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

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