DTRM: A New Reputation Mechanism to Enhance Data Trustworthiness for High-Performance Cloud Computing

Hui Lin[1], Jia Hu[2*], Chuanfeng Xu[1], Jianfeng Ma[3], Mengyang Yu[1]

[1]Fujian Provincial Key Laboratory of Network Security and Cryptology, College of Mathematics and Informatics, Fujian Normal University, Fuzhou, 350007 China.
[2]College of Engineering, Mathematics and Physical Sciences, University of Exeter, Exeter, EX4 4QF, UK.
[3] School of Cyber Engineering, Xidian University, Xi'an 710071, China

Abstract: Cloud computing and the mobile Internet have been the two most influential information technology revolutions, which intersect in mobile cloud computing (MCC). The burgeoning MCC enables the large-scale collection and processing of big data, which demand trusted, authentic, and accurate data to ensure an important but often overlooked aspect of big data -- data veracity. Troublesome internal attacks launched by internal malicious users is one key problem that reduces data veracity and remains difficult to handle. To enhance data veracity and thus improve the performance of big data computing in MCC, this paper proposes a Data Trustworthiness enhanced Reputation Mechanism (DTRM) which can be used to defend against internal attacks. In the DTRM, the sensitivity-level based data category, Metagraph theory based user group division, and reputation transferring methods are integrated into the reputation query and evaluation process. The extensive simulation results based on real datasets show that the DTRM outperforms existing classic reputation mechanisms under bad mouthing attacks and mobile attacks.

Keywords: Cloud Computing; Reputation Mechanism; Trustworthiness; Data Veracity

1. Introduction

Mobile Cloud Computing (MCC) combines cloud computing and mobile computing to provide mobile users with data storage and processing services in clouds that perform resource-intensive computing [1, 2]. The MCC infrastructure involves a set of cloud resources accessed remotely by the users equipped with different devices through the Internet [3]. A typical MCC architecture as shown in Fig. 1 [3], consists of a mobile client network and a cloud service platform. The mobile client network includes mobile devices, base transceiver station (BTS) and a mobile network. The cloud service platform includes cloud application servers, cloud controllers, data centres etc., to offer data-rich services such as queries of electronic medical records. As a highly promising information technology trend, MCC enables the large-scale collection and processing of big data for emerging applications [4].

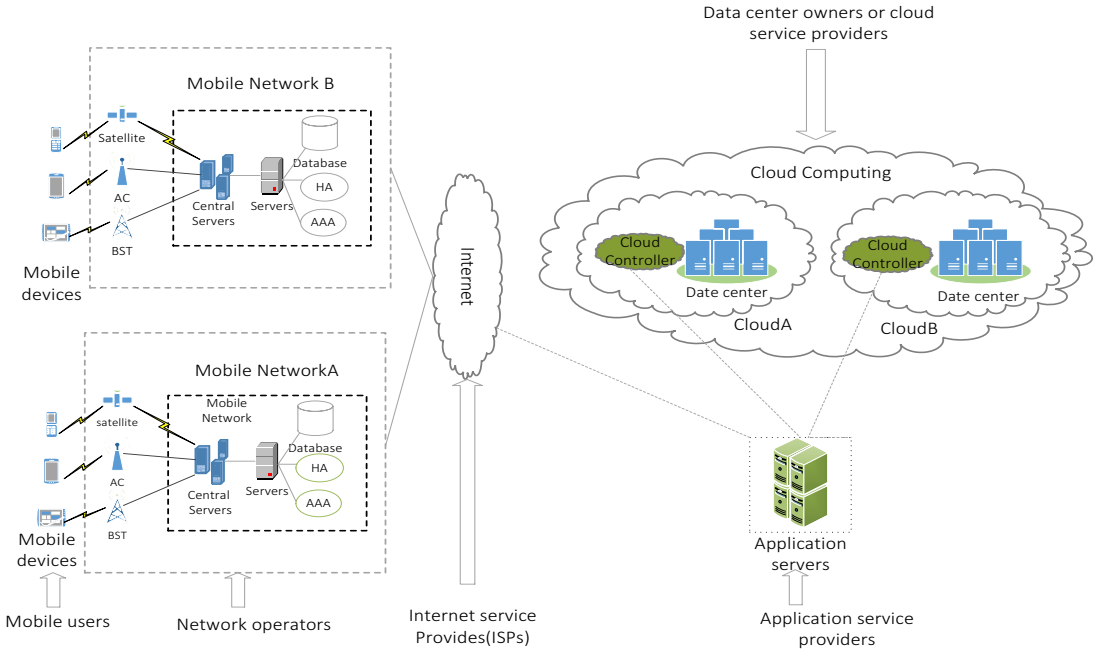

Fig. 1. Mobile Cloud Computing Infrastructure

In the era of big data, data can be produced by online and offline transactions, social networks, sensors and through our daily life activities [5, 6]. The proper processing of big data can result in informative, intelligent and relevant decision making to influence our daily behaviors, scientific developments, and the planning and policies [7]. In order to avoid making decisions based on the analysis of uncertain and imprecise 'dirty' data, it is crucial to verify and clean the data, which leads to the designation of a fourth V in big data: Veracity [8]. Data veracity refers to the quality or trustworthiness of the data and addresses the confidentiality, integrity, and availability of the data [9]. Data veracity includes two aspects: data certainty defined by their statistical reliability; and data trustworthiness defined by a number of factors including data origin, collection and processing methods such as infrastructure and facility [10, 11].

With the ubiquitous access to the Internet enabled by MCC, it is not uncommon that big data contains biases, noises, and abnormalities, which pose a big threat to data veracity. Moreover, there are many security issues that can affect data veracity such as external denial-of-service, credential stealing, remote code injection, data integrity attacks, internal attacks, and supply chain attacks [8]. Consequently, the availability, confidentiality, and integrity of both the original data and the data analytics results are threatened by these attacks,

e.g., the degraded availability of a big data system, the compromised confidentiality of the data and analytics, and the violated integrity of the data and analytic results.

High-quality and trustworthy data can not only ensure the veracity of information but also improve the performance of big data computing. It is highly desirable to clean data before analyzing it and using it to make decisions. As an effort to tackle the aforementioned challenges, this paper focuses on enhancing the data veracity in MCC by designing a new reputation mechanism. The major contributions of this work include the following:

- We propose a new Data Trustworthiness enhanced Reputation Mechanism (DTRM) to defend against the bad mouthing attacks and internal mobile attacks for enhancing data veracity in MCC.
- The DTRM develops three key security schemes including a sensitivity-level based data category scheme, a metagraph based user group division strategy, and a reputation transferring method.
- Simulation experiments based on real datasets demonstrate that the DTRM improves the performance of the reputation mechanism compared to the state-of-the-art including the ATrust [12] and TSCM [13] mechanisms.

The remainder of this paper is organized as follows. Section 2 presents a brief review of the related work. Section 3 describes the adversary models. Section 4 presents the implementation details of the DTRM. Section 5 analyzes the cost and evaluates the performance of the DTRM. Finally, Section 6 concludes this paper.

2. Related Work

As an important aspect of big data, data veracity has been investigated in some related papers in the literature [8,9,14-19]. For example, Kepner et al. [8] introduced a technique called Computing on Masked Data (CMD) to improve data veracity while allowing a wide range of computations and queries to be performed with low overhead. The CMD combines efficient cryptographic encryption methods with an associative array representation of big data. Lozano et al. [9] identified the challenges and proposed an approach and a corresponding framework for automated veracity assessment of Open Source Information. The framework describes necessary components and shows how a veracity assessment network is gradually built up and expanded from direct and transitive veracity assessments.

Bodnar et al. [14] proposed a veracity assessment model for information dissemination on social media networks that combines natural language processing and machine learning algorithms to mine textual content generated by users. Agarwal et al. [15] proposed a crowdsourcing based solution to solve the big data veracity problem that uses the sentiment analysis method to deal with identifying the sentiment expressed in a piece of text. Ashwin et al. [16] proposed three indices named as topic diffusion, geographic dispersion, and spam index to measure the veracity of Twitter topics from tweets themselves. These measures are tested using tweets about oil companies as validators. Debattista et al. [17] defined the veracity of Big Data as "conformity with truth or facts", and described eight Linked Data quality metrics and two techniques to improve and maintain quality and address Big Data's veracity challenge.

Since this paper focuses on using a high-performance reputation mechanism to enhance the data veracity in MCC, in the following, we mainly review the existing research results regarding reputation mechanisms in MCC. Kim et al. [20] proposed a trust management mechanism for reliable data integration, management and applications in MCC. The mechanism suggested a method to quantify a one-dimensional trusting relationship based on the analysis of telephone call data from mobile devices. Shen et al. [21] developed an integrated reputation management platform, Harmony, for collaborative cloud computing. Harmony incorporates an integrated reputation management component, a multi-QoS-oriented resource selection component and a price-assisted resource control component to enhance their mutual interactions for efficient and trustworthy resource sharing among clouds. Zhang et al. [22] presented a general framework to jointly design incentive mechanisms and reputation schemes in social cloud systems. The proposed framework combined a repeated game framework-based incentive mechanism with a differential reputation-based reward/punishment scheme to incentivize users to contribute their resources. Chang et al. [23] studied the MCS network trustworthiness problem, and proposed a cloud based trust management scheme (CbTMS) that combines Characteristics Checking Scheme (PCS) and Trust Credit Assessment (TCA) to detect suspicious Sybil nodes, reduce the negatively influence on the effectiveness of sensing data in MCS network and enhance entire MCS network performance. Kantarci et al. [13] proposed a reputation-based

Sensing-as-a-Service scheme, namely, Trustworthy Sensing for Crowd Management (TSCM) to ensure trustworthiness in crowd sensing management for MCS systems. Palaghias et al. [24] presented opportunistic sensing system MobTrust to reliably derive and quantify trust relationships for MCS systems by combining the extracted real-world social graph, the estimated social relations with the contextual information provided by the detected social interactions. Lin et al. [25] proposed a reliable recommendation and privacy preserving based cross-layer reputation mechanism that integrates the cross-layer design with recommendation reputation reliability evaluation mechanism and the privacy preserving scheme to protect the security and privacy against internal attacks.

3. The Adversary Model

In the MCC architecture, the application cloud server informs all mobile clients about their assigned data tasks and distributes tasks to mobile clients who meet the requirements of applications. This paper focuses on the internal security threats [25-27] that can affect data trustworthiness. The internal threats are launched by an inside attacker who is a legal and certified mobile client. The internal attacks may compromise certain users and gain full control of them. Once mobile clients are compromised, the attacker can gain access to all stored information, including public keys and private keys. The attacker could also reprogram the captured mobile clients to behave in a malicious manner. Therefore, the traditional encryption and authentication techniques may no longer be effective. The specific internal attacks considered in this paper are as follows [18, 28-30]:

Mobile attacks: Malicious mobile client can move position to disguise as a normal client to start a new round of interactions with others.

Bad mouthing attacks: MCC openness can allow any participant to contribute data, which means that attackers can provide erroneous and malicious sensing data as well as recommended opinions for their own benefit.

For example, when a participant intends to request (or provide) a service from (or to) another participant (including unknown participant) or to query a participant's reputation from other participants, it will send a request message to its neighbors. When an adversary receives the request, it will launch bad mouthing attacks to provide disinformation for their own benefit. The adversary may also launch mobile attacks by moving to a new position,

where the adversary will be identified as a new participant, enabling it to launch another round of attacks.

4. A Data Trustworthiness enhanced Reputation Mechanism (DTRM)

In this section, we elaborate on the proposed Data Trustworthiness enhanced Reputation Mechanism (DTRM), which integrates the reputation mechanisms [18, 25, 27] with the mobile crowd sensing (MCS) [15], metagraph theory [31], data category and user group division technologies [32] to enhance data trustworthiness, defend against the insider threat and enhance the big data veracity in MCC. The DTRM is implemented in mobile sensing devices and cloud service providers to perform bidirectional reputation evaluation. In the rest of the paper, the terms "participant", "mobile client", and "user" are used interchangeably.

In DTRM, the big data to be processed are collected by the mobile sensing devices and are classified into different categories based on the sensitivity level (*SL*) of data. The data sensitivity level reflects the confidentiality and the privacy of the data. The higher is the sensitivity level of data, the greater is the need for the confidentiality and privacy protection. Therefore, a high reputation is required for a device to access the data in a category with a high sensitivity level. In this work, the sensitivity level of sensing data is decided by the data owner, fixed and divided into five grades from 1 to 5. Also, we use the Metagraph theory [31], a graphical data structure for representing a collection of directed set-to-set mappings, to divide all users into different groups according to the relevancy and familiarity between them. We assume that the users or groups belong to the same trust domains. The proposed scheme could also be adapted to other communications networks and computing systems [33-35]. The relevancy and familiarity are computed through the received sensing information [36] such as services, networks and applications and so on from the mobile sensing devices. The details of the DTRM are described as follows.

4.1 Metagraph based User Group Division

In DTRM, we use Metagraph theory [31] for representation and calculation of different possible kinds of trust relations between persons and groups and the transition of trust from group to person and vice versa. In this paper, we focus on how to represent and evaluate the trust relationship between users and groups through the Metagraph.

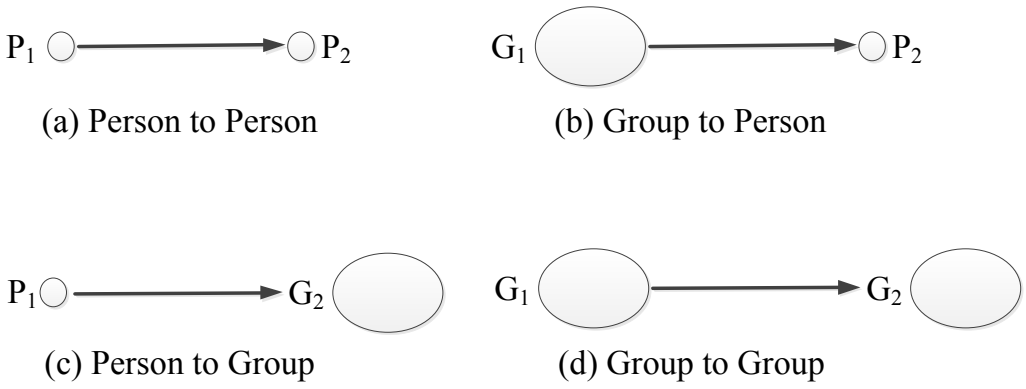

Fig. 2. Trust Relationship in DTRM

In DTRM, trust relations are assumed as person-person, person-group, group-person, and group-group which are shown in Figure 2. At the same time, since persons might possess different positions within a group, they will have different membership degrees. The higher a degree in the group, the more likely the behavior of the person will be based on the standards and norms of the group. Therefore, in order to calculate the reputation value of a person based on group membership, this criterion should also be considered. The membership degree is defined as the following function [31]:

$$f_m: (Person,\ Group,\ Motivation,\ Attraction,\ Position) \rightarrow [0,1] \quad (1)$$

where f_m is the function which returns the numerical value of the person's membership degree in the group. *Person* and *Group* are the given person and group. *Motivation* shows the level of the person's participation in group interactions and responsibilities. *Attraction* represents the amount of the person's connection to the group. A person with low attraction value may easily leave the group. *Position* represents the power and influence of the person in the group which is dependent on social relations created in the group by that person as well as the amount of his capability and expertise in performing tasks in group.

Definition 1. The generating set of a metagraph is the set of elements $X = \{x_1, x_2, \quad, x_n\}$, which represents variables of interest occurring in the edges of the metagraph.

Definition 2. An edge e in a metagraph is a pair $e=<V_e,W_e> \in E$ (where E is the set of edges) consisting of an invertex $V_e \subset X$ and an outvertex $W_e \subset X$, each of which may contain any number of elements. The different elements in the invertex (outvertex) are co-inputs (co-outputs) of each other.

Definition 3. A metagraph $S = < X, X, E >$ is then a graphical construct specified by its

generating set X and a set of edges E defined on the generating set. X is characterized by its membership function f_X: $X \rightarrow [0, 1]$. For each $x \in X$, f(x) illustrates the truth value of the statement of "x belongs to X". E is defined as a function $f_E : E \rightarrow [0, 1]$, where E is an edge set and the membership value of an edge is also called certainty factor (CF) of the edge. For simplicity, assign x_i denoting $(x_i, f_X(x_i))$ and e denoting (e ,$CF\ e$), i.e.,(e , f_E (e)).

The metagraph based trust concepts in DTRM are considered as follows:

(1) Trust values among groups are represented using metagraph.

(2) Generating set X represents users and their membership degrees in their corresponding groups.

(3) Edges between two groups indicate the existence of trust between them. For example the edge $e=<V_e,W_e> \in E$ represents the trust of group V_e toward group W_e.

(4) The label of edge $e=<V_e,W_e> \in E$ is a couple of values <t;c>: the first component is the reputation value of group V_e toward group W_e, while the second component is the quality of the reputation value assignment (i.e. a confidence value), both of these components are in the range [0, 1].

(5) A high reputation value means that the trustee has gained a good feedback, whereas a confidence value close to 1 indicates that the trustor estimates the correlated reputation value with precision.

As an example, consider the metagraph $S = < X, X, E >$ in Figure 3. The sets X and X are $X = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7\}$ and $X = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7\}$, respectively.

Where, $x_1 = (x_1, f_X(x_1)), x_2 = (x_2, f_X(x_2)),$ $, x_7 = (x_7, f_X(x_7))$ and

$$G_1 = \{x_1, x_2\},\ f_X(x_1) = 0.8,\ \ f_X(x_2) = 0.9$$

$$G_2 = \{x_2, x_3\},\ f_X(x_2) = 0.9,\ \ f_X(x_3) = 0.7$$

$$G_3 = \{x_4, x_5\},\ f_X(x_4) = 0.8,\ \ f_X(x_5) = 0.6$$

$$G_4 = \{x_6, x_7\},\ f_X(x_6) = 0.7,\ \ f_X(x_7) = 0.6$$

The set of edges is $E = \{e_1, e_2, e_3, e_4\}$. Where, $e_1 =< \{x_1, x_2\}, \{x_4\} >, e_2 =< \{x_2, x_3\}, \{x_5\} >,$ $e_3 =< \{x_4, x_5\}, \{x_6, x_7\} >, e_4 =< \{x_5\}, \{x_7\} >$. The edge e_1 between groups G_1 and G_2 is labeled as < 0.7; 0.6>. It shows that there exist a trust relationship between group G_1 and group G_2 and the reputation value of group G_1 to group G_2 is 0.7, and it is estimated with precision 0.6.

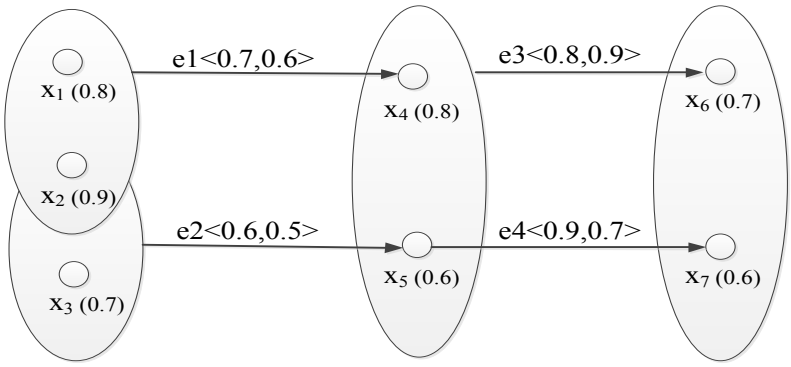

Fig. 3. Metagraph Based Trust Relationship Example

4.2. Direct Reputation Computation

The direct reputation computation is run at each user that stores its historical opinion towards the others in the relevant local database. When a user wants to request (or provide) a service from (or to) another user (including unknown user), it will send a request message to those trustworthy neighboring users or groups built by the method proposed in subsection 4.1. Each user receiving the request will first execute the direct reputation computation function to evaluate the requestor's direct reputation and judge whether it is a malicious user and whether to provide the service.

Suppose x and y are two users that want to interact presently, and x wants y to provide service. Let $G=\{G_1,G_2, \ldots, G_n\}$ and $SL=\{sl_1,sl_2, \ldots, sl_m\}$ ($sl_i \in [0,1]$, $i=1,\ldots,m$) be the users group set and data sensitivity level set, respectively. The detailed process of the direct reputation computation is described as follows.

(1) The user x who want y to provide service sends a *REQUEST* message to y;

(2) User y receives the *REQUEST* message, and then judges whether x belong to the same group;

a) If x and y belong to a same group, then the direct reputation of x toward y, $R_{y:x}^{\text{Direct}}$, can be computed as:

$$R_{y:x}^{\text{Direct}} = f_m^{cur}(x) * \left[\frac{1}{|\text{SL}|} * \sum_{j=i}^{|\text{SL}|} \left(\frac{IA_s^j}{IA_{total}^j} * \varphi_j \right) \right] \tag{2}$$

where $f_m^{cur}(x)$ is the current membership degree of x in the group computed according to Eq. (1). IA_s^j, IA_{total}^j denote the number of successful and total access or gain attempts to the data with sensitivity level j, respectively. φ is the weight

factor that determine how much the accessed data sensitivity of the interaction affect $R_{y:x}^{\text{Direct}}$. In DTRM, the historical records of accessed data sensitivity level will also influence the direct reputation computation. High and low sensitive data are the data with high data sensitivity level and low data sensitivity level, respectively. For a user used to access the low sensitive data, its sudden access to a higher sensitive data is noteworthy. Therefore, we define and compute the accessed data sensitivity factor φ as:

$$\begin{cases} \varphi = E(\mu_t) \\ \mu_t = \dfrac{\sum_{j=i}^{|\mathrm{SL}|} N_j}{\sum_{j=1}^{|\mathrm{SL}|} N_j}, (t = 1 \quad N_{slot}) \end{cases} \tag{3}$$

where μ_t is the rate between the number of accesses to the data sensitivity level higher than the current required level *i* and the total number of accesses to all levels. N_j represents the number of times that *x*'s historical accessed data sensitivity level is confirmed as *j*, and N_{slot} denotes the number of the time slots.

b) If *x* and *y* belong to a different group, suppose $x \in \mathrm{G}_1$ and $y \in \mathrm{G}_2$,then the direct reputation of *x* toward *y,* $R_{y:x}^{\text{Direct}}$, can be computed as:

$$R_{y:x}^{\text{Direct}} = Trust(x,y) * \tau * \left[\frac{1}{|\mathrm{SL}|} * \sum_{j=i}^{|\mathrm{SL}|} \left(\frac{IA_s^j}{IA_{total}^j} * \varphi_j \right) \right] \tag{4}$$

where $Trust(x,y)$ is the reputation value of *x* toward *y* and it can be computed as

$$Trust(x,y) = f_m^{cur}(x) * trust(\mathrm{G}_1, \mathrm{G}_2) * f_m^{cur}(y) \tag{5}$$

In equation (5), $trust(\mathrm{G}_1, \mathrm{G}_2)$ is the reputation value of group G_1 toward group G_2, and we assume that the reputation values between groups are pre-computed and stored in the local database. τ is the time factor that determine how much the interaction time affect $R_{y:x}^{\text{Direct}}$. We then formally define the τ as:

$$\tau = \tau_{y:x,T_n} * \beta_{T_n} \tag{6}$$

where β_{T_n} is the density of the historical interaction until time T_n and $\tau_{y:x,T_n}$ is the

weight factor, which determines how much the distribution of the interactions affects the $R_{y:x}^{\text{Direct}}$ at time T_n. $\tau_{y:x,T_n}$ and β_{T_n} can be computed as follows.

$$\beta_{T_n} = 1 - e^{\wedge}(-\frac{\sum_{i=1}^{|SL|} N_i}{m * n}) \tag{7}$$

$$\tau_{y:x,T_n} = \sum_{l=1}^{n} (\frac{T_l}{m} * \frac{l}{n}) \tag{8}$$

where N_i is the number of times the historical accessing behaviors or interactions are confirmed on the sensitivity level *i*. *m* and *n* are the number of time slots and cycle *T* respectively, e.g., in this paper, *T* is equal to 10 seconds, *m* is 5, so one time slot equals 2 seconds.

(3) When *y* gets the direct reputation computation result it will identify whether *x* is trustworthy by comparing the direct reputation result with the threshold as follows. Here, we assume $sl_i(x)$ be the *i*th sensitivity level of data that user *x* want to access currently, and Th_{min}, Th_{max} be lower and upper bound for the reputation.

a) If $R_{y:x}^{\text{Direct}} \geq sl_i(x) * Th_{max}$ then *x* is considered as a trustworthy node and *y* sends *Accept* message to *x*;

b) If $R_{y:x}^{\text{Direct}} < sl_i(x) * Th_{min}$ then *x* is considered as a malicious user and *y* will notify all other groups and other users in the same group.

c) If $sl_i(x) * Th_{min} < R_{y:x}^{\text{Direct}} < sl_i(x) * Th_{max}$ then *y* executes the recommended reputation query, recommended reputation computation and final reputation computation.

4. 3. Recommended Reputation Computation

If the direct reputation computation in section 4.2 cannot lead to a decision, *y* will first execute the recommended reputation query using Algorithm 1 to query *x*'s reputation from other users.

Algorithm 1: Recommended Reputation Query

Input: Users *x* and *y's* information
Output: *x's* reputation
1. Begin
2. Do
3. *Y* firstly select those users and groups that have registered and stored in the local database to end the *Query* message*;*

4. Wait (3-5 seconds);
5. Any user *k* in group G receives the *Query* message, *k* computes *y*'s direct reputation ;
6. If *y*'s direct reputation > *Threshold* then
7. *k* retrieves its direct opinion about *x* on local reputation database and sends *Reply* message to *y* as *Reply* (*k*,*G*, $R_{k:x}^{\text{Direct}}$, *time*, *location*);
8. Else
9. *k* drops the *Query* message;
10. End if
11. If *y* doesn't receive any return opinions then
12. *Y* broadcasts the *Query* message to other users and groups don't registered and stored in the local database;
13. Wait (3-5 seconds);
14. Repeat steps 5-10;
15. End if
16. While (received return opinions < *Threshold*)
17. End

Afterwards, y will compute the integrated recommended reputation combining the received replies of recommended reputations to the query, which will be described in the following.

First, *y* will choose the trustworthy and reliable recommenders from those have returned the opinion by computing the familiarity and relevancy between *y* and the recommender *u FR(u,y)* as:

(1) If the candidate recommenders *u,* named direct recommender*, x* and *y* belong to a same group, but *u* is a new group member and never interacts with *y*. *FR(u,y)* is given by:

$$FR(u,y) = f_m^{cur}(u) \tag{9}$$

(2) If the candidate recommenders *u,* named transferring recommender, belong to a different group G_1 and *y* belongs to group G_2, and there has not the direct history trust relationship between two groups. We assume the recommended reputation transferring path is $\text{RPath} = \{pr_p \mid p = 1 \quad P\}$, and compute the *FR(u,y)* as*:*

$$\begin{cases} FR(u,y) = f_m^{cur}(u) * Trust(RPath) * f_m^{cur}(y) \\ Trust(RPath) = \dfrac{1}{|\text{P-1}|} * \prod_{i=1}^{P-1} \dfrac{c_{i,i+1} + t_{i,i+1}}{2} \end{cases} \tag{10}$$

where $Trust(RPath)$ is the reputation value of the transferring path RPath. $c_{i,i+1}$ and $t_{i,i+1}$ are the reputation value and confidence value of group $G_i \in$ RPath toward group $G_{i+1} \in$ RPath respectively, as defined in subsection 4.1.

(3) *Y* finds out the trustworthy and reliable recommenders and builds a recommender set **R** by comparing each recommender *u*'s *FR(u,y)* with a threshold, one by one.

(4) For a transferring recommender, if there are many recommend opinion coming from

different paths, the most reliable path denoted as *MRPath* is chosen based on the rules below. Here, we assume $RP(i),(i=1,\ \ ,n)$ is the set of the recommend paths and each path includes *j* groups.

$$\begin{cases} MRPath = Max(\zeta_1 * Trust(RP(i)) + \zeta_2 * SL_{RP(i)}), \\ \zeta_1 + \zeta_2 = 1, \zeta_1, \zeta_2 \in [0,1] \end{cases} \tag{11}$$

where ζ_1 and ζ_2 are the weight factors corresponding to the reputation and data sensitivity level of path $RP(i)$ respectively. Then we define and compute the $SL_{RP(i)}$ as:

$$SL_{RP(i)} = Min(SL_j^i),\ (1 \le j \le n) \tag{12}$$

where SL_j^i is the data sensitivity level of group G_i in the *j*-th path in $RP(i)$.

Second, suppose *y* receives *m* (*m*>1) direct recommended opinions and *n* (*n*>1) transferring recommended opinions, then the final integrated recommended reputation, $R_{y:x}^{\text{Rec}}$, can be defined and computed as follows.

$$\begin{cases} R_{y:x}^{\text{Rec}} = \eta_1 * R^{Dir\text{-}Rec} + \eta_2 * R^{Tran\text{-}Rec} \\ R^{Dir\text{-}Rec} = \dfrac{1}{m} * \sum_{j=1}^{m} \left(FR(j,y) * R_{j:x}^{\text{Direct}} \right) \\ R^{Tran\text{-}Rec} = \dfrac{1}{n} * \sum_{k=1}^{n} \left[FR(k,y) * \text{MRPath}(k) * R_{k:x}^{\text{Direct}} \right] \\ \eta_1 + \eta_2 = 1, \eta_1, \eta_2 \in [0,1] \end{cases} \tag{13}$$

where $R^{Dir\text{-}Rec}$ and $R^{Tran\text{-}Rec}$ are the integrated direct recommended reputation and transferring recommended reputation, respectively. η_1, η_2 are the weight factors, which determine how much the integrated direct recommended reputation $R^{Dir\text{-}Rec}$ and transferring recommended reputation $R^{Tran\text{-}Rec}$ affect the final integrated recommended reputation evaluation, respectively.

4.4. Final Reputation Computation

After finding the direct and final integrated recommended reputation, the final reputation $R_{y:x}^{\text{Final}}$ can be computed as:

$$\begin{cases} R_{y:x}^{Final} = \alpha_1 * R_{y:x}^{\text{Direct}} + \alpha_2 * R_{y:x}^{\text{Rec}} \\ \alpha_1 + \alpha_2 = 1, \alpha_1, \alpha_2 \in [0,1] \end{cases} \tag{14}$$

where α_1, α_2 are the weight factors for the direct reputation and final integrated recommended reputation, respectively.

5. Performance Evaluation

The real Weibo-Net-Tweet dataset [37] was used in the simulation for evaluating the performance of the propoosed reputation mechanism in MCC. In the Weibo-Net-Tweet dataset, to begin with, 100 mobile clients were randomly selected as seed mobile clients, and then their followees and followees of followees were collected. The Weibo-Net-Tweet dataset includes in total 1.7 million mobile clients and 0.4 billion following relationships among them. For each mobile client, the dataset collected 1,000 most recent microblogs. The process inlcudes in total 1 billion microblogs. Each mobile client's profile contains name, gender, verification, #bi-following, #followers, #followees, and #microblogs (# stands for "the number of"). We focus on the retweet behaviors and develop a Java-based simulator to implement our reputation mechanism and measure its performance and accuracy. We use a three-server cluster to deploy the mobile cloud platform; each server has 16 GB memory and 12 XeonTM CPUs with 24 cores. We use three laptops as the mobile terminals, each with Intel(R) Core 2 Duo T5870 2.00GHz CPU.

In the simualtion, we define a good participant as a participant who always sends genuine sensing reports. However, an adversary does not necessarily send false sensing reports each time. They may launch on-off attacks by sending correct reports in order to gain reputation and then only send false reports randomly or at a specific time.

The values for security parameters $\zeta_1, \zeta_2, \eta_1, \eta_2, \alpha_1, \alpha_2$ are 0.6, 0.4, 0.5, 0.5, 0.6, 0.4, which are empirical values obtained from multiple experiments. Of which, ζ_1, ζ_2 are the weight factors in Eq. (11) associated with the reputation value and data sensitivity level of a path. η_1 and η_2 are the weight factors in Eq. (13) used to determine how much the direct recommended reputation and transferring recommended reputation affect the final integrated recommended reputation, respectively. α_1, α_2 are the weight factors in Eq. (14) used to determine how much the direct reputation and final integrated recommended reputation affect the final reputation, respectively. Each data point depicted in the following figures is the average of the results obtained from 100 runs of simulation experiments with a simulation time of 100 s each.

The performance of the proposed DTRM is compared to the ATrust [12] and TSCM [13] because they are the similar and latest related mechanisms. The following performance

metrics are evaluated when internal bad mouthing attacks and mobile attacks are present:

Data Trustworthiness Rate (DTR): The rate of the trustworthy report data to the total amount of report data provided by the mobile clients.

Malicious client Detection Rate (MDR): The accuracy of detecting and identifying malicious mobile clients.

False Positive Rate (FPR): The ratio of the number of false reports on malicious mobile clients to the total number of reports on malicious mobile clients.

Reputation Evaluation Accuracy rate (REA): The accuracy of reputation evaluation of a mobile client.

5.2.1 Data Trustworthiness Rate (DTR)

First, we investigate how the DTRM performs in an honest network and a hostile network, respectively. In the honest network, all the mobile clients are good participants, while in the hostile network, the mobile clients may be adversaries who give false information with a random probability.

The DTR of the DTRM in the honest network is shown in Fig. 4 (a). In this simulation, we compare the DTR of the DTRM with the W-DTRM. The results show that the adoption of the new data category, user group division and reputation transferring methods makes the DTR of the DTRM higher. First, in the DTRM, the data category scheme classifies the data into different categories based on the sensitivity level of the data, and only the opinions provided by the user having high enough reputation value may be accepted, which improves the DTR. Second, metagraph based user group division strategy divides all mobile clients into different groups according to the relevancy and familiarity between them, and only the opinions from the relevant and familiar mobile clients may be adopted, which makes the source of the opinions more trustworthy and enhances the DTR of the reputation mechanism. Third, the reputation transferring method can effectively solve the reputation loss problem caused by the mobile clients' movement and defend against the internal mobile attacks, and thus will also make the sensing data more trustworthy.

We also analyze the impact of the malicious attacks on the data trustworthiness, and compare the DTR of the DTRM with those of the ATrust and TSCM. Comparing to the results in Fig. 4 (a), in Fig. 4 (b) and (c) where the bad mouthing attacks and mobile attacks

are present, the DTR of DTRM decreases by 10%, and 12%, respectively. In addition, Fig. 4 (b) and (c) show that the DTR of the DTRM is higher than the other two mechanisms. The reason is that the ATrust and TSCM lack effective classification schemes for mobile clients' relevancy and familiarity and data sensitivity level, therefore, some irrelevant, unfamiliar and unreliable mobile clients can also provide the sensing data or opinions, which decreases the effectiveness and credibility of the data and opinions. Furthermore, neither the ATrust or TSCM consider the reputation loss problem caused by the mobile clients' movement, which also makes their DTRs worse than that of the DTRM.

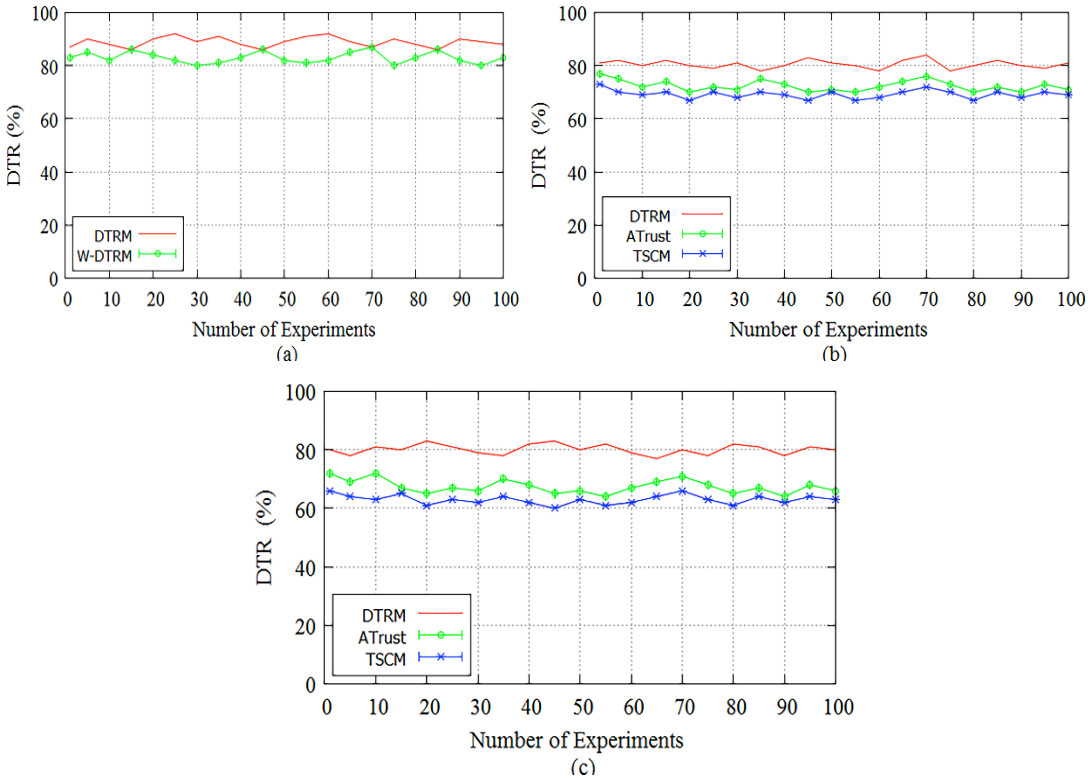

Fig. 4. Data Trustworthiness Rate (a) in an honest network, (b) with bad mouthing attacks, and (c) with mobile attacks

5.2.2 Malicious client detection rate (MDR)

Next, we analyze the malicious client detection rate under two hostile network environments with bad mouthing attacks and mobile attacks, respectively.

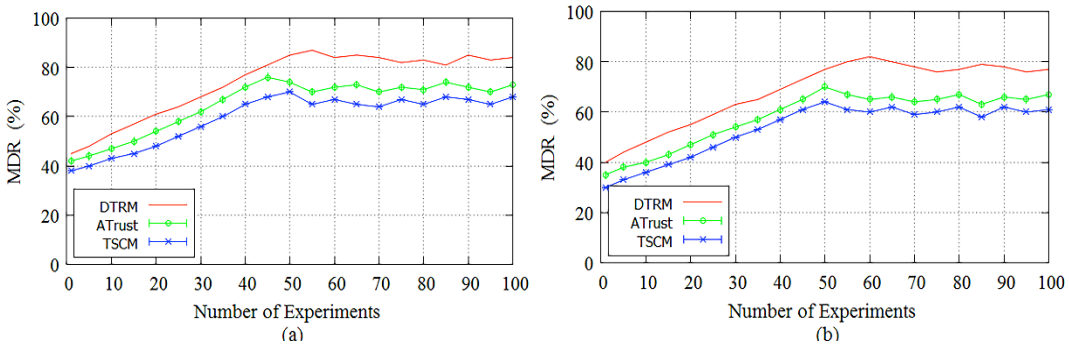

Fig.5. Malicious client detection rate (a) with bad mouthing attacks and (b) with mobile attacks

In Fig. 5 (a) and (b), as expected, the MDR increases with the simulation rounds. It is observed that the MDR of the DTRM is highest among the three mechanisms. This is because that the combination of data category, user group division and reputation transferring scheme improves the accuracy, efficiency, and reliability of the reputation evaluation and thus enhances the MDR. Although the other mechanisms also adopt related methods to improve the accuracy and reliability of the reputation evaluation, they either consider the improvement of the direct reputation evaluation or the improvement of the recommended reputation evaluation in isolation. Moreover, neither of the other two mechanisms consider the impact of the mobile clients' movement on the accuracy and reliability of the reputation evaluation.

5.2.3 False Positive Rate (FPR)

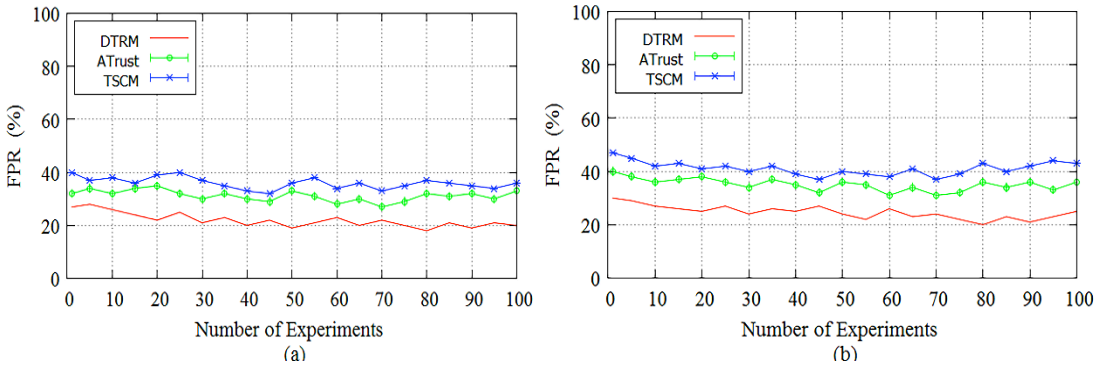

Fig. 6. False Positive Rate (a) with bad mouthing attacks and (b) with mobile attacks

We also evaluate the effectiveness and reliability of the three reputation mechanisms by comparing their FPR performances. The results in Fig. 6 (a) and (b) show that the FPR of all the three mechanisms are less than 40% and 55% respectively, and the FPR of the DTRM is higher than the other two mechanisms. The reason is that the user group division in DTRM can effectively enhance the reliability of the reputation opinion providers and thus improve the accuracy of the reputation evaluation. Meanwhile, the data category ensures that only

those opinions with similar data sensitivity levels and related historical data access categories are accepted, which also improves the accuracy of the reputation evaluation. Moreover, the reputation transferring makes the mobile clients' reputation be valid along with the mobile clients' moving, which decreases the possibility of the malicious mobile clients' reputation re-initialization and further raised the accuracy of the reputation evaluation. ATrust and TSCM did not consider improving the effectiveness and reliability of opinion providers and solving the reputation loss problem during the mobile clients' moving.

5.2.4 Reputation Evaluation Accuracy Rate (REA)

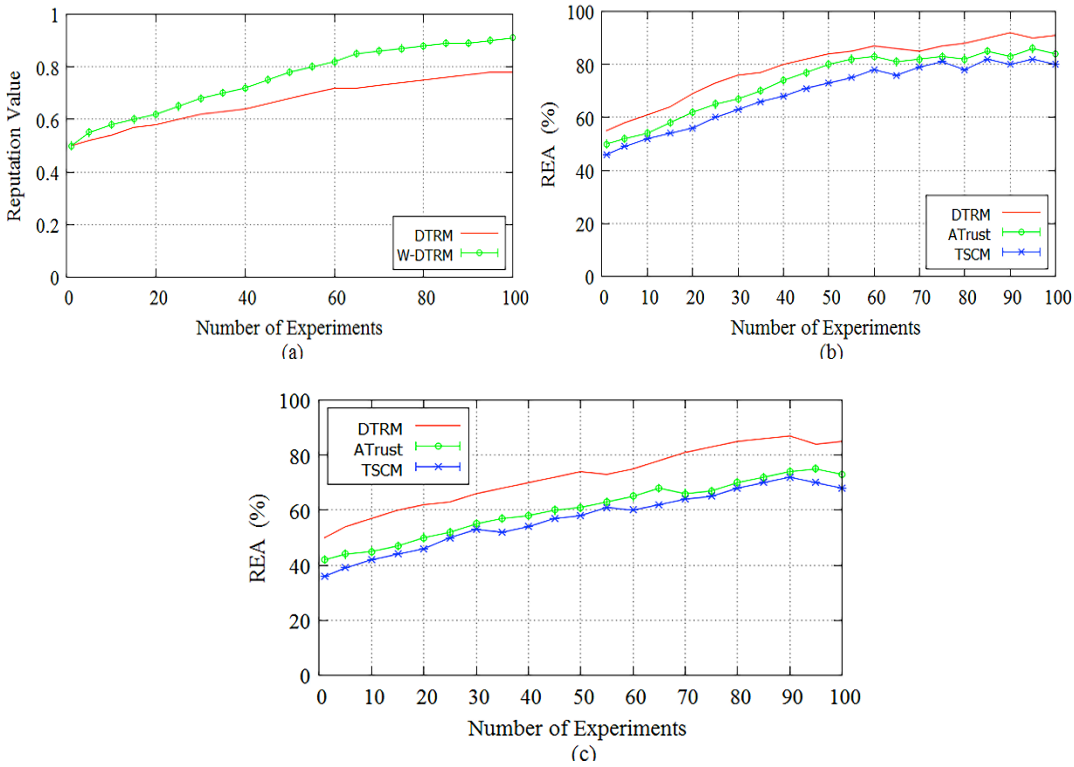

Fig. 7. Reputation evaluation accuracy rate (a) in an honest network, (b) with bad mouthing attacks, and (c) with mobile attacks

Finally, we analyze the reputation evaluation accuracy rate in an honest network and a hostile network, respectively. The results in Fig. 7 (a) show that under the honest network environment, the mobile clients' reputation of the W-DTRM increases faster than the DTRM. The DTRM implements fine-grained reputation evaluation where only the mobile clients that have high relevancy and familiarity with the request can provide their opinions and update their reputation. For W-DTRM, it adopts a coarse-grained reputation evaluation that allows any mobile clients to provide opinions and update their reputation if their opinions happen to be true, which makes the mobile clients' reputation increase faster than under the DTRM

mechanism.

In Fig. 7 (b) and (c), we compare the REA of the three mechanisms with the adversaries present. The percentage of adversaries is set at 30%. In Fig. 5 (b) and (c), as expected, the REA of the DTRM is higher than the other two mechanisms. The reason is that the ATrust or TSCM lacks fine-grained reputation evaluation scheme, therefore, some irrelevant, unfamiliar and unreliable mobile clients' opinions may be accepted and then their reputation will be evaluated or updated, which makes their REAs worse than that of the DTRM. Furthermore, neither the ATrust or TSCM considers the reputation loss problem and they re-initialize the reputation of the adversary moving to a new area as a new client, which also makes their REAs worse than the DTRM.

6. Conclusions

This paper investigated the challenging and important problem of enhancing data trustworthiness for enhancing the veracity of big data, and thus improving the performance of the big data computing. A novel Data Trustworthiness enhanced Reputation Mechanism (DTRM) was proposed to defend against internal attacks, which combines sensitivity-level based data category scheme, metagraph theory based user group division and reputation transferring methods. The real dataset based simulation experiments have demonstrated that the proposed DTRM performs better than those of the existing ATrust and TSCM schemes in terms of the data trustworthiness rate, the malicious client detection rate, the false positive rate, and the reputation evaluation accuracy rate under bad mouthing attacks and mobile attacks. The proposed DTRM scheme can be used as an effective tool to enhance big data veracity for high-performance cloud computing.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (61602360, 61772008), the Pilot Project of Fujian Province (formal industry key project) (2016Y0031), the Foundation of Science and Technology on Information Assurance Laboratory (KJ-14-109) and the Fujian Provincial Key Lab of Network Security and Cryptology Research Fund.

References

[1] A.N. Khan, M.L.M. Kiah, S.U. Khan, S.A. Madani, Towards secure mobile cloud computing: a survey, Future Gener. Comput. Syst. 29 (5) (2013) 1278-1299.

[2] M.B. Mollah, M.A.K. Azad, A. Vasilakos, Security and privacy challenges in mobile cloud computing: Survey and way ahead, J. Netw. Comput. Appl. 84 (2017) 38–54.

[3] H.T. Dinh, C. Lee, D. Niyato, P. Wang, A survey of mobile cloud computing: architecture, applications, and approaches, Wirel. Commun. Mob. Com. 13 (18) (2013): 1587-1611.

[4] L. A. Tawalbeh, R. Mehmood, B. Elhadj, H. Song, Mobile cloud computing model and big data analysis for healthcare applications, IEEE Access, 4 (2016): 6171-6180.

[5] H. Wang, Z. Xu, W. Pedrycz, An overview on the roles of fuzzy set techniques in big data processing: Trends, challenges and opportunities, Knowl-Based Syst. 118 (2017): 15-30.

[6] C.P. Chen, C.Y. Zhang, Data-intensive applications, challenges, techniques and technologies: a survey on big data, Inf. Sci. 275 (2014) 314–347.

[7] S.K. Pal, S.K. Meher, A. Skowron, Data science, big data and granular mining, Pattern. Recogn. Lett. 67 (2015) 109–112.

[8] J. Kepner, V. Gadepally, P. amichaleas, N. Schera, M.varia, Computing on masked data: a high performance method for improving big data veracity, in: High Performance Extreme Computing Conference (HPEC), IEEE, 2014, pp. 1-6.

[9] U. Franke， M. Rosell， V. Vlassov, Towards Automatic Veracity Assessment of Open Source Information, in: 2015 IEEE International Congress on Big Data (BigData Congress), IEEE, 2015, pp.199-206.

[10] Z. Yan, W.X. Ding, X.X. Yu, H.Q. Zhu, R. H. Deng, Deduplication on Encrypted Big Data in Cloud, IEEE Trans. Big Data. 2(2) 2016 138-150.

[11] S. Yin, O. Kaynak, Big data for modern industry: challenges and trends, Proceedings of the IEEE. 103(2) (2015) 143-146.

[12] M. Pouryazdan, B. Kantarci, T. Soyata, H. Song, Anchor-assisted and vote-based trustworthiness assurance in smart city crowdsensing, IEEE Access, 4 (2016): 529-541.

[13] B. Kantarci, H.T. Mouftah, Trustworthy sensing for public safety in cloud-centric internet of things, IEEE Internet of Things Journal, 1(4) (2014) 360-368.

[14] T. Bodnar, C. Tucker, K. Hopkinson, S.G. Bilén, Increasing the veracity of event detection on social media networks through user trust modeling, in: 2014 IEEE International Conference on Big Data (Big Data), IEEE, 2014, pp. 636-643.

[15] B. Agarwal, A. Ravikumar, .S. Saha, A Novel Approach to Big Data Veracity using Crowdsourcing Techniques and Bayesian Predictors, in: Proceedings of the 9th Annual ACM India Conference, ACM, 2016, pp.153-160.

[16] M. Kumar, N. K. Rath, S. K. Rath, Analysis of microarray leukemia data using an efficient MapReduce-based K-nearest-neighbor classifier, Journal of biomedical informatics, 60 (2016): 395-409.

[17] J. Debattista, C. Lange, S. Scerri, S. Auer, Linked'Big'Data: towards a manifold increase in big

data value and veracity, in: 2015 IEEE/ACM 2nd International Symposium on Big Data Computing (BDC), IEEE, 2015, pp.92-98.

[18] H. Lin, J. Hu, Y.L. Tian, L. Xu, Toward better data veracity in mobile cloud computing: A context-aware and incentive-based reputation mechanism, Information Sciences, 387 (2017) 238-253.

[19] D. Vatsalan, Z. Sehili, P. Christen, E. Rahm, Privacy-Preserving Record Linkage for Big Data: Current Approaches and Research Challenges, Handbook of Big Data Technologies, Springer International Publishing, 2017, pp. 851-895.

[20] M. Kim, O. P. Sang, Trust management on user behavioral patterns for a mobile cloud computing, Cluster Computing, 16 (4) (2013) 725–731.

[21] H. Shen, G. Liu, an efficient and trustworthy resource sharing platform for collaborative cloud computing, IEEE Transactions on Parallel and Distributed Systems, 25(4) (2014) 862-875.

[22] Y. Zhang, M. Schaar, Incentive provision and job allocation in social cloud systems, IEEE Journal on Selected Areas in Communications, 31(9) (2013) 607-617.

[23] S.H. Chang, Z.R. Chen, Protecting Mobile Crowd Sensing against Sybil Attacks Using Cloud Based Trust Management System, Mobile Information Systems, 2016 (2016) 1-10.

[24] N. Palaghias, N. Loumis, S. Georgoulas, K. Moessner, Quantifying trust relationships based on real-world social interactions, in: 2016 IEEE International Conference on Communications (ICC), IEEE, 2016, pp.1-9.

[25] H. Lin, L. Xu, Y. Mu, W. Wu, A reliable recommendation and privacy-preserving based cross-layer reputation mechanism for mobile cloud computing, Future Generation Computer Systems, 52 (2015) 125-136.

[26] D. Zissis, D. Lekkas, Addressing cloud computing security issues, Future Generation Computer Systems, 28 (3) (2012) 583–592.

[27] H. Lin, J. Hu, J.F. Ma, L. Xu, L. Yang, CRM: a new dynamic cross-layer reputation computation model in wireless networks, Computer Journal, 58.4 (2014) 656-667.

[28] H. Wang, Z. Xu, H. Fujita, S. Liu, Towards felicitous decision making: An overview on challenges and trends of Big Data, Information Sciences, 367 (2016) 747-765.

[29] E. Bertino, E. Ferrari, Big data security and privacy, A Comprehensive Guide Through the Italian Database Research Over the Last 25 Years, Springer, 2018, pp. 425- 439.

[30] J. Zhang, Z. Zhang, H. Guo, Towards Secure Data Distribution Systems in Mobile Cloud Computing, IEEE Transactions on Mobile Computing, 99 (2017): 1-12.

[31] M. Ezhei, B. T. Ladani, Gtrust: A group based trust model, The ISC International Journal of Information Security, 5(2) (2014) 155-169.

[32] X. Wang, W. Cheng, P. Mohapatra, T. Abdelzaher, Artsense: Anonymous reputation and trust in participatory sensing，in: 2013 Proceedings IEEE INFOCOM, IEEE, 2013, pp. 2517-2525.

[33] Y. Wu, G. Min, K. Li, B. Javadi, Modeling and analysis of communication networks in multicluster systems under spatio-temporal bursty traffic, IEEE Transactions on Parallel & Distributed Systems, 23(5)(2012) 902-912.

[34] Y. Wu, G. Min, D. Zhu, An analytical model for on-chip interconnects in multimedia embedded systems, ACM Transactions on Embedded Computing Systems, 13(1s) (2013) 1-29.

[35] W. Miao, G. Min, Y. Wu, H. Wang, J. Hu, Performance modelling and analysis of software-Defined Networking under Bursty Multimedia Traffic, ACM Transactions on Multimedia Computing Communications & Applications, 12(5s) (2016)1-77.

[36] C. Huang, G. Min, Y. Wu, Y. Ying, K. Pei, Z. Xiang, Time series anomaly detection for trustworthy services in cloud computing systems, IEEE Transactions on Big Data, 99(2017)1-1.

[37] J. Zhang, J. Tang, J. Li, Y. Liu, C. Xing, Who influenced you? predicting retweet via social influence locality, ACM Transactions on Knowledge Discovery from Data (TKDD), 9(3) (2015)1-25.