover.

B. Impact of Number of Vehicles

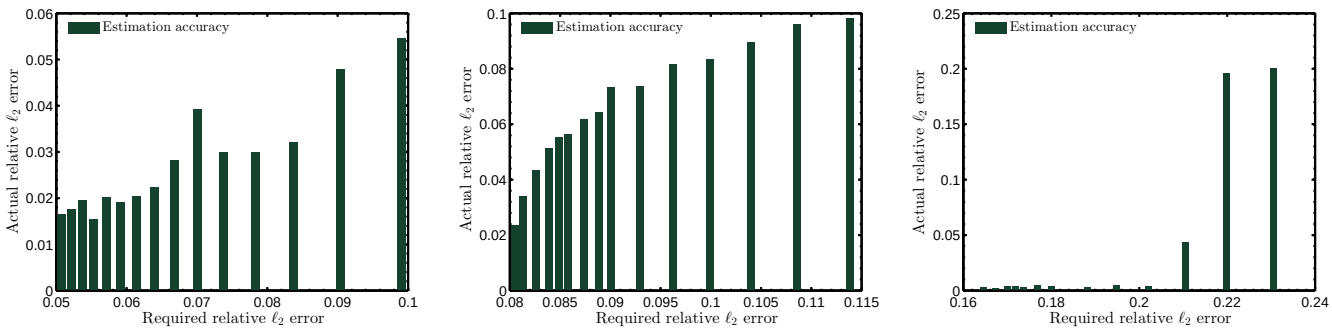
We examine the impact of the number of vehicles to the data gathering performance, which can reflect the scalability of our approach. We vary the vehicle numbers from 100 to $1,000$ at a step of 100. We set the number of APs as 80, and the required maximum error as 20%. Parameters h and m are determined according to the required maximum error and the property of the vehicular network. Besides, in the spray and wait routing algorithm, we set 8 copies for each packet. The comparison of the five schemes for all three different datasets of data readings is reported in Fig. 5.

We can see that our compressive sensing based approach consistently achieves much better estimation accuracy than other schemes over a wide range of number of vehicles. Importantly, our approach can successfully guarantee that the achieved estimation error is smaller than the required maximum error. With the SPEED dataset, for example, when there are 600 vehicles in the network, the estimation error of our approach is over 87.5% smaller than that of any other schemes. In general cases, our approach has an estimation a 30% higher estimation accuracy than that of *SW+KNN* and *DF+KNN*. With the SPEED dataset, we can also see that as the number of vehicles becomes larger, the estimation accuracy of our approach can noticeably be improved. This is mainly because the needed number of hops h is not linearly increasing with the number of vehicles. As the number of vehicles increases, there are more chances for a measurement to be collected.



(a) IDEAL (b) POSITION (c) SPEED

Fig. 6. The estimation error against the number of APs for all the three datasets.



(a) IDEAL (b) POSITION (c) SPEED

Fig. 7. The actual estimation error against the required maximum estimation error for all three datasets.

C. Impact of Number of APs

We next study the impact of the number of APs in the network. As shown in Theorem 1, we need more seeds than the number of measurements to ensure that at least m measurements can be collected. Increasing the number of APs can increase σ, and hence to improve the data gathering performance. In this scenario, we use $1,000$ vehicles with the same error requirement 10%. We vary the number of APs from 10 to 80. We report the comparison of the five schemes in Fig. 6.

We can see that our approach can achieve good data gathering performance even with a small number of APs. As the number of APs increases, with the same overhead, our approach achieves a higher estimation accuracy. Notice that with the SPEED dataset, the estimation accuracy degrades largely when the number of APs is too small, e.g., 10 in the simulation. This is because in this case the number of collected measurements is too small, and hence the approach fails to meet the estimation error requirement.

D. Impact of Error Requirement

The required maximum estimation error decides the virtual sparsity K' that we can use. Therefore, it influences the cost of our approach. In this simulation, we set the number of vehicles as $1,000$ and the number of APs as 80. Then, we plot the achieved relative error against the required maximum estimation error for different datasets in Fig. 7.

We have two observations. First, it is clear that our approach can guarantee to meet the error requirement. Second, as the required error increases, the actual estimation error of our approach also increases. This shows that our approach is able to leverage the tolerance on estimation error for reducing the transmission overhead.

VI. Related Work

In this section we briefly review related work and highlight the differences of our work from the related work.

A. Compressive Sensing in Mobile Networks

A few works have used compressive sensing in mobile networks. In [17], a mobile cooperative network is tasked with building a map of the spatial variations of a parameter of interest. By using the compressive sensing they build a map of the parameter with a small number of measurements. There is a specific work on collecting OFDM channel information in [18], where a compressive sensing based channel estimation technique exploiting the channel's delay-Doppler sparsity is proposed. In [6], Li et al. build a traffic sensing system with probe vehicles for metropolitan scale traffic sensing, and an efficient compressive sensing based algorithm is proposed for finding the best estimate traffic condition matrix.

All these works do not consider the communication issue when applying compressive sensing for data recover. In this work we consider the reduction of communication cost in a vehicular network where contacts are scarce.

B. Compressive Sensing in Static Networks

Compressive sensing is becoming a new paradigm for data gathering in sensor networks for it can provide universal sampling with decentralized simple encoding and low overhead. In [7], a universal compressive wireless sensing scheme is proposed, in which sensed data is measured by synchronized amplitude-modulated analog transmissions to the fusion center in a single hop network. In [8], Luo et al. present the complete scheme to apply compressive sensing to data collection in large scale sensor networks. The capacity gain brought by a compressive sensing scheme is validated by both analysis and simulation results. Besides, it also discusses the different method to sparsify the sensed data. In [10], a compressive sensing based approach is designed for counting and positioning targets from multiple categories in sensor networks. They rigorously justify the validity of the problem formulation and propose a GMP algorithm to complement the recover algorithms. In [19], the authors aim at minimizing the energy consumption in data collection with compressive sensing and formulates a mixed-integer programming for recovering.

The significant differences vehicular networks from sensor networks, such as high node mobility and network connection unavailability suggest that existing solutions for sensor networks are inapplicable to the problem studied in this paper.

C. Monitoring with Vehicular Networks

Vehicles as powerful mobile sensors can be used in various ways. A good survey on urban vehicular sensing platforms is offered in [20]. MobEye [21] is a protocol for vehicular urban sensing, which opportunistically diffuses sensed data summaries and creates indexes for querying of sensed data. CarTel [22] a data management system proposed for querying and collecting data from mobile vehicles, enables the application development with data collected. VTrack [23] uses less accurate sensors rather energy-hungry GPS for estimating road traffic delay.

VII. Conclusion

In this paper we focus on monitoring with vehicular networks, in which vehicles act as powerful mobile sensors. Scarce contacts in vehicular networks raise the main challenge for efficient monitoring. Through analyzing datasets of real vehicular data readings in the taxi traces in Shanghai, China, we unveil the strong correlation in vehicular data readings. Based on this observation, we have proposed `CSM`. To provide the guarantee on estimation accuracy, we first characterize the relationship between estimation error (ℓ_2 error) and sparsity property of a dataset. Then, we determine the minimum number of seeds and the transmission hop length for compressive measurements in the network. The objective is to reduce communication cost while guaranteeing the required estimation accuracy. We analytically derive the expected delay and the communication cost of our approach. Trace-driven simulations based on real vehicular GPS traces show that `CSM` successfully achieves the required estimation accuracy with low communication cost.

References

[1] J. Eriksson, L. Girod, B. Hull, R. Newton, S. Madden, and H. Balakrishnan, "The pothole patrol: using a mobile sensor network for road surface monitoring," in *Proc. ACM MobiSys*, 2008.

[2] P. Mohan, V. N. Padmanabhan, and R. Ramjee, "Nericell: rich monitoring of road and traffic conditions using mobile smartphones," in *Proc. ACM SenSys*, 2008.

[3] T. Spyropoulos, K. Psounis, and C. S. Raghavendra, "Spray and wait: an efficient routing scheme for intermittently connected mobile networks," in *Proc. ACM SIGCOMM workshop on DTN*, 2005.

[4] V. Erramilli, M. Crovella, A. Chaintreau, and C. Diot, "Delegation forwarding," in *Proc. ACM MobiHoc*, 2008.

[5] D. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, April 2006.

[6] Z. Li, Y. Zhu, H. Zhu, and M. Li, "Compressive sensing approach to urban traffic sensing," in *Proc. IEEE ICDCS*, 2011.

[7] W. Bajwa, J. Haupt, A. Sayeed, and R. Nowak, "Compressive wireless sensing," in *Proc. ACM IPSN*, 2006.

[8] C. Luo, F. Wu, J. Sun, and C. W. Chen, "Compressive data gathering for large-scale wireless sensor networks," in *Proc. ACM MobiCom*, 2009.

[9] J. Wang, S. Tang, B. Yin, and X.-Y. Li, "Data gathering in wireless sensor networks through intelligent compressive sensing," in *Proc. IEEE INFOCOM*, 2012.

[10] B. Zhang, X. Cheng, N. Zhang, Y. Cui, Y. Li, and Q. Liang, "Sparse target counting and localization in sensor networks based on compressive sensing," in *Proc. IEEE INFOCOM*, 2011.

[11] E. J. Cands, "The restricted isometry property and its implications for compressed sensing," *Comptes Rendus Mathematique*, vol. 346, no. 9C10, 2008.

[12] E. Candes and T. Tao, "Decoding by linear programming," *IEEE Transactions on Information Theory*, vol. 51, no. 12, dec. 2005.

[13] E. Candes, J. Romberg, and T. Tao, "Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information," *IEEE Transactions on Information Theory*, vol. 52, no. 2, pp. 489 – 509, Feb. 2006.

[14] W. Wang, M. Garofalakis, and K. Ramchandran, "Distributed sparse random projections for refinable approximation," in *Proc. ACM IPSN*, 2007.

[15] H. Zhu, S. Chang, M. Li, K. Naik, and S. Shen, "Exploiting temporal dependency for opportunistic forwarding in urban vehicular networks," in *Proc. IEEE INFOCOM*, 2011, pp. 2192 –2200.

[16] H. Zhu, L. Fu, G. Xue, Y. Zhu, M. Li, and L. Ni, "Recognizing exponential inter-contact time in vanets," in *Proc. IEEE INFOCOM*, 2010, pp. 1 –5.

[17] Y. Mostofi, "Compressive cooperative sensing and mapping in mobile networks," *IEEE Transactions on Mobile Computing*, vol. 10, no. 12, dec. 2011.

[18] G. Taubock and F. Hlawatsch, "A compressed sensing technique for ofdm channel estimation in mobile environments: Exploiting channel sparsity for reducing pilots," in *Proc. IEEE ICASSP*, 2008.

[19] L. Xiang, J. Luo, and A. Vasilakos, "Compressed data aggregation for energy efficient wireless sensor networks," in *Proc. IEEE SECON*, 2011.

[20] U. Lee and M. Gerla, "A survey of urban vehicular sensing platforms," *Computer Networks*, vol. 54, no. 4, 2010.

[21] U. Lee, E. Magistretti, M. Gerla, P. Bellavista, and A. Corradi, "Dissemination and harvesting of urban data using vehicular sensing platforms," *IEEE Transactions on Vehicular Technology*, vol. 58, no. 2, pp. 882 –901, 2009.

[22] B. Hull, V. Bychkovsky, Y. Zhang, K. Chen, M. Goraczko, A. Miu, E. Shih, H. Balakrishnan, and S. Madden, "Cartel: a distributed mobile sensor computing system," in *Proc. ACM SenSys*, 2006.

[23] A. Thiagarajan, L. Ravindranath, K. LaCurts, S. Madden, H. Balakrishnan, S. Toledo, and J. Eriksson, "Vtrack: accurate, energy-aware road traffic delay estimation using mobile phones," in *Proc. ACM SenSys*, 2009.