Computer Networks 120 (2017) 157–169

Contents lists available at ScienceDirect

Computer Networks

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

FIDC: A framework for improving data credibility in mobile crowdsensing

Tongqing Zhou[a,*], Zhiping Cai[a,c], Kui Wu[b], Yueyue Chen[a], Ming Xu[a]

[a] College of Computer, National University of Defense Technology, Changsha, Hunan 410073, China
[b] Department of Computer Science, University of Victoria, Victoria, BC, Canada
[c] School of Computer and Software, Nanjing University of Information Science & Technology, Nanjing, Jiangsu 210044, China

ARTICLE INFO

Article history:
Received 9 October 2016
Revised 26 February 2017
Accepted 7 April 2017
Available online 8 April 2017

Keywords:
Mobile crowdsensing
Data credibility
Provenance information
Data clustering
Logical reasoning

ABSTRACT

Mobile crowdsensing has become a popular paradigm to collaboratively collect sensing data from pervasive mobile devices. Since the devices used for mobile crowdsensing are owned and controlled by individuals with unpredictable reliability, varied capabilities, and unknown intentions, data collected with mobile crowdsensing may be untrustworthy. In particular, a mobile crowdsensing system is subject to collusion attacks where a group of malicious participants collaboratively send fake information to mislead the system. Defending against collusion attacks requires stronger defense mechanisms not available in existing works. In this paper, we propose a new framework for improving data credibility, named FIDC, in mobile crowdsensing to alleviate the threats posed by collusion attacks. FIDC seamlessly integrates two types of correlations: the spatial correlation of sensing data and the correlation between sensing data and provenance knowledge. While both correlations have been adopted separately in previous crowdsensing systems, the exploitation of an joint effort in FIDC poses a special technical challenge to fine-tune the performance. Evaluated extensively with a public mobile crowdsensing data for temperature monitoring, FIDC outperforms existing methods with respect to false detection accuracy and overall data credibility.

1. Introduction

The past few years have witnessed the massive prevalence of human-carried smart devices. These devices are equipped or connected with a rich set of powerful embedded sensors, such as GPS, wireless interface, and air quality monitor. Such advancements lead to a new sensing paradigm, known as mobile crowdsensing (MCS) [1] or participatory sensing [2], where individuals use their own mobile devices to perform sensing tasks, and collect environmental data for specific applications running in the cloud-based platform. So far, a broad spectrum of MCS applications have been developed, including environment monitoring [3], city management [4], network measurement [5], and many more.

A major concern of MCS is on the credibility of collected sensing data [6]. MCS relies on mobile devices of individuals with unknown trustworthiness, varied capabilities, and different intentions to perform sensing tasks. In fact, it has been reported that participants may submit measurements with random values to get rewarded with minimal effort [7]. Even worse, dishonest individuals may inject deliberately fabricated data to mislead the system. As shown in Fig. 1, dishonest participants could corrupt the collected data, which results in incorrect data analysis results. If the data credibility problem remains, the MCS system would eventually become a Garbage-in-Garbage-out (GIGO) system or a system serving for illegitimate purposes. For example, in a noise monitoring application, a real estate agent may submit false noise readings with lower values regarding a specific region to promote the sale of their own properties. Overall, it is critical to ascertain data credibility in nearly all MCS applications.

While extensive research has been devoted to addressing the data credibility problem [8–10], the problem is kept largely open when the system is under collusion attacks, i.e., a group of malicious participants work together to mislead the system into making a wrong decision [11]. Specifically, for those schemes relying on the correlation characteristics of collected data to discover abnormal data [8], collusively contributed false data can be neither filtered out as outlier nor identified with majority voting. Consequently, exploiting data characteristics alone is not able to guarantee data credibility. On the other hand, building trust on the provenance (i.e., the derivation history) is suitable for the evaluation of binary observation [9,10], but not effective for decimal data, which is a more common data format under collusion attacks. For example, it is easy to find support for

* Corresponding author.
E-mail addresses: zhoutongqing@nudt.edu.cn, zhoutongqing1991@163.com (T. Zhou), zpcai@nudt.edu.cn (Z. Cai), wkui@uvic.ca (K. Wu), yueyuechen@nudt.edu.cn (Y. Chen), xuming@nudt.edu.cn (M. Xu).

http://dx.doi.org/10.1016/j.comnet.2017.04.015
1389-1286/

observation "high temperature", but difficult to find support for data "temperature = 20°C". No single solution works well under potential collusive threat, and this motivates our work.

In this work, by jointly exploiting data characteristics and provenance knowledge, we propose a novel framework to improve data credibility, named FIDC, for MCS applications. Specifically, FIDC is designed to mitigate a set of data falsification attacks which aim at compromising data aggregation process of existing systems. FIDC takes advantage of spatial correlation of sensing data and credibility metric regarding provenance[1] information. Intuitively, spatial correlation could be explored to provide discrete groups for provenance based credibility evaluation. On the other hand, provenance knowledge includes those information independent to the collected data, which makes provenance-based credibility assessment immune to collusion attacks. Hence, we could integrate spatial correlation and data-provenance correlation in a data distilling-and-filtering manner.

The major technical challenges of the integration include: (1) how to prepare proper discrete data groups by analyzing spatial correlation, and (2) how to evaluate the credibility of these groups (instead of data points) with provenance knowledge. In view of the above challenges, FIDC first introduces a clustering algorithm to exploit spatial correlation, which would formally separate data into different groups. Further, FIDC refers to provenance of two dimensions: participant provenance (reputation) and context provenance (co-located events), and leverages these information to calculate a credibility score for each group and distinguish the corrupted part. In this way, the integration is properly organized to improve the overall credibility of collected data and effectively defend against collusion attacks.

The main contributions of this paper include:

1) We propose FIDC to defend against the potential data falsification threats. In FIDC, both spatial correlation of data dimension and correlation between sensing data and provenance knowledge (w.r.t. user reputation and context) are studied to improve overall data credibility.
2) A clustering algorithm is utilized to analyze correlation characteristics of collected data; participants reputation together with context information are introduced to constitute a credibility metric to guide the false filtering process.
3) We extensively evaluate our proposed framework with synthetic traces of temperature measurements. Results show that FIDC achieves high credibility of sensing data under the collusion attacks.

2. Related work

MCS is a new sensing paradigm functionally extending the idea of traditional wireless sensor networks (WSNs). With data collection as the core mission, reliability issues of collected data in WSNs [12,13] have been well studied [14]. For example, Zhu et al. [15] propose a vote-based solution to detect injected false data packets by checking endorsements of the co-located nodes. In [16], a clustering algorithm is used to detect anomaly. While solutions in WSNs are instructive to the study of data credibility in MCS, influences of human involvement and corresponding threats (e.g., collusion attack) must be carefully considered. In addition to the studies in the context of WSNs, some recent works have focused on assuring data credibility for MCS, which can be roughly categorized into model-based schemes and false detection-based schemes, as shown in Fig. 2.

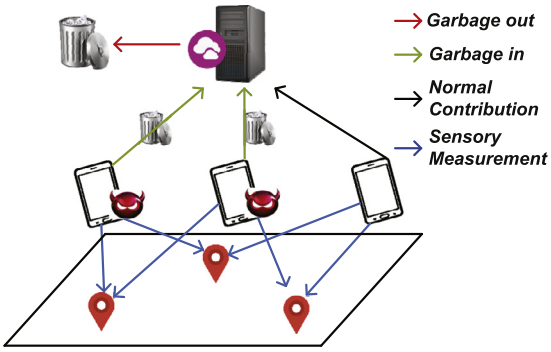

Fig. 1. An example of MCS application with dishonest participants involved in data collection. The contributions of dishonest participants are essentially garbage data. As a result, the output using the garbage data may be useless or misleading, turning the application to a Garbage-in-Garbage-out (GIGO) system.

2.1. Model-based schemes

Model-based schemes build a system to assess the credibility of collected data. Social factors are introduced to estimate data credibility in [17–19]. In these solutions, social relationship is used to describe how dependable one data source is [17], or initiate a voting on collected data among the participants by providing an interaction network [18]. In fact, having all participants in a single social group is not feasible and limits the amount of participants for an application.

In [9] and [10], provenance information is first introduced to assist the trust assessment process. Provenance is a set of user information and contextual factors that describe the origin of the collected data. Modeling and evaluating the corresponding provenance can yield an comprehensive understanding of data credibility. As one type of provenance, users' reputation information often acts as a metric of the trustworthiness of the sensing data [20,21]. To calculate participants' reputation, empirical models (e.g., Gompertz function) are adopted to estimate one's cooperative level based on their behavior in the history. In view of the context provenance, Wang et al. [10] point out that multiple events observed during a short period or at the same location share logical relations. So they propose to evaluate data credibility based on the support of co-located events. However, it is not easy to set a trust threshold to formally distinguish false and normal contribution, because contribution with more support could also be abnormal. Consequently, such solutions are not adequate for autonomous false detection.

2.2. False detection-based schemes

False detection-based solutions try to improve data credibility through identifying and discarding the false data. Techniques in this category include TPM-based schemes, location attestation-based schemes, and data analysis-based schemes.

In [22] and [23], Trusted Platform Module (TPM) is adopted to ensure that data sensed by a mobile sensor and reported to an application server are indeed captured by authentic and authorized devices within the system. In other words, sensing data that fail to pass the authenticity check are considered false. However, the embedded trust module is not readily available for most mobile devices, and malicious participants can cause distortion of the measurements by deliberately initiating sensing action.

As location being a common tag for sensor measurements, validating location can achieve a certain degree of reliability of the

[1] Provenance knowledge could be in many different forms. Its intuitive meaning refers to some extra knowledge known before hand. For instance, in an air pollution monitoring system, the provenance knowledge could be a news report of gas leak in a region.