PURE: Blind Regression Modeling for Low Quality Data with Participatory Sensing

Shan Chang, *Member, IEEE*, Hongzi Zhu, *Member, IEEE*, Wei Zhang, Li Lu, *Member, IEEE*, and Yanmin Zhu, *Member, IEEE*

Abstract—Participatory regression modeling is a cost-efficient mechanism to establish the relationships among multiple dimensions of sensory data collected from volunteers. Getting an accurate model estimate is challenging for two main reasons. First, with the concern of confidentiality of individual private data, the original data are nearly unavailable; second, low quality data with outliers are inherently embedded in the collected data. In this paper, we propose an innovative scheme, PURE, which can accurately estimate the global regression model without the need for knowing local private data (referred to as *blind regression modeling*) even when there is a large portion of outliers embedded. The wisdom of PURE is to let individual participants peer judge and further improve the global estimate via negotiations. Meanwhile, during the whole process, all information is exchanged in an aggregated way. By design, PURE is secure and can well protect individual privacy. Furthermore, PURE is a lightweight protocol suitable for mobile devices. Extensive trace-driven simulation results show that PURE can achieve an outstanding accuracy gain of two orders of magnitude even with random outliers near a ratio of 50 percent compared with the state-of-the-art least square estimator.

Index Terms—Participatory sensing, blind regression modeling, data confidentiality, low quality data

1 Introduction

Participatory sensing is a revolutionary paradigm, where ordinary people are empowered to voluntarily collect and share sensory data about their surrounding environments using mobile devices (e.g., smartphones and tablets). Bunches of appealing participatory sensing applications have been proposed recently, e.g., spanning intelligent transportation [1], air quality monitoring [2], grocery bargain hunting [3], [4], data delivery [5], and social networking [6]. The typical system model of a participatory sensing application is illustrated in Fig. 1, where the server is present to assign sensing tasks to and collect data from participants. The common task of a large majority of such applications is to conduct *blind* participatory regression modeling, which is to establish the statistical relationship among multiple dimensions of those voluntary sensory data without the data confidentiality being violated. For example, GreenGPS [1] is a fuel-saving navigation service which relies on voluntary data collected from individuals to construct the linear model between fuel costs and routing decisions.

- *S. Chang is with the School of Computer Science & Technology, Donghua University, Shanghai 201620, P.R. China. E-mail: changshan@dhu.edu.cn.*
- *H. Zhu, Y. Zhu, and W. Zhang are with the Department of Computer Science and Technology, Shanghai Jiao Tong University, Shanghai 200000, P.R. China. E-mail: {hongzi, yzhu}@cs.sjtu.edu.cn, zhangseuedu@sjtu.edu.cn.*
- *L. Lu is with the School of Computer Science and Engineering, University of Electronic Science and Technology, Chengdu, P.R. China. E-mail: luli2009@uestc.edu.cn.*

Manuscript received 17 Nov. 2014; revised 15 Mar. 2014; accepted 22 Apr. 2015. Date of publication 28 Apr. 2015; date of current version 16 Mar. 2016. Recommended for acceptance by J. Chen.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TPDS.2015.2427805

Solving the blind participatory regression modeling problem, however, is very challenging for three reasons. First, individual sensory data might be private and sensitive and should be strongly protected; otherwise, people would be reluctant to take part in such participatory sensing applications. For examples in GreenGPS, drivers might not want to expose their location and velocity information to others. With the concern of confidentiality of individual private data, participants would not directly send their original data to a central server for regression modeling. Thus, it would be very difficult to estimate an accurate regression model without knowing the data. Second, as the sensory data are collected from untrained ordinary people with various mobile devices, errors tend to be inevitable (e.g., keypunch errors, misplaced decimal points and wrong data representation). Therefore, the data quality is usually very low with a large portion of outliers. Furthermore, there may be malicious participants who deliberately contribute falsified data, misleading the server to conclude a biased regression result. Without dealing with those outliers, the final statistical results might be of little use. Li and Cao [8] proposed a privacy-aware incentive mechanism for mobile sensing, which promotes participants providing high quality data while prevents the possible privacy leakage. However, it cannot guarantee the elimination of the gross error, generated unintentionally. Last, the limited power and computation capabilities of mobile devices (e.g., power and computation) as well as the communication cost for uploading sensory data also pose an urgent demand for lightweight participatory regression modeling scheme.

In the literature, several schemes [9], [10], [26] have been proposed to address data confidentiality issues when mining in distributed databases. However, the quality of data is ensured by the administrator of such databases, who can pre-delete suspicious outliers. Thus, such schemes usually

1045-9219 © 2015 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

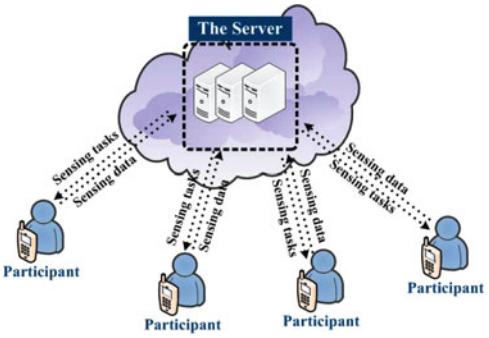


Fig. 1. The typical system model of a participatory sensing application.

do not consider the influence of incorrect data on the statistics. Several distributed privacy-preserving outlier detection schemes have also been proposed [7], [17], where outliers can be found by checking the pairwise distances of all data samples. However, those data far away from the majority of data may still follow the general pattern (relation) in the data. In other words, an distance-based outlier is not necessarily a regression outlier. In the area of wireless sensor networks, the data aggregation [12], [13], privacy-preserving [15], [16], and outlier detection [14] problems have been widely investigated, however, they are usually considered as independent problems and are investigated separately. Moreover, most work on privacy-preserving data aggregation problem can only calculate some preliminary statistics results [15], [16], such as *sum* and *max/min*, instead of complex regression models. As a result, there exists no successful solution, to the best of our knowledge, to addressing the blind participatory regression modeling problem with low-quality data.

In this paper, we propose an innovative scheme, PURE, effectively tackling the challenges of linear blind participatory regression modeling problem, where the multiple dimensions of data are assumed to follow a linear model. The core idea of PURE is to let individual participants not only collect sensory data but also help establish the global optimal estimation of the linear regression model. More specifically, participants with consistent observations would build their local regression models. Then, those local models are distributed over all participants for peer reviewing using the peers' own observations. After that, one of those local models is chosen by the server as the current global model estimate based on the peer-reviewing results. Given the global estimate, the server lets all participants judge their own observations according to the current global estimate. Upon request, each participant adjusts the weights of its observations so that those which are far away from the current global estimate will have lower weights, and generates its comment about the current estimate accordingly. By collecting those "comments" in an aggregated way, the server can refine the global estimate. After a few rounds of such "negotiations", all participants can achieve an agreement about the current global estimate, which leads to the optimal global estimate in terms of the distance to the majority of observations.

As, in PURE, each participant reports only intermediate results rather than private raw data, we theoretically prove that PURE can ensure data confidentiality of participants and defend against collusion attacks. Moreover, PURE is a lightweight protocol designed for mobile devices, which only involves simple computation on local observations and requires a very limited number of interactions with the server. We evaluate the performance of PURE through extensive simulations using realistic traces and the results demonstrate the robustness of PURE in the presence of outliers even near a ratio of 50 percent. On average, PURE can achieve an accuracy gain of two orders of magnitude for random outliers and three times for normal distributed outliers compared with the mostly used least square (LS) estimator [11].

The remainder of this paper is organized as follows. In Section 2, we introduce problem formulation and preliminaries. Section 3 describes the models and design goals. In Section 4, we elaborate the design of PURE. Section 5 presents the security analysis. In Section 6, we show the performance evaluation. We review related work in Section 7. Section 8 concludes and outlines the directions for future work.

2 Problem Formulation and Preliminaries

2.1 Background of Linear Regression

We first introduce the basic regression modeling problem in a participatory sensing application, where m participant nodes $\{N_1, N_2, \ldots, N_m\}$ (e.g., mobile devices) and a server S are involved. Each participant N_i for $i = 1, 2, \ldots, m$ collects a number of its own readings (observations) about p independent variables $c_i^1, c_i^2, \ldots, c_i^p$ and one dependent variable y_i.The kth observation of N_i can be denoted as a tuple of $(c_{i,k}^{(1)}, c_{i,k}^{(2)}, , c_{i,k}^{(p)}, y_{i,k})$. We have the following assumption:

Assumption 1. *Participants are independent (i.e., they measure those variables on their own). Measurements performed by the same participant at different time are also independent (i.e., previous measurements have no effect on later ones) and obey the same distribution.*

The server S gathers n observations from each participant to illuminate any underlying association between variables by fitting equations to the observed variables, according to a specific model. If a linear model is adopted, we have the definition as follows:

Definition 1. *A linear regression model relates the dependent or "response" variables $y_{i,k}$ to explanatory variables $\mathbf{x}_{i,k}^T = (1, x_{i,k}^{(1)}, x_{i,k}^{(2)}, , x_{i,k}^{(p)})$ for $i = 1, . , m$ and $k = 1, . , n$, such that*

$$y_{i,k} = \mathbf{x}_{i,k}^T \boldsymbol{\beta} + \epsilon_{i,k}, \tag{1}$$

where $\boldsymbol{\beta}^T = [\beta^{(0)}, \ldots, \beta^{(p)}]$ is the coefficient vector, $x_{i,k}^{(t)} = f_t(c_{i,k}^{(t)})$ for $t = 1, \ldots, p$ and error $\epsilon_{i,k}$ is a random variable with expectation of zero. Note that as the explanatory variables $x_{i,k}^{(t)}$ for $t = 1, \ldots, p$ can be any known function $f_t(\cdot)$ of the independent variable $c_i^{(t)}$, the regression model is linear with respect to the regression coefficient vector $\boldsymbol{\beta}$.

2.2 Linear Blind Regression Modeling with Low Quality Data

We first describe the definition of regression outliers as follows,

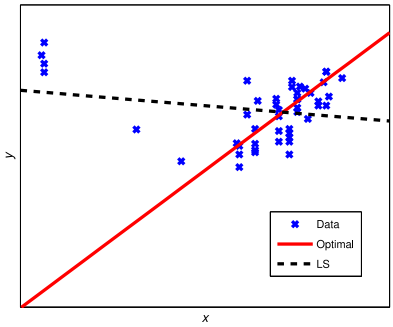


Fig. 2. The derived model using LS is severely biased from the majority of data when there are outliers.

Definition 2. *An observation is called to be a regression outlier if it deviates from the relation followed by the majority of the data.*

Remark 1. In participatory sensing applications, different participants may have different fractions of outliers in their observations. The total number of outliers, however, should be limited up to 50 percent; otherwise, it would be impossible to distinguish between "good" and "bad" data even if the server had all original data.

Before the definition of our problem, we introduce two properties, i.e., *low-data-quality-tolerant* and *blind*, as follows,

Property 1. *A participatory regression modeling is low-data-quality-tolerant if the derived relation still fits the majority of the data even if the portion of regression outliers reaches up to 50 percent of all observations.*

Property 2. *A participatory regression modeling is blind if the original observations of each participant cannot be obtained or inferred by any other participant and the server as well during the estimation of regression model.*

In this paper, we aim to achieve an accurate linear regression model which is resilient to low quality data and ensure data confidentiality of participants at the same time. We define our problem as follows,

Definition 3. *The problem of blind linear regression modeling with low quality data is referred to as, given the original private observations, finding the optimal linear regression estimate so that it has both the blind and low-data-quality-tolerant properties.*

In general, it is very difficult to address the above problem. Approaches based on high-quality data cannot be directly adopted as outliers can significantly bias the regression results. For example in Fig. 2, the gray dashed line depicts the regression result with the LS estimator on a data set of 47 two-dimension observations with outliers. It can be seen that the derived model is severely pulled away from the majority of data by outliers. Moreover, the original observations are not available due to the data confidentiality concern, making the problem even harder.

2.3 Background of M-Estimation

Given all observations, the unknown coefficient vector $\boldsymbol{\beta}$ can be estimated as $\hat{\boldsymbol{\beta}}^T = [\hat{\beta}^{(0)}, \ldots, \hat{\beta}^{(p)}]$, by using various regression estimators, e.g., Least Squares estimator. Thus, the expected value of $y_{i,k}$, called the *fitted value*, is $\hat{y}_{i,k} = x_{i,k}^T \hat{\boldsymbol{\beta}}$ and the residual can be calculated as $r_{i,k} = y_{i,k} - \hat{y}_{i,k}$.

In this paper, we exploit a particular M-estimator [30] to achieve the optimal model coefficients by minimizing the estimating function defined as

$$\sum_{i=1}^{m}\sum_{k=1}^{n}\rho\left(\frac{r_{i,k}^{\hat{\boldsymbol{\beta}}}}{s}\right), \tag{2}$$

where $\rho(\cdot)$ is an objective function, $r_{i,k}^{(\hat{\boldsymbol{\beta}})}$ denotes the residual calculated with $\hat{\boldsymbol{\beta}}$ at participant N_i using the kth observation, and s is the dispersion of residuals which is used to normalize $r_{i,k}^{(\hat{\boldsymbol{\beta}})}$. Specifically, s is defined as the solution to

$$\frac{1}{m \times n}\sum_{i=1}^{m}\sum_{k=1}^{n}\rho\left(\frac{r_{i,k}^{(\hat{\boldsymbol{\beta}})}}{s}\right) = K, \tag{3}$$

where K is set to $E(\rho(u))$, which is the expected value of $\rho(u)$ in which u has a standard normal distribution [31].

With outlier observations, the objective function $\rho(u)$ should be chosen so that larger residuals (from potential outliers) will receive smaller influence on the estimation. Moreover, $\rho(u)$ should strictly increase when the absolute value of u is smaller than a threshold a and be a constant otherwise. We adopt the Tukey bisquare [32] as the objective function, defined as

$$\rho(u) = \begin{cases} \frac{u^2}{2} - \frac{u^4}{2a^2} + \frac{u^6}{6a^4}, & if|u| \le a \\ \frac{a^2}{6}, & if|u| > a, \end{cases} \tag{4}$$

where a is chosen so that $\rho(a) = 2K$. With this setting, estimate of s can tolerate up to 50 percent outliers [30]. Particularly, setting $a = 1.547$ (accordingly $K = 0.199$) satisfies both the requirements on a and K [33].

In order to achieve the minimum of (2), we have the following partial differentiation equations with respect to each of the $p+1$ parameters of $\boldsymbol{\beta}$,

$$\sum_{i=1}^{m}\sum_{k=1}^{n}\mathbf{x}_{i,k}^{(j)}\varphi\left(\frac{r_{i,k}^{(\hat{\boldsymbol{\beta}})}}{s}\right) = 0, j = 0, \ldots, p, \tag{5}$$

where $\varphi(u) = \frac{\partial\rho(u)}{\partial u}$.

As no closed form solution to (5) exists, *Iteratively Reweighted Least Squares* (IRLS) is required to find an approximate solution. In specific, the procedure is that, given an initial estimate $\hat{\boldsymbol{\beta}}_{(0)}$, residuals $r_{i,k}^{(\hat{\boldsymbol{\beta}}_{(0)})}$ are calculated and $\hat{\boldsymbol{\beta}}_{(1)}$ can be determined by solving (5). Then, $\hat{\boldsymbol{\beta}}_{(1)}$ can be used in the second iteration and get $\hat{\boldsymbol{\beta}}_{(2)}$. This procedure continues until a convergence criterion has been met. With this M-estimator, when the initial estimate $\hat{\boldsymbol{\beta}}_{(0)}$ can tolerate high ratio of outliers, the resolved regression model can also resistant to the same portion of outliers [32].

3 Models and Design Goals

3.1 Models

We consider typical participatory sensing application scenarios where participants are mobile device users and they can communicate with the server and other users via WiFi

and 3G/4G. We characterize the participants and server from the following perspectives:

- *The server* is greedy but rational. The first goal of the server is regression modeling. The server also wishes to obtain private data of participants as many as possible.
- *All participants* are curious about the content of private data. They try to infer the private information of others by observing intermediate results during regression modeling. They may collude (or with server) to share information in order to deduce more private information. We assume that the number of participants in collusion is limited.
- *Malicious participants* are selfish and rational. They may contribute falsified data, misleading the server to conclude a regression model as they wish. We assume that the number of malicious participants is also limited.

Note that, we do not distinguish whether outliers are caused by normal errors from honest participants or deliberately corrupted data from malicious ones.

3.2 Design Goals

An efficient and practical scheme to addressing the problem defined in Definition 3 needs to meet the following requirements:

- *Strong data confidentiality.* In a participatory sensing application, leaking private information will lead to reductions in participants and frustrate the application. As a result, such a regression scheme should strongly protect the data confidentiality of participants so that other participants including the server cannot obtain the original observations of one particular participant.
- *Good modeling accuracy.* As the existence of outliers embedded in observations is prevalent, it is essential to achieve good modeling accuracy. Such a scheme should be resilient to low-quality data and achieve good modeling accuracy even when almost half of the data are outliers.
- *Low communication cost.* In typical participatory sensing scenarios, participants are mobile device users and exchange messages with the server via wireless communications which may incur extra communication fees and power consumption. Therefore, such a scheme should have a low communication overhead.

4 Design of PURE

4.1 Design Overview

In the case of participatory sensing, an effective linear regression modeling scheme should have both low-data-quality-tolerant and blind properties. We propose an innovative scheme, called PURE, which completely satisfies this rigid requirement with low computation and communication costs. The core idea of PURE is to let participants not only collect sensory data but also to be involved in the whole process of modeling decision. As no original data are available, it is impossible to directly get the global optimal regression model at the server. As a result, PURE conducts an iterative procedure and lets participants and the server negotiate about the fitness of the current global estimate for individual local observations in a few number of rounds until the final agreement requirement is met. During all negotiations, participants only report aggregated results to the server, which well preserves the local data confidentiality. To achieve this, PURE integrates three effective stages:

Collecting effective local estimates. To protect the privacy of participants, instead of naively sending all observations to the server, each participant estimates a local regression model and reports the estimated model to the server. Those local estimates can be further used to achieve the initial global estimate. In the case where outliers are embedded in local observations, those locally estimated models might severely deviate from the true model. In order to reduce such effect, we check the data consistency of each participant and collect effective local estimates only from those participants whose observations follow the same trend.

Establishing an initial global estimate. Given all local estimates, to perform the M-estimation, the server needs to select one preferable local estimate as the initial global estimate so that it can be as "close" to the optimal global model as possible. To this end, the server distributes all collected local estimates over all participants and asks all participants to peer review on each local estimate by checking the distances between each local estimate and their own observations. Finally, the local model having the minimal median of such distances over all local models is chosen as the initial global estimate.

Refining the global estimate. As the initial global estimate may not best fit all observations, the server coordinates an iterative negotiation with all participants to further refine the global estimate. In each iteration, given the current global estimate, all participants first judge the quality of their own observations. Observations which are far away from the current model are potential outliers and will be assigned smaller weights. Then, participants report the server with the corrected residuals of their own observations to the current global estimate in an aggregated way. With those corrected residuals, the server can refine the current estimate and start the next iteration. After a few rounds of such refinement, when all participants can achieve an "agreement" on the weights of observations, the global optimal estimation is achieved.

By design, PURE is robust to outliers and can strongly protect the privacy of each participant. We conduct intensive security analysis and extensive trace-driven simulations to demonstrate the efficacy of PURE design, which are elaborated in Sections 5 and 6, respectively.

4.2 Collecting Effective Local Estimates

In order to avoid participants with obvious outliers to generate local estimates, we first check the data consistency of each participants. To this end, each participant N_i for $i = 1, \ldots, m$ first collects a group of $p+2$ independent observations of its readings and gets $p+2$ tuples of the response and explanatory variables $\{y_{i,k}, x_{i,k}^{(1)}, x_{i,k}^{(2)}, , x_{i,k}^{(p)}\}$ for $k = 1, \ldots, p+2$. Then, N_i draws a subgroup of $p+1$ observations out of the group (without loss of generality, assume

the index of the observations in the subgroup is $1, \ldots, p+1$) to solve the following system of equations

$$\begin{cases} y_{i,1} &= \theta_i^{(0)} + \theta_i^{(1)} x_{i,1}^{(1)} + \theta_i^{(p)} x_{i,p}^{(p)} \\ &\vdots \\ y_{i,p+1} &= \theta_i^{(0)} + \theta_i^{(1)} x_{i,p+1}^{(1)} + \theta_i^{(p)} x_{i,p+1}^{(p)}, \end{cases} \tag{6}$$

and gets a hyperplane of $p+1$ dimensions with coefficients $\boldsymbol{\theta}_i = \{\theta_i^{(0)}, \theta_i^{(1)}, \ldots, \theta_i^{(p)}\}$. As there are $C_{p+2}^{p+1} = p+2$ of such hyperplanes with corresponding coefficients denoted as $\boldsymbol{\theta}_{i,1}$,$\boldsymbol{\theta}_{i,2}$,...,$\boldsymbol{\theta}_{i,p+2}$, N_i checks the consistency of those hyperplanes by measuring the *cosine similarity* between vectors $\boldsymbol{\theta}_{i,u}$ and $\boldsymbol{\theta}_{i,v}$ for $u, v \in \{1, \ldots, p+2\}$, defined as

$$Sim(\boldsymbol{\theta}_{i,u}, \boldsymbol{\theta}_{i,v}) = \frac{\boldsymbol{\theta}_{i,u} \cdot \boldsymbol{\theta}_{i,v}}{\parallel \boldsymbol{\theta}_{i,u} \parallel \times \parallel \boldsymbol{\theta}_{i,v} \parallel}. \tag{7}$$

If $Sim(\boldsymbol{\theta}_{i,u}, \boldsymbol{\theta}_{i,v}) \geq \sigma$ where σ is a predefined threshold for any pair of $\boldsymbol{\theta}_{i,u}$ and $\boldsymbol{\theta}_{i,v}$, then N_i is considered as a *consistent participant*.

Remark 2. It should be noted that those participants with their observations following the same trend could be consistent, even though this local trend may be severely bias from the global trend.

After checking local observations, consistent participants are required to report their local estimated models to the server. More specifically, if N_c is consistent, N_c estimates the regression model fitting all $p+2$ observations using the LS estimator and generates locally estimated model $\hat{\boldsymbol{\theta}}_c = \{\hat{\theta}_c^{(0)}, \ldots, \hat{\theta}_c^{(p)}\}$ and reports $\hat{\boldsymbol{\theta}}_c$ to the server.

After the above consistency checking, inconsistent participants would be cancelled out from involving in the selection of initial estimate of global regression model. This can significantly reduce the complexity of the problem. However, whether a good initial estimate can be found from only consistent participants is not clear. We have the theorem as follows,

Theorem 1. *Given that all observations are independent and observations from the same participant obey the same distribution, if the number of explanatory variables is p, the portion of outliers among observations of one participant is at most ε, and the probability that at least one participant has no outliers in its observations (called a "clean" participant) is η, then the number of participants m involved in the participatory sensing should be no less than $\frac{\log(1-\eta)}{\log(1-(1-\varepsilon)^{p+2})}$, and the expectation of the number of clean participants is at least $m(1-\varepsilon)^{p+2}$.*

Proof. For a specific participant, the probability that an observation is not an outlier is at least $1-\varepsilon$ and the probability that all $p+2$ observations are not outliers is at least $(1-\varepsilon)^{p+2}$. The probability that a group is "contaminated" (i.e., the local observations of the participant contain at least one outlier) is at most $1-(1-\varepsilon)^{p+2}$. Then the probability that all m participants have their observations contaminated is at most $(1-(1-\varepsilon)^{p+2})^m$, which means the probability η that at least one group is clean is at least $1-(1-(1-\varepsilon)^{p+2})^m$. Thus, we have $m \geq \frac{\log(1-\eta)}{\log(1-(1-\varepsilon)^{p+2})}$. As participants are independent and the probability that one participant is clean is at least $(1-\varepsilon)^{p+2}$, the distribution of the number of clean participants can be lower bounded by a binomial distribution $\mathbb{B}(m, (1-\varepsilon)^{p+2})$. Then the expectation of the number of clean participants is at least $m(1-\varepsilon)^{p+2}$ and this concludes the proof. □

This means as long as we have a sufficiently large number of participants, with high probability, we will have at least one clean participant submitting its local estimated model to the server for further processing. For example, when nine explanatory variables are involved in the model and the outlier ratio of all participants cannot exceed 30 percent, if the probability that at least one participant has no outliers is 0.95 (i.e., with high probability), then the number of participants should be no less than 150 and the expected number of clean participants is at least three.

4.3 Establishing an Initial Global Estimate

In order to determine which local consistent estimate can best fit the majority of observations among all participants and therefore is more appropriate to be chosen as the initial estimate of the global regression model, the server distributes all local estimates to all participants, who assist the server in determining the best initial estimate.

Specifically, for each consistent estimate $\hat{\boldsymbol{\theta}}_c$ for $c = 1, \ldots, \pi$, participant N_i calculates the residuals using its $p+2$ observations as follows,

$$\mathbf{r}_i^{(\hat{\boldsymbol{\theta}}_c)} = \begin{bmatrix} r_{i,1}^{(\hat{\boldsymbol{\theta}}_c)} \\ \vdots \\ r_{i,p+2}^{(\hat{\boldsymbol{\theta}}_c)} \end{bmatrix} = \begin{bmatrix} y_{i,1} \\ \vdots \\ y_{i,p+2} \end{bmatrix} - \begin{bmatrix} 1 & \ldots & x_{i,1}^{(p)} \\ \vdots & \ddots & \vdots \\ 1 & \ldots & x_{i,p+2}^{(p)} \end{bmatrix} \begin{bmatrix} \hat{\theta}_c^{(0)} \\ \vdots \\ \hat{\theta}_c^{(p)} \end{bmatrix}, \tag{8}$$

where $r_{i,k}^{(\hat{\boldsymbol{\theta}}_c)}$ denotes the residual calculated with model coefficients $\hat{\boldsymbol{\theta}}_c$ at participant N_i using the kth observation.

To determine the best initial estimate, one straightforward solution for the server is to collect all residuals calculated with a specific local estimate and calculate the least mean of square of all residuals and then choose the local estimate with the least mean of square of residuals as the best initial estimate. The problem with this solution, however, is very obvious. First, the least mean of square of residuals is very sensitive to outliers; second, if the number of local estimates is sufficient large, then it is possible for the server or other adversaries to infer the original observations of all participants. For example, if $\pi \geq p+1$, then it is easy to solve the following system of equations and get $y_{i,k}$ and $\mathbf{x}_{i,k}^T$:

$$\begin{bmatrix} r_{i,k}^{(\hat{\boldsymbol{\theta}}_1)} \\ \vdots \\ r_{i,k}^{(\hat{\boldsymbol{\theta}}_\pi)} \end{bmatrix} = \begin{bmatrix} y_{i,k} \\ \vdots \\ y_{i,k} \end{bmatrix} - \begin{bmatrix} \hat{\theta}_1^{(0)} & \hat{\theta}_1^{(1)} & \ldots & \hat{\theta}_1^{(p)} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\theta}_\pi^{(0)} & \hat{\theta}_\pi^{(1)} & \ldots & \hat{\theta}_\pi^{(p)} \end{bmatrix} \begin{bmatrix} 1 \\ x_{i,k}^{(1)} \\ \vdots \\ x_{i,k}^{(p)} \end{bmatrix}. \tag{9}$$

In PURE, we choose the Least Median of Squares (LMS) of residuals to evaluate local estimates, which is a robust measure of central tendency and can tolerate an outlier ratio of 50 percent [22]. In order to find the median of squares of residuals of one specific consistent estimate $\hat{\boldsymbol{\theta}}_c$ and

meanwhile to prevent the private value $(y_{i,k}, \mathbf{x}_{i,k}^T)$ from being revealed by others, the residuals should be submitted to the server in a way that (9) cannot be constructed by any others.

To this end, N_i randomly re-arranges the sequence of the residuals of its $p+2$ observations. In specific, for each estimate $\hat{\boldsymbol{\theta}}_c$, N_i conducts a random permutation scheme on corresponding set of residuals, i.e., transforms the ordered set $\{r_{i,1}^{(\hat{\boldsymbol{\theta}}_c)}, r_{i,2}^{(\hat{\boldsymbol{\theta}}_c)}, \ldots, r_{i,p+2}^{(\hat{\boldsymbol{\theta}}_c)}\}$, to $\{r_{i,\sigma_{c,1}}^{(\hat{\boldsymbol{\theta}}_c)}, r_{i,\sigma_{c,2}}^{(\hat{\boldsymbol{\theta}}_c)}, \ldots, r_{i,\sigma_{c,p+2}}^{(\hat{\boldsymbol{\theta}}_c)}\}$, where there is a bijection from $\{1, 2, , p+2\}$ to $\{\sigma_{c,1}, \sigma_{c,2}, , \sigma_{c,p+2}\}$. Then N_i sends the re-arranged residual sets to the server instead of the original ones. In this way, for each $\hat{\boldsymbol{\theta}}_c$, the server obtains all residuals of N_i in another order. It is hard for the server to deduce the relationships between the original and rearranged residual pairs. We give a detail analysis on the difficulty of recovering the original observations in Section 5.

After getting all residuals from every participant, it is easy for the server to conclude the median of squares of residuals for each local estimate $\hat{\boldsymbol{\theta}}_c$. The server chooses the local estimate with the LMS of residuals, denoted as $\hat{\boldsymbol{\theta}}_\delta$, as the initial estimate of the global regression model.

4.4 Refining the Global Estimate

As introduced in Section 2.3, in order to solve (5) (particularly, $n = p+2$) and get the optimal global estimation, an iterative IRLS procedure is required. We elaborate the procedure in this section.

To reduce the significance of outliers to the global estimation, we define a 'weight' function as $w(u) = \frac{\varphi(u)}{u}$, yielding $w_{i,k} = w(\frac{r_{i,k}}{s})$, and then substitute this in to (5). We have

$$\sum_{i=1}^{m}\sum_{k=1}^{p+2} x_{i,k}^{(j)} w_{i,k}(y_{i,k}(y_{i,k} - \mathbf{x}_{i,k}^T\boldsymbol{\beta}))\frac{1}{s} = 0, j = 0, \ldots, p$$

$$\sum_{i=1}^{m}\sum_{k=1}^{p+2} x_{i,k}^{(j)} w_{i,k} y_{i,k} = \sum_{i=1}^{m}\sum_{k=1}^{p+1} x_{i,k}^{(j)} w_{i,k}\mathbf{x}_{i,k}^T\boldsymbol{\beta}, j = 0, \ldots, p.$$

Define

$$X_i = \begin{bmatrix} \mathbf{x}_{i,1}^T \\ \vdots \\ \mathbf{x}_{i,p+2}^T \end{bmatrix}, Y_i = \begin{bmatrix} y_{i,1} \\ \vdots \\ y_{i,p+2} \end{bmatrix}, W_i = \begin{bmatrix} w_{i,1} & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & w_{i,p+2} \end{bmatrix},$$

then

$$\sum_{i=1}^{m}(X_i)^T W_i X_i \boldsymbol{\beta} = \sum_{i=1}^{m}(X_i)^T W_i Y_i$$

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{m}(X_i)^T W_i X_i\right)^{-1}\sum_{i=1}^{m}(X_i)^T W_i Y_i.$$

Given the initial global estimate $\hat{\boldsymbol{\beta}}_{(0)} = \hat{\boldsymbol{\theta}}_\delta$, $w_{i,k}^{(0)}$ can be determined by $r_{i,k}^{(\hat{\boldsymbol{\beta}}_{(0)})}$ and thus $\hat{\boldsymbol{\beta}}$ can be iteratively calculated as:

$$\hat{\boldsymbol{\beta}}_{(q)} = \left(\sum_{i=1}^{m}(X_i)^T W_i^{(q-1)} X_i\right)^{-1}\sum_{i=1}^{m}(X_i)^T W_i^{(q-1)} Y_i, \text{for } q > 0.$$

In the above procedure, the server first calculates the parameters of s by requiring residuals $\boldsymbol{r}_i^{\hat{\boldsymbol{\theta}}_\delta}$ from N_i for $i \in [1, m]$, and solving (3). Then, the server distributes s to all participants. Note that sharing $r_{i,k}^{(\hat{\boldsymbol{\theta}}_\delta)}$ with the server only provides one equation in (9), hence it would not violate the private data $(y_{i,k}, \mathbf{x}_{i,k}^T)$ of N_i. In order to calculate each $\hat{\boldsymbol{\beta}}_{(q)}$ for $q > 0$, the server needs to know $\mathbf{x}_{i,k}$ and $r_{i,k}^{(\hat{\boldsymbol{\beta}}_{(q-1)})}$ of all participants. Instead of sending those local data to the server, with s, each participant N_i can locally compute its own $(X_i)^T W_i^{(q-1)} X_i$ and $(X_i)^T W_i^{(q-1)} Y_i$. In addition, if N_i reports those results to the server in each iteration, then it is possible for the server to retrieve the original observations when q is sufficiently large. To avoid this, participants need to further securely aggregate their local results. To this end, we use a secure summation scheme based on the slicing technical. For example, we explain the key idea of data slicing in calculating $\sum_{i=1}^{m}(X_i)^T W_i X_i$ as follows:

First, each participant N_i dynamically selects l_{in} other participants nearby to distribute its local result. It is assumed that any pair of participants nearby can achieve a unique pairwise key used for secure data transmission.

Second, N_i slices its data into $l_{in}+1$ random slices, i.e., $(X_i)^T W_i X_i = \sum_{j=1}^{l_{in}+1} A_i^{(j)}$, where $A_i^{(j)}$ is a matrix of dimension $(p+2)\times(p+2)$.

Third, N_i keeps one of $A_i^{(i)}$ to itself while sending each other slice $A_i^{(j)}, j \neq i$ to the corresponding participant N_j it has selected. Meanwhile, N_i can also receive l_{out} slices $A_j^{(i)}$ from l_{out} different participants. Then N_i recalculates its local matrix using its own slice $A_i^{(i)}$ and l_{out} slices received from others, i.e., $A_i^{(i)} + \sum_{j=1}^{l_{out}} A_j^{(i)}$, and sends it to server.

Finally, server adds up all the received values. It is easy to check that the result is the sum of $(X_i)^T W_i X_i$ for all i.

With this iteration procedure, two factors are crucial to the accuracy of the resolved model, i.e., the initial estimate $\hat{\boldsymbol{\beta}}_{(0)}$ and the convergence of iterations. As to the initial estimate $\hat{\boldsymbol{\beta}}_{(0)}$, we use $\hat{\boldsymbol{\theta}}_\delta$ obtained from the above section which is expected to be close to the optimal. With regard to the convergence of iterations, it is guaranteed with such an M-estimation using the proposed weight function [32]. In PURE, the iteration stops when $Sim(\hat{\boldsymbol{\beta}}_{(q)}, \hat{\boldsymbol{\beta}}_{(q-1)})$ is larger than a threshold. Intuitively, by increasing the number of iterations, the approximate solution can be arbitrarily close to the optimal solution. However, this will also cause a large number of interactions between the server and participants, which leads to unpleasant regression costs in terms of delay and network overhead. We will further study the tradeoff between model accuracy and regression costs in the performance evaluation section.

4.5 Computation Complexity on Mobile Device of Participants

As mobile devices owned by participants are resource-constraint, we analyze the computation complexity mainly on the participant side. In PURE, the main computation is multiplication, the computation complexity can be represented as the number of multiplication operations. We analyze the computation complexity on each step of PURE as follows:

4.5.1 Checking Consistency of Local Observations

In this step, N_i firstly solves $(p+2)$ systems of linear equations using (6) to compute $(p+2)$ vectors $\boldsymbol{\theta}_{i,u}$ ($1 \leq u \leq p+2$, each one is in dimension of $(p+1)$), and then calculates the similarity among them. These two operations take $O((p+2)(p+1)^3 + (p+2)(p+1)^2)$ multiplications at maximum. In addition, the local regression modeling needs $O((p+2)^3)$ multiplications.

4.5.2 Determining $\hat{\theta}_\delta$

In this step, N_i needs to conduct π matrix (each in dimension of $(p+2) \times (p+2)$) multiplications, which takes $O(\pi(p+2)^3)$ multiplications.

4.5.3 Refining the Global Model

In this step, N_i should iteratively determine the accuracy of current global model and present the "comment" on it. In each iteration, there are three matrix multiplications (each matrix in size of $(p+2, p+2)$), and seven multiplications for each observation. Thus, the computation complexity should be $O(3(p+2)^3 + 7(p+2))$.

In fact, p indicates the number of explanatory variables $\{x_{i,k}^{(1)}, x_{i,k}^{(2)}, , x_{i,k}^{(p)}\}$ of which the number is restricted by applications and usually less than one hundred.

5 Security Analysis

In this section, we analyze the security of PURE under three typical attack types.

5.1 Observation Recovery Attacks

In the stage of deciding $\hat{\theta}_\delta$, given the residuals about those local estimates reported from each participant, the server tries to recover the original observations. Note that for each local estimate, a participant conducts a random permutation on the corresponding residuals and reports a new order of residuals to the server. To recover one observation of N_i, e.g., $y_{i,k}$ and $\mathbf{x}_{i,k}^T$, as shown in (9), the server needs to know at least p corresponding residuals, e.g., $r_{i,k}^{(\hat{\boldsymbol{\theta}}_j)}, j = 1, \ldots, p$. As the probability for the server to correctly guess the position of the corresponding residual $r_{i,k}^{(\hat{\boldsymbol{\theta}}_j)}$ in the random permutation of residuals $\{r_{i,\sigma_{c,1}}^{(\boldsymbol{\theta}_j)}, r_{i,\sigma_{c,2}}^{(\boldsymbol{\theta}_j)}, \ldots, r_{i,\sigma_{c,p+2}}^{(\boldsymbol{\theta}_j)}\}$ is $\frac{1}{p+2}$, the probability that the server can correctly guess all p residuals and recover $y_{i,k}$ and $\mathbf{x}_{i,k}^T$ would be a small probability of $(\frac{1}{p+2})^p$. For example, even when $p = 3$, the probability is only 0.008.

Nevertheless, the server can conduct a brute-force search attack. Specifically, the server first re-arranges all received residual permutations on p local estimates and guesses $p+2$ observations of N_i one by one by solving (9). Then, it estimates the regression model $\hat{\boldsymbol{\theta}}_c^{'}$ fitting the recovered observations. Last, it compares $\hat{\boldsymbol{\theta}}_c^{'}$ with the local model $\hat{\boldsymbol{\theta}}_c$. If $\hat{\boldsymbol{\theta}}_c^{'} = \hat{\boldsymbol{\theta}}_c$, the server believes that all recovered observations are correct; otherwise, it repeats the whole process. The efficiency of such attack relates to the probability $\mathbb{P}(R)$ that the server correctly recovers all $p+2$ observations, which is $\mathbb{P}(R) = \prod_{k=1}^{p+2}(\frac{1}{k})^p = [\frac{1}{(p+2)!}]^p$. For example, when $p = 3$, $\mathbb{P}(R)$ is 0.48×10^{-8}, which is very slim.

5.2 Collusion Attacks

Malicious participants can collude to gather slices sent by participant N_i to retrieve $(X_i)^T W_i^{(q)} X_i$ which can be used to recover the local observations of N_i. In PURE, with the proposed summation scheme, to recover $(X_i)^T W_i^{(q)} X_i$, malicious participants have to collect all l_{in} slices $A_i^{(j)}$ sent by N_i as well as all l_{out} slices $A_j^{(i)}$ received by N_i. However, it is hard to know all $A_j^{(i)}$ as they are encrypted. Given M colluding participants, the probability $\mathbb{P}(M)$ that M adversaries can deduce $(X_i)^T W_i^{(q)} X_i$ equals to the probability that all participants connecting to N_i have colluded, i.e., $\mathbb{P}(M) = \frac{C_M^{l_{out}+l_{in}}}{C_m^{l_{out}+l_{in}}}$. As $M \ll m$, $\mathbb{P}(M)$ is extremely small.

5.3 Data Manipulation Attacks

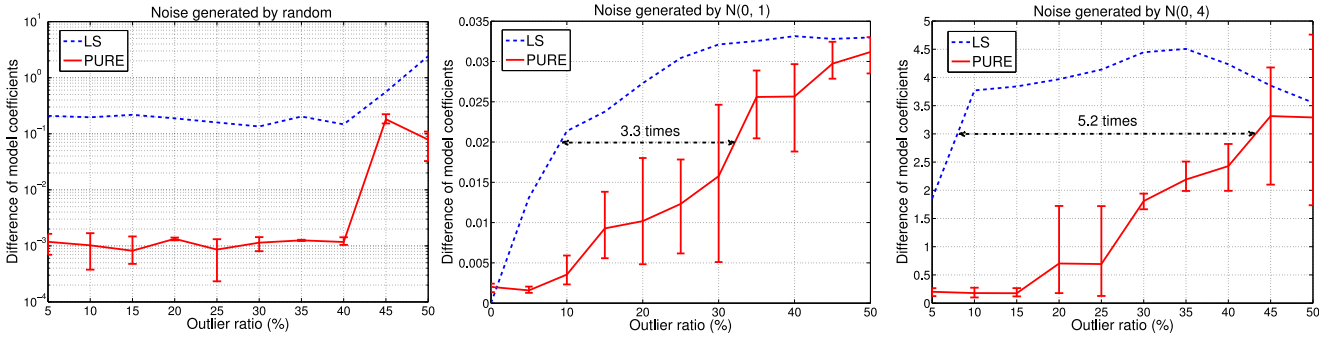
The regression result can be influenced by data manipulation. Malicious participants can falsify observations or intermediate results to mislead the server, incurring a bias regression model as they wish. In PURE, the false data, however, can be treated as common outliers as long as those false data are less than 50 percent of all observations.

6 Prototype Implementation

To validate the practical feasibility of PURE, we have implemented the PURE protocols on 36 Android smartphones (owned by graduate students and faculties of our laboratory, and undergraduate students in the teaching class of authors) and a HP Z230 desktop computer (equipped with an Inter Core i7 3.2 GHz CPU and 8 GB RAM, running on Windows 8). We use all smartphones to serve as different participants in a participatory sensing task and set the desktop as the server to collect sensory data from participants and coordinate the regression modeling procedure accordingly. Data communication between participants and the server is based on the WLAN of our laboratory or 3G network.

With this prototype system, we conduct a small scale experiment based on the data set of *vending time* [37], which consists of 25 original observations. For each observation, three measures about the time required to service a vending machine, the number of products stocked and the distance walked by the route driver are recorded, respectively (p is 2 as defined in Definition 1).

To make the scenario more practical with outlier observations, we further conduct two data operations as follows. First, in PURE, for each participant, $p+2$ observations should be collected, which means that, for 36 participants, 144 independent observations are needed. Due to the lack of sufficient original observations, 119 synthetic observations are generated by first applying an LS estimator to the 25 original observations and then randomly producing additional observations according to the derived model. It should be noted that, as the original observations has no outliers, synthetic observations derived in this way are perfectly homogenous with those original observations. We random divide the 144 observations into 36 groups



(a) Random noise (b) Normal noise under $N(0,1)$ (c) Normal noise under $N(0,4)$

Fig. 3. Difference of model coefficients $\mathbb{S}$ vs. outlier ratio.

(each contains four observations) and distribute them to all participants. Second, as the original and synthetic observations are rather ideal for a linear model, noise should be injected into the data. We choose three different types of noise to generate, and superpose them on observations, i.e., *random noise*, *normal distributed noise satisfying* $N(0,1)$ *and* $N(0,4)$, respectively (please refer to Section 7.1 for more details on noise generation). We vary the outlier ratio ε from 5 to 50 percent at an interval of 5 percent. For this purpose, under each setting of noise (outlier ratio ε and types of noise), the server generates $144 \cdot \varepsilon$ tokens and randomly distributes them to all participants (i.e., each participant gets at most four tokens). For each token, a participant randomly picks one observation and produces an outlier observation by superposing the noise generated according to the given noise type.

In PURE, given an outlier rate ε, the number of explanatory variables p, and the number of participants m involved in the participatory sensing, the probability η that at least one group of obseravtions is clean is at least $1-(1-(1-\varepsilon)^{p+2})^m$ (please refer to Theorem 1). Hence, consider the maximum outlier rate 50 percent, $p=2$ and $m=36$, η is over 90 percent under all noise settings, which guarantee that with high probability, one clean group can be obtained. There are two thresholds should be per-decided before the experiment. First, we set the constant threshold σ, in the stage of collecting effective local estimates, for participants to check its consistency to 0.9. Second, we set the convergence criteria of the iterations for the server to stop iterations (defined in Section 4.4) to 0.001. For each noise setting and outlier ratio, the server performs linear regression modeling adopting PURE and LS, respectively. The accuracy of an estimate $\boldsymbol{\beta}$ is evaluated by calculating the Euclidean distance $\mathbb{S}$ between $\boldsymbol{\beta}$ and the ground truth $\boldsymbol{\beta}^*$, which is obtained using an LS estimator on all original observations without any noise. We run the experiment 10 times and calculate the averages.

Fig. 3 plots the average difference of model coefficients between the estimated model and the global optimal $\boldsymbol{\beta}^*$ as a function of outlier ratio under different noise settings. In Fig. 3a, it can be seen that, under the random noise setting, PURE can achieve excellent accuracy even when ε increases to 40 percent. The estimated model derived by PURE can be two orders of magnitude closer to the optimal than that derived by LS. Figs. 3b and 3c illustrate the cases in normal distributed noise settings with variance equal to one and four, respectively. It can be seen that PURE outwits LS in all settings. For instance, PURE can tolerate more than three times of outliers than LS when the variance of noise is one and more than five times when the variance of noise is increased to four.

We also measure the average running time on each smartphone and count the average number of iterations for PURE to finalize the modeling progress. For example, the average running time for a Galaxy Nexus 3 (G3) smartphone (equipped with a 1.2 GHz dual-core CPU and 1 GB RAM, running on Android 4.2) to first check its local consistency and to review the intermediate global estimate received from the server in each iteration is 0.9 μs and 4.33 μs, respectively, and that for the server to determine $\hat{\theta}_\delta$ is about 0.38 ms. With the rigid convergence criteria is set as $\mathbb{S}(\hat{\boldsymbol{\beta}}_{(q)}, \hat{\boldsymbol{\beta}}_{(q-1)}) < 0.001$, the average number of iterations for PURE to converge with three noise types is 2.7, 2.4 and 3.1, respectively. From the results of this experiment, we have the experience that PURE is not only robust to a large number of outliers but also lightweight in terms of computation and communication costs. We further extensively investigate the performance of PURE on more data sets in the following section.

7 Performance Evaluation

7.1 Methodology

We examine the performance of PURE via trace-driven simulations. We use four well-known data sets as follows:

1) *Car price* [34]. This data set includes 804 observations. Each observation has eight measures about the influence function of the retail price of cars.
2) *Car mile per gallon (mpg)* [34]. In this data set, each observation has eight measures about the influence function of the mpg of cars. The number of observations is 398.
3) *Body fat* [35]. In this data set, each observation has fourteen measures about the influence function of the percentage of body fat. The number of observations is 252.
4) *Octane rating* [36]. Each observation has five measures about the influence function of the three raw materials to the octane rating of a particular petrol. There are 82 observations.

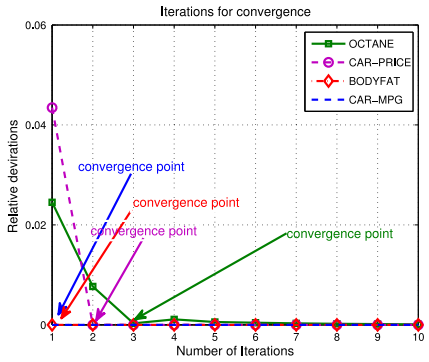


Fig. 4. Model deviations vs. the number of iterations.

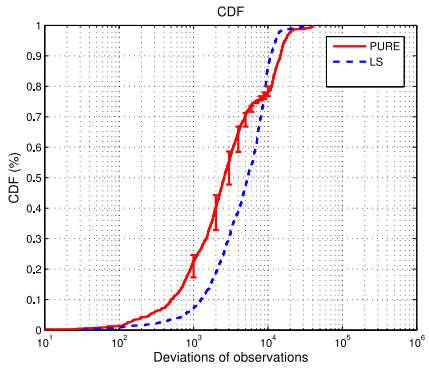


Fig. 5. CDF of residuals of observations in *Car price* data set.

Based on the *Car mpg*,*Body fat* and *Octane rating* data sets, we also synthesize three data sets, where various degrees of noise are added. Specifically, all the noises are random vectors that all dimensions have independent and identical distributions. Thus we have the following three types of noises:

1) *Random noise*. Variables obey uniform distribution within the range $[0, v_{max} - v_{min}]$, where v_{max} and v_{min} refer to the maximum measure and the minimum measure in the data set, respectively;
2) *Normal distributed noise with* $N(0, 2)$. Variables obey normal distribution with a mean of zero and a variance of two;
3) *Normal distributed noise with* $N(0, 4)$. Variables obey normal distribution with a mean of zero and a variance of four.

We evaluate the accuracy of an estimate $\hat{\boldsymbol{\beta}}$ by calculating the Similarity Sim between $\hat{\boldsymbol{\beta}}$ and the global optimal estimate $\hat{\boldsymbol{\beta}}_{optimal}$ according to (7). We compare PURE with the state-of-art LS-based private data preserving participatory regression modeling schemes [9], [10].

7.2 Number of Iterations Needed for Convergence

As a larger number of iterations means more interactions between participants and the server, this results to a longer delay and larger communication cost in modeling. In this experiment, we examine how PURE converges in refining the global estimate. We use all four data sets and randomly separate the observations in a data set into groups according to the number of explanatory variables of an observation p, i.e., each group has $p + 2$ observations. In data set of *Car price*, *Car mpg*, *Body fat* and *Octane rating*, the number of observations in each group is therefore 9, 9, 15 and 6, respectively. We vary the number of iterations from one to 10 at an interval of one and run the experiment 10 times and get the average.

Fig. 4 plots the relative deviations of $\hat{\boldsymbol{\beta}}_{(q)}$ and $\hat{\boldsymbol{\beta}}_{(q-1)}$, defined as $\frac{\|\hat{\boldsymbol{\beta}}_{(q)} - \hat{\boldsymbol{\beta}}_{(q-1)}\|}{\|\hat{\boldsymbol{\beta}}_{(q-1)}\|}$ as a function of the number of iterations, where q is the number of iterations and $\hat{\boldsymbol{\beta}}_{(0)}$ is the initial global estimate $\hat{\boldsymbol{\theta}}_{\delta}$. It can be seen that, for all data sets, the value of deviation drops down very quickly as the number of iterations increases. Given the rigid convergence criteria the deviation is less than 0.001, for original data set *Car price*, and *Car mpg*, *Body fat*, *Octane rating* with 20% $N(0, 2)$ noise, the number of iterations before PURE converges is 2, 1, 1, and 3, respectively.

7.3 Modeling Accuracy under Different Datsets

We first conduct LS and PURE (run 10 times and get the average) estimator on the four datasets. Fig. 5 plots the cumulative distribution function (CDF) of deviations of observations in *Car price*, i.e., $|y_i - \hat{y}_i|$, to LS estimator and PURE. It can be seen that *Car price* is a low-quality dataset which leads to a severe bias of the model derived using LS estimator. For example, more than 50 percent observations have larger than 3,000 deviations with the LS model. It can also be seen that the regression model estimated by PURE are more fitted to the majority of observations (e.g., only 30 percent deviations of observations are larger than 3,000).

Fig. 6 plots the CDF of relative deviations of observations in other three datasets. It can be seen that 1) all three datasets are relative high-quality that LS estimator can build an accurate regression model where more than 90 percent observations have smaller than 0.45 relative deviations; 2) for high-quality dataset, PURE has the same performance with LS.

7.4 Robustness under Different Outlier Ratios

In this experiment, we further examine the robustness of PURE against outliers under different outlier ratios ε, defined as the ratio of the portion of outliers to the total amount of data, and of different types. In specific, we choose the three high quality datasets, and use the global optimal model $\boldsymbol{\beta}^*$ estimated by LS estimator as the ground truth. In PURE, given an outlier rate ε and the number of explanatory variables p, the number of observation groups should be at least $\frac{\log(1-\eta)}{\log(1-(1-\varepsilon)^{p+2})}$, so that at least one clean

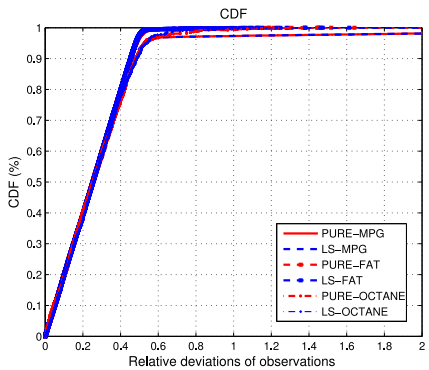


Fig. 6. CDF of relative residuals of observations in *Car mpg*, *Body fat* and *Octane* data sets.

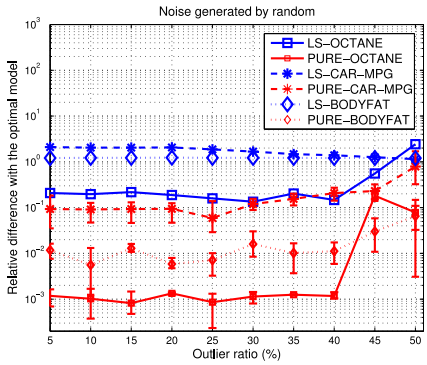


Fig. 7. Difference from the optimal model vs. outlier ratio with random noise.

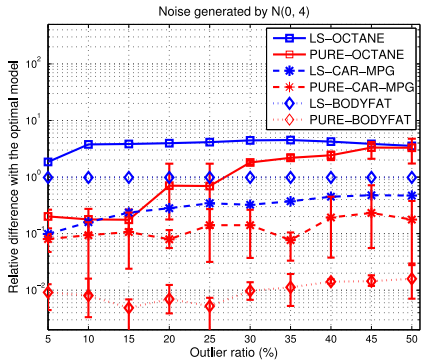


Fig. 9. Difference from the optimal model vs. outlier ratio with normal noise under $N(0, 4)$.

group is included with probability η. Due to the limited number of observations in the original datasets, for large ε, we need first to generate extra observations according to the optimal model $\boldsymbol{\beta}^*$. We also add small uniformly distributed random noise in the range $[0, 1]$ on the generated observations to imitate the real data. We set $\sigma = 0.9$, $\eta = 0.9$ and vary the outlier ratio ε from 5 to 50 percent at an interval of 5 percent. For each ε, we generate a trace accordingly to satisfy the requirement on the amount of observations for 10 times. We then randomly divide the trace into groups as introduced in the above experiment and run the experiment 10 times and get the average over all traces.

Fig. 7 plots the relative difference of model coefficients between the estimated model and the global optimal $\boldsymbol{\beta}^*$ as a function of outlier ratio under the random noise setting. Note that the plot is on a linear-log scale. We can see that at a given outlier ratio, PURE can achieve perfect accuracy even when ε increases to 40 percent. The estimated model derived by PURE can be two orders of magnitude closer to the optimal than that derived by LS. Figs. 8 and 9 plot the relative difference of model coefficients as a function of outlier ratio under the normal distributed noise settings with variance equal to two and four, respectively. It can be seen that PURE outwits LS in all settings. In addition, when the variance of noise is increased from two to four, the performance gaps between PURE and LS become even larger. LS estimator is sensitive to outliers. When the outliers have a big difference to normal data, LS gets bad performances even under a low outlier ratio (e.g., 5 percent), whereas PURE maintains an good accuracy under different ε, which can be two orders of magnitude than LS. Overall, we find that PURE is robust to not only the number of outliers embedded in the data but also the randomness of outliers.

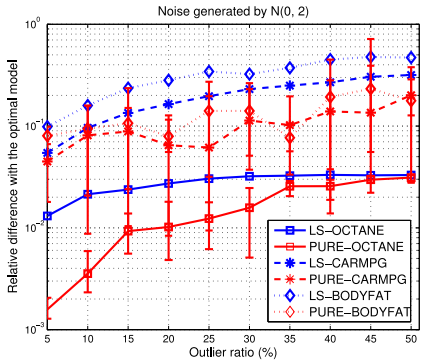


Fig. 8. Difference from the optimal model vs. outlier ratio with normal noise under $N(0, 2)$.

7.5 Performance Comparison

We compare PURE with several state-of-art linear regression modeling schemes with regard to the following aspects: capability to protect participants' privacy, capability to tolerate outliers, capability to resist a malicious server, computational cost on the participant side, and communication cost. We present the results in Table 1.

Xing et al. [9] have proposed an LS-based privacy-preserving regression modeling approach, named M-PERM, where tree-structure aggregations are conducted. Specifically, individual participants are organized into hierarchical clusters. A participant first locally computes its private aggregation result, and then transmits the result upward to the aggregation point in the current cluster. In such a way, private data are aggregated layer by layer until to the server. In M-PERM, the messages should be securely transmitted within the network. In order to achieve this, symmetric encryption and key distribution schemes, which increases the computational cost. At each level of the hierarchy, each participant only needs to transmit its aggregation result once and therefore the transmission cost is low.

Ahmadi et al have proposed a scheme [10] where participants only need to compute some features about their sensitive observations and submit the derived features instead of the original data to the server in one packet. The most computationally expensive operation requires one matrix multiplication $\mathcal{M}^T\mathcal{M}$, ($\mathcal{M}$: an $n \times p$ matrix, where n is the number of observations, p is the number of explanatory variables). Hence, both the computational and communication cost are low.

The above two schemes are focus on secure and distributed realizations of LS estimator. In order to protect the private data of participants, participants locally compute a set of aggregation results base on their observations, and provide them instead of original observations to the sever. After obtaining all aggregation results, the server can conclude the regression model, which is exactly the same as the model derived from traditional LS estimator. Meanwhile, others including the server cannot deduce the original observation according to the aggregation results. As the original data are kept secret at each participant, the privacy protection of data is strong.

TABLE 1
Performance Comparison

Schemes	Privacy protection	Outlier tolerance	Against malicious server	Computational cost	Communication cost
PURE	Strong	Yes	Yes	Low	Medium
M-PERM	Strong	No	Yes	High	Low
H. Ahmadi's	Strong	No	Yes	Low	Low
PoolView	Medium	No	No	Low	Low

Nevertheless, because this scheme is based on the LS estimator, it has weak capability to deal with outliers.

Several privacy-preserving techniques based on data alteration have been proposed for participatory sensing applications. For example, PoolView [24] can protect private data by injecting particular noise into the original data set. Specifically, the server determines a noise model and shares the structure and probability distributions of all parameters of the model with all participants. By choosing random values for these parameters from the specified distribution, one participant can generate its own private noise which is used for perturbing its original data. The server collects all perturbed data and adopts a LS-estimator to obtain the trend of the perturbed data. Since the distribution characteristics of noise are statistically known, it is possible to subtract the average of noise from the sum of perturbed data, yielding an approximated trend of the data. Because PoolView adopts a LS-estimator, it can hardly handle outliers. In general, PoorView can defend against traditional filtering attacks such as PCA and spectral filtering. However, the strength of privacy protection depends on the noise model generated by the server. For instance, a malicious server may distribute a noise model with a deliberate set of parameters so that, using the parameter distributions sent by the server, a participant can only generate a very small set of noise streams. In this way, the private data is vastly less blurred and therefore can be easily exposed to the sever. As noise is locally generated by participants and all perturbed data are transmitted to the server in one time, the computational and communication costs are low.

8 Related Work

Privacy-preserving data aggregation issues have been widely investigated in wireless networks. PriSence is a privacy preserving data aggregation solution in people centric urban sensing systems based on the idea of data slicing and mixing [15]. iPDA is an integrity-protecting private data aggregation scheme in which data integrity is achieved through redundancy by constructing disjoint aggregation paths [16]. Ozdemir and Yang [25] proposed a polynomial regression based secure data aggregation protocol. Most of the privacy-preserving aggregation schemes only focus on calculating additive or non-additive aggregation functions such as *sum*, and *max*/*min*. Existing solutions cannot be used for private regression estimation directly.

Privacy-preserving outlier detection has been studied in distributed systems. Most schemes deal with the distance-based outliers, which are defined as data points having further distance than a number of other data points. However, in regression modeling, such a data point can be a "leverage" point. Furthermore, the proposed secure solution in [7], [17] requiring pair-wise comparison between data and homomorphic encryptions on observations, which leads to low efficiency. Density-based outliers detection is investigated to identify local outliers [18], [19]. Group outliers may break down those schemes.

The random-value perturbation techniques are used for protecting privacy of data by masking the sensitive data using random noise. Agrawal et al. [20] proposed a perturbation based method in which privacy-preserving multidimensional aggregations on data are partitioned across multiple agents. Kargupta et al. challenged the utility of such technique in privacy protection [22]. They pointed out that some random-data distortions preserve little data privacy. Huang et al. proposed two data reconstruction methods based on data correlations [23]. They claimed that when the correlations are high, the original data can be reconstructed more accurately, i.e., more private information can be disclosed. Evfimievski et al. developed an approach, named "amplification", to limit the privacy breaches when tackles with the problem of mining association rules [21]. PoolView provides privacy guarantees on stream data in participatory sensing applications, where participants cooperatively measure aggregate phenomena of interest [24]. The core idea is to add random noise with a known distribution to the user's data, after which a reconstruction algorithm is used to estimate the distribution of the original data. This perturbation method is resilient to traditional filtering techniques, such as Kalman filter, and Spectral filtering. However, using such perturbation techniques in privacy preserving regression will introduce error in the regression model, which leads to the inaccuracy of modeling.

Studies which focus on the privacy-preserving accurate regression model construction are more related to our work. Du et al. [26] studied the Secure two-party Multivariate Linear Regression and Secure two-party Multivariate Classification problems. Sanil et al. [28] addressed the problem that multiple data owners have data on certain subjects but on different sets of attributes of those entities. They proposed an algorithm that enables data owners to conduct a linear regression analysis with complete records without disclosing the values of their own attributes. Recently, two most relevant works have been proposed [9], [10]. In these schemes, data privacy issues are addressed in participatory sensing and regression coefficients estimation is achieved. A series of data transformation and aggregation operations are operated at the participatory nodes (and clusters), which help keeping the data of participants private while not introducing any additional error to model construction. In [9], [27], secure regression model fitting and diagnosing are investigated.

However, it is worth pointing out that these schemes are based on the least square estimation. The correctness

of them relies on the assumption that original data are collected correctly by participants, and no gross errors (which occur as the result of mistakes) are involved in the original data set. Any unusual observations may break down the estimation, since least square estimation is very sensitive to outliers.

9 Conclusion and Future Work

In this paper, we have proposed PURE scheme of blind regression modeling under low quality data in participatory sensing. PURE can provide strong protection on data confidentiality by only exchanging aggregated information and achieve extraordinary accuracy with a large portion of random outliers by refining the global model in iterations. Both security analysis and extensive simulation results demonstrate the efficacy of PURE. In future work, we will establish a prototype system of participatory sensing on our campus, and further examine the feasibility of PRUE based on the real deployment.

Acknowledgments

This research is supported by the National Natural Science Foundation of China (Grant Nos. 61300199, 61173171, 61472068, 61370205, 61472254, 61170238, 61202375, 61472255 and 61420106010), 973 Program (2014CB340303), the Fundamental Research Funds for the Central Universities (2232014D3-21), China Postdoctoral Science Foundation funded project (2014M550466), STCSM (12ZR1414900), and CCF-Tencent Open Fund.

References

[1] R. Ganti, N. Pham, H. Ahmadi, S. Nangia, and T. Abdelzaher, "GreenGPS: A participatory sensing fuel-efficient maps application," in *Proc. 8th Int. Conf. Mobile Syst., Appl. Serv.*, 2010, pp. 151–164.

[2] D. Mendez, A. Perez, M. Labrador, and J. Marron, "P-Sense: A participatory sensing system for air pollution monitoring and control," in *Proc. IEEE Int. Conf. Pervasive Comput. Commun. Workshops*, 2011, pp. 344–347.

[3] L. Deng and L. Cox, "LiveCompare: Grocery bargain hunting through participatory sensing," in *Proc. 10th Workshop Mobile Comput. Syst. Appl.*, 2009, pp. 1–6.

[4] S. He, D. Shin, J. Zhang, and J. Chen, "Toward optimal allocation of location dependent tasks in crowdsensing," in *Proc. INFOCOM*, 2014, pp. 745–753.

[5] H. Zhou, J. Chen, H. Zhao, W. Gao, and P. Cheng, "On exploiting contact patterns for data forwarding in duty-cycle opportunistic mobile networks," *IEEE Trans. Veh. Technol.*, vol. 62, no. 9, pp. 4629–4642, Nov. 2013.

[6] E. Miluzzo, N. Lane, S. Eisenman, and A. Campbell, "CenceMe-injecting sensing presence into social networking applications," in *Proc. 2nd Eur. Conf. Smart Sens. Context*, 2007, pp. 1–28.

[7] J. Vaidya and C. lifton, "Privacy-preserving outlier detection," in *Proc. 4th IEEE Int. Conf. Data Mining*, 2004, pp. 233–240.

[8] Q. Li and G. Cao, "Providing privacy-aware incentives for mobile sensing," in *Proc. IEEE Int. Conf. Pervasive Comput. Commun.*, 2013, pp. 76–84.

[9] K. Xing, Z. Wan, P. Wu, and H. Zhu, "Mutual privacy-preserving regression modeling in participatory sensing," in *Proc. INFOCOM*, 2013, pp. 3139–3147.

[10] H. Ahmadi, N. Pham, R. Ganti, T. Abdelzaher, S. Nath, and J. Han, "Privacy-aware regression modeling of participatory sensing data," in *Proc. 8th ACM Conf. Embedded Netw. Sens. Syst.*, 2010, pp. 99–112.

[11] M. Kutner, C. Nachtsheim, J. Neter, and W. Li. *Applied Linear Statistical Models*. New York, NY, USA: McGraw-Hill, 2005.

[12] J. Cui and F. Valois, "Data aggregation in wireless sensor networks: Compressing or forecasting?" in *Proc. IEEE Wireless Commun. Netw. Conf.*, 2014, pp. 2892–2897.

[13] S. He, J. Chen, D. K.Y. Yau, and Y. Sun, "Cross-layer optimization of correlated data gathering in wireless sensor networks," *IEEE Trans. Mobile Comput.*, vol. 11, no. 11, pp. 1678–1691, Nov. 2012.

[14] S. Guo, H. Zhang, Z. Zhong, J. Chen, Q. Cao, and T. He, "Detecting faulty nodes with data errors for wireless sensor networks," *ACM Trans. Sens. Netw.*, vol. 10, no. 3, Article 40, 2014.

[15] J. Shi, Y. Zhang, Y. Liu, and Y. Zhang, "PriSense: Privacy-preserving data aggregation in people-centric urban sensing systems," in *Proc. INFOCOM*, 2010, pp. 1–9.

[16] W. He, H. Nguyen, X. Liu, K. Nahrstedt, and T. Abdelzaher, "iPDA: An integrity-protecting private data aggregation scheme for wireless sensor networks," in *Proc. IEEE Military Commun. Conf.*, 2008, pp. 1–7.

[17] Z. Zhou, L. Huang, Y. Wei, and Y. Yun, "Privacy preserving outlier detection over vertically partitioned data," in *Proc. Int. Conf. E-Bus. Inform. Syst. Security*, 2009, pp. 1–5.

[18] W. Jin, A. Tung, and J. Han, "Mining top-n local outliers in large databases," in *Proc. 7th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2001, pp. 293–298.

[19] M. Breuning, H. Kriegel, R. Ng, and J. Sander, "LOF: Identifying density-based local outliers," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2000, pp. 93–104.

[20] R. Agrawal, R. Srikant, and D. Thomas, "Privacy preserving OLAP," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2005, pp. 251–262.

[21] A. Evfimievski, J. Gehrke, and R. Srikant, "Limiting privacy breaches in privacy preserving data mining," in *Proc. 22nd ACM SIGMOD-SIGACT-SIGART Symp. Principles Database Syst.*, 2003, pp. 211–222.

[22] H. Kargupta, S. Datta, Q. Wang, and K. Sivakumar, "Random-data perturbation techniques and privacy-preserving data mining," *Knowl. Inform. Syst.*, vol. 7, no. 4, pp. 387–414, 2005.

[23] Z. Huang, W. Du, and B. Chen, "Deriving private information form randomized data," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2005, pp. 37–48.

[24] R. Ganti, N. Pham, Y. Tsai, and T. Abdelzaher, "PoolView: Stream privacy for grassroots participatory sensing," in *Proc. 6th ACM Conf. Embedded Netw. Sens. Syst.*, 2008, pp. 281–294.

[25] S. Ozdemir and X. Yang, "Polynomial regression based secure data aggregation for wireless sensor networks, in *Proc. GLOBECOM*, 2011, pp. 1–5.

[26] W. Du, Y. Han, and S. Chen, "Privacy-preserving multivariate statistical analysis: Linear regression and classification," in *Proc. SIAM Int. Conf. Data Mining*, 2004, pp. 222–233.

[27] A. Karr, X. Lin, A. Sanil, and J. Reiter, "Secure regression on distributed databases," *J. Comput. Graphical Statist.*, vol. 14, no. 2, pp. 1–18, 2005.

[28] A. Sanil, A. Karr, X. Lin, and J. Reiter, "Privacy preserving regression modelling via distributed computation," in *Proc. 10th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2004, pp. 677–682.

[29] F. Hampel, "Beyond location parameters: Robust concepts and methods," *Bulletin Int. Statist. Inst.*, vol. 46, pp. 375–382, 1975.

[30] P. Huber, *Robust Statistics*. New York, NY, USA: Wiley, 1981.

[31] P. J. Rousseeuw and A. M. Leroy, *Robust Regression and Outlier Detection*. Hoboken, NJ, USA: Wiley, 1987.

[32] V. Yohai, "High breakdown-point and high efficiency robust estimates for regression," *Annals Statist.*, vol. 15, no. 2, pp. 642–656, 1987.

[33] P. Rousseeuw and V. Yohai, "Robust regression by means of S-estimators," in *Proc. Robust Nonlinear Time Series Anal.*, 1984, pp. 256–272.

[34] [Online]. Available: http://www.dcc.fc.up.pt/ ltorgo/Regression /DataSets.html, 2014.

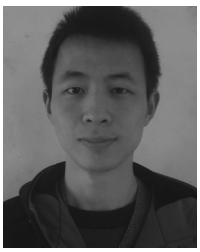
[35] StatLib—Datasets Archive. [Online]. Available: http://lib.stat.cmu.edu/datasets/, 2005.

[36] REGRESSION Linear Regression Datasets. [Online]. Available: http://people.sc.fsu.edu/ jburkardt/datasets/regression/x17.txt, 2011.

[37] The ROUSSEEUW datasets. [Online]. Available: http://www.uni-koeln.de/themen/statistik/data/rousseeuw, 2008.

Shan Chang (M'08) received the BS degree in computer science and technology from the Xián Jiaotong University in 2004 and the PhD degree in computer software and theory from the Xián Jiaotong University in 2013. From 2009 to 2010, she was a visiting scholar with the Department of Computer Science and Engineering, the Hong Kong University of Science and Technology. She was also a visiting scholar with BBCR research lab, the Electrical and Computer Engineering Department, University of Waterloo from 2010 to 2011. Since 2013, she has been an assistant professor with the Department of Computer Science and Technology, Donghua University, Shanghai. Her research interests include security and privacy in mobile networks and wireless sensor networks. She is a member of the IEEE Computer Society, and IEEE Communication Society.

Hongzi Zhu (M'06) received the PhD degree in computer science from Shanghai Jiao Tong University in 2009. He is now an associate professor at the Department of Computer Science and Engineering in Shanghai Jiao Tong University. His research interests include vehicular networks, mobile computing, and smart computing. He is a member of the IEEE, the IEEE Computer Society and the Communication Society.



Wei Zhang received the bachelor of science degree in computer science from Southern East University in 2012 and is working towards the master's degree from Shanghai Jiao Tong University in 2015. His research interests include vehicular networks and mobile computing.

Li Lu (M'06) received the BS and MS degrees at Zhejiang University in 2000 and 2003, respectively and the PhD degree in information security from the Chinese Academy of Sciences in 2007. He was a post-doctoral fellow in the Department of Computer Science and Engineering at the Hong Kong University of Science and Technology from 2008 to 2010. He is an associate professor in the School of Computer Science and Engineering, the University of Electronic Science and Technology of China. His research interests include applied cryptography, network security, pervasive computing, and sensor networks. He is a member of the IEEE and IEEE Computer Society.

Yanmin Zhu (M'02) received the PhD degree in computer science from the Department of Computer Science and Engineering at the Hong Kong University of Science and Technology in 2007. He is an associate professor in the Department of Computer Science and Engineering at Shanghai Jiao Tong University. His research interests include vehicular networks, sensor networks, and mobile computing. Prior to joining Shanghai Jiao Tong University, he was a research associate with the Department of Computing at the Imperial College London. He is a member of the IEEE and the IEEE Communications Society.

▷ **For more information on this or any other computing topic, please visit our Digital Library at** www.computer.org/publications/dlib.