Journal of Network and Computer Applications 77 (2017) 123–134

Contents lists available at ScienceDirect

Journal of Network and Computer Applications

journal homepage: www.elsevier.com/locate/jnca

# Compressive sensing based data quality improvement for crowd-sensing applications

Long Cheng[a], Jianwei Niu[a,*], Linghe Kong[b], Chengwen Luo[c], Yu Gu[d], Wenbo He[e], Sajal K. Das[f]

[a] State Key Lab of Virtual Reality Technology and Systems, Beihang University, China
[b] Department of Computer Science and Engineering, Shanghai Jiao Tong University, China
[c] College of Computer Science and Software Engineering, Shenzhen University, China
[d] Watson Health Cloud, IBM Watson Health, USA
[e] School of Computer Science, McGill University, Canada
[f] Department of Computer Science, Missouri University of Science and Technology, USA

ARTICLE INFO

*Keywords:*
False data detection and correction
Crowd-sensing
Spatio-temporal compressive sensing

ABSTRACT

Crowd-sensing enables to collect a vast amount of data from the crowd by allowing a wide variety of sources to contribute data. However, the openness of crowd-sensing exposes the system to malicious and erroneous participations, inevitably resulting in poor data quality. This brings forth an important issue of false data detection and correction in crowd-sensing. Furthermore, data collected by participants normally include considerable missing values, which poses challenges for accurate false data detection. In this work, we propose DECO, a general framework to detect false values for crowd-sensing in the presence of missing data. By applying a tailored spatio-temporal compressive sensing technique, DECO is able to accurately detect the false data and estimate both false and missing values for data correction. Through comprehensive performance evaluations, we demonstrate the efficacy of DECO in achieving false data detection and correction for crowd-sensing applications with incomplete sensory data.

## 1. Introduction

The increased computational power and sensing capabilities of mobile devices (e.g., smartphones and tablets), along with cloud computing technology have made possible a new pervasive data collection paradigm - crowd-sensing (also known as participatory sensing) (Christin et al., 2011). This new data collection paradigm leverages individuals to collect and share sensory data from surrounding environments using their data collection devices such as smartphones, thus achieving cost-effective and large-scale data gathering (Reddy et al., 2010). Authors in Kuznetsov et al. (2010) and Grosky et al. (2007) give a broader definition: crowd-sensing refers to any mechanism by which individuals in the general public collect, share and analyze local sensory data. For example, people may share temperature sensors from their homes, or entities share private sensor networks for environmental monitoring. In this work, we use the *broad-sense definition* to refer to the crowd-sensing. Many crowd-sensing applications have emerged in recent years, including environment, transportation and civil infrastructure monitoring (Dutta et al., 2009; Kanjo, 2010), health and fitness monitoring (Lin et al., 2012), urban and social sensing (Ahn et al., 2010), radiomap construction in WiFi fingerprinting (Jun et al., 2013; Luo et al., 2014), and automatic inference of indoor semantics (Luo et al., 2015). Crowd-sensing also finds a wide range of applications for industrial sensing intelligence (Muntés-Mulero et al., 2013), such as for large-scale monitoring in modern industrial plants, targeting at improved productivity and increased workplace safety (Huo et al., 2015).

The inherent openness of crowd-sensing systems enables ubiquitous data collection by allowing anyone to contribute data. However, it also exposes the systems to malicious and erroneous participations. The sensory data contributed by crowd are not always reliable, since they can submit fake data to earn rewards without performing the actual sensing task (Talasila et al., 2013). Malicious users may purposely contribute false data for their own benefits. For example, in the real-time traffic monitoring, selfish users may report the false traffic jam alerts so as to divert the traffic on roads ahead for themselves. A leasing agent may intentionally generate fictitious low noise readings to promote the rental housing in a particular region (Huang et al., 2010). In addition, attackers may compromise the mobile devices to provide faulty sensor readings (Saroiu and

[*] Corresponding author.
*E-mail addresses:* chenglong@buaa.edu.cn (L. Cheng), niujianwei@buaa.edu.cn (J. Niu), linghe.kong@sjtu.edu.cn (L. Kong), chengwen@szu.edu.cn (C. Luo), yugu@us.ibm.com (Y. Gu), wenbohe@cs.mcgill.ca (W. He), sdas@mst.edu (S.K. Das).

http://dx.doi.org/10.1016/j.jnca.2016.10.004
Received 5 August 2016; Received in revised form 19 September 2016; Accepted 4 October 2016
Available online 11 October 2016
1084-8045/ © 2016 Elsevier Ltd. All rights reserved.

Wolman, 2010). Another category of false data (i.e., unintentional false data) stems from the failures of certain algorithms or built-in sensors on mobile devices. For instance, location, as one of the crucial contexts for crowd-sensing, is often inaccurately estimated in real-world systems (Jun et al., 2013). As a result, the same openness characteristic of crowd-sensing can threaten its success and impact the quality of services. In particular, the false data problem is one of the critical issues that affect the proper operation of crowd-sensing systems.

Techniques have been developed to achieve data integrity and correctness (Amintoosi and Kanhere, 2013; Wang et al., 2013; Kurasawa et al., 2014). However, no system has been presented as a general approach to detect and correct false data for crowd-sensing. There are a few existing solutions such as introducing the reputation management (Amintoosi and Kanhere, 2013; Wang et al., 2013) or providing hardware-based security to avoid cheating in crowd-sensing (Akshay Dua and Bulusu, 2009). The reputation based false data avoidance monitors the behaviour of participants and assign them reputation scores. However, reputation based approach is still vulnerable to collusion and Sybil attacks. On the other hand, even the participating users are trustworthy, it is still difficult to guarantee the correctness of all collected data, such as the unintentional false data. More recently, Kurasawa et al. (2014) pointed out that data collected by crowd usually include considerable missing values in practical crowd-sensing systems. They proposed a method to estimate missing values using a recursive regression model. The incompleteness of sensory data poses several challenging issues for accurate false data detection. Different from Kurasawa et al. (2014), the main objective of this work is to detect false values in crowd-sensing in the presence of non-negligible missing data. Our idea is to employ the spatio-temporal compressive sensing (ST-CS) technique (Roughan et al., 2012) to reconstruct the sensory data given an incomplete and partially inaccurate dataset. We check data consistency with co-located participants, and detect potential false data from misbehaving or erroneous participants.

In this work, we present a generalized false data detection and correction (DECO) framework, which is designed to detect incorrect data and perform possible correction with high probability in crowd-sensing environment. The contributions from this work are summarized as follows:

- Distinctive from existing works, we focus on false data detection considering the presence of considerable missing data in crowd-sensing. To address this challenge, we propose to exploit ST-CS technique, which can achieve an effective data reconstruction for high data-loss scenarios.
- Considering the spatial proximity of participants cannot be directly derived from the potentially inaccurate reported location information in practical crowd-sensing systems, we present a method to infer spatial adjacency of participants based on multidimensional sensor readings.
- We develop a general false data detection and correction algorithm by applying a tailored ST-CS technique for crowd-sensing. To the best of our knowledge, there are few other efforts applying ST-CS techniques for false data correction in crowd-sensing.
- Experimental case study and empirical evaluations done based on public dataset demonstrate the efficacy of DECO in achieving false data detection and correction for crowd-sensing applications with incomplete sensory data.

The rest of this paper is organized as follows. We survey previous work in Section 2. Section 3 describes the system model and motivations behind this work. Section 4 elaborates the design of DECO framework in details. Section 5 provides evaluation results by applying DECO in crowd-sensing-based WiFi fingerprinting and crowd-sensing environment monitoring applications. Finally, conclusions are drawn in Section 6. A short conference paper (Cheng et al., 2015) containing some preliminary results of this paper has appeared in IEEE/ACM IWQoS 2015.

## 2. Related work

Crowd-sensing has attracted extensive attentions in recent years. A large part of existing research efforts focus on proposing different crowd-sensing applications. The CarTel system (Bret et al., 2006) collects, processes, delivers, analyzes, and visualizes data from sensors located on mobile units (i.e., mobile phones and in-car embedded devices), which can be used for traffic mitigation, road surface monitoring and hazard detection. CommonSense (Dutta et al., 2009) is a crowd-sensing system collecting air quality data. LiveCompare (Deng and Cox, 2009) can facilitate price comparison of grocery items through participants using their camera phones to snap a photograph of the price tag of their product of interest. Authors in Kanjo (2010) proposed NoiseSPY, a participatory sound sensing system that allows users to collaboratively explore a city-scale noise levels in real-time. BeWell (Lin et al., 2012) assists individuals in maintaining a healthy lifestyle by keeping track of their everyday behaviors. MetroTrack (Ahn et al., 2010) presents a mobile-event tracking system to track mobile targets through collaboration among local sensing devices. Crowd-sensing-based WiFi fingerprinting has also received considerable attention during the past several years due to its potential efficacy to reduce the cost of radiomap construction (Rai et al., 2012; Yang et al., 2012; Wang et al., 2012; Kong et al., 2015; Luo et al., 2014). Recently, crowd-sensing-based industrial intelligence (Huo et al., 2015) has been proposed for large-scale collaborative monitoring to improve efficiency and security industrial environment. Authors in Huo et al. (2015) proposed the concept of "workers as sensors", which monitor industrial working spaces, e.g., measuring the concentration of toxic gas and reporting emergency events in real time to administrators.

Privacy preserving and incentive mechanism in crowd-sensing have attracted considerable attention in the literature. Privacy concern matters since sensor data contributed by crowd normally includes personally identifiable spatial-temporal stamps (Christin et al., 2011). The authors in De Cristofaro and Soriente (2013) introduce a privacy-enhanced infrastructure for crowd-sensing. The success of crowd-sensing is strongly dependent on users' enthusiasm for participating to provide sufficient and reliable sensory data (Luo and Tham, 2012). During the data collection, a user may consume his own private resources including device battery, computation power, privacy and manual effort. Therefore, many crowd-sensing incentive mechanisms are designed to encourage the general public to provide quality data (Lee and Hoh, 2010; Restuccia and Das, 2014; Luo et al., 2014).

Despite a plethora of research on crowd-sensing, there are a number of challenges in developing a practical crowd-sensing system. In particular, providing data correctness and trustworthiness is an important aspect for the proper functions of knowledge inference and incentive distribution in crowd-sensing. To motivate the voluntary collection of high quality data, reputation management (Huang et al., 2010; Amintoosi and Kanhere, 2013; Wang et al., 2013) has been introduced in crowd-sensing systems. In Reddy et al., the authors proposed five metrics (timeliness, capture, relevancy, coverage and responsiveness) to evaluate the quality of data and participants from a crowd-sensing campaign. However, the existing state-of-the-art data quality improvement solutions (Min et al., 2013; Vergara-Laurens et al., 2014; Kurasawa et al., 2014) lack general means to detect, validate and correct the gathered sensory data. Authors in Nam et al. (2010) and Ahmadi et al. (2010) presented privacy-preserving mechanisms for ensuring privacy of location-tagged crowd-sensing data while allowing accurate data reconstruction at the server side. LOCATE (Boutsis and Kalogeraki, 2013) is a middleware that aims to provide privacy preservation for crowd-sensing systems so that leak of sensitive data is prevented. These works mainly focus on manually perturbed data reconstruction. On the contrary, our work targets at a general

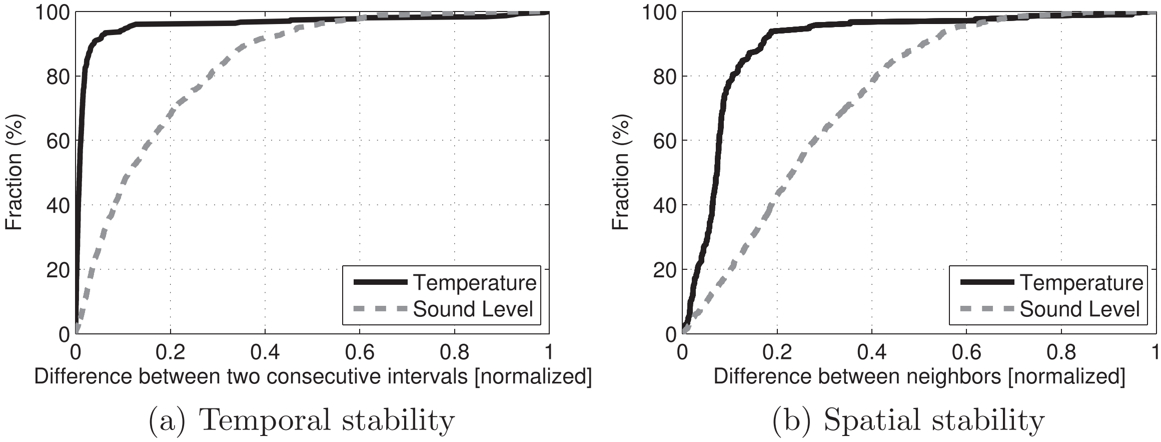


**Fig. 1.** Spatio-temporal stability in sensory data collected by smartphones.

framework for detecting potentially malicious or erroneous sensory data after data have been collected.

## 3. Preliminaries

In this section, we introduce the system model and motivations behind this work.

### *3.1. System model*

We consider a typical client-server crowd-sensing architecture, where a larger number of mobile devices are tasked into community-based data gathering. The sensory data collected by participants are reported (e.g., through WiFi or cellular networks) to a central application server. A task normally specifies multiple modalities of sensory data to be collected based on the application requirements (Christin et al., 2011), and an individual data collection device may be involved into multiple concurrent sensing tasks (Das et al., 2010; Song et al., 2014). In this work, we consider the collected data in a crowd-sensing system are *multidimensional time-series sensor readings* (Ahmadi et al., 2010; Groat et al., 2012). For example, in our prior project (Yow et al., 2014) studying social networks through mobile phone sensing, in addition to phone usage and co-location information, other context data were collected as well, such as the magnetic, audio, accelerometer and gyroscope sensory data. Although it was a single crowdsourcing task in Yow et al. (2014), we collected the multidimensional context data to cross-validate the obtained variables of interest. We assume the server could associate sensor data reports from the same participants. In many cases, participants in crowd-sensing are anonymous in terms of physical persons to avoid private information leakage. However, they may expose their user names to a crowd-sensing system since participants will probably receive incentive rewards after the data contributions. In addition, it is possible to link multiple data submissions from a participant according to the network-layer information such as IP addresses.

Generally, data types collected in crowd-sensing applications can be classified into the following two categories: 1) *public sensory data* such as for environmental monitoring or intelligent transportation applications (Dutta et al., 2009; Kanjo, 2010); and 2) *user behavioral sensory data* for human-centric applications such as social network (Ahn et al., 2010) and personalized health and fitness (Lin et al., 2012). In particular, this paper focuses on the public sensory data (e.g., environmental parameters, traffic variables or spacial information) in crowd-sensing environment.

### *3.2. Data representation*

Assume a crowd-sensing system with $N$ participants, multidimensional time-series data are generated by each participant and then reported to the centralized server. Let us assume that time is divided into continuous slots (e.g., five minutes per time unit) and the data reporting time span includes $T$ time slots. A participant $i$ produces a data record at time $t$ with $K$ different sensor types, where each sensor reading can be denoted by $s(i, t, k)$, where $i \in [1, N]$, $t \in [1, T]$ and $k \in [1, K]$.

To facilitate the description of our false data detection and correction approach, we use an $N \times K \times T$ matrix (i.e., $S_{N\times K\times T}$) to represent the collected data in a crowd-sensing system. For the dimension $k$ out of $K$-dimensional sensory dataset, we define an $N \times T$ sensory data matrix $S(k)$, which records the raw sensor readings collected from $N$ participants for $T$ time slots.

We assume $N$ participants are in a reasonably large area (e.g., in a university). In case of very large-scale sensing areas (e.g., city scale), to reduce the computation complexity of false data detection, we can subdivide the area into smaller grids and group the collected data for each sub-area with participants' locations in that sub-area only. Then, false data detection and correction will be performed for each sub-area.

### *3.3. Spatio-temporal correlation in sensory data*

Existing work has revealed that sensory data normally exhibit spatio-temporal correlation in either crowd-sensing environment (Nam et al., 2010; Kurasawa et al., 2014; Rallapalli et al., 2010) or traditional wireless sensor networks (Kong et al., 2013). That is, the sensory values or the value changes in one dimension from the same participant are usually similar at adjacent time slots, and sensory values from physically correlated participants are similar for a particular time instant. To confirm this empirically, we conducted experiments using 10 smartphones to collect temperature and sound level measurements in an open space for 60 min. The definitions of spatio-temporal stabilities and low-rank feature (i.e., redundancy) can be found in Section 4.2.

Fig. 1 plots the cumulative distribution function (CDF) curves of the spatio-temporal stability for temperature and sound level measurements collected by smartphones. From Fig. 1(a), we observe that over 90% of temperature changes between adjacent time slots are less than 5%. Sound level result shows a high variance relatively, around 35% of sound level changes are larger than 20%. Fig. 1(b) shows similar trends in spatial stability as the results in temporal stability in Fig. 1(a). These results indicate that temperature data in smartphone sensing show high spatio-temporal stability. While the sound level measurements are more sensitive to local noises (possibly due to smartphone microphones' heterogeneity and the noisy environment), showing less spatio-temporal correlation compared with the temperature.

Fig. 2 illustrates the distribution of singular values after performing the singular value decomposition over the sensory data matrix. It

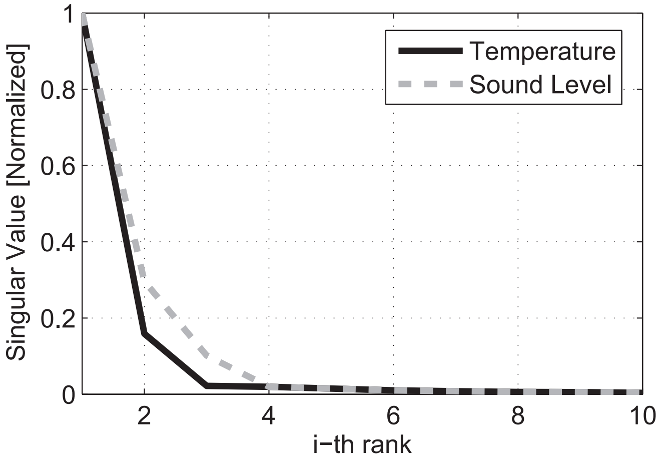


**Fig. 2.** Low-rank feature in sensory data collected by smartphones.

shows both temperature and sound level data exhibit obvious low-rank structures. These experimental results demonstrate that there exists inherent low-rank features and certain spatio-temporal correlation in real-world sensory data.

### 3.4. False data in crowd-sensing

Only in an ideal environment, participants provide accurate and complete sensing readings (Kurasawa et al., 2014; Vergara-Laurens et al., 2014). Unfortunately, on one hand, there exists inexperienced and malicious participants, which may provide corrupted sensory data in crowd-sensing systems. On the other hand, sensor readings are liable to be biased due to many reasons such as hardware heterogeneity and failure. As a result, the sensory data matrix normally contains missing and false sensor values, which motivates us to propose the DECO framework in this work.

To detect the false data, we can exploit data reconstruction techniques to rebuild the sensory data matrix $\mathbb{S}(k)$ based on the imperfect data matrix $S(k)$. Given the reconstructed sensory data matrix, by comparing difference between $\mathbb{S}(k)$ and $S(k)$, it is possible to detect data inconsistencies and likely to infer real data values in crowd-sensing systems. Therefore, the *key objective* in DECO is to develop an efficient data interpolating technique to reconstruct the sensory data matrix that approximates the real data values as close as possible.

### 3.5. Spatio-temporal compressive sensing

Compressive sensing (CS) (Candes and Tao, 2006) has attracted considerable attention as a generic methodology for recovering the unknowns based on partial observations. Spatio-temporal compressive sensing (ST-CS) has been proposed to reconstruct missing values for Internet traffic measurements (Roughan et al., 2012) and wireless sensor networks (Kong et al., 2013). The main idea is that many signals or datasets that are collected from real-world applications exhibit certain structure or redundancy, e.g., neighboring rows or columns in a sensory data matrix often have values close to each other. By utilizing this prior knowledge, ST-CS accurately reconstructs missing values in these real-world datasets.

## 4. Deco framework design

In this section, we first provide an overview of the DECO framework, then detail the underlying core components that comprise the DECO architectural framework. Specially, we address two challenges in applying ST-CS for false data detection and correction: 1) how to accurately derive the spatial adjacency of participants? and 2) how to properly apply ST-CS for data reconstruction in crowd-sensing environment?

### 4.1. Overview

The DECO framework is illustrated in Fig. 3. In typical crowd-sensing environment, sensory data are collected and uploaded to the central data server by a large number of participants over wide spans of space and time. DECO is designed as an enhancement layer for false data detection and correction in various crowd-sensing systems. DECO not only improves the data quality, but also provides useful information for application-layer modules such as reputation management and incentive distribution. Essentially, DECO improves the quality of service provided by a crowd-sensing system to the end users.

The data characterization module analyzes the low-rank structure and spatio-temporal properties in each data dimension (e.g., temperature, humidity, noise level, pressure, and location according to different sensor types) based on a training sensory dataset. The spatial constraint estimation module estimates the proximity of participants based on any context condition available in the dataset being detected. The key idea of DECO is to employ the ST-CS technique (Roughan et al., 2012) to reconstruct the sensory data given an incomplete and partially inaccurate dataset, in the event that the sensory data being reconstructed exhibit low-rank structure and spatio-temporal properties. Otherwise, data interpolation methods such as Delaunay Triangulation (Kong et al., 2013) and K-Nearest Neighbor can be used to rebuild the sensory data matrices. Since it has been shown that ST-CS can achieve an effective reconstruction even for high data-loss scenarios, in this work, we focus our investigation on the ST-CS based data reconstruction.

### 4.2. Data characterization

To apply the ST-CS technique for data reconstruction, we first characterize spatial and temporal dependencies for each data type in real-world sensor datasets (assume we have certain number of

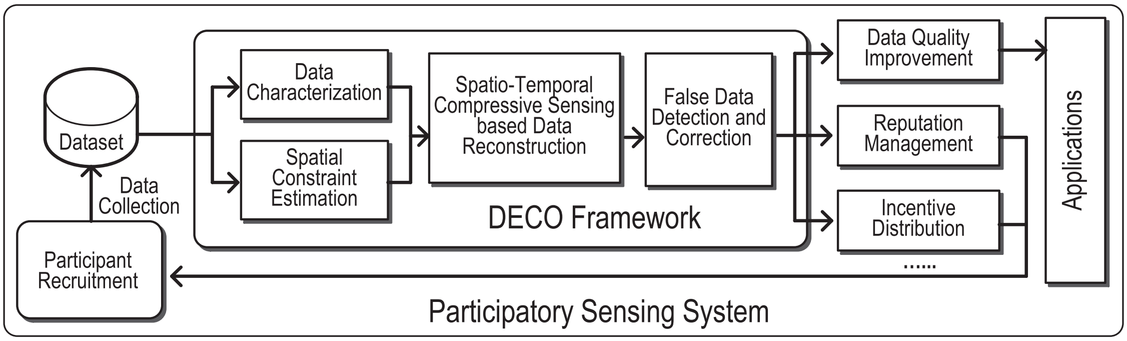


**Fig. 3.** DECO framework for crowd-sensing.

trustable participants for data collection in the bootstrap phase), using the following low-rank structure, temporal stability, and spatial stability metrics. For those data types (e.g., environmental parameters and location-dependent information) that exhibit pronounced low-rank structure (i.e., redundancy) and spatio-temporal stability, the ST-CS technique can be applied for efficient matrix reconstruction (Roughan et al., 2012).

### 4.2.1. Low-rank structure

As mentioned in Section 3.2, for the dimension $k$ (according to sensor types) in a sensory dataset, an $N \times T$ sensory data matrix $S(k)$ represents the data collected in a specific geographical area with $N$ participants during a given $T$ time period. Singular Value Decomposition (SVD) decomposes this $N \times T$ matrix into three matrices (Kong et al., 2013).

$$S(k) = U\Sigma V^{*}, \tag{1}$$

Where $U$ is an $N \times N$ unitary matrix, $\Sigma$ is an $N \times T$ diagonal matrix with non-negative real numbers on the diagonal, $V$ is a $T \times T$ unitary matrix, and $V^{*}$ denotes the conjugate transpose of $V$. The diagonal entries $\sigma_i$ of $\Sigma$ are known as the singular values of $S(k)$. Typically, these singular values are sorted in descending order. The rank of $S(k)$, denoted by $r$, is defined as the number of its non-zero singular values. If $r \ll min(N, T)$, we say $S(k)$ exhibits low-rank structure.

### 4.2.2. Temporal stability

We measure the temporal stability of $S(k)$ by calculating the normalized difference values $\Delta s(i, t, k)$ between adjacent time slots for the $k$th dimension data uploaded by participant $i$ during $T$ time period ($i \in [1, N]$, $t \in [1, T]$ and $k \in [1, K]$):

$$\Delta s(i, t, k) = \frac{|s(i, t, k) - s(i, t-1, k)|}{max_{\forall i, \forall t}\{|s(i, t, k) - s(i, t-1, k)|\}}, \tag{2}$$

where $max_{\forall i, \forall t}\{|s(i, t, k) - s(i, t-1, k)|\}$ is the maximal difference between any two consecutive time slots in $S(k)$.

### 4.2.3. Spatial stability

The rationale of spatial stability is that, in real-world sensor datasets, sensor readings measured by geographically nearby participants at the same time slot may be close in value. We first define the adjacency matrix $H$,

$$H = (h(i, j))_{N \times N} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ are neighbors;} \\ 0 & \text{otherwise,} \end{cases} \tag{3}$$

where $i, j \in [1, N]$. Both rows and columns in $H$ represent participants, and $h(i, j)$ represents whether participants $i$ and $j$ are neighbors or not. In the data characterization phase, we assume that data are collected by trustable participants in a reasonably large area. Thus, neighborhood in $H$ can be derived by the Euclidean distance between two participants (i.e., $i$ and $j$ are neighbors if their distance is less than a threshold $d$) or Bluetooth scanning to discover co-located participants (Yow et al., 2014).

The spatial stability of $S(k)$ at a specific time $t$ is measured by computing the normalized difference $\blacktriangledown s(i,t,k)$between the sensor value uploaded by a participant $i$ and the average value of $i$'s all neighbors:

$$\blacktriangledown s(i,t,k) = \frac{s(i,t,k) - (\sum_{j=1}^{N} s(j,t,k)h(i,j) / \sum_{j=1}^{N} h(i,j))}{max_{\forall i, \forall t} s(i,t,k) - min_{\forall i, \forall t} s(i,t,k)}, \tag{4}$$

where $\sum_{j=1}^{N} h(i, j)$ is the number of one-hop neighbors of node $i$. $max_{\forall i, \forall t} s(i, t, k)$ and $min_{\forall i, \forall t} s(i, t, k)$ are the maximum and minimum values in $S(k)$, respectively. ($\sum_{j=1}^{N} s(j, t, k)h(i, j) / \sum_{j=1}^{N} h(i, j)$) represents the average value of $i$'s all neighbors.

## 4.3. ST-CS for sensory matrix reconstruction

In the case where crowd sensory data exhibit a spatio-temporal structure, ST-CS leverages this structure to rebuild the sensory data matrix. Here, we briefly introduce the ST-CS technique, and refer interested readers to Kong et al. (2013) for more details.

Let us assume an $N \times T$ sensory data matrix $S(k)$ is being detected. $S(k)$ may contain missing and false values. We define an $N \times T$ missing index matrix $B(k)$, which indicates whether a data sampling in $S(k)$ is missing or not.

$$B(k) = (b(i, t, k))_{N \times T} = \begin{cases} 0 & \text{if } x(i, t, k) \text{ is missing,} \\ 1 & \text{otherwise.} \end{cases}$$

The objective in DECO is to accurately estimate $\mathbb{S}(k)$, which can be decomposed by SVD, and re-written as follows:

$$\mathbb{S}(k) = \mathcal{L}\mathcal{R}^{*},$$

where $\mathcal{L} = U\Sigma^{1/2}$, $\mathcal{R} = V\Sigma^{1/2}$, and $\mathcal{R}^{*}$ denotes the conjugate transpose of $\mathcal{R}$. Through theoretical derivations, the ST-CS matrix reconstruction problem is formulated as the following optimization problem:

$$min\{\|B(k)\cdot(\mathcal{L}\mathcal{R}^{*}) - S(k)\|_F^2 + \lambda(\| \mathcal{L} \|_F^2 + \| \mathcal{R}^{*} \|_F^2) \quad + \|\mathbb{H}\mathcal{L}\mathcal{R}^{*}\|_F^2 \\ + \|\mathcal{L}\mathcal{R}^{*}\mathbb{T}\|_F^2\},$$

where $\lambda$ is the Lagrange multiplier, and $\|\cdot\|_F^2$ is the Frobenius (Euclidean) norm. $B(k)$ and $S(k)$ are known. $\mathbb{H}$ and $\mathbb{T}$ are the spatial and temporal constraint matrices, which will be introduced in the following subsections. Note that $\|\mathbb{H}\mathcal{L}\mathcal{R}^{*}\|_F^2$, $\|\mathcal{L}\mathcal{R}^{*}\mathbb{T}^{*}\|_F^2$, and $\|B(k)\cdot(\mathcal{L}\mathcal{R}^{*}) - S(k)\|_F^2$ need to be set equal in the similar order of magnitude, otherwise, they may overshadow the others during optimization (Roughan et al., 2012). Then, by tuning $\lambda$, $\mathcal{L}$ and $\mathcal{R}$ can be estimated in this optimization problem, and $\mathbb{S}(k)$ is consequently estimated.

## 4.4. Deriving spatial constraint

For the purpose of characterizing spatial stability for different types of sensor readings, we can use ground truth co-location information to derive the spatial adjacency matrix $H$. However, real-world crowd-sensing produces inaccurate and uncertain sensory data as well as missing values. It poses challenges in accurately estimating spatial adjacency matrix $H$ in practical environment. For example, GPS traces are likely to be obfuscated for privacy-preserving on the participant-side prior to sharing them (Ahmadi et al., 2010; Nam et al., 2010). Bluetooth device discovery in mobile scenarios normally contains missing values since each inquiry scan incurs about 12 s delay (Yow et al., 2014). Malicious adversaries may deliberately upload forged location data (Talasila et al., 2013). In other words, we cannot directly derive $H$ from the location information in crowd-sensing datasets. However, applying ST-CS for sensory data reconstruction requires a good approximated spatial adjacency matrix $H$.

To address the above challenges, we propose to infer participants' proximity based on multidimensional sensor readings in crowd-sensing systems. We classify the sensory data into 1) *spatially-dependent* (location of the samples, WiFi AP signatures, and Bluetooth signatures) and 2) *non-spatial* (environmental variables) information/attributes. If values of spatially-dependent variables are similar, it is more likely that the two participants are nearby each other, e.g., NearMe (Krumm and Hinckley, 2004) compares the WiFi signatures to estimate the proximity of mobile devices to one another. As shown in Section 3.3, real-world environmental measurements made at nearby locations may be closer in value than measurements made at locations farther apart, *but not* vice versa. However, intuitively, if non-spatial values are remarkably different, it is likely that the two participants are far away.

The rationale of our spatial adjacency discovery is that, spatially-dependent information provides positive clues for proximity estimation. While non-spatial information can be used as non-adjacent (negative) indicators, which potentially improve the estimation accuracy of $H$. Assume at time slot $t$ in data dimension $k$, participants $i$ and $j$

have sensor readings $s(i, t, k)$ and $s(j, t, k)$, respectively. There are $M$ dimensional sensory data that we take into account for estimating $H$. We define a general proximity function to estimate the adjacency of participants $i$ and $j$ as follow, which is independent of specific applications:

$$h(i, j) = \max\left\{0, \sum_{k=1}^{M} \omega_k \cdot Similarity(i, j, t, k)\right\}, \tag{5}$$

where $\omega_k$ is the weight coefficient of the $k$th dimensional data ($\sum_{k=1}^{M} \omega_k = 1$) in the proximity function. $Similarity(i, j, t, k)$ is the function measuring the similarity of the $k$th dimensional data reported by participants $i$ and $j$ at time slot $t$. If $\sum_{k=1}^{M} \omega_k \cdot Similarity(i, j, t, k)$ is a negative value, $h(i, j)$ is set to 0.

Specifically, we use Pearson's correlation coefficient to measure the similarity for WiFi AP signatures (it also applies for calculating the similarity of two Bluetooth signatures). Suppose the two WiFi signatures measured by participants $i$ and $j$ at time $t$ are,

$$s(i, t, \text{“}wifi\text{”}) = \{(ap_1^{(i)}, rss_1^{(i)}), (ap_2^{(i)}, rss_2^{(i)})\ldots\},$$
$$s(j, t, \text{“}wifi\text{”}) = \{(ap_1^{(j)}, rss_1^{(j)}), (ap_2^{(j)}, rss_2^{(j)})\ldots\},$$

where $ap$ denotes the AP MAC address, and $rss$ represents the associated signal strength. Let $n_i$ and $n_j$ denote the number of APs scanned by participants $i$ and $j$ at time $t$, respectively. Since some APs in $s(i, t, \text{“}wifi\text{”})$ may not have corresponding APs in $s(j, t, \text{“}wifi\text{”})$, and vice versa, we add virtual APs with fixed signal strength (e.g., −110 dB m) in order to let them have the exactly same AP sequences. Let $n_{ij}$ denote the length of the AP sequence after inserting the virtual APs. Then, we calculate the Pearson's correlation coefficient $\sigma$ $(-1 \le \sigma \le 1)$, which is set as $Similarity(i, j, t, \text{“}wifi\text{”})$.

$$\sigma = \frac{\sum_1^{n_{ij}} (rss_n^{(i)} - \overline{rss^{(i)}})(rss_n^{(j)} - \overline{rss^{(j)}})}{\sqrt{\sum_1^{n_{ij}} (rss_n^{(i)} - \overline{rss^{(i)}})^2}\sqrt{\sum_1^{n_{ij}} (rss_n^{(j)} - \overline{rss^{(j)}})^2}}, \tag{6}$$

where $\overline{rss^{(i)}}$ denotes the average value of signal strength in $i$'s AP sequence after inserting the virtual APs.

For absolute location data, such as GPS coordinates, we use Euclidean distance to measure their similarity. That is, if the GPS distance between two participants $i$ and $j$ is less than a range $d$ at time $t$, then $Similarity(i, j, t, \text{“}gps\text{”}) = 1$, otherwise, the value is set 0.

We then exploit the non-spatial attributes (which are usually scalars such as temperature and noise level) to calibrate the estimation of $h(i, j)$. For non-spatial information, we define the similarity function of the $k$th dimensional data reported by participants $i$ and $j$ at time slot $t$ as follow ($Similarity(i, j, t, k)$ is abbreviated as $Sim$):

$$Sim = -\frac{\sum_{\tau=t-\theta}^{t+\theta} |(s(i, \tau, k) - s(j, \tau, k))|}{(2\theta + 1)(max_{\forall i, \forall t} s(i, t, k) - min_{\forall i, \forall t} s(i, t, k))} \tag{7}$$

where $(\sum_{\tau=t-\theta}^{t+\theta} |(s(i, \tau, k) - s(j, \tau, k))|/(2\theta + 1)$ is the mean value of absolute differences during the $(2\theta + 1)$ time period, and $\theta$ is an adjustable parameter specifying the time window length $(\theta \ge 0)$. Considering the temporal stability of non-spatial sensing data, we average the sensor readings over a time window for robust estimation. Since non-spatial information is used as a negative indicator, $Sim$ is always a negative value.

Note that the spatially-dependent sensory data dominates the estimation of spatial adjacency matrix $H$. Therefore, their weight coefficients should be larger than those of non-spatial data. By combining the similarities (either positive or negative) from $M$ dimensional sensory data, we calculate the value of every $h(i, j)$, which ranges from 0 to 1. Then, the spatial adjacency matrix $H$ is transformed to the spatial constraint $\mathbb{H}$ as follows.

$$\mathbb{H} = (\hbar(i, j))_{N\times N} = \begin{cases} 0 & \text{if } \sum_{j=1}^{N} h(i, j) == 0; \\ 1 & \text{else if } i == j; \\ -\dfrac{h(i, j)}{\sum_{j=1}^{N} h(i, j)} & \text{otherwise,} \end{cases} \tag{8}$$

where $\hbar(i, j)$ is an element in $\mathbb{H}$. We assign $-\frac{h(i,j)}{\sum_{j=1}^{N} h(i,j)}$ to each neighboring node in order to make the sum of elements in each row to be 0. Finally, we derive $\mathbb{H}$, which is applied as the spatial constraint into the ST-CS matrix reconstruction.

### 4.5. Temporal constraint

Since the temporal stability is an inherent feature of real-world sensory data, the temporal constraint matrix $\mathbb{T}$ is relatively easy to obtain. We set $\mathbb{T} = Toeplitz(0, 1, -2, 1)_{T\times T}$ (Rallapalli et al., 2010), which denotes the *Toeplitz* matrix with central diagonal given by ones, the first upper diagonal given by minus two, the second upper diagonal given by ones, and the others given by zeros, e.g.,

$$\mathbb{T} = \begin{bmatrix} 1 & -2 & 1 & 0 & \cdots \\ 0 & 1 & -2 & 1 & \vdots \\ 0 & 0 & 1 & -2 & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots \end{bmatrix}_{T\times T}. \tag{9}$$

The additional temporal constraints capture the temporal stability properties in crowd-sensing datasets, which is expected to filter out noises and errors in ST-CS matrix reconstruction.

### 4.6. False data detection and correction algorithm

Fig. 4 shows the flowchart of false data detection and correction in DECO framework. Assume the $k$th dimensional sensory data matrix $S(k)$ is being detected, which exhibits low-rank structure and spatio-temporal stability. We derive the spatial constraint $\mathbb{H}$ based on the other $K - 1$ dimension data, i.e., selecting $M$ dimensional sensory data to estimate $\mathbb{H}$ $(M \le K - 1)$, which has been discussed in Section 4.4.

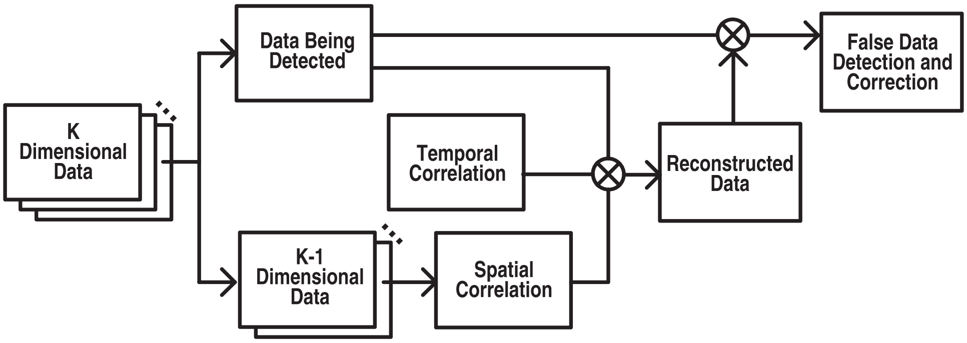


**Fig. 4.** Flowchart of false data detection and correction algorithm.

Next, we reconstruct sensory data matrix by applying the ST-CS method with $\mathbb{H}$ and $\mathbb{T}$ constraints, and obtain $\mathbb{S}(k)$. Then, we check data inconsistencies between $S(k)$ and $\mathbb{S}(k)$, and finally we identify the false sensor readings in $S(k)$.

However, improperly utilizing ST-CS for data reconstruction could lead to low accuracy and high false positives. This is because, neighboring sensor readings normally have mutual influence in ST-CS based data reconstruction, i.e., false data from one participant may have negative influence on the data estimation for his/her neighbors. As a result, good quality sensor readings may be misdeemed as false data. Therefore, conservatively, we need to first identify potentially untrusted participants in the $K$-dimensional sensory dataset, which can be inferred based on their low trust levels in reputation and trust assessment (Christin et al., 2014) or high proportion of outliers in the dataset (Vergara-Laurens et al., 2014). Later on, we will explain DECO also works without knowing untrusted participants. Let $U$ denote the untrusted participant set, where $u$ represents an untrusted participant in $U$ ($\forall\ u \in U$). We assume data contributed by participant $u$ may contain erroneous values (either partial or whole data). We employ DECO to efficiently detect potential false data, and estimate the corresponding values for these untrusted participant in $U$.

**Algorithm 1.** False data detection & correction for $S(k)$.

```
Input:
  $S(k)$, $B(k)$, Untrusted participant set $U$ ($\forall\ u \in U$);
Output:
  $\mathbb{F}(k)$: false data index matrix for $S(k)$;
  $\mathbb{S}(k)$: sensory data matrix with correction;
Procedure:
1: Derive spatio-temporal constraints $\mathbb{H}$ and $\mathbb{T}$;
2: for $\forall\ u$ in $U$ do
3:     $b(u, t, k) \leftarrow 0, \forall\ t \in [1, T], b(u, t, k) \in B(k)$;
       //Mark $u$'s data as missing values in $B(k)$
4: end for
5: Apply ST-CS for matrix reconstruction using the updated $B(k)$,
   i.e., solving
   $\min\{\|B(k)\cdot(\mathcal{L}\mathcal{R}^{*}) - S(k)\|_F^2 + \lambda(\|\mathcal{L}\|_F^2 + \|\mathcal{R}^{*}\|_F^2) + \|\mathbb{H}\mathcal{L}\mathcal{R}^{*}\|_F^2 +;$
   $\|\mathcal{L}\mathcal{R}^{T}\mathbb{T}^{*}\|_F^2\}$
6: $\mathbb{S}(k) \leftarrow \mathcal{L}\mathcal{R}^{*}$; //Obtain the reconstructed matrix
7: $\mathbb{F}(k) \leftarrow (0)_{N\times T}$; //Initialization
8: for $\forall\ u$ in $U$ do
9:     for $\forall\ t = 1$ to $T$ do
10:        if $|s(u, t, k) - \hat{s}(u, t, k)| > \xi_k$ then
11:            $f(u, t, k) = 1$; //$\forall\ \hat{s}(u, t, k) \in \mathbb{S}(k)$
12:        end if
13:    end for
14: end for
15: $\mathbb{F}(k) \leftarrow (f(u, t, k))_{N\times T}$;
16: return $\mathbb{F}(k)$ and $\mathbb{S}(k)$;
end Procedure
```

The proposed false data detection and correction algorithm is described in Algorithm 1, which will be repeated sequentially for each dimension in the $K$-dimensional sensory dataset ($k \in [1, K]$). First, we derive spatio-temporal constraints $\mathbb{H}$ and $\mathbb{T}$ (Line 1). For any untrusted participant $u$, we mark his/her sensor readings as missing values in $B(k)$ (Lines 2–4), to avoid untrusted data misleading the data reconstruction. We then rebuild $\mathbb{S}(k)$ by applying ST-CS matrix reconstruction using the updated $B(k)$ (Lines 5–6). An individual threshold $\xi_k$, a data type specific parameter, is needed for detecting potential false values in the $k$th dimensional data ($k \in [1, K]$). For every participant $u$ at each time slot, a sensor reading is marked as a potential false value if $s(u, t, k)$ in $S(k)$ is notably different compared to the corresponding value $\hat{s}(u, t, k)$ in $\mathbb{S}(k)$ (Lines 8–14). Finally, we obtain the false data index matrix $\mathbb{F}(k)$, in which each nonzero element indicates a possible false value (Line 15). Since data reconstruction fills in the gaps of any missing values in the dataset, the reconstructed $\mathbb{S}(k)$ naturally provides estimated values for those potentially false data (and missing data) in $S(k)$.

The rationale of the false data detection in DECO design is that, by checking data consistency with co-located participants over a reasonably long time period, a misbehaving or erroneous participant has a very small possibility to convince the false data. Actually, DECO can be extended for screening the sensory data matrix $S(k)$ without prior knowledge of the untrusted participant set $U$. For example, we can sequentially check each participant by using the aforementioned method, and figure out potential data inconsistencies. Interestingly, sensory data matrix reconstruction enables quantitative description about the data quality of each participant. The false data index matrix $\mathbb{F}(k)$ provides useful statistical information for reputation management and incentive distribution, which are two important functions that might affect the successful deployment of a crowd-sensing system.

## 5. Performance evaluation

### 5.1. Case study i: crowd-sensing-based wifi fingerprinting

In this section, we present our testbed experiment when applying DECO for data quality improvement in a crowd-sensing-based WiFi fingerprinting system.

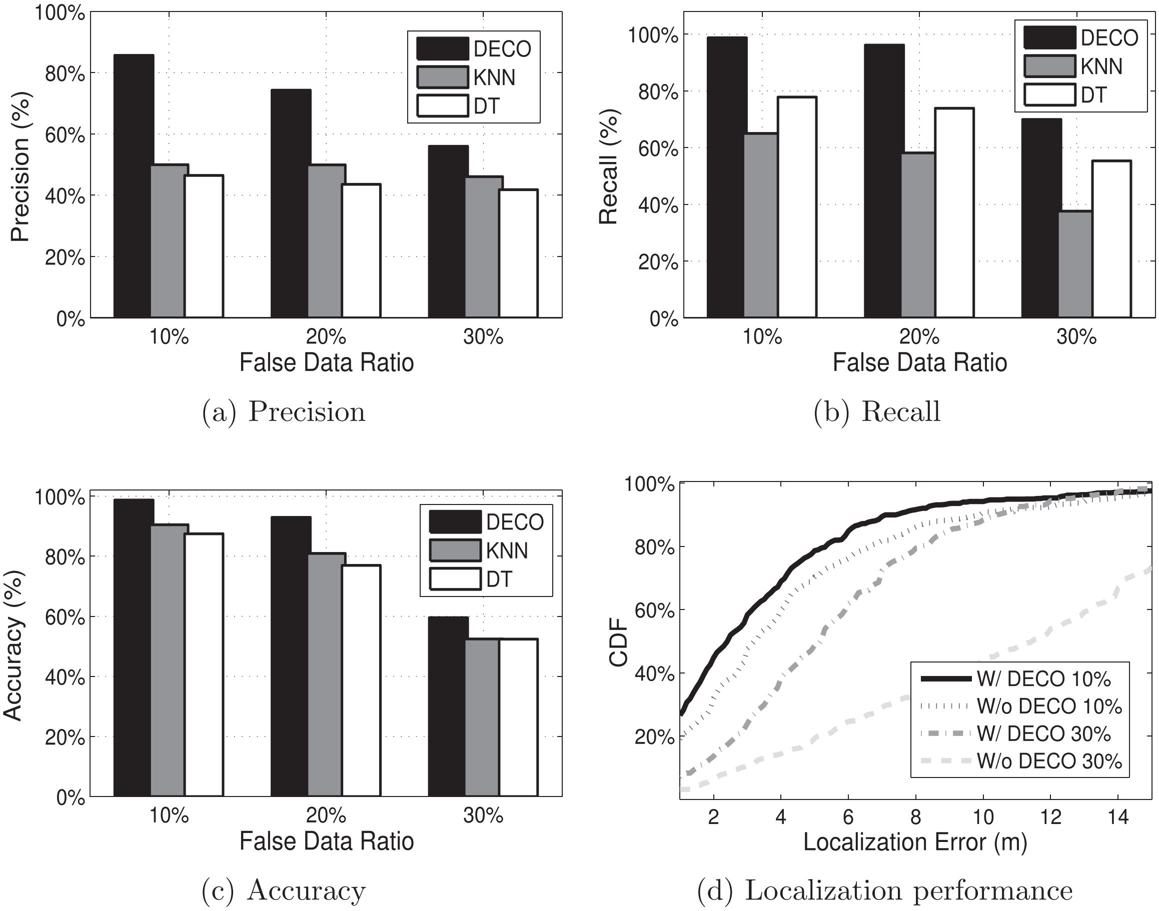
#### 5.1.1. Motivation

WiFi fingerprinting is considered a promising indoor localization approach with rapidly increased deployments of WiFi access points (Jun et al., 2013; Luo et al., 2016a). Typically, it is composed of two phases: an offline training phase and an online localization phase. During the training phase, dedicated site surveyors collect RSS values from multiple WiFi APs at different reference points to construct the radiomaps, which are used for localization in the online phase. However, the widespread use of WiFi fingerprint-based indoor localization is still limited due to the labor-intensive training phase to construct the radiomaps.

The idea of crowd-sensing-based WiFi fingerprinting is to utilize casual users to collect WiFi fingerprints, which enables training data to be crowdsourced without explicit effort of site surveyors (Luo et al., 2014, Luo et al., 2016b). However, radiomap construction with crowd-sensing introduces a new challenge: the fingerprinting system is exposed to malicious and erroneous users, and there is no data quality guarantee of the crowdsourced radiomap. Therefore, efficient data validation method that is able to detect incorrect values and perform possible data quality improvement is essential in crowd-sensing-based indoor localization systems.

#### 5.1.2. Experimental setup

We conduct an experiment using 10 Samsung Galaxy S4 smartphones for crowd-sensing-based WiFi fingerprinting. Before the experiment, all smartphones are synchronized by the ClockSync application. Users equipped with smartphones walk around in a campus building over 120 min totally. Each smartphone running an Android service in the background opportunistically collects WiFi fingerprints, and uploads the collected data to the localization server. To obtain the proximity estimation of smartphones, we also collect other sensory data available in the phones, including bluetooth neighbor scans, temperature, humidity and sound level measurements. We apply the method introduced in Section 4.4 to infer smartphones' proximity based on other multidimensional sensor readings, and derive the spatial constraint $\mathbb{H}$. The weight coefficient settings in spatial adjacency



(a) Precision

(b) Recall

(c) Accuracy

(d) Localization performance

**Fig. 5.** Impact of false data ratio.

estimation are: $\omega_{bluetooth} = 0.7$, and non-spatial attributes equally share the remaining 0.3. To obtain the ground truth to construct the WiFi fingerprint radiomap, we ask participants to manually tap their locations whenever they are passing the predefined reference locations. An average of 35 WiFi APs could be detected in our experiment. At the server side, we implement the proposed DECO system in Matlab.

*Threat Model*: In this experiment, we consider that any participant may act maliciously and may upload fake WiFi fingerprints to the system. We also consider the cases that participants stop providing data due to the lack of interest or motivation in the data collection campaign or some inexperienced participants fail to upload the collected data.

In order to emulate the potential deterioration of data quality in the system, we set two adjustable parameters $R_m$ and $R_f$ to control the amount of missing and false data during the data collection. Users could set these two parameters through the data collection software interface. In the bootstrap phase, a smartphone sets itself as an untrusted participant with a probability of $R_f$, and then sets the scanned RSS with random false values ranging from −100 db m to −30 db m at each time slot (since in our experiment, the collected RSS values normally range from −100 db m to −30 db m). For trusted participants, at each time slot, the Android data collection service sets the RSS values to be missing value *Nil* with a probability of $R_m$.

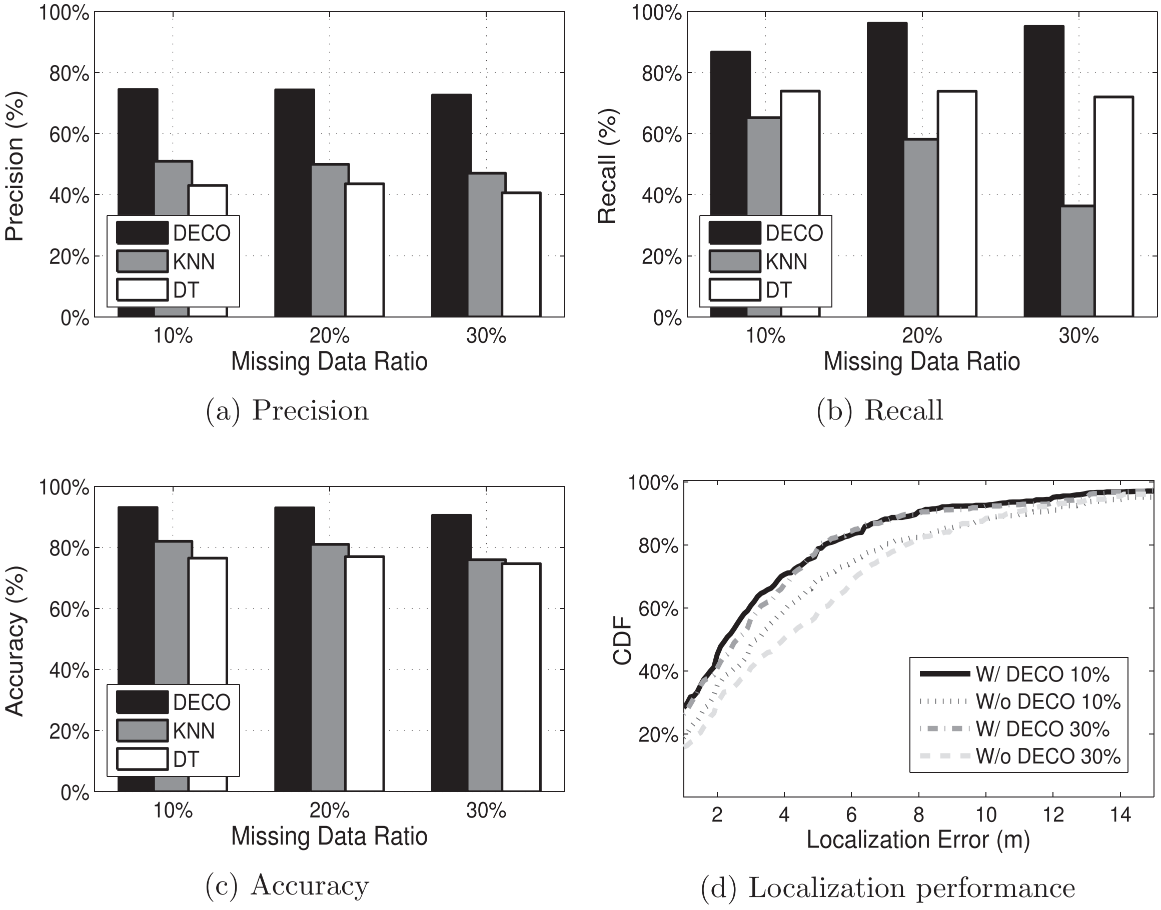
We compare the performance of DECO against widely known data interpolation methods: K-Nearest Neighbors (KNN) (Cover and Hart, 1967) and Delaunay Triangulation (DT) (Kong et al., 2010; Vergara-Laurens et al., 2014). We measure the detection precision, recall and accuracy using different data interpolation methods. These metrics are defined as follows, Precision=TP/(TP+FP), Recall=TP/(TP+FN) and Accuracy=(TP+TN)/(TP+TN+FP+FN), where TP, FP, TN, FN are true positive, false positive, true negative and false negative respectively. In this experiment, if the difference between the estimated RSS values and the original values are larger than averagely 20% of the original values (i.e., the threshold $\xi_k$ in Algorithm 1), this RSS scan is marked as a false WiFi fingerprint. To measure the effectiveness of data correction, we apply the basic location determination method in WiFi fingerprinting (Bahl and Padmanabhan, 2000) and compare the final localization error with/without DECO's data correction of the radiomap. In Bahl and Padmanabhan (2000), the location is determined by averaging over 3 nearest neighbors based on RSS values in the radiomap.

### 5.1.3. Impact of false data ratio

We first investigate the impact of false data ratio on DECO's performance and the final localization accuracy with fixed ratio of missing values. In this experiment, we set $R_m = 20\%$, and vary the false data ratio $R_f$ from 10% to 30%.

Fig. 5 shows the false data detection performance. As the false data ratio increases, the results of precision and recall ratio in Fig. 5a and b drop accordingly, resulting the degradation of final false data detection accuracy as shown in Fig. 5(c). However, as DECO utilizes the spatio-temporal constraints and takes the whole data structure into consideration (Candes and Tao, 2006), it outperforms the local interpolation methods KNN and DT, and achieves higher false data detection accuracy in all cases.

Fig. 5(d) shows the final localization error with and without DECO's false data correction. The false data introduced by malicious users significantly degrades the localization performance. As the false data ratio increases from 10–30%, the average localization error is increased from 4.6 to 10.8 m. After performing data correction with DECO, the localization error is significantly reduced to 3.2 m and 5.3 m, respectively. DECO efficiently detects and corrects false fingerprints and generates high-quality radiomaps, which improves localization accu-



(a) Precision

(b) Recall

(c) Accuracy

(d) Localization performance

**Fig. 6.** Impact of missing data ratio.

racy in our crowd-sensing-based WiFi fingerprinting system.

### 5.1.4. *Impact of missing data ratio*

Next, we study the impact of missing data ratio on the system performance given fixed percentage of false values. We set $R_f = 20\%$ and vary the missing data percentage $R_m$ from% 10 to 30%.

The amount of missing data affects the performance of false data detection and correction. As the data become more sparse, the data reconstruction becomes less accurate. Fig. 6 shows the evaluation results in this scenario. As the missing data ratio increases from 10 to 30%, DECO remains a high detection accuracy while the performance of KNN and DT drops more significantly. This shows that DECO remains robust even with a large portion of missing data.

The final corrected fingerprints improve the localization performance. As shown in Fig. 6(d), DECO reduces the localization from 3.8 to 3.1 m with 10% missing data, and from 5.2 to 3.2 m with 30% missing data, which shows that DECO is able to efficiently detect and correct false fingerprints that introduced by crowd-sensing participants.

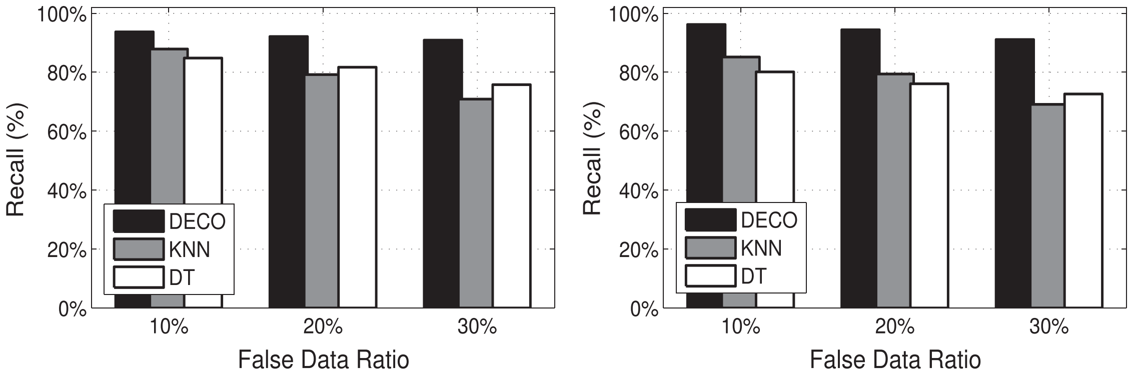
## 5.2. *Case study ii: crowd-sensing for environment monitoring*

To evaluate the potential of DECO for crowd-sensing with a large number of data contributors, in this section, we compare the performance of DECO against KNN and DT based on a synthetic dataset in the context of crowd-sensing based environment monitoring application scenario.

### 5.2.1. *Datasets*

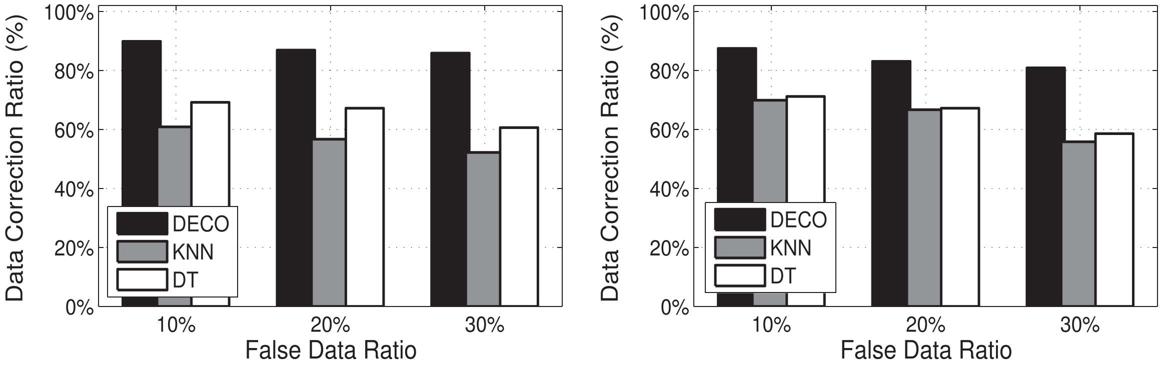
Since there is no public dataset specially for crowd-sensing (Kurasawa et al., 2014), we generate a synthetic dataset based on the frequently used GreenOrbs (Mo et al., 2009) dataset in our evaluation. In GreenOrbs, 450 static sensor nodes were deployed to gather temperature, light, and humidity measurements once every 10 min. We generate a scenario where sensor nodes (either smartphone built-in sensors or external sensors that can be connected to a smartphone) are freely moved in a public area as follows: at each time slot (every 10 min in GreenOrbs dataset), a sensor node either moves to another position or remains at the same place. If moving, we randomly choose one of its one-hop neighbors as the node's substitute. That is, the sensor node will move to the substitute's position and obtain sensor readings same as the substitute's at the next time slot. This gives us up to 450 participants for data collection over an extended period. The low-rank property and spatio-temporal stability in GreenOrbs dataset have been shown in Kong et al. (2013), which means ST-CS can be applied in this dataset.

We use the smartphone proximity information in Section 5.1 to emulate the practical proximity estimation in crowd-sensing. For each pair of ground truth co-located smartphones $i$ and $j$ ($i$ and $j$ encounter each other), we extract the spatially-dependent information at the encounter time. After that, we obtain a collection of spatially-dependent sensory data from the smartphone sensing. Then, we randomly assign one measurement (i.e., WiFi and Bluetooth signatures) to each pair of ground truth co-located nodes in the synthetic dataset. Together with the non-spatial measurements in GreenOrbs, we apply the proposed proximity estimation method and finally obtain the estimated spatial constraints at each time slot. We observed that WiFi and Bluetooth signatures dominate the spatial adjacency estimation. The weight coefficient settings in spatial adjacency estimation are: $\omega_{wifi} = 0.4$, $\omega_{bluetooth} = 0.4$, and non-spatial attributes (i.e., any two



(a) Temperature data (b) Light data

**Fig. 7.** Recall rate.



(a) Temperature data (b) Light data

**Fig. 8.** False data correction.

attributes among temperature, light and humidity) share the remaining 0.2. For KNN, the parameter $K$ is set as the number of one-hop neighbors in GreenOrbs dataset.

*5.2.2. Experimental setup*

DECO framework is designed to accurately detect and correct false data for incomplete sensory datasets with considerable missing values. To verify the effectiveness of DECO, we randomly remove 20% sensor readings from the dataset, thus generating an incomplete sensory dataset. Then, we select 10–30% nodes as the untrusted participants, and randomly inject false data into the data reports of these untrusted participants. The injected false data values are randomly selected within the range of the ground truth sensor readings.

The experimental evaluation focuses on two metrics. 1) Recall rate: the ratio of false sensor readings discovered by detection methods to the total number of false sensor readings. 2) Data correction ratio: the ratio of the correct estimation in value to the total number of false data. In this experiment, if the difference between the estimated value and the original value (after injecting false data in the dataset) is larger than 20% of the original value, a sensor reading is marked as a false value. When calculating the data correction ratio, if the difference between the estimated value and ground truth is within 10% of the ground truth value, it is considered a correct estimation.

*5.2.3. Evaluation results*

Fig. 7 depicts the recall rate results for temperature and light false data detection. From Fig. 7(a), DECO achieves 93% recall rate on average. As the false data percentage increases from 10 to 30%, detection rates for all methods also decrease. However, both KNN and DT show poor performance when the false data percentage becomes large. Since KNN is a local interpolation method, which simply averages values of the $\mathcal{K}$ one-hop neighbors to estimate the missing value, KNN's performance drops quickly when the false data percentage is larger. While even at high false data percentage, DECO provides above 90% detection rate. Fig. 7(b) shows the same trend as observed in Fig. 7(a).

In addition to detecting the false data, DECO is able to estimate both false and missing data at the same time. Fig. 8 reports the correction rate when comparing DECO with KNN and DT. As the false data percentage increases, all schemes experience decreased data correction rate. Overall, DECO shows obvious advantages over the other methods. This is because, KNN and DT interpolate data based on only the spacial correlation among neighboring nodes without considering the temporal correlation. Even at high false data percentage, the data recovery ratio in DECO is less than 20% for both temperature and light data.

## 6. Conclusion

In this work, we presented DECO, a false data detection and correction framework tailored for crowd-sensing with missing data. We showed there exists inherent low-rank features and spatio-temporal correlations in real-world sensory data. Then we developed a false data detection and correction algorithm by applying the spatio-temporal compressive sensing technique. Through comprehensive performance comparisons, we demonstrated that DECO is well suited for data quality improvement in crowd-sensing with considerable missing data. The proposed algorithm effectively identifies false data and outperforms the state-of-the-art methods in false data correction.

## Acknowledgements

This work was supported in part by the 973 Program (2013CB035503), National Natural Science Foundation of China (61190125, 61300174, 61303202, 61572060, 61602319, 61672349), China Postdoctoral Science Foundation (2013M530511, 2014T70026), and Open Foundation of State Key Laboratory of Networking and Switching Technology (Beijing University of Posts and Telecommunications, SKLNST-2013-1-02) and CERNET Innovation Project 2015 (NGII20151004).

## References

Ahmadi, Hossein, Pham, Nam, Ganti, Raghu, Abdelzaher, Tarek, Nath, Suman, Han, Jiawei, 2010. Privacy-aware regression modeling of participatory sensing data. In: SenSys'10. pp. 99–112.

Ahn, Gahng-Seop, Musolesi, Mirco, Lu, Hong, Olfati-Saber, Reza, Campbell, Andrew T., 2010. Metrotrack: predictive tracking of mobile events using mobile phones. In: DCOSS'10, pp. 230–243, 2010.

Amintoosi, Haleh, Kanhere, Salil S., 2013. A reputation framework for social participatory sensing system. Mob. Netw. Appl., 1–13.

Bahl, P., Padmanabhan, V.N., 2000. RADAR: an in-building rf-based user location and tracking system. In: INFOCOM'00.

Boutsis, I., Kalogeraki, V., 2013. Privacy preservation for participatory sensing data. In: PerCom'13. pp. 103–113.

Candes, E.J., Tao, T., 2006. Near-optimal signal recovery from random projection: universal encoding strategies? IEEE Trans. Inf. Theory 52 (12), 5406–5425.

Cheng, Long, Kong, Linghe, Luo, Chengwen, Niu, Jianwei, Gu, Yu, He, Wenbo, Das, Sajal K., 2015. False data detection and correction framework for participatory sensing. In: IWQoS'15.

Christin, Delphine, Reinhardt, Andreas, Kanhere, Salil S., Hollick, Matthias, 2011. A survey on privacy in mobile participatory sensing application. J. Syst. Softw. 84 (11), 1928–1946.

Christin, D., Pons-Sorolla, D. Rodriguez , Hollick, M., Kanhere, S.S., 2014. Trustmeter: A trust assessment scheme for collaborative privacy mechanisms in participatory sensing applications. In: ISSNIP'14.

Cover, T., Hart, P., 1967. Nearest neighbor pattern classification. IEEE Trans. Inf. Theory 31 (1), 21–27.

Das, Tathagata, Mohan, Prashanth, Padmanabhan, Venkata N., Ramjee, Ramachandran, Sharma, Asankhaya, 2010. Prism: platform for remote sensing using smartphones. In: Proceedings MobiSys'10. pp. 63–76.

De Cristofaro, E., Soriente, C., 2013. Participatory privacy: enabling privacy in participatory sensin. IEEE Netw. 27 (1), 32–36.

Deng, Linda, Cox, Landon P., 2009. Livecompare: grocery bargain hunting through participatory sensing. In: HotMobile'09.

Dua, Akshay, Bulusu, Nirupama, Feng, Wu-Chang, Hu, Wen, 2009. Towards trustworthy participatory sensing. In: HotSec'09.

Dutta, Prabal, Aoki, Paul M., Kumar, Neil, Mainwaring, Alan, Myers, Chris, Willett, Wesley, Woodruff, Allison, 2009. Common sense: participatory urban sensing using a network of handheld air quality monitors. In: SenSys'09. pp. 349–350.

Groat, M.M., Edwards, B., Horey, J., He, Wenbo, Forrest, S., 2012. Enhancing privacy in participatory sensing applications with multidimensional data. In: PerCom'12. pp. 144–152.

Grosky, W.I., Kansal, A., Nath, S., Liu, Jie, Zhao, Feng, 2007. Senseweb: an infrastructure for shared sensing. IEEE Multimed. 14 (4), 8–13.

Huang, Kuan Lun, Kanhere, Salil S., Hu, Wen, 2010. Are you contributing trustworthy data? The case for a reputation system in participatory sensing. In: Proceedings MSWIM'10. pp. 14–22,.

Hull, Bret, Bychkovsky, Vladimir, Zhang, Yang, Chen, Kevin, Goraczko, Michel, Miu, Allen, Shih, Eugene, Balakrishnan, Hari, Madden, Samuel, 2006. Cartel: a distributed mobile sensor computing system. In: SenSys'06. pp. 125–138.

Huo, Zhiqiang, Shu, Lei, Zhou, Zhangbing, Chen, Yuanfang, Li, Kailiang, Zeng, Junlin, 2015. Data collection middleware for crowdsourcing-based industrial sensing intelligence. In: Proceedings of the ACM International Workshop on Mobility and MiddleWare Management in HetNets.

Jun, Junghyun, Gu, Yu, Cheng, Long, Lu, Banghui, Sun, Jun, Zhu, Ting, Niu, Jianwei, 2013. Social-loc: Improving indoor localization with social sensing. In: SenSys'13. pp. 1–14,.

Kanjo, Eiman, 2010. Noisespy: a real-time mobile phone platform for urban noise monitoring and mapping. Mob. Netw. Appl. 15 (4), 562–574.

Kong, Linghe, He, Liang, Liu, Xiao-Yang, Gu, Yu, Wu, Min-You, Liu, Xue, 2015. Privacy-preserving compressive sensing for crowdsensing based trajectory recovery. In: ICDCS'15.

Kong, Linghe, Jiang, Dawei, Wu, Min-You, 2010. Optimizing the spatio-temporal distribution of cyber-physical systems for environment abstraction. In: ICDCS. pp. 179–188.

Kong, Linghe, Xia, Mingyuan, Liu, Xiao-Yang, Wu, Min-You, Liu, Xue, 2013. Data loss and reconstruction in sensor networks. In: INFOCOM'13. pp. 1654–1662.

Krumm, John, Hinckley, Ken, 2004. The nearme wireless proximity server. In: UbiComp'04. pp. 283–300.

Kurasawa, H., Sato, H., Yamamoto, A., Kawasaki, H., Nakamura, M., Fujii, Y., Matsumura, H., 2014. Missing sensor value estimation method for participatory sensing environment. In: PerCom'14. pp. 103–111.

Kuznetsov, Stacey, Paulos, Eric, 2014. Participatory sensing in public spaces: Activating urban surfaces with sensor probes. In: DIS'10. pp. 21–30.

Lee, Juong-Sik, Hoh, Baik, 2010. Dynamic pricing incentive for participatory sensin. Perv. Mob. Comput. 6 (6), 693–708.

Lin, Mu, Lane, Nicholas D., Mohammod, Mashfiqui, Yang, Xiaochao, Lu, Hong, Cardone, Giuseppe, Ali, Shahid, Doryab, Afsaneh, Berke, Ethan, Campbell, Andrew T., Choudhury, Tanzeem, 2012. Bewell+: multi-dimensional wellbeing monitoring with community-guided user feedback and energy optimization. In: Proceedings of the Conference on Wireless Health. pp. 1–8.

Luo, Chengwen, Hong, Hande, Cheng, Long, Chan, Mun Choon, Li, Jianqiang, Ming, Zhong, 2016a. Accuracy-aware wireless indoor localization: feasibility and applications. J. Netw. Comput. Appl. 62, 128–136.

Chengwen Luo, Long Cheng, Mun Choon Chan, Yu Gu, Jianqiang Li, Zhong Ming.Pallas: Self-bootstrapping Fine-grained Passive Indoor Localization Using WiFi Monitors. In IEEE Transactions on Mobile Computing, 2016b

Luo, C., Hong, H., Cheng, L., Sankaran, K., Chan, M.C., 2015. imap: automatic inference of indoor semantics exploiting opportunistic smartphone sensing. In: Sensing, Communication, and Networking (SECON), 2015 Proceedings of the 12th Annual IEEE International Conference on. pp. 489–497.

Luo, Chengwen, Hong, Hande, Chan, Mun Choon, 2014. Piloc: a self-calibrating participatory indoor localization system. In: IPSN'12. pp. 143–154.

Luo, Tie, Kanhere, Salil, Das, Sajal, Tan, Hwee-Pink, 2014. Optimal prizes for all-pay contests in heterogeneous crowdsourcing. In: MASS'14.

Luo, Tie, Tham, Chen-Khong, 2012. Fairness and social welfare in incentivizing participatory sensing. In: SECON'12. pp. 425–433,.

Min, Hong, Scheuermann, Peter, Heo, Junyoung, 2013. A hybrid approach for improving the data quality of mobile phone sensing. Int. J. Distrib. Sens. Netw..

Mo, Lufeng, He, Yuan, Liu, Yunhao, Zhao, Jizhong, Tang, Shao-Jie, Li, Xiang-Yang, Dai, Guojun, 2009. Canopy closure estimates with greenorbs: sustainable sensing in the forest. In: SenSys'09. pp. 99–112.

Muntés-Mulero, Victor, Paladini, Patricia, Manzoor, Jawad, Gritti, Andrea, Larriba-Pey, Josep-Lluís, Mijnhardt, Frederik, 2013. Crowdsourcing for industrial problems. In: Citizen in Sensor Networks. Springer. pp. 6–18.

Pham, Nam, Ganti, Raghu K., Uddin, Yusuf S., Nath, Suman, Abdelzaher, Tarek, 2010. Privacy-preserving reconstruction of multidimensional data maps in vehicular participatory sensing. In: Proceedings EWSN'10. pp. 114–130.

Rai, Anshul, Chintalapudi, Krishna Kant, Padmanabhan, Venkata N., Sen, Rijurekha, 2012. Zee: zero-effort crowdsourcing for indoor localization. In: MobiCom'12.

Rallapalli, Swati, Qiu, Lili, Zhang, Yin, Chen, Yi-Chao, 2010. Exploiting temporal stability and low-rank structure for localization in mobile networks. In: MobiCom'10. pp. 161–172.

Reddy, Sasank, Burke, Jeff, Estrin, Deborah, Hansen, Mark, Srivastava, Mani. A framework for data quality and feedback in participatory sensing. In: SenSys'07. pp. 417–418.

Reddy, Sasank, Estrin, Deborah, Srivastava, Mani, 2010. Recruitment framework for participatory sensing data collections. In: Pervasive'10.

Restuccia, Francesco, Das, Sajal, 2010. Fides: A trust-based framework for secure user incentivization in participatory sensing. In: WoWMoM'14.

Roughan, M., Zhang, Yin, Willinger, W., Qiu, Lili, 2012. Spatio-temporal compressive sensing and internet traffic matrices. IEEE/ACM Trans. Netw. 20 (3), 662–676.

Saroiu, Stefan, Wolman, Alec, 2010. I am a sensor, and i approve this message. In: HotMobile'10. pp. 37–42.

Song, Z., Liu, C., Wu, J., Ma, J., Wang, W., 2014. Qoi-aware multi-task-oriented dynamic participant selection with budget constraint. IEEE Trans. Veh. Technol..

Talasila, M., Curtmola, R., Borcea, C., 2013. Improving location reliability in crowd sensed data with minimal efforts. In: Proceedings WMNC'13.

Vergara-Laurens, IJ., Mendez, D., Labrador, M.A., 2014. Privacy, quality of information, and energy consumption in participatory sensing systems. In: PerCom'14. pp. 199–207.

Wang, He, Sen, Souvik, Elgohary, Ahmed, Farid, Moustafa, Youssef, Moustafa, Choudhury Romit Roy, 2012. No need to war-drive: unsupervised indoor localization. In: MobiSys'12. pp. 197–210.

Wang, Xinlei Oscar, Cheng, Wei, Mohapatra, Prasant, Abdelzaher, Tarek, 2013. Artsense: anonymous reputation and trust in participatory sensing. In: INFOCOM'13. pp. 2517–2525.

Yang, Zheng, Wu, Chenshu, Liu, Yunhao, 2012. Locating in fingerprint space: wireless indoor localization with little human intervention. In: Mobicom'12. pp. 269–280.

Yow, W., Quin, Li, Xiaoqian, Hung, Wan-Yu, Goldring, Megan, Cheng, Long, Gu, Yu, 2014. Predicting social networks and psychological outcomes through mobile phone sensing. In: ICC'14. pp. 3925–3931.

**Long Cheng** is a postdoctoral researcher at Beihang University. He got his Ph.D. in the State Key Lab of Network and Switching Technology, Beijing University of Posts and Telecommunications, China, 2012. His research interests include network security and forensics, wireless sensor networks, cyber-physical systems, mobile and pervasive computing.

**Jianwei Niu** received his M.S. and Ph.D. degrees in 1998 and 2002 in computer science from Beijing University of Aeronautics and Astronautics (BUAA, now renamed as Beihang University). He was a visiting scholar at School of Computer Science, Carnegie Mellon University, USA from Jan 2010 to Feb 2011. He is a pro- fessor in the School of Computer Science and Engineering, BUAA. He is an IEEE member and has published more than 100 referred papers and filed more than 20 patents. He has been on various chairs and TPC members for many international conferences. He served as the Program Chair of IEEE SEC 2008. He received New Century Excellent Researcher Award from Ministry of Education of China in 2009. His current research interests include mobile and pervasive computing.

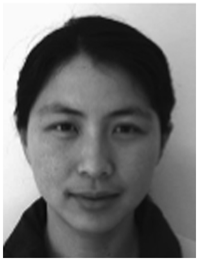
**Linghe Kong** is currently an associate professor with the Department of Computer Science and Engineering at Shanghai Jiao Tong University.

From 2013 to 2016, he was a Postdoctoral Fellow at Columbia University, McGill University, and Singapore University of Technology and Design. He received his Ph.D. degree in Computer Science from Shanghai Jiao Tong University 2012, Master degree in Telecommunication from TELECOM

SudParis 2007, and B. Eng. degree in Automation from Xidian University 2005. His research interests include wireless communications, sensor networks, and mobile computing.

**Chengwen Luo** currently is an assistant professor in College of Computer Science and Software Engineering, Shenzhen University, China. Before joining SZU, he was a postdoctoral re- searcher in CSE, The University of New South Wales (UNSW), Australia. He received his Ph.D. degree from School of Computing, National University of Singapore (NUS), Singapore. Dr. Luo is the author and co-author of several research papers in top venues of mobile computing and WSN such as ACM SenSys, ACM/IEEE IPSN, etc. His research interests include mobile and pervasive computing, indoor localization, wireless sensor networks, and security aspects of Internet of Things

**Yu Gu** currently is a research scientist at Watson Health Cloud, IBM Watson Health, USA. He recieved the PhD degree from the University of Minnesota, Twin Cities in 2010. Dr. Gu is the author and co-author of over 100 papers in premier journals and conferences. His publications have been selected as graduate-level course materials by over 20 universities in the United States and other countries. His research includes Networked Embedded Systems, Wireless Sensor Networks, Cyber-Physical Systems, Wireless Networking, Real-time and Embedded Systems, Distributed Systems, Vehicular Ad-Hoc Networks and Stream Computing Systems. Dr. Gu is a member of ACM and IEEE.



**Wenbo He** received the Ph.D. degree from University of Illinois at Urbana-Champaign, Champaign, IL, USA, in 2008. He is currently an Assistant Professor in School of Computer Science at McGill University. She was an Assistant Professor in Department of Electrical Engineering at University of Nebraska-Lincoln from 2010 to 2011. She was with the Department of Computer Science at University of New Mexico from 2008 to 2010. Her research focuses on Pervasive Computing, and Privacy-preserving Techniques, etc. During August 2000 to January 2005, she was a software engineer in Cisco Systems, Inc. Dr. He received the Mavis Memorial Fund Scholarship Award from College of Engineering at UIUC in 2006, and the C.W. Gear Outstanding Graduate Award in 2007 from the Department of Computer Science at UIUC. She is also a recipient of the Vodafone Fellowship from 2005 to 2008, and the NSF TRUST Fellowship in 2007 and 2009. She is a member of the IEEE.

**Sajal K. Das** is currently the head and chair professor in Computer Science Department at Missouri University of Science and Technology, and the founding director of the Center for Research in Wireless Mobility and Networking (CReW-MaN). He was a university distinguished scholar professor of computer science and engineering at the University of Texas at Arlington. During 2008–2011, he served the US National Science Foundation (NSF) as a program director in the Division of Computer Networks and Systems. His current research interests include wireless and sensor networks, mobile and pervasive computing, cloud computing, cyberphysical systems and smart heath care, security and privacy, biological and social networks, applied graph theory and game theory. He has published extensively in these areas and made fundamental contributions. He holds five US patents and coauthored books on Smart Environments: Technology, Protocols, and Applications (Wiley, 2005), Handbook on Securing Cyber-Physical Critical Infrastructure: Foundations and Challenges (Morgan Kaufmann, 2012), and Mobile Agents in Distributed Computing and Networking (Wiley, 2012). He is a recipient of the IEEE Computer Society Technical Achievement Award for pioneering contributions to sensor networks and mobile computing; IEEE Region 5 Outstanding Engineering Educator Award; and Eight Best Paper Awards including those at IEEE SmartGridComm'12, QShine'09, EWSN'08, IEEE PerCom'06, and ACM MobiCom'99. He serves as the founding editor-in-chief of Pervasive and Mobile Computing journal, and also as an associate editor ACM/Springer Wireless Networks, Journal of Parallel and Distributed Computing, and Journal of Peer-to-Peer Networking. He is the cofounder of IEEE PerCom, WoWMoM and ICDCN conferences. He is a senior member of the IEEE.