Leveraging crowd to improve data credibility for mobile crowdsensing

Michael Shell, *Member, IEEE,* John Doe, *Fellow, OSA,* and Jane Doe, *Life Fellow, IEEE*

***Abstract*—The abstract goes here.**

***Index Terms*—mobile crowdsensing, data credibility, cluster analysis, reputation mechanism.**

I. Introduction

THE demand for more pervasive sensing of physical world and the proliferation of human-carried smart devices with rich set of embedded sensors have given rise to a new sensing-oriented application paradigm– mobile crowdsensing(MCS). In a MCS application, public crowd instead of mote-class sensors perform participatory or opportunistic sensing, and collect interested data for further aggregation and analysis in cloud-based platform. A broad spectrum of applications have been developed based on MCS, including environment monitoring(e.g. noise pollution), smart traffic(e.g. congestion monitoring), city management(e.g. network measurement), etc.

A major challenge for the adoption of such applications is that the credibility of collected data cannot be guaranteed, because participants have the motivation to submit false data to earn money without actually executing the task or merely mislead the analysis result. For instance, vehicle drivers may generate false speed information to mislead the estimation of congestion situation and divert the traffic ahead to empty their own road. Such misbehavior will lead to deviation of data aggregation and further influence the availability of the application.

Many works has been done to handle false data and improve the reliability of the collected data. In traditional wireless sensor networks, inner-cluster endorsement[1] and statistic analysis[2] are introduced to detect false data injection attack launched by compromised sensors and aggregators, respectively. However, the endorsement scheme requires a prior knowledge on the number of malicious sensors, while the statistic scheme technically solves a MITM(Man-In-The-Middle) attack. Another category of solution try to increase the credibility with location attestation obtained from infrastructure-based[3] or neighbor-assisted[4] verification. Unfortunately, the requirement of infrastructure and neighbor support may not be feasible, while the real-time verification process is time-consuming, and most importantly, this solution cannot handle the possible submission of false sensory data with a valid location.

This work propose a new approach CRDaFi to improve the credibility of data in MCS without third-party involvement or extra-knowledge requirement. With the crowd-collected data related to a specific location, CRDaFi performs sensory data clustering with the guidance of a reputation mechanism, through which false data can be filtered out. Then the credibility of this set of data is obtained through a statistic analysis phase.

The rest of the letter is organized as follows. In Section II, the problem will be formulated and related assumptions will be described. And then the CRDaFi scheme will be introduced in Section III with its components described in detail. In Section IV, simulation results that indicate the efficiency of the proposed scheme will be provided. Finally, conclusions will be presented in Section V.

II. Problem Statement

The MCS application considered in this letter consists of a cloud-based platform and a set of users $U = \{u_1, ..., u_n\}$ who performs sensing task T at a location L. Without loss of generality, we only consider tasks that require numeric data(e.g. noise). In task T, every user u_i collect the requiring information, denoted as $S_i = \{s_i^1, ..., s_i^n\}$ and submit it with L_i to the platform before pre-defined time deadline t_e, thus the submitted data of u_i is a tuple in key-value syntax, denoted as $d_i =< L_i, S_i >$.

After the completion of T, the platform will obtain a data set $D = \{d_1, ..., d_n\}$, based on which some aggregation function f will be computed. However, as some false data exists in D, the aggregation result will deviate from the true value. Here we say data d_i is false if location L_i or sensory data S_i is invalid, and the validity of L_i and S_i is defined as follows:

1) announced location L_i of u_i is valid if it is the same with its current location.
2) sensory data S_i is valid if it reflects the real physical scenario of the corresponding location L.

Based on the validity of L_i and S_i, data d_i can be classified into 4 categories showed in Table I. Note that existed location attestation-based schemes try to improve reliability by identifying data in category B[1] and C, ignoring possibly false data in category A, which is more often the case(common) for a malicious user to fabricate.

We aim to increase the credibility of D, denoted as $\Re_D$, through identifying and filtering invalid sensory data(data in category A and C). Credibility $\Re_D$ can be defined as:

M. Shell is with the Department of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, 30332 USA e-mail: (see http://www.michaelshell.org/contact.html).

J. Doe and J. Doe are with Anonymous University.

Manuscript received April 19, 2005; revised September 17, 2014.

[1] Indeed, providing valid sensory data(category B) at invalid location is not feasible as a user is unable to contribute valid data of one location when he is not really there.

TABLE I
MY CAPTION

d_i L_i	T	F
T	Normal data	class A
F	class B	class C

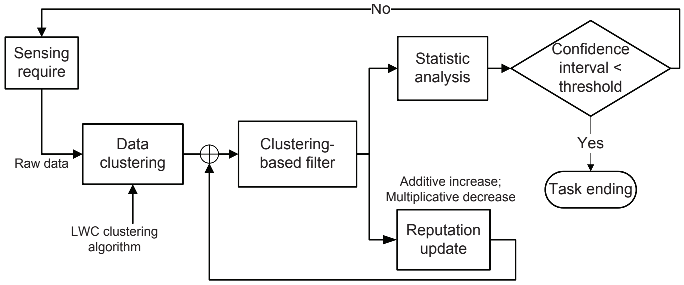

Fig. 1. Framework of CRDaFi scheme

$$\Re_{\mathrm{D}} = (\frac{|\mathrm{f}(D) - f(\tilde{D})|}{\min(\mathrm{f}(D), f(\tilde{D}))})^{-1}$$

where $\tilde{D} = D - D_f$, and D_f is the set of false data. Obviously, the less false data in D, the more similar $f(D)$ and $f(\tilde{D})$ will be, and D will have a higher credibility.

Finally, we assume that location L refers to a sensing area around L instead of a specific spot. We further assume that malicious users would try to deviate the aggregation function result as much as possible to achieve misleading effect. Taking average function as example, if we have N normal user with data value equals to M, and N malicious user with value equals to $2M$, then the aggregation result will be 1.5 times of the actual one. However, we do not make any assumption or set any limitation on the number of malicious users in U.

III. Proposed Scheme

IV. Simulation results

Experiments part.

V. Conclusion

The conclusion goes here.

References

[1] S. Zhu, S. Setia, S. Jajodia, and P. Ning, "An interleaved hop-by-hop authentication scheme for filtering of injected false data in sensor networks," in *Security and privacy, 2004. Proceedings. 2004 IEEE symposium on*. IEEE, 2004, pp. 259–271.

[2] B. Przydatek, D. Song, and A. Perrig, "Sia: Secure information aggregation in sensor networks," in *Proceedings of the 1st international conference on Embedded networked sensor systems*. ACM, 2003, pp. 255–265.

[3] N. Sastry, U. Shankar, and D. Wagner, "Secure verification of location claims," in *Proceedings of the 2nd ACM workshop on Wireless security*. ACM, 2003, pp. 1–10.

[4] M. Talasila, R. Curtmola, and C. Borcea, "Link: Location verification through immediate neighbors knowledge," in *Mobile and Ubiquitous Systems: Computing, Networking, and Services*. Springer, 2012, pp. 210–223.