DECO: False Data Detection and Correction Framework for Participatory Sensing

Long Cheng*†, Linghe Kong‡, Chengwen Luo§, Jianwei Niu*, Yu Gu¶, Wenbo He‡ and Sajal Das‖
*State Key Lab of Virtual Reality Technology and Systems, Beihang University, Beijing 100191, China
†State Key Lab of Networking & Switching Tech., Beijing Univ. of Posts & Telecomm., Beijing, China
‡McGill University, Canada §National University of Singapore, Singapore
¶IBM Research Austin, USA ‖Missouri University of Science and Technology, USA

***Abstract*—Participatory 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 participatory sensing exposes the system to malicious and erroneous participations, inevitably resulting in poor data quality. This brings forth the important issues of false data detection and correction in participatory 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 participatory 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. We validate our design through an experimental case study.**

I. INTRODUCTION

Participatory sensing is to leverage 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 [1]. Many participatory sensing applications have emerged in recent years, including environment, transportation and civil infrastructure monitoring [2], health and fitness monitoring [3], context sensing [4] and radiomap construction in WiFi fingerprinting [5], [6]. The inherent openness of participatory 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 participants are not always reliable as they can submit fake data to earn rewards without performing the actual sensing task [7]. 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 [8]. In addition, attackers may compromise the mobile devices to provide faulty sensor readings [9]. 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, locations, as the crucial context for participatory sensing, are often inaccurately estimated in real-world systems [5]. Therefore, the same openness characteristic of participatory 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 participatory sensing systems.

978-1-4673-7113-1/15/$31.00 2015 IEEE

Techniques have been developed to achieve data integrity and correctness [10]–[12]. However, no system has been presented as a general approach to detect and correct false data for participatory sensing. More recently, Kurasawa et al. [12] pointed out that data collected by participants usually include considerable missing values in practical participatory sensing systems. The incompleteness of sensory data poses several challenging issues for accurate false data detection. 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 participatory 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 participatory sensing. To address this challenge, we propose to exploit spatio-temporal compressive sensing (ST-CS) technique, which can achieve an effective data reconstruction for high data-loss scenarios.
- Considering that the spatial proximity of participants cannot be directly derived from the potentially inaccurate locations reported in participatory 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 participatory sensing. To the best of our knowledge, there are few other efforts applying ST-CS techniques for false data correction in participatory sensing environment.

The rest of this paper is organized as follows. Section II describes the preliminaries. Section III elaborates the design of DECO framework in details. Section IV provides evaluation results by applying DECO in participatory sensing-based WiFi fingerprinting. Finally, conclusions are drawn in Section V.

II. PRELIMINARIES

A. System Model

We consider a typical client-server participatory sensing architecture, where a larger number of mobile devices are tasked into community-based data gathering. The sensory data collected by participants are uploaded (*e.g.*, through WiFi or cellular networks) to a central application server. A sensing task normally specifies multiple modalities of sensory data to be collected based on the application requirements [13], and an individual data collection device may be involved into multiple concurrent sensing tasks [14], [15]. In this work, we consider that the collected data in a participatory sensing system are *multi-dimensional time-series sensor readings* [16], [17].

B. Data Representation

Assume a participatory sensing system with N participants, multidimensional time-series data are generated by each participant and then reported to the centralized server. Let us also 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 participatory sensing system. For 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.

C. False Data in Participatory Sensing

Only in an ideal environment, participants provide accurate and complete sensor readings [12], [18]. Unfortunately, on one hand, there exists inexperienced and malicious participants, which may provide corrupted sensory data to participatory 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 participatory 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.

D. Spatio-Temporal Compressive Sensing

Compressive sensing (CS) [19] 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 [20], wireless sensor networks [21] and trajectory data [22]. 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 is able to accurately reconstruct missing values in real-world datasets.

III. DECO FRAMEWORK DESIGN

A. Overview

The DECO framework is illustrated in Fig. 1. In typical participatory 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 can be deployed as an enhancement layer for false data detection and correction in various participatory sensing systems.

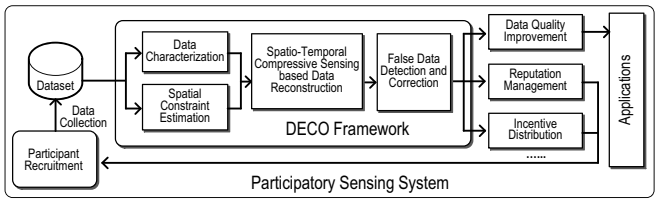


Fig. 1. DECO framework for participatory sensing

In Fig. 1, 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 [20] 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 [21] 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.

DECO not only improves the data quality, but also provides useful information for various application-layer modules such as reputation management and incentive distribution. Essentially, DECO improves the quality of service provided by a participatory sensing system to the end users.

B. 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 sensory datasets, using the low-rank structure, temporal stability, and spatial stability metrics. Due to space limit, detailed definitions of these three metrics are referred to [20], [21]. 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.

C. ST-CS for Sensory Matrix Reconstruction

In the case where participatory 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 [20] 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:

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

Through theoretical derivations, the ST-CS matrix reconstruction problem is formulated as the following optimization problem:

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

where λ 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 $||\mathbb{H}\mathscr{L}\mathscr{R}^*||_F^2$, $||\mathscr{L}\mathscr{R}^*\mathbb{T}^*||_F^2$, and $||B(k)\cdot(\mathscr{L}\mathscr{R}^*) - S(k)||_F^2$ need to be set equal in the similar order of magnitude, otherwise, they may overshadow the others during optimization [20]. Then, by tuning λ, $\mathscr{L}$ and $\mathscr{R}$ can be estimated in this optimization problem, and $\mathbb{S}(k)$ is consequently estimated.

D. Deriving Spatial Constraint

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 $H(t)$ at a particular time slot t,

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

where $i, j \in [1, N]$. i and j are neighbors if their distance is less than a threshold d. Both rows and columns in $H(t)$ represent participants, and $h(i,j,t)$ represents whether participants i and j are neighbors or not at time slot t.

Participatory sensing produces inaccurate and uncertain sensory data as well as missing values. It poses challenges in accurately estimating spatial adjacency matrix $H(t)$. For example, GPS traces are likely to be obfuscated for privacy-preserving on the participant-side prior to sharing them [16]. Malicious adversaries may deliberately upload forged location data [7]. Therefore, we cannot directly derive $H(t)$ from the location information in participatory sensing datasets. While applying ST-CS for sensory data reconstruction requires a good approximated $H(t)$.

To address the above challenge, we propose to infer participants' proximity based on multidimensional sensor readings in participatory sensing systems. We classify the sensory data into 1) *spatially-dependent* (*e.g.*, location of the samples, WiFi AP signatures, and Bluetooth signatures) and 2) *non-spatial* (*i.e.*, environmental variables) information/attributes. If values of spatially-dependent variables are similar, it is more likely that the two participants are nearby each other. 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 $H(t)$. 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 $H(t)$. We define a general proximity function to estimate the adjacency of participants i and j as follow, which is independent of specific applications:

$$h(i,j,t) = \min\{0, \quad \sum\nolimits_{k=1}^{M} \omega_k \cdot Similarity(i,j,t,k)\},$$

where ω_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, $h(i,j,t)$ is set to 0.

Specifically, we use Pearson's correlation coefficient [23] (giving a value between 0 and 1) to measure the similarity for spatially-dependent sensory data, such as WiFi AP signatures and Bluetooth signatures.

We then exploit the non-spatial attributes (which are usually scalars such as temperature and noise level) to calibrate the estimation of $h(i,j,t)$. 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):

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

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 θ is an adjustable parameter specifying the time window length ($\theta \geq 0$). 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 $H(t)$. 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 $h(i,j,t)$, which ranges from 0 to 1. Then, the spatial adjacency matrix $H(t)(\forall t \in [1,T])$ is transformed to the overall spatial constraint $\mathbb{H}$ as follows.

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

where $\hbar(i,j,t)$ is an element in $N \times N \times$ T matrix $\mathbb{H}$, and the sum of elements in each row is 0. Finally, we can derive $\mathbb{H}$, which is applied as the spatial constraint into the ST-CS matrix reconstruction.

E. 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. According to [24], we set $\mathbb{T} = Toeplitz(0,1,-2,1)_{T\times T}$, 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.*,

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

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

F. False Data Detection and Correction Algorithm

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. Subsequently, 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 [25] or high proportion of outliers in their reported data [18]. Let U denote the untrusted participant set, where u represents an untrusted participant in U ($\forall u \in U$). We employ DECO to efficiently detect potential false data, and estimate the corresponding values for these untrusted participant in U.

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

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 $\min\{||B(k) \cdot (\mathscr{L}\mathscr{R}^*) - S(k)||_F^2 + \lambda(||\mathscr{L}||_F^2 + ||\mathscr{R}^*||_F^2) + ||\mathbb{H}\mathscr{L}\mathscr{R}^*||_F^2 + ||\mathscr{L}\mathscr{R}^T\mathbb{T}^*||_F^2\}$;
6: $\mathbb{S}(k) \leftarrow \mathscr{L}\mathscr{R}^*$; //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

using the updated $B(k)$ (Lines 5-6). An individual threshold ξ_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. 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 participatory sensing system.

IV. PERFORMANCE EVALUATION

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

A. Background

WiFi fingerprinting is considered a promising indoor localization approach with rapidly increased deployments of WiFi

access points [5]. 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 participatory 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 [6]. However, radiomap construction with participatory 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 participatory sensing-based indoor localization systems.

B. Experimental Setup

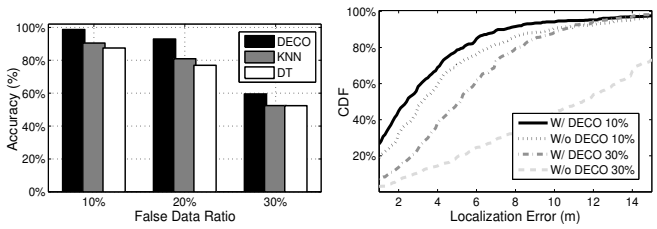
We conducted an experiment using 10 Samsung Galaxy S4 smartphones for participatory sensing-based WiFi fingerprinting. Before the experiment, all smartphones are synchronized by the ClockSync application. Users equipped with smartphones walk around in a campus building over 120 minutes totally. Each 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. Since the WiFi fingerprint data are awaiting for validation, we apply the method introduced in Section III-D to infer smartphones' proximity based on other multidimensional sensor readings, and derive the spatial constraint $\mathbb{H}$. The weight coefficient settings in spatial adjacency 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 technique 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 dbm to -30 dbm at each time slot. 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) [26] and Delaunay Triangulation (DT) [18], [27]. We measure the detection *accuracy* using different data interpolation methods. Here, the accuracy is the proportion of true results (both true positives and true negatives) among the total number of data examined. 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 ξ_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 [28] and compare the final localization error with/without DECO's data correction of the radiomap.

C. Evaluation Results



(a) Accuracy (b) Localization performance

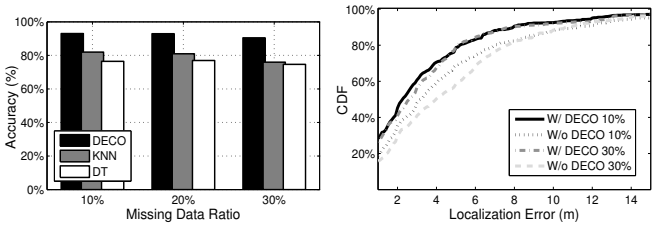
Fig. 2. Impact of false data ratio

1) 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%∼30%.

Fig. 2 shows the false data detection performance. As the false data ratio increases, the results of false data detection accuracy in Fig. 2(a) drop accordingly. However, as DECO utilizes the spatio-temporal constraints and takes the whole data structure into consideration, it outperforms the local interpolation methods KNN and DT, and achieves higher false data detection accuracy in all cases.

Fig. 2(b) 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% to 30%, the average localization error is increased from 4.6m to 10.8m. After performing data correction with DECO, the localization error is significantly reduced to 3.2m and 5.3m, respectively. DECO efficiently detects and corrects false fingerprints and generates high-quality radiomaps.

2) 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%.



(a) Accuracy (b) Localization performance

Fig. 3. Impact of missing data ratio

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. 3 shows the evulation 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 large portion of missing data.

The final corrected fingerprints improves the localization performance. As shown in Fig 3(b), DECO reduces the localization from 3.8m to 3.1m with 10% missing data, and from 5.2m to 3.2m with 30% missing data, which shows that DECO is able to efficiently detect and correct false fingerprints data that introduced by crowdsourcing users in the presence of a large portion of missing data.

V. CONCLUSION

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

ACKNOWLEDGEMENTS

This work was supported in part by the National Natural Science Foundation of China (61170296, 61190125, 61300174, 61303202), 973 Program (2013CB035503), China Postdoctoral Science Foundation (2013M530511, 2014T70026, 2014M560334), and Open Foundation of State Key Lab of Networking & Switching Tech. (Beijing Univ. of Posts & Telecomm., SKLNST-2013-1-02).

REFERENCES

[1] S. Reddy, D. Estrin, and M. Srivastava, "Recruitment framework for participatory sensing data collections," in *Pervasive '10*, 2010.

[2] P. Dutta, P. M. Aoki, N. Kumar, A. Mainwaring, C. Myers, W. Willett, and A. Woodruff, "Common sense: participatory urban sensing using a network of handheld air quality monitors," in *SenSys '09*, 2009.

[3] M. Lin, N. D. Lane, M. Mohammod, X. Yang, H. Lu, G. Cardone, S. Ali, A. Doryab, E. Berke, A. T. Campbell, and T. Choudhury, "Bewell+: multi-dimensional wellbeing monitoring with community-guided user feedback and energy optimization," in *Proceedings of the conference on Wireless Health*, 2012, pp. 1–8.

[4] C. Luo, H. Hong, L. Cheng, K. Sankaran, and M. C. Chan, "iMap: Automatic inference of indoor semantics exploiting opportunistic smartphone sensing," in *SECON'15*, 2015.

[5] J. Jun, Y. Gu, L. Cheng, B. Lu, J. Sun, T. Zhu, and J. Niu, "Social-loc: Improving indoor localization with social sensing," in *SenSys '13*, 2013.

[6] C. Luo, H. Hong, and M. C. Chan, "Piloc: A self-calibrating participatory indoor localization system," in *IPSN '12*, 2014, pp. 143–154.

[7] M. Talasila, R. Curtmola, and C. Borcea, "Improving location reliability in crowd sensed data with minimal efforts," in *Proc. WMNC '13*, 2013.

[8] K. L. Huang, S. S. Kanhere, and W. Hu, "Are you contributing trustworthy data? the case for a reputation system in participatory sensing," in *Proc. MSWIM '10*, 2010, pp. 14–22.

[9] S. Saroiu and A. Wolman, "I am a sensor, and i approve this message," in *HotMobile '10*, 2010, pp. 37–42.

[10] H. Amintoosi and S. Kanhere, "A reputation framework for social participatory sensing systems," *Mobile Networks and Applications*, pp. 1–13, 2013.

[11] X. Oscar Wang, W. Cheng, P. Mohapatra, and T. Abdelzaher, "Artsense: Anonymous reputation and trust in participatory sensing," in *INFOCOM '13*, 2013, pp. 2517–2525.

[12] H. Kurasawa, H. Sato, A. Yamamoto, H. Kawasaki, M. Nakamura, Y. Fujii, and H. Matsumura, "Missing sensor value estimation method for participatory sensing environment," in *PerCom '14*, 2014.

[13] D. Christin, A. Reinhardt, S. S. Kanhere, and M. Hollick, "A survey on privacy in mobile participatory sensing applications," *J. Syst. Softw.*, vol. 84, no. 11, pp. 1928–1946, Nov. 2011.

[14] W. Yow, X. Li, W.-Y. Hung, M. Goldring, L. Cheng, and Y. Gu, "Predicting social networks and psychological outcomes through mobile phone sensing," in *ICC '14*, 2014, pp. 3925–3931.

[15] Z. Song, C. Liu, J. Wu, J. Ma, and W. Wang, "Qoi-aware multi-task-oriented dynamic participant selection with budget constraints," *IEEE Transactions on Vehicular Technology*, 2014.

[16] H. Ahmadi, N. Pham, R. Ganti, T. Abdelzaher, S. Nath, and J. Han, "Privacy-aware regression modeling of participatory sensing data," in *SenSys '10*, 2010, pp. 99–112.

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

[18] I. Vergara-Laurens, D. Mendez, and M. Labrador, "Privacy, quality of information, and energy consumption in participatory sensing systems," in *PerCom '14*, 2014, pp. 199–207.

[19] E. Candes and T. Tao, "Near-optimal signal recovery from random projections: Universal encoding strategies?" *IEEE Transactions on Information Theory*, vol. 52, no. 12, pp. 5406–5425, 2006.

[20] M. Roughan, Y. Zhang, W. Willinger, and L. Qiu, "Spatio-temporal compressive sensing and internet traffic matrices," *IEEE/ACM Transactions on Networking*, vol. 20, no. 3, pp. 662–676, 2012.

[21] L. Kong, M. Xia, X.-Y. Liu, M.-Y. Wu, and X. Liu, "Data loss and reconstruction in sensor networks," in *INFOCOM'13*, 2013, pp. 1654–1662.

[22] L. Kong, L. He, X.-Y. Liu, Y. Gu, M.-Y. Wu, and X. Liu, "Privacy-preserving compressive sensing for crowdsensing based trajectory recovery," in *ICDCS'15*, 2015.

[23] J. Krumm and K. Hinckley, "The nearme wireless proximity server," in *UbiComp'04*, 2004, pp. 283–300.

[24] S. Rallapalli, L. Qiu, Y. Zhang, and Y.-C. Chen, "Exploiting temporal stability and low-rank structure for localization in mobile networks," in *MobiCom '10*, 2010, pp. 161–172.

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

[26] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. 31, no. 1, pp. 21–27, 1967.

[27] L. Kong, D. Jiang, and M.-Y. Wu, "Optimizing the spatio-temporal distribution of cyber-physical systems for environment abstraction," in *ICDCS*, 2010, pp. 179–188.

[28] P. Bahl and V. Padmanabhan, "RADAR: an in-building rf-based user location and tracking system," in *INFOCOM '00*, 2000.