Mutual Privacy-Preserving Regression Modeling in Participatory Sensing

Kai Xing*, Zhiguo Wan†, Pengfei Hu*, Haojin Zhu ‡,
Yuepeng Wang*, Xi Chen§ Yang Wang*, Liusheng Huang*,
*University of Science and Technology of China, Anhui, China 230027
Email: {kxing,angyan,lshuang}@ustc.edu.cn,
{pfh, wangyuep}@mail.ustc.edu.cn
†Tsinghua University, Beijing, China 100084
Email: wanzhiguo@tsinghua.edu.cn
‡Shanghai Jiao Tong University, Shanghai, China
Email: zhu-hj@sjtu.edu.cn
§State Grid Info & Telecom Company Ltd., Beijing 100761
Email: xichen@sgcc.com.cn

***Abstract*—As the advancement of sensing and networking technologies, participatory sensing has raised more and more attention as it provides a promising way enabling public and professional users to gather and analyze private data to understand the world. However, in these participatory sensing applications both data at the individuals and analysis results obtained at the users are usually private and sensitive to be disclosed, e.g., locations, salaries, utility usage, consumptions, behaviors, etc. A natural question, also an important but challenging problem is how to keep both participants and users data privacy while still producing the best analysis to explain a phenomenon.**

In this paper, we have addressed this issue and proposed M-PERM, a mutual privacy preserving regression modeling approach. Particularly, we launch a series of data transformation and aggregation operations at the participatory nodes, the clusters, and the user. During regression model fitting, we provide a new way for model fitting without any need of the original private data or the exact knowledge of the model expression. To evaluate our approach, we conduct both theoretical analysis and simulation study. The evaluation results show that the proposed approach produces exactly the same best model as if the original private data were used without leakage of the fitted model to any participatory nodes, which is a significant advance compared with the existing approaches [1-5]. It is also shown that the data gathering design is able to reach maximum privacy protection under certain conditions and be robust against collusion attack. Furthermore, compared with existing works under the same context (e.g., [1-5]), to our best knowledge it is the first work showing that not only the model coefficients estimation but also a series of regression analysis and model selection methods are reachable in mutual privacy preserving data analysis scenarios such as participatory sensing.

I. Introduction

As the increasing popularity of mobile devices and sensing technologies, participatory sensing [6], [7] has been viewed as a promising way that first time enables the possibility in the collection of not only traditional sensing data (e.g., temperature, humidity, acoustic signals, etc.) but also individual exposure and activities [8-14]. However, this new data gathering way may also bring in privacy concerns into the networks. For example, in some existing participatory sensing applications [9-11], [15-22] each participant's data may be private and sensitive to be disclosed to others, e.g., location presence, behavior, attitude, or even income. Therefore, a natural question, also an important and challenging problem is how to protect every participant's data privacy without affecting the usage of data to explain a phenomenon to users.

However, many existing privacy preserving approaches, e.g., [23-25], heavily rely on a trusted infrastructure (usually the sink/base station) and crypto approaches. However, in participatory sensing networks the data are usually collected/analyzed by users instead of a trusted sink/base station, which may possibly disclose the data privacy of each participant at the user.

Furthermore, existing privacy preserving data aggregation approaches can only provide limited information about the data, e.g., summation, average, maximum, minimum, etc., which is far from satisfying users' needs and further greatly limits users' ability to conduct complex data analysis in privacy preserving participatory sensing applications. For example, current privacy preserving regression modeling approaches for participatory sensing either provide only model coefficients estimation by assuming the preknown knowledge of model expression and thus lack the ability to find the best model to fit the data [1-3], or have to leak the model coefficients to every participant during model fitting [4], [5]. Therefore, it is imperative to provide an effective solution to conduct complex data analysis and model fitting with mutual privacy protected in participatory sensing applications.

To address the privacy issue and to facilitate multi-variate data aggregation and analysis, we propose a series of privacy preserving data gathering and model fitting designs for mutual privacy preserving participatory sensing. The objective is to learn from the multi-variate data collected from participants, conduct data analysis, generalize to a model relating these variables, fit the data and predict the future, at the same time protect both participants and users' privacy. The main contributions of this paper are outlined as follows:

978-1-4673-5946-7/13/$31.00 ©2013 IEEE

- We propose a systematic mutual privacy preserving design to facilitate multi-variate data aggregation and analysis for privacy sensitive participatory sensing.
- The proposed mutual privacy preserving data aggregation protocol are proved to be able to achieve maximum privacy protection under certain conditions.
- Theoretical analysis shows that our approach is robust against collusion attack.
- The proposed regression model fitting approach produces exactly the same best model as if the original private data were used without requiring the knowledge of exact model expression or leaking the fitted model information to participants, which is a significant advance compared with the existing approaches [1-3].
- To our best knowledge, it is the first work showing that not only model coefficients estimation but also regression analysis and model fitting are doable for mutual privacy preserving participatory sensing.
- The evaluation results show that our approach is efficient in both computation and communication.

The rest of the paper is organized as follows: Section II presents the related works. The background knowledge, models and assumptions are introduced in Section III. Section IV is devoted to the development of our privacy preserving regression model fitting design. Section V provides our security analysis and discusses the enhanced security design with randomization. Section VI reports our evaluation results, followed by the conclusions in Section VII.

II. Related Work

In this section, we summarize the most relevant existing research.

[26] presents the first algorithm for provably secure hierarchical in-network data aggregation. [27] provides a state-of-the-art survey of privacy-preserving techniques for WSNs. [23] first introduces the privacy-preserving data aggregation in wireless sensor networks, and presents two privacy-preserving data aggregation schemes for additive aggregation functions. [24] proposes a family of secret perturbation-based schemes that can protect sensor data confidentiality without disrupting additive data aggregation. [25] presents the design and evaluation of PriSense, a new solution to privacy preserving data aggregation in people centric urban sensing systems based on the concept of data slicing and mixing. Multidimensional privacy-preserving data aggregation is proposed in [28]. Groat *et al.* [29] presents KIPDA, a privacy-preserving aggregation method for maximum and minimum aggregation functions.

Verifiable privacy-preserving range query in two-tiered sensor networks is proposed in [30], which uses bucketing scheme to mix the data and employ encoding numbers to prevent the storage nodes from dropping data. [31] proposes SDAP, a Secure Hop-by-hop Data Aggregation Protocol for sensor networks, which is based on the principles of divide-and-conquer and commit-and-attest. [32] proposes a framework for a series of secure information aggregation in large sensor networks, e.g., the computation of the median and the average of the measurements, the estimation of the network size, and the minimum and maximum sensor reading. [33] presents iPDA, an integrity-protecting private data aggregation scheme. In iPDA, data privacy is achieved through data slicing and assembling technique and data integrity is achieved through redundancy by constructing disjoint aggregation paths/trees. [34] proposes a scheme that can detect ill-performed aggregation without knowing the actual content of sensory data, and therefore allow sensory data to be kept concealed. [35] proposes a private data aggregation protocol that protects individual sensed values during the data aggregation process and is robust to data-loss. [36] develops a navigation service, called GreenGPS, that allows drivers to find the most fuelefficient routes for their vehicles between arbitrary endpoints. [37] specially focuses on spatial and temporal privacy of users in participatory sensing applications. [38] studies three representative scenarios-personal sensing, designated sensing, and community sensing with respect to their privacy and security implications.

The most related work are proposed in [1-3], where [1] presents an algorithm that enables users, who want to conduct a linear regression analysis with complete records without disclosing values of their own attributes, to compute the exact regression coefficients under the complete regression model; [2] studies multivariate statistical analysis methods under secure 2-party Computation (S2C) framework, including regression coefficients estimation under the complete regression model; and [3] addresses data privacy issues in participatory sensing and provides regression coefficients estimation of the complete regression model based on least square estimation.

It is worth pointing out that the correctness of the most related approaches in [1-3] relies on the assumption that the complete model (the global regression expression that includes all the variables in a data set) is the best regression model. Unfortunately, this assumption usually does not necessarily hold in reality. It is very likely that only a portion of the variables in the data set are related to the dependent variable; or some variables are highly correlated and can be represented by each other. In other words, the complete model is biased.

The most related works are proposed by Alan F. Karr, e.g., [4], [5]. Specifically, a series approaches for secure regression modeling and statistical analysis are proposed. However, the model coefficients have to be disclosed to every participant during model fitting for least square estimation and regression analysis.

Our approach differs from [1-5] in that 1. it has the power to find the best regression model without requiring the knowledge of exact model expression or leaking the fitted model to participatory nodes; 2. it is the first work showing that not only model coefficients estimation but also complex regression analysis (including model fitting) is doable for mutual privacy preserving participatory sensing; 3. it has the ability to achieve maximum privacy protection under certain conditions.

III. Preliminaries, Models and Assumptions

A. Background Knowledge

Regression model fitting is widely used in many types of data analysis and decision making process, e.g., air/water quality prediction, pollution monitoring, user-behavior prediction, etc.

In this section, we start from bi-variate curve fitting in order to ease readers' reading, then discuss multi-variate curve fitting and regression model fitting.

1) Least Squared Estimate (LSE) for Coefficient Estimation in Deterministic Models: Given a model

$$b = \mathcal{C}_0 + \mathcal{C}_1 \cdot \mathtt{f}_1(x_1) + \mathcal{C}_2 \cdot \mathtt{f}_2(x_2) + \cdots + \mathcal{C}_w \cdot \mathtt{f}_w(x_w) \quad (1)$$

and p samples $\{(x_{i,1}, x_{i,2}, \cdots, x_{i,w}, b_i), i = 1, 2, \cdots, p\}$, and the function set $\{\mathtt{f}_0, \mathtt{f}_1, \mathtt{f}_2, \cdots, \mathtt{f}_w\}$, where $\mathtt{f}_0(\cdot) = 1$, the objective of model coefficient estimation is to find the optimal $(w+1)$-dimensional vector $\boldsymbol{\mathcal{C}} = (\mathcal{C}_0, \mathcal{C}_1, \cdots, \mathcal{C}_w)^T$ to best fit the data with the minimum estimation error Δ.

Based on Least Squared Estimate (LSE), we can formulate the fitting problem as

$$\arg\min_{\boldsymbol{\mathcal{C}}} \boldsymbol{\Delta} \quad (2)$$

where

$$\Delta = \|\boldsymbol{\delta}\|^2 = \sum_{i=1}^{p} \delta_i^2 = \sum_{i=1}^{p} \left(b_i - \sum_{j=0}^{w} \hat{\mathcal{C}}_j \mathtt{f}_j(x_{i,j}) \right)^2$$

and $\hat{\mathcal{C}}_j$ is the estimated coefficient of variable x_j.

Specifically, Δ is called **unexplained variation**. We also define the **total variation** Δ' as

$$\Delta' = \sum_{i=1}^{p} (b_i - \bar{b})^2 = \sum_{i=1}^{p} b_i^2 - p\bar{b}^2$$

Let

$$\frac{\partial \boldsymbol{\Delta}}{\partial \mathcal{C}_k} = 0 \quad (3)$$

By solving Eq.3, we have

$$\boldsymbol{G}\boldsymbol{\mathcal{C}} = \boldsymbol{d} \quad (4)$$

where

$$\boldsymbol{G} = [g_{ij}]_{(w+1)\times(w+1)} \quad (5)$$

$$g_{ij} = \sum_{k=1}^{p} \mathtt{f}_j(x_{k,j}) \mathtt{f}_i(x_{k,i}) \quad (6)$$

and

$$\boldsymbol{d} = (d_1, d_2, \cdots, d_{w+1})^T \quad (7)$$

$$d_i = \sum_{k=1}^{p} b_k \mathtt{f}_i(x_{k,i}) \quad (8)$$

where $i, j \in \{0, 1, \cdots, w\}$.

Based on Eq.5-Eq.8, the model coefficients can be estimated via LSE.

2) Regression Model Fitting: In regression model fitting, if the user has known the exact variables of the model, as shown in Eq.9, it is easy to apply the least square method to find the best estimate of coefficients $\boldsymbol{\mathcal{C}} = (\mathcal{C}_0, \mathcal{C}_1, \cdots, \mathcal{C}_w)^T$ to fit the data.

However, it is very likely that

- there are only a portion of the variables in $\{x_1, x_2, \cdots, x_w\}$ are related to the dependent variable b;
- some variables $\{x_i, \cdots, x_j, \cdots, x_k\}$ may be highly correlated.

In other words, given a data set with w variables, there may exist some models with less variables that can better explain the data and provide more accurate fitting than the complete model in Eq.9. Namely the complete model may overfit and lead to a biased result. Note that it is often the case in reality that there is limited knowledge about the exact variables of the model in a given data set. From this point of view, Least Squared Estimator is far from enough for regression modeling. In order to find the best model, we introduce regression model fitting via regression analysis. In the following, we introduce some basic definitions of regression analysis.

Without loss of generality, we denote the complete model

$$b_{compl} = \mathcal{C}_0 + \mathcal{C}_1 \cdot \mathtt{f}_1(x_1) + \mathcal{C}_2 \cdot \mathtt{f}_2(x_2) + \cdots + \mathcal{C}_w \cdot \mathtt{f}_w(x_w) \quad (9)$$

where $\{x_1, x_2, \cdots, x_w\}$ represents w variables, and the reduced model

$$b_{reduced} = \mathcal{C}_0 + \mathcal{C}_1 \cdot \mathtt{f}_1(x_1) + \mathcal{C}_2 \cdot \mathtt{f}_2(x_2) + \cdots + \mathcal{C}_q \cdot \mathtt{f}_q(x_q) \quad (10)$$

Based on Eq.5 and Eq.6, we define

$$\boldsymbol{L} = [l_{ij}]_{(w)\times(w)} \quad (11)$$

$$l_{ij} = g_{ij} - \frac{g_{0,i} \cdot g_{0,j}}{p} \quad (12)$$

where $i, j \in \{1, \cdots, w\}$,

According to [39], L is usually an invertible matrix. Let

$$L^{-1} = H = [h_{ij}]_{(w)\times(w)} \quad (13)$$

B. Network Model

We consider a participatory sensing network consisting of a user and m clusters[1]. To facilitate our analysis, we assume each cluster contains c participatory sensing nodes, one of which could serve as a data aggregation point. Let $P^{(i)}$ denote the data aggregation point of the i-th cluster, and $\{V^{(i1)}, V^{(i2)}, \cdots, V^{(ic)}\}$ the participatory sensing nodes in this cluster. Each node $V^{(ij)}$ is preinstalled a random seed before deployment.

We assume that 1. any pair of nodes within the same cluster share a unique pairwise key; 2. the user node also shares a unique pairwise key with each aggregation point. We assume that the messages are securely transmitted within the network, which can be achieved via conventional symmetric encryption and key distribution schemes.

[1]Note that our scheme can adapt to general cluster formation schemes. Due to page limits, we refer the readers to [40], [41] for cluster formation.

C. Security Model

In this paper, we assume that participatory sensing nodes follow a semi-honest model [42]. Specifically speaking, they are honest and follow the protocol properly except that they may record intermediate results and try to deduce the private information of other nodes. We also assume there are limited number of nodes in collusion.

We adopt a powerful attack model: the attack can compromise any node, data aggregation point or even the sink node. The purpose of attacker is to obtain other nodes' private data based on the information from compromised nodes. However, we assume there are limited number of compromised nodes in the network.

We further define the following privacy design goals:

Node-wise Privacy The data of each sensing node should be privately kept to itself. The other nodes including the user or data aggregation points, cannot learn enough information to find out the private tuples while performing the aggregation over the sensing networks.

User-wise Privacy The aggregation results and the fitted regression model should not be disclosed to any one except the user himself. None of data aggregation points or individual nodes should know the analysis result.

IV. Privacy Preserving Regression Model Fitting

In this section, we first introduce the data aggregation scheme, then propose our regression model fitting design based on the aggregated results via statistic tests.

A. Data Aggregation for Regression Model Fitting

Our data aggregation procedure can be summarized in three steps: data aggregation at each node; data aggregation at each cluster; overall data aggregation at the user.

1) Data Aggregation at Each Node: Before aggregation, each node $V^{(ij)}$ first collects u tuples of its readings before aggregation,

$$\{(x_{k,1}^{(ij)}, x_{k,2}^{(ij)}, \cdots, x_{k,w}^{(ij)}, b_k^{(ij)}) \mid k = 1, 2, \cdots, u\}$$

, then locally computes its private data $\mathbf{\Theta}^{(ij)}$ based on Eq.6 and Eq.8, where

$$\mathbf{\Theta}^{(ij)} = \begin{bmatrix} \boldsymbol{g}_0 \\ \vdots \\ \boldsymbol{g}_n \\ \boldsymbol{d} \\ \sum_{k=1}^{u} b_k^2 \end{bmatrix} \tag{14}$$

and $\boldsymbol{g}_i$ is the ith column of $\boldsymbol{G}$. The size of $\mathbf{\Theta}^{(ij)}$ is w^2+3w+3.

2) Data Aggregation at Each Cluster: During aggregation, the data aggregation point $P^{(i)}$ generates c different positive numbers $\rho_1, \rho_2, \cdots, \rho_c$, and distributes to $V^{(ij)}$. Then node $V^{(ij)}$ generates a random vector $\boldsymbol{r}^{(ij)} = (r_1, r_2, \cdots, r_{c-w^2-3w-3})^T$, and computes

$$s_{(k)}^{(ij)} = \begin{bmatrix} 1 & \rho_k & \cdots & \rho_k^c \end{bmatrix} \begin{bmatrix} \mathbf{\Theta}^{(ij)} \\ \boldsymbol{r}^{(ij)} \end{bmatrix} \tag{15}$$

$V^{(ij)}$ keeps $s_{(j)}^{(ij)}$ to itself and sends $s_{(k)}^{(ij)}$ to $V^{(ik)}$, where $(k \neq j)$. Then node $V^{(ij)}$ computes the summation (Note $S^{(ij)}$) of the data it receives from other nodes within the cluster,

$$S^{(ik)} = \sum_{j=1}^{c} s_{(k)}^{(ij)} \tag{16}$$

After receiving $S^{(i1)}, S^{(i2)}, \cdots, S^{(ic)}$, the data aggregation point $P^{(i)}$ can build c linear equations displayed as follows.

$$\begin{bmatrix} 1 & \rho_1 & \cdots & \rho_1^c \\ 1 & \rho_2 & \cdots & \rho_2^c \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \rho_c & \cdots & \rho_c^c \end{bmatrix} \mathbf{\Phi}^{(i)} = \begin{bmatrix} S^{(i1)} \\ S^{(i2)} \\ \\ \vdots \\ S^{(ic)} \end{bmatrix} \tag{17}$$

According to equations (15) and (16), we have

$$S^{(ik)} = \begin{bmatrix} 1 & \rho_k & \cdots & \rho_k^c \end{bmatrix} \begin{bmatrix} \sum_{j=1}^{c} \mathbf{\Theta}^{(ij)} \\ \sum_{j=1}^{c} \boldsymbol{r}^{(ij)} \end{bmatrix} \tag{18}$$

By solving Eq.17 and Eq.18, the aggregation point $P^{(i)}$ can have the aggregated summation of $\mathbf{\Theta}^{(ij)}$ of cluster i.

$$\mathbf{\Phi}^{(i)} = \begin{bmatrix} \sum_{j=1}^{c} \mathbf{\Theta}^{(ij)} \\ \sum_{j=1}^{c} \boldsymbol{r}^{(ij)} \end{bmatrix} \tag{19}$$

where j represents node $V^{(ij)}$ in cluster i, $j \in \{1, 2, \cdots, c\}$. According to Eq.19, only the upper part of $\mathbf{\Phi}^{(i)}$ is useful. We have

$$\mathbf{\Theta}^{(i)} = \sum_{j=1}^{c} \mathbf{\Theta}^{(ij)} \tag{20}$$

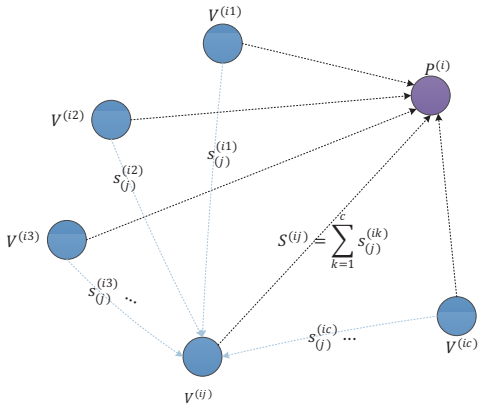


Fig. 1. Data aggregation at each cluster

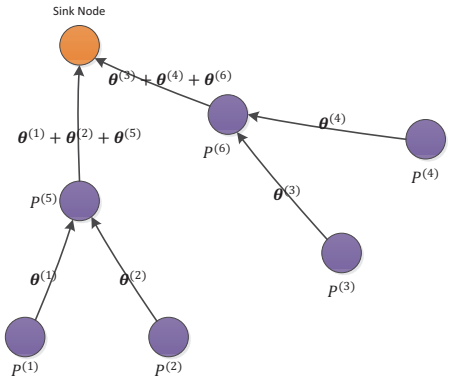


Fig. 2. Data aggregation at the user

3) Overall Data Aggregation at the User: Through a routing tree shown in figure 2, every data aggregation point sends its summations $\mathbf{\Theta}^{(i)}$ to the user. So at the sink node, according to (20), the user can obtain a vector of all summations

$$\mathbf{\Theta} = \sum_{i=1}^{m} \mathbf{\Theta}^{(i)} = \sum_{i=1}^{m}\sum_{j=1}^{c} \mathbf{\Theta}^{(ij)} \tag{21}$$

B. Multi-variate Regression Model Fitting

In this section, we conduct regression model fitting based on partial F-test and a variation of t-test [43]. We first test the complete regression model based on partial F-test

1) Is Regression Analysis Doable without Leaking the Fitted Model to Participatory Nodes?: In order to fit a good regression model and analyze its goodness, existing approaches [4], [5] require the fitted model be available to every participatory node. Otherwise regression analysis, e.g., LSE based analysis, etc., are not doable. However, this causes the leakage of the fitted model and thus breaks user-wise privacy.

Therefore a question comes for mutual privacy preserving regression modeling: *is regression analysis doable without leaking the fitted model information to participatory nodes*? By investigating all kinds of regression analysis methods, we find that most regression analysis methods heavily rely on the computation of the unexplained variation $\Delta = \sum_{i=1}^{p}\left(b_i - \hat{b_i}\right)^2$. Therefore the computability of Δ, or more specifically $\hat{b_i}$ is the key for mutual privacy preserving regression modeling.

However, without leaking the model information Δ seems not computable in [1-5] since the estimated response variable $\hat{b}$ is not computable at the user. However, according to Lemma 1, we prove that Δ could be computable without the need of computing $\hat{b}$, which can further lead to other regression analysis methods available for mutual privacy preserving participatory sensing.

***Lemma** 1:* Δ is computable without the assistance of computing $\hat{b}$ at each individual participatory node.

Proof: Due to the page limit, we omit the proof here. For more details, please refer to [44]. ■

***Corollary** 1:* Regression analysis is generally doable without the need of leaking the fitted model information to any participatory node.

2) Feasibility Test of Regression Modeling: Note that not all data sets could be fitted by a regression model, it is important to conduct a feasibility test in order to know whether it is proper to conduct regression modeling for a data set. However, none of existing privacy preserving regression modeling approaches [1-3] is able to launch such test.

In our approach, we launch the feasibility test of regression modeling for a data set based on F-test. According to [43], whether the variables of a given data set could fit to a regression model or not can be determined by F-test,

$$H_0 : \mathcal{C}_1 = \mathcal{C}_2 = \cdots = \mathcal{C}_w = 0, 1 < q < w \tag{22}$$

which claims that none of the independent variables $x_{q+1}, x_{q+2}, \cdots, x_w$ affect b, versus the alternative hypothesis

H_1 : At least one of $\mathcal{C}_{q+1}, \mathcal{C}_{q+2}, \cdots, \mathcal{C}_w$ doesn't equal to zero

which claims that at least one of the variables $x_1, x_2, \cdots, x_w$ affects b in a regression model.

Intuitively, a large value of F-test indicates that there is at least one of the variables $\{x_1, x_2, \cdots, x_w\}$ that makes the sum of squared error of the complete model Δ_{compl} substantially reduced. Otherwise, none of the independent variables are significantly correlated to the dependent variable.

If there is at least one of $x_1, x_2, \cdots, x_w$ significantly affects b, we need to further identify which of $x_1, x_2, \cdots, x_w$ is related to b.

3) Significance of Individual Variables: In reality it is very likely that only a portion of the variables in a data set are related to the response variable; or some variables are highly correlated and can be represented by others. However, existing privacy preserving regression modeling approaches [1-3] simply include all the variables of the data set in their models. In other words, the fitted regression models may probably lead to biased fitting to the data.

In our approach, we compute the significance of each independent variables with t-test, based on which we can further determine which variable could better be included in the model.

$$t = \frac{\hat{\mathcal{C}_j}}{\sqrt{h_{jj} \cdot \frac{\Delta}{p-w-1}}} \tag{23}$$

where h_{jj} is given by Eq.13, and Δ is the unexplained variation of the complete model.

By applying the t-statistic given above, we have

$$H_0 : \mathcal{C}_j = 0 \quad \text{and} \quad H_1 : \mathcal{C}_j \neq 0$$

Specifically, we reject H_0 in favor of H_1 at the probability of a Type I error equal to α if and only if either of the following conditions hold.

4) Regression Model Fitting: In this section, we show the procedure to find the best model via testing all reasonable regression models and compare them on the basis of well known criterions in regression analysis, C statistic and Adjusted R.

Without loss of generality, suppose that Eq.10 is a model to be tested, we have

$$\bar{R}^2 = (R^2 - \frac{q-1}{p-1})(\frac{p-1}{p-q}) \tag{24}$$

where

$$R^2 = 1 - \frac{\text{Unexplained variation}}{\text{Total variation}}$$

Based on Eq.9 and Eq.10, we define the unexplained variation of the complete model and the model to be tested as Δ_{compl} and $\Delta_{reduced}$ respectively; and the total variation of the complete model and the model to be tested as Δ'_{compl} and $\Delta'_{reduced}$, respectively.

We have

$$C = \frac{\Delta_{reduced}}{\Delta_{compl}/(p-(w+1))} - (p-2q-2) \tag{25}$$

Based on [39], the goodness of a fitted model could be evaluated by C statistic and Adjusted R, and the best regression model should follow with the least C statistic and the peak high Adjusted R^2. With this observation, the regression model fitting algorithm is given as follows.

Algorithm 1 Mutual Privacy Preserving Regression Model Fitting

Input: the data set, the function set $\{\mathtt{f}_0, \mathtt{f}_1, \mathtt{f}_2, \cdots, \mathtt{f}_w\}$, and the aggregation result Θ
Output: the fitted model.

1: **function** $Model$=**M-PERM**(data set, function set, Θ)
2: $i = 1$
3: **loop**until $i = w + 1$
4: Test the models of i variables
5: Record the model with the smallest C statistic and the peak high Adjusted R value. ▷ Find the best models consist of i variables.
6: $i = i + 1$
7: **end loop**
8: Find the model with the smallest C statistic and the peak high Adjusted R value among all the recorded models. ▷ Find the best model.
9: Return the model found
10: **end function**

V. Security Analysis

In this section, we analyze the privacy and confidentiality issues of our proposed scheme and show that under certain conditions our scheme is capable of achieving maximum privacy protection in the network.

***Lemma** 2:* Node $V^{(ij)}$ cannot deduce the private data of other participatory nodes.

[2]There could be multiple models with similar peak high Adjusted R

Proof: According to the computation process at node $V^{(ij)}$, only during the protocol execution

$$\overset{receive}{\Rightarrow} s^{(i1)}(j), s^{(i2)}(j), \ldots \overset{compute}{\Rightarrow} S^{(ij)}$$

there exists probability that node $V^{(ij)}$ can receive the message from the node $V^{(ik)}$. However, $f^{(ik)}(j)$ is a cumulated sum of c float number. According to Eq.15, it is impossible to resplit $f^{(ik)}(j)$ into c float number and deduce Θ^{ik}. Therefore, node $V^{(ij)}$ cannot deduce the private data of other nodes.

***Corollary** 2:* The aggregation point can not deduce the private data of any particular node.

***Lemma** 3:* **Node-wise Maximum Privacy Protection** Given node $V^{(ij)}$, and $u > w+4$, where u denotes the number of sample tuples aggregated at individual node, w denotes the number of variables in the data set. If the samples and the variables at node $V^{(ij)}$ are independent of each other, the privacy-preserving regression model fitting scheme proposed in section IV could achieve node-wise maximum privacy protection.

Proof: According to Eq.14, Eq.15, Eq.16, and the slicing scheme proposed in Section IV-A2, each node generates $(c - w^2 - 3w - 3)$ random numbers and the vector $\boldsymbol{\Theta}^{(ij)}$ is obtained from $w \cdot u$ unknown values. Then, the equation (15) contains $(c + w \cdot u - w^2 - 3w - 3)$ unknown values. If $u > w+4$, we have $(c + w \cdot u - w^2 - 3w - 3) > (w \cdot u) > (w^2 + 4w) \geq (w+1)^2 + 1$. Namely even if all the other (c-2) nodes are in collusion in the cluster, the unknown values of a node cannot be revealed to the attacker, because the number of unknown values is larger than the number of equations according to Eq.4.

Besides, note that nodes outside the cluster i do not communicate with node $V^{(ij)}$, there is no way to reveal the private data of node $V^{(ij)}$ unless node $V^{(ij)}$ itself is cracked.

Therefore the proposed privacy preserving scheme can achieve node-wise maximum privacy protection.

***Corollary** 3:* The user can not deduce the private data of any particular node.

***Lemma** 4:* **User-wise Maximum Privacy Protection** The final aggregation results and the fitted regression model cannot be obtained by either data aggregation point or individual nodes unless the attacker obtains all m keys between the user and the m data aggregation points or cracks all m data aggregation points.

Proof: The lemma holds trivially.

***Corollary** 4:* The participatory nodes can not deduce the aggregation results and the fitted regression model obtained by the user.

VI. Evaluation Study

A. Experiment Settings

Our experiment study is conducted based on MATLAB. There are 4 data sets we have used in the evaluation study, three of which are well-known data sets, where *attitude* [45] concerns the attitude of the clerical employees of a large financial organization; *auto* [46] concerns city-cycle fuel consumption in miles per gallon; *salary* [47] concerns discrimination

in salaries for 52 tenure-track professors in a small college. The other data set *random* is a self-generated data set with 6 variables in Matlab, $\{\mathtt{f}_1(x_1), \mathtt{f}_2(x_2), \cdots, \mathtt{f}_6(x_6)\}$, where each sample value of every variable is randomly selected in the range of $[0, 100]$ with a variation value randomly selected from $[0, 10]$, namely $\mathtt{f}_i(x_i) = x_i + \varepsilon$. We also set the response variable $b = \mathtt{f}_1(x_1) + \mathtt{f}_2(x_2) + \mathtt{f}_3(x_3)$. Specifically, we generate 180 sample tuples in the simulation.

In the settings, we randomly deploy N participatory sensing nodes in an area of 100×100 grids, where N depends on the participatory nodes in the data sets. All the results are averaged over 20 runs.

Due to the limited number of observations in the data set, we set the number of clusters to 1 in the data sets *attitude*, *auto*, and *salary*, and the number of clusters to 2 in the data set *random*. The size of the training data of *attitude*, *auto*, *salary*, and *random* is set to 31, 360, 43, and 120 respectively, and the size of the test data of *attitude*, *auto*, *salary*, and *random* is set to 10, 32, 9, and 60 respectively.

B. Accuracy Analysis

To evaluate our approach, we compare our approach with the one proposed [3], and the best model given in SAS by regular regression analysis of the original data.

To facilitate the comparison between our approach, the one proposed [3], and the best model given in SAS by regular regression analysis of the original data, we use M-PERM, [3], and Regression to represent them, respectively.

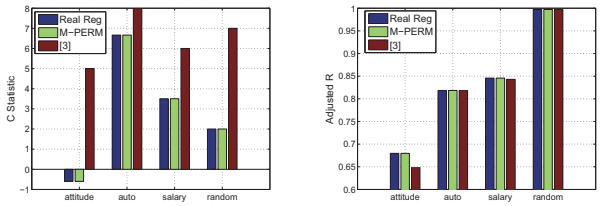
Fig.3(a) and Fig.3(b) provide the comparison results of C statistics and Adjusted Rs of the models given by M-PERM, [3], and regular regression in SAS with the original data. According to the results shown in Fig.3(a) and Fig.3(b), it is obvious to observe that the model given by M-PERM is the same as the best model given by regular regression analysis in SAS with the original data. The C statistics of M-PERM of the four data sets are all substantially better than the one given by [3], and all its Adjusted R values are better than that of [3]. Generally speaking, M-PERM is superior to [3], and can achieve the same best model as if the original private data were used by regular regression analysis in SAS.

Fig.4 shows the C statistics of different models checked by M-PERM for each of the four data sets. Taking Fig.4(a) (the data set of attitude) as an example, the model with 2 variables selected by t-test (namely the 2 variables of most significance to the response variable) reaches the smallest C statistic, which indicates that this model should be the best model according to C statistic. From Fig.4, we can see that M-PERM can always find the best model according to C statistic.

Fig.5 shows the Adjusted R values of different models checked by M-PERM for each of the four data sets. Taking Fig.5(a) (the data set of attitude) as an example, the model with 2 variables selected by t-test (namely the 2 variables of most significance to the response variable) reaches the highest Adjusted R value, which indicates that this model should be the best model according to Adjusted R. Taking Fig.5(b) (the data set of auto) as an example, started from 2 variables, the models with 2 or more variables selected by t-test (namely the 2 or more variables of most significance to the response variable) begin to have similar peak-high Adjusted R values, which indicates that these models with 2 or more variables could be good models to fit the data according to Adjusted R. However, the best model should be further determined by other criterion. For example, C statistic in Fig.4 provides us the exact best model.

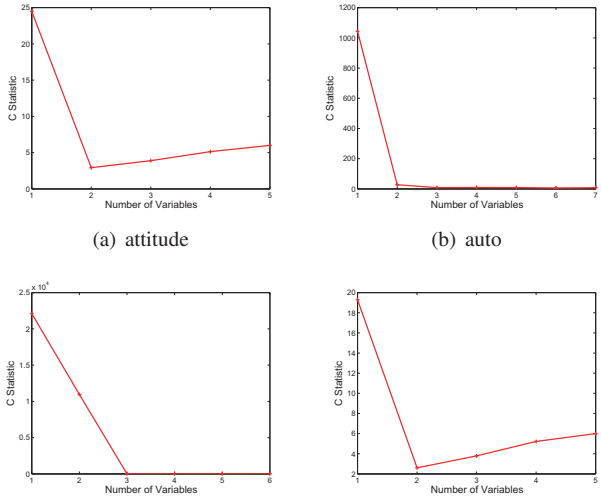
Fig.6 shows the standard error of the models given by M-PERM and [3] of 20 runs over the four data sets. In this figure, we can see that M-PERM always achieves smaller standard error than that of [3]. Fig.7 shows the reconstructed values of M-PERM and [3]of the four data sets, compared with the real value of the test data. In this figure, we can see that M-PERM always better fits the data and generally performs better than [3].

Generally speaking, we can see that M-PERM has the capability to find the best regression model as if the original private data were used by regular regression analysis in SAS. It always substantially performs better than [3].



(a) C statistic (b) Adjusted R

Fig. 3. C statistics and Adjusted Rs of the models given by M-PERM, [3], and regular regression with the original data



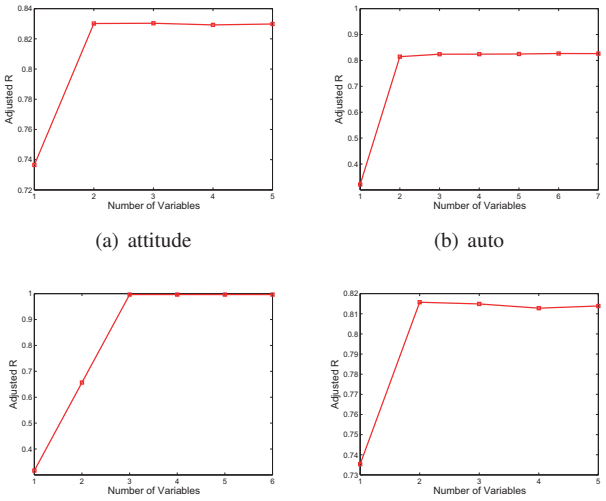
(a) attitude (b) auto

(c) random (d) salary

Fig. 4. C statistics of different models for the four data sets.

C. Communication Overheads

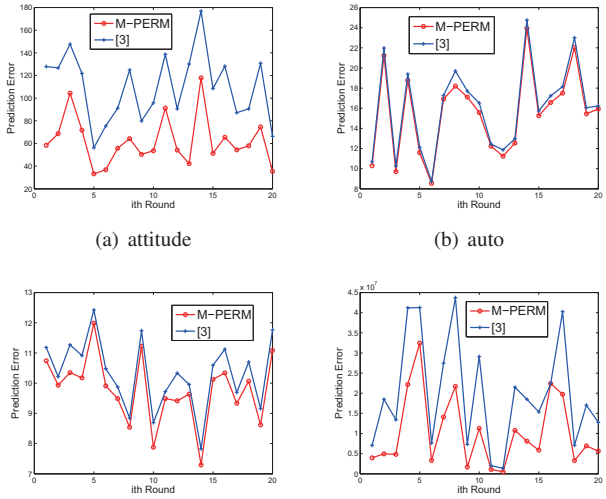
In this section, we study the communication overhead of M-PERM and compare it with [3]. Given the network of N nodes,



(a) attitude (b) auto

(c) random (d) salary

Fig. 5. Adjusted Rs of different models for the four data sets.



(a) attitude (b) auto

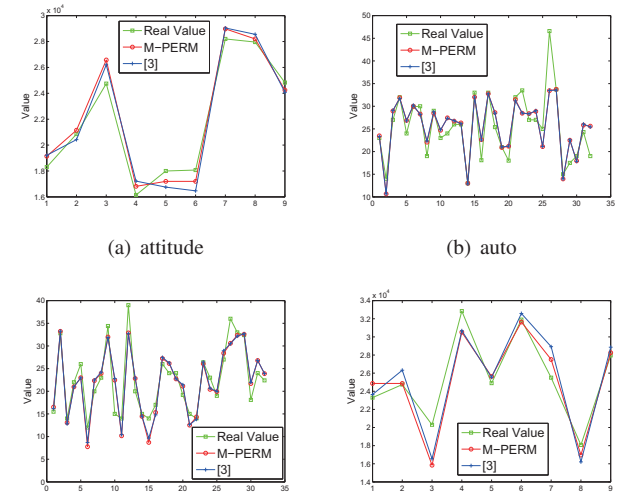
(c) random (d) salary

Fig. 6. Prediction errors of M-PERM and [3] for the four data sets.

with m clusters and c nodes in each cluster, we have $N = m \times c$ The communication overhead of [3] is $O(N^{3/2}(w^2+w+1))$.

For M-PERM, the communication overhead within each cluster is $O(c^{5/2})$, the communication overhead is $O(mN^{1/2}(w^2+3w+3))$ Therefore, the overall communication overhead of M-PERM is $O(Nc^{3/2}) + mN^{1/2}(w^2+3w+3)$. The communication overhead of M-PERM should be smaller than that of [3] by setting a proper cluster size in the network. In other words, when the cluster size is fixed, the communication overhead of M-PERM increases slower than that of [3].

Fig.8 shows the communication overhead of M-PERM and [3] when the cluster size is fixed to 30. As shown in the figure, as the number of nodes increases, the communication overhead of [3] quickly goes beyond the communication overhead of M-PERM. Namely M-PERM has a better performance of

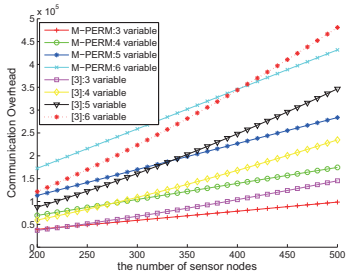


(a) attitude (b) auto

(c) random (d) salary

Fig. 7. Predicted values of M-PERM and [3] over the test data of the four data sets compared with the real data values.

communication overhead than that of [3].



Schemes	Communication Overhead
Method in [3]	$O(N^{3/2} \cdot (w^2+w+1))$
M-PERM	$O(Nc^{3/2}) + mN^{1/2}(w^2+3w+3)$

Fig. 8. Communication Overhead

VII. Conclusion

In this paper, we have addressed mutual privacy preserving issues in participatory sensing and the challenges for regression analysis in user-defined participatory data analysis. Different from existing approaches, our approach has the capability to find the best model as if the original private data were used without requiring the knowledge of the exact model expression or leaking any regression results. This is a significant advance compared with the existing approaches [1-5]. The analysis also indicates that our data gathering design could reach maximum privacy protection under certain conditions and be robust against collusion attack. Furthermore, compared with existing works under the same context (e.g., [1-3]), our work first time shows that not only model coefficients estimation but also regression analysis and model fitting are reachable for mutual privacy preserving participatory sensing.

VIII. Acknowledgments

This research is supported by NSFC under grant 61170267,61003218,61272444,61003223, Jiangsu NSF under grant BK2011358, RFDP under grant 20113402120008, and National 973 Program under grant 2011CB302905.

REFERENCES

[1] A. Sanil, A. Karr, X. Lin, and J. Reiter, "Privacy preserving regression modelling via distributed computation," in *ACM SIGKDD'04*, 2004, pp. 677–682.

[2] W. Du, Y. Han, and S. Chen, "Privacy-preserving multivariate statistical analysis: Linear regression and classification," in *Proceedings of the 4th SIAM International Conference on Data Mining*, vol. 233, 2004.

[3] H. Ahmadi, N. Pham, R. Ganti, T. Abdelzaher, S. Nath, and J. Han, "Privacy-aware regression modeling of participatory sensing data," in *ACM SenSys*, 2010, pp. 99–112.

[4] A. F. Karr, A. F. Karr, X. Lin, X. Lin, A. P. Sanil, A. P. Sanil, J. P. Reiter, and J. P. Reiter, "Secure regression on distributed databases," *J. Computational and Graphical Statist*, vol. 14, pp. 263–279, 2004.

[5] A. F. Karr, X. Lin, A. P. Sanil, and J. P. Reiter, "Secure statistical analysis of distributed databases using partially trusted third parties. manuscript in preparation," in *In Statistical Methods in Counterterrorism*. Springer, 2005.

[6] J. Burke, D. Estrin, M. Hansen, A. Parker, N. Ramanathan, S. Reddy, and M. B. Srivastava, "Participatory sensing," in *In: Workshop on World-Sensor-Web (WSW06): Mobile Device Centric Sensor Networks and Applications*, 2006, pp. 117–134.

[7] A. Dua, N. Bulusu, W.-C. Feng, and W. Hu, "Towards trustworthy participatory sensing," in *Proceedings of the 4th USENIX conference on Hot topics in security*, 2009, pp. 8–13.

[8] T. S. Lena, V. Ochieng, M. Carter, J. Holguín-Veras, and P. L. Kinney, "Elemental Carbon and PM2.5 Levels in an Urban Community Heavily Impacted by Truck Traffic," in *Environmental Health Perspectives*, no. 10, 2002.

[9] J. Corburn, "Confronting the Challenges in Reconnecting Urban Planning and Public Health," in *American Journal of Public Health*, no. 4, 2004, pp. 541–549.

[10] M. P. G. Oliveira, E. B. Medeiros, and J. Clodoveu A. Davis, "Planning the Acoustic Urban Environment: a GIS-CenteredApproach," in *ACM GIS'99*, 1999.

[11] J. C. et al, "Overview: Mapping for changełthe emergence of a new practice," in *Participatory learning and action*, 2006, pp. 13–20.

[12] K. Xing, X. Cheng, J. Li, and M. Song, "Location-centric storage and query in wireless sensor networks," in *Wireless Networks*, vol. 16, no. 4, 2010, pp. 955–967.

[13] S. Ji and Z. Cai, "Distributed data collection in large-scale asynchronous wireless sensor networks under the generalized physical interference model," *Networking, IEEE/ACM Transactions on*, vol. PP, no. 99, p. 1, 2012.

[14] S. Ji, R. Beyah, and Z. Cai, "Snapshot/continuous data collection capacity for large-scale probabilistic wireless sensor networks," in *INFOCOM, 2012 Proceedings IEEE*. IEEE, 2012, pp. 1035–1043.

[15] K. Xing and X. Cheng, "From time domain to space domain: Detecting replica attacks in mobile ad hoc networks," in *IEEE INFOCOM*, 2010, pp. 1–9.

[16] P. Li, P. Fan, K. Xing, H. Wang, Z. Jiang, and F. Wang, "Tussle between aps in a location-dependent pricing game," in *IEEE Transactions on Wireless Communications*, vol. 11, no. 1, 2012, pp. 338–345.

[17] S. Ji and Z. Cai, "Distributed data collection and its capacity in asynchronous wireless sensor networks," in *INFOCOM, 2012 Proceedings IEEE*. IEEE, 2012, pp. 2113–2121.

[18] C. D., G. J., R. A., H. M., and K. S.S., "Privacy-preserving collaborative path hiding for participatory sensing applications," in *IEEE MASS'11*, 2011, pp. 341–350.

[19] K. Vu, R. Zheng, and J. Gao, "Efficient algorithms for k-anonymous location privacy in participatory sensing," in *IEEE INFOCOM'12*, march 2012, pp. 2399 –2407.

[20] M. Groat, B. Edwards, J. Horey, W. He, and S. Forrest, "Enhancing privacy in participatory sensing applications with multidimensional data," in *IEEE PerCom'12*, march 2012, pp. 144 –152.

[21] H. Zhu, X. Lin, R. Lu, P. Ho, and X. Shen, "Slab: A secure localized authentication and billing scheme for wireless mesh networks," *Wireless Communications, IEEE Transactions on*, vol. 7, no. 10, pp. 3858–3868, 2008.

[22] H. Zhu, X. Lin, R. Lu, Y. Fan, and X. Shen, "Smart: A secure multilayer credit-based incentive scheme for delay-tolerant networks," *Vehicular Technology, IEEE Transactions on*, vol. 58, no. 8, pp. 4628–4639, 2009.

[23] W. He, X. Liu, H. Nguyen, K. Nahrstedt, and T. Abdelzaher, "Pda: Privacy-preserving data aggregation in wireless sensor networks," in *IEEE INFOCOM*, 2007, pp. 2045–2053.

[24] T. Feng, C. Wang, W. Zhang, and L. Ruan, "Confidentiality protection for distributed sensor data aggregation," in *IEEE INFOCOM'08*. IEEE, 2008, pp. 56–60.

[25] J. Shi, R. Zhang, Y. Liu, and Y. Zhang, "Prisense: Privacy-preserving data aggregation in people-centric urban sensing systems," in *Proceedings of 29th IEEE International Conference on Computer Communications*, 2010, pp. 1–9.

[26] H. Chan, A. Perrig, and D. Song, "Secure hierarchical in-network aggregation in sensor networks," in *Proceedings of the 13th ACM conference on Computer and Communications Security*, 2006, pp. 278–287.

[27] N. Li, N. Zhang, S. Das, and B. Thuraisingham, "Privacy preservation in wireless sensor networks: A state-of-the-art survey," *Ad Hoc Networks*, vol. 7, no. 8, pp. 1501–1514, 2009.

[28] X. Lin, R. Lu, and X. Shen, "Mdpa: multidimensional privacy-preserving aggregation scheme for wireless sensor networks," *Wireless Communications and Mobile Computing*, vol. 10, no. 6, pp. 843–856, 2010.

[29] M. Groat, W. He, and S. Forrest, "Kipda: k-indistinguishable privacy-preserving data aggregation in wireless sensor networks," pp. 2024–2032, 2011.

[30] B. Sheng and Q. Li, "Verifiable privacy-preserving range query in two-tiered sensor networks," in *IEEE INFOCOM*, 2008, pp. 46–50.

[31] Y. Yang, X. Wang, S. Zhu, and G. Cao, "Sdap: A secure hop-by-hop data aggregation protocol for sensor networks," *ACM Transactions on Information and System Security*, vol. 11, no. 4, pp. 1–43, 2008.

[32] B. Przydatek, D. Song, and A. Perrig, "Sia: Secure information aggregation in sensor networks," in *Proceedings of the 1st international conference on Embedded networked sensor systems*, 2003, pp. 255–265.

[33] W. He, H. Nguyen, X. Liuy, K. Nahrstedt, and T. Abdelzaher, "ipda: An integrity-protecting private data aggregation scheme for wireless sensor networks," in *IEEE Military Communications Conference,*, 2008, pp. 1–7.

[34] C. Wang, G. Wang, W. Zhang, and T. Feng, "Reconciling privacy preservation and intrusion detection in sensory data aggregation," in *Proceedings of 31th IEEE International Conference on Computer Communications*, 2011, pp. 336–340.

[35] M. Conti, L. Zhang, S. Roy, R. Di Pietro, S. Jajodia, and L. Mancini, "Privacy-preserving robust data aggregation in wireless sensor networks," *Security and Communication Networks*, vol. 2, no. 2, pp. 195–213, 2009.

[36] R. Ganti, N. Pham, H. Ahmadi, S. Nangia, and T. Abdelzaher, "Greengps: A participatory sensing fuel-efficient maps application," in *Proceedings of the 8th international conference on Mobile systems, applications, and services*, 2010, pp. 151–164.

[37] K. Huang, S. Kanhere, and W. Hu, "Preserving privacy in participatory sensing systems," *Computer Communications*, vol. 33, no. 11, pp. 1266–1280, 2010.

[38] D. Christin, M. Hollick, and M. Manulis, "Security and privacy objectives for sensing applications in wireless community networks," in *Computer Communications and Networks (ICCCN), 2010 Proceedings of 19th International Conference on*, 2010, pp. 1–6.

[39] B. L. Bowerman, *forcasting and time series: an applied approach*. Duxbury Press, 2000.

[40] S. Bandyopadhyay and E. J. Coyle, "An energy efficient hierarchical clustering algorithm for wireless sensor networks," in *IEEE INFOCOM'03*, vol. 3, 2003, pp. 1713 – 1723.

[41] O. Younis, M. Krunz, and S. Ramasubramanian, "Node clustering in wireless sensor networks: Recent developments and deployment challenges," *IEEE Network Magazine*, vol. 20, pp. 20–25, 2006.

[42] O. Goldreich, *Foundations of Cryptography: Volume II, Basic Applications*. Cambridge: Cambridge University Press, 2004.

[43] G. Casella and R. L. Berger, *Statistical Inference*. Duxbury Press, 2001.

[44] K. Xing and P. Hu, "Tech. report: Mutual privacy-preserving regression modeling in participatory sensing," in *Tech Report*, 2012, http://staff.ustc.edu.cn/ kxing/Publications/TechReport/pfhu.pdf.

[45] "Data sets in R," http://127.0.0.1:23163/library/datasets/html/attitude.html.

[46] "S. weisberg (1985). applied linear regression, 2nd edition," http://archive.ics.uci.edu/ml/datasets/Auto+MPG.

[47] "Professor Salaries," http://data.princeton.edu/wws509/datasets/#salary.