QnQ: A Reputation Model to Secure Mobile Crowdsourcing Applications from Incentive Losses

Shameek Bhattacharjee, Nirnay Ghosh, Vijay K. Shah, and Sajal K. Das
Department of Computer Science, Missouri University of Science and Technology, Rolla, USA
{shameek, ghoshn, vksr38, sdas}@mst.edu

Abstract*—A major limitation of mobile Crowd Sourcing (CS) applications is the generation of false (or spam) contributions due to selfish and malicious behaviors of users, or wrong perception of an event. Such false contributions induce loss of revenue through disbursement of undue incentives and also negatively affects the application's operational reliability. In this work, we propose a reputation model, called QnQ, to segregate different user classes such as honest, selfish, or malicious based on their reputation scores. The resultant score is then used as an indicator to decide an incentive for a user. Unlike existing works, QnQ ensures fairness to different user behaviors by *unifying 'quantity' (degree of participation) and 'quality' (accuracy of contribution)*. Specifically, QnQ utilizes evidences from a rating feedback mechanism to propose an event-specific expected truthfulness metric by considering total feedback volume, probability mass for positive evidence, and the discounted probability mass of uncertain evidence. To classify an event as true or not, a generalized linear model is used to transform its truthfulness into *quality of information (QoI)*. Finally, the QoIs of various events in which a user participates, are aggregated to compute a user's reputation score. For evaluation of QnQ through experimental study, we consider a vehicular crowdsourcing application. QoI performance of our model is compared with Jøsang's belief model, while reputation and incentive leakage is compared with Dempster-Shafer based reputation model. Experimental results demonstrate that QnQ is able to better capture subtle differences in user behaviors by unifying both quality and quantity, and significantly reduces undue incentives in presence of rogue contributions.

***Index Terms*—Participatory sensing, Trust, Reputation, Secure Crowd sourcing, Security Economics, Vehicular Crowd Sensing,**

I. Introduction

Sophistication in mobile devices (e.g., smartphones, tablets) and their widespread adoption have given rise to a novel interactive sensing paradigm, known as *Participatory Sensing (PS)* or *Crowd Sourcing (CS) [5]*. In CS systems, a crowd of citizens voluntarily submit certain observations termed as *contributions* (viz., report, image, audio) about their environment to a CS application server, which thereafter fuses such contributions to conclude a summarized statistic (*or information*) and publishes it to support improved decision making.

An important category of CS applications is vehicular traffic management and monitoring [3]. In such applications, a user's contributions are equivalent to 'reports' about various road conditions that they might have observed. Based on certain correlations among such reports, the CS application decides whether a certain traffic 'event' has occurred, and publishes this 'information' as a broadcast notification on the smartphone application. Such information improves driving experiences through dynamic route planning and re-routing of traffic in busy cities. Two notable examples of real vehicular CS applications include Google's *Waze* and *Nericell* [15]. Other practical examples of CS applications are *FourSquare*, and *Yelp* which help users to find best destinations in their geographical proximity for food, entertainment, and other attractions or events of interest. The real benefit of CS applications is that fine grained and precise sensory observations can be obtained *quickly* without depending on the deployment of expensive and dedicated infrastructures [17]. However, the major drawback is its "open" nature (accessible to all) which may expose such applications to false contributions [9] [21].

Most of the CS applications need to use various incentive mechanisms to motivate the users to keep contributing regularly, and thus preserve their viability [13]. It has been noted that in most of these mechanisms, the deciding factor of incentive is the user's *degree of participation* (i.e. "quantity" or how much they contribute). However, *selfish users* may take advantage of this loophole and intermittently generate false contributions to boost their participation for gaining undue incentives [17], incurring revenue losses to the CS system. Furthermore, there could be *malicious users* who attempt to cripple the CS applications by generating a large number of bogus contributions in collusion [21]. Recently, such colluding attack was launched against Waze in Israel, by which fake traffic jam reports were created to orchestrate traffic re-routing and unnecessary roadblocks [19]. Occasionally, false contributions may also be generated owing to wrong perception. Regardless of the motive, false contributions incur loss of revenue due to unnecessary disbursement of incentives and also tarnishes operational reliability of the CS application.

In our preliminary work, we studied a real data set from Waze [3], and established that the 'quantity' rather than 'quality' of contributions decides incentives (details presented in Section II-B). We argue that besides the quantity, there is also a simultaneous need for assessing *quality of information (QoI)* generated from user contributions. This *QoI is essentially a measure of the trustworthiness* of the summary statistic and is equivalent to its *trust score*. Additionally, user reputation based on his level of truthful cooperation is required to determine:

978-1-5386-0683-4/17/$31.00 ©2017 IEEE

(i) if a user is selfish or malicious, (ii) the incentive received by the user, and (iii) acceptance of future reports from him.

A. Motivation

Apart from expensive ground truth based monitoring (or truth discovery), an easier way to assess QoI is to allow other users in the proximity to provide a feedback rating (viz., positive, negative, uncertain) for each published information [11] [17]. Based on such feedbacks, QoI and reputation are quantified. Now, let us synthesize some of the limitations of existing QoI and reputation models which are usually based on Beta and Dirichlet distributions (such as Jøsang's belief model [11], Dempster-Shafer reputation [22]), or their variants (refer to Section II for further details):

First, existing works only utilize the proportion of positive feedbacks in the QoI measure. However, we show that accurate QoI scoring should also include the effect of total number of feedbacks (i.e., feedback mass) that a published information has received. This step is important to weaken the success rates of malicious ratings. Second, unlike most existing works we consider a dynamic discounting of uncertain feedbacks to ensure that the QoI measure is null invariant (i.e., not influenced by high uncertainty or inconclusive feedbacks). Third, our model is able to propose a reputation score for each user that unifies both his degree of participation (quantity) as well the quality of each contribution. Such a reputation score provides dual benefits: (i) it segregates different types of users, viz., *honest, selfish,* and *malicious*, and (ii) it forms a basis to judiciously incentivize or penalize based on varying behaviors. Finally, our design principle is free from the *cold start* problem and learns from the evidences, rather than from history.

B. Contributions of the Paper

This paper proposes a novel model, called QnQ, for trust and reputation scoring in a CS system in presence of malicious and selfish users. First, based on the feedbacks received over a particular published information (or event), we calculate the Bayesian inference based belief, disbelief, and uncertainty masses. Thereafter, we model the expected truthfulness of the published event as a regression score using generalized *Richard's equation* and *Kohlsrausch relaxation function* as the weights to the belief and uncertainty masses, respectively. This step weakens the effect of malicious and incorrect feedbacks while being null invariant. Subsequently, we map the expected truthfulness to a QoI measure using the logit link function that quantifies the possibility of the event's occurrence.

Next, we keep track of the QoI measures of all the published events for which a given user had contributed (through reports), and then calculate a raw user reputation score by aggregating them. Eventually, we compute a normalized user reputation score by normalizing the aggregated score (within the interval [-1, +1]) through a logistic distribution function. This normalized user reputation score is utilized for classification and judicial disbursement of incentives based on both quality and quantity of his contribution.

Finally, we conduct extensive performance evaluation of the proposed QnQ model using a vehicular crowdsourcing system as a proof-of-concept. We demonstrate that QnQ outperforms Jøsang's belief and Dempster-Shafer (D-S) based reputation models in terms of classification and incentives. Experimental results show that QnQ is able to give a reputation score, that rewards both quality and quantity and reduces undue incentives in presence of dishonest users while ensuring fairness.

II. Limitations of Existing Work

This section reviews the state-of-the-art for QoI and user reputation scoring models, followed by discussion on certain important limitations of existing literature for CS scenarios under selfish and malicious users.

A. Quality of Information (QoI)

Research in QoI is broadly classified under: (i) improvement of quality and (ii) estimation of quality. Improving QoI is achieved either through incentive mechanisms [23] or selection of appropriate sensing agents [6]. The incentive mechanisms in CS motivate users to continue furnishing contributions (reports) in lieu of monetary or non-monetary rewards (viz., entertainment, educational opportunities). In contrast, estimation of quality aims at assessing the 'veracity' of the information, once it is received from the users [20]. The veracity assessment may be either on the individual reports or on the inferred information statistic. Broadly, the QoI is assessed by modeling evidence obtained from ground truth or feedback mechanisms.

Ground truth monitoring [10] compares user contributions with a ground truth generated by a mobile trusted agent or a watchdog physical infrastructure. If the contribution matches with the agent or the watchdog's input, it is considered good else bad. Beta distribution [12] is often used to model such binary evidences into a QoI score. However, availability of ground truth is not immediate, often not guaranteed, and sometimes not feasible. Additionally, acquiring ground truth often requires deployment of dedicated, specialized infrastructure, or agents undermining the real benefits of crowdsourcing.

Some real CS applications, viz., FourSquare, Waze, Yelp, use a rating feedback mechanism, whereby other consumers of the service provide positive, negative or neutral ratings on the published information. The estimation of QoI is achieved based on the feedbacks received. In most cases, estimation of QoI is achieved using Jøsang's belief models [11] that computes the QoI based on the ratio of positive feedback to the total feedback with some fixed weight to the ratio of uncertain feedbacks. The benefits of using a feedback rating paradigm is that it is easy, fast, less expensive and really exudes the essence of a true mobile crowd sourcing and participatory sensing paradigm. Nevertheless, we observe the following inherent weaknesses in Beta reputation and Jøsang's belief models:

Confidence of the Feedback Community: Jøsang's belief model for expected Bayesian belief (E^J) is given as: $E^J = b + a.u$, where $b = \frac{r+1}{r+s+t+3}$; $d = \frac{s+1}{r+s+t+3}$;

Table I: Limitations of Jøsang's Belief Model

Issues	Examples	Jøsang's QoI	Comment
Confidence of Rating Community	E1:$\langle 7, 3, 2, 2\rangle$ E2:$\langle 70, 30, 20, 20\rangle$	0.55 0.57	Negligble difference in QoI. Does not account for feedback mass
Not Null Invariant	E3:$\langle 105, 5, 0, 100\rangle$ E4:$\langle 25, 5, 0, 20\rangle$	0.51 0.53	QoI is quite high in $E3$ even as most ratings are "undecided"

$u = \frac{t+1}{r+s+t+3}$; and r, s, and t denote the number of positive, negative, and uncertain ratings. $a = 0.5$ is the relative atomicity which is equal to the reciprocal of the cardinality of inference state space $\{true, false\}$ [11]. However, this model fails to capture the confidence of the feedback community, which makes the resultant expected belief (QoI in our case) more vulnerable to manipulation by malicious raters who provide positive ratings (respectively negative) to false (respectively true) contributions. This may influence the QoI score in favor of the adversaries, as shown in Table I, where each event is denoted as $E : \langle N, r, s, t\rangle$, such that, N is the total number of received ratings.

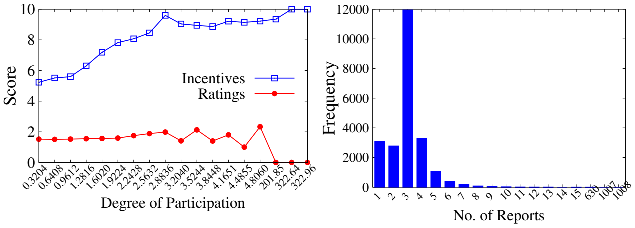
For event $E1$, 3 out of 7 total feedbacks are good, whereas $E2$ has 30 such feedbacks out of 70. Jøsang's belief model measures almost the same QoI for both examples. From an adversary's perspective, it is easy to compromise/manage 3 good raters in $E1$ and maintain the same fraction of positive ratings as $E2$. However, it is harder to maintain the same fraction when the crowd is large (as in $E2$), and the adversary has to manipulate 30 raters. Hence, given the same fraction of positive feedbacks, any event with more feedbacks should be considered as more trustworthy. If this feature is not incorporated, QoI becomes more vulnerable to feedback manipulation attacks, viz. *ballot stuffing* (See Sec. III-B).

Not Null Invariant to Uncertainty: Jøsang's belief models do not offer *null invariance* property. By this problem, QoI of an event can also be increased due to high proportion of undecided feedbacks, which may be either intentionally generated, or be a result of legitimate uncertainty. Either way such events should not unduly increase the trust (QoI) score.

For Example, event $E3$ in Table I has 105 feedbacks, out of which 100 are uncertain; however it achieves almost the same QoI as $E4$ which in contrast has only 20 uncertain ratings. For most conservative services, it may be risky or unwise to give as high a QoI score to $E3$ as to $E4$. Thus, the QoI measure needs a mathematical provision for controlling the impact of high uncertainty on the QoI.

B. User Reputation Scores

Traditionally, reputation scoring models in crowd sourcing use either Beta or Dirichlet distributions as theoretical basis for probabilistic trust modeling [16]. Jøsang's belief and D-S based models are the state-of-the-art approaches that exploit either of these distributions to model evidences into trust or reputation scores. Recent works [9], [2] [16], use a *Gompertz* function based deterministic time-based reputation management, rather than evidence based scoring. Moreover, they assume that a prior accurate reputation exists, cannot unify quality and quantity of participation, has no provision to handle uncertainty, and cannot thwart the effect of rogue ratings. Most of the recent works do not consider active dishonest reports or ratings and do not consider economic incentives attached with the reputation dynamics. In summary, the state-of-the-art works still depend on either Beta or Dirichlet distribution-based Jøsang belief and D-S based-reputation models [22] as the core evidential reputation scoring component. Hence, this work seeks to compare our results with these models. Some limitations of these models identified by us are as follows:



(a) Quality vs Quantity (b) Report Generation Frequency

Figure 1: Study on Waze Study

Sacrificing Participation for Quality: D-S model [22] does not simultaneously capture degree of participation and quality into the reputation score. This limitation of D-S model is depicted in Table II. Although users 1 and 2 have the same reputation, the latter has much higher number of contributions. Even with 52 additional good contributions, user 2 ends up with a score almost similar to User 1. This is grossly unfair as it undermines the higher participation of users. If the maximum reputation is attainable with lower contributions, users may not be motivated enough to participate more, and CS application will underachieve its possible potential.

Table II: Sacrificing Participation for Quality

User	Participation	Good	Bad	Dempster Score
1	9	9	0	0.99
2	61	61	0	1.00
3	20	18	2	0.99

Sacrificing Quality for Participation: In [3], we studied a real data set from Waze and identified that quality may be sacrificed for participation. Fig. 1b shows that the majority of the users have generated around three reports over the span of one week. However, there are a few users who have generated a very high number of reports (around 600 to 1000). Additionally, it is evident from Fig. 1a that the incentive of the users gradually increases with higher participation rate. Conversely, the ratings assigned to the users with high participation are very low and even drops to zero while maximum incentive is received. Thus, the reputation score of a user needs to *unify* both degree and quality of participation.

III. System and Threat Models

In this section, we present the system and threat models.

A. System Model and Design

As depicted in Fig. 2, our system model captures a particular city area which may consist of U users, all equipped with smart mobile devices and subscribed to a vehicular CS application.

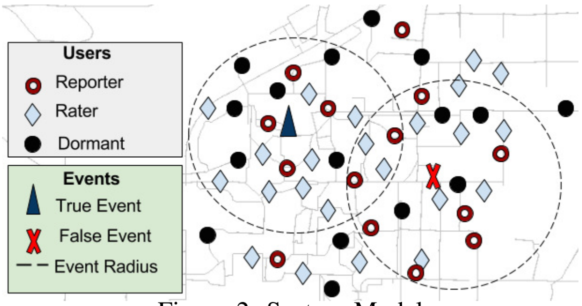


Figure 2: System Model

Two important components of this system are:

Report: A report is an alert furnished by an user after he perceives an incident (viz., accident, jam, road closure). However, due to the presence of selfish and malicious users, there may be reports generated for incidents that have never occurred.

Event: An event E_k, $k \in \{1, \cdots, K\}$ is a summarized information which is published on a live map after the CS application receives a certain number of similar reports, where K is the total number of events observed over the system lifetime. If reports from two different users indicate similarity in terms of location, time epoch, and incident type, they indicate the 'same' event.

In our model, there exists two types of users:

Reporter: A reporter is a user who has propensity to generate reports and has reported at least one incident. Any such user is liable to have a reputation score which reflects the overall quality of reports contributed as well as the degree of participation. Incentives are decided based on this score. To remove bias from his feedbacks, a reporter is not allowed to rate a published event for which he had generated a report.

Rater: A rater is a user who provides feedback on his perceived usefulness of an event through one of these categories: *Useful* (α), *Not Useful* (β), and *Not Sure* (γ) (e.g. in Foursquare).

For a published event, a rater is allowed to submit only one rating. In our system model, we view the act of providing ratings as an obligation, and it is not rewarded. Hence, for majority of the normal users, there is no selfish incentive to provide false ratings. However, false ratings could be motivated by malicious intents. Collecting as many feedbacks as possible is important in absence of the ground truth. This may be achieved by auto-activation of a pop-up rating query during an active rater's navigation through the adjacent region of the event, provided that he has not reported for the event under scrutiny. The rating query asks rater to click one of three possible options: *Useful*, *Not useful*, and *Not sure*, to gather a subjective judgment about the published event. Our model can be easily extended to 5-star rating systems, by visualizing (4, 5), (1, 2) and (3) as 'useful', 'not useful' and 'not sure'.

B. Threat Model and Assumptions

Among the dishonest reporters, there are selfish and malicious ones. A selfish user is a legitimate user who generates true and false reports intermittently, with certain probabilities for maximizing his incentives. Such selfish user acts alone or at best collude in a small scale. Malicious users are either actual user devices compromised and controlled by an adversary or are hardware emulators that masquerade themselves as real devices to the CS system. A proportion of the compromised devices can be made to act in collusion as reporters while the remaining as raters. Malicious raters act in the following ways:

Ballot stuffing: A rater submits positive feedbacks to an incorrect (false) published event generated by dishonest reporters.

Bad mouthing: A rater submits negative feedbacks to a legitimate published event generated by honest reporters.

Note that, hardware emulators can further generate numerous sybil interfaces to magnify the problem of false reports and ratings. However, such sybil entities could be identified by existing methods [21]. Hence, we assume that it is only the hardware emulators or compromised devices which pose a real threat of colluding attacks. In general, the adversary has a constrained attack budget by which it can compromise/deploy only a limited number of devices/emulators that act as either reporters and raters. This is evident from [21], where authors generated 1000+ sybil (virtual) interfaces to collude a Waze-like application, but were compelled to deploy only 10 emulators (physical systems) due to budgetary constraint. However, in most cases, with larger rater populations, the rating mechanism becomes less likely to get sabotaged. For any rating-based system, the number of raters is always higher compared to the number of reporters generating reports/reviews, which is evident from the *Epinions* dataset [14]. It shows that the number of feedbacks is roughly three to four times the number of reviews (reports) for any item.

The adversary uses this constrained attack budget to manipulate a fraction δ_{mal} of the reporters and raters to generate false reports and ratings. It will be a significant fraction for scenarios with limited number of legitimate users. However, in presence of a significant crowd of independent users, δ_{mal} will be low, and it will not be possible to sabotage the entire proportion of genuine feedbacks. Since Crowd Sensing applications are meant for urban spaces, we assume that for majority of times, substantial number of authentic raters are likely to be present in the vicinity of an event, thus reducing the proportion of false ratings to the total feedbacks.

IV. QnQ: Proposed Reputation Scoring Model

Now we present the modules of the proposed reputation scoring model, called QnQ.

A. Posteriori Probability Masses

The first step is to derive the expressions for the posteriori probability masses associated with feedbacks: *Useful*, *Not Useful*, and *Not Sure*. The probability masses are estimated for each event E_k based on the available evidence (i.e., supporting each rating type), using a classical Bayesian approach For simplicity, we drop E_k from all the notations. Let $\bar{\omega} = \{\omega_\alpha, \omega_\beta, \omega_\gamma\}$ be the three tuple probability parameter to be estimated. Here, $\omega_\alpha, \omega_\beta, \omega_\gamma$ are the unknown probabilities of observing a *Useful*, *Not Useful*, or *Not Sure* feedback, respectively. We denoted $H(\bar{\omega})$ as the hypothesis, such that it has three possibilities of either taking α , β or γ. Formally, $P(H(\bar{\omega}) = \alpha|\bar{\omega}) = \omega_\alpha$, $P(H(\bar{\omega}) = \beta|\bar{\omega}) = \omega_\beta$, $P(H(\bar{\omega}) = \gamma|\bar{\omega}) = \omega_\gamma$. Let F_α, F_β, and F_γ be the random variables denoting the number of feedbacks η_α, η_β, and η_γ, received for each feedback category, respectively, such that $N = \eta_\alpha + \eta_\beta + \eta_\gamma$. The *evidence* vector, denoted as $F(N) = \{F_\alpha, F_\beta, F_\gamma\}$, should be modeled as a multi-nomial distribution given by:

$$P(F(N)|\bar{\omega}) = \frac{N!}{\eta_\alpha!\eta_\beta!\eta_\gamma!}\omega_\alpha^{\eta_\alpha}\omega_\beta^{\eta_\beta}\omega_\gamma^{\eta_\gamma} \tag{1}$$

The posteriori hypothesis of positive outcome based on the evidence vector and assumed prior is given as:

$$P(H(\bar{\omega}) = \alpha|F(N)) = \frac{P(H(\bar{\omega}) = \alpha, F(N))}{P(F(N))} \tag{2}$$

Similarly, the posteriori hypothesis of negative and uncertain outcomes can be represented by replacing α with β and γ respectively in Eqn. (2).

Solving the above (see [4]), belief, disbelief, and uncertainty probability masses are derived as follows: $P(H(\bar{\omega}) = \alpha|F(N)) = \frac{\eta_\alpha+1}{N+3} = b$, $P(H(\bar{\omega}) = \beta|F(N)) = \frac{\eta_\beta+1}{N+3} = d$, and $P(H(\bar{\omega}) = \gamma|F(N)) = \frac{\eta_\gamma+1}{N+3} = u$, respectively. These are the posteriori probability masses for *Useful*, *Not Useful*, and *Not Sure* feedbacks as perceived by the raters, respectively. Note that, when $\eta_\alpha = \eta_\beta = \eta_\gamma = 0$, all the possibilities are equiprobable under no information (non-informative prior).

B. Expected Truthfulness of an Event

Contemporary research regards trust as a way of choice under uncertainty and risk [8]. Hence, it is natural that trustworthiness of an event should account for uncertain evidences apart from the positive evidences [11]. Thus, we propose w_b and w_u as the coefficients (or weights) of belief and uncertainty masses respectively, where weights control the extent to which positive and uncertain probability masses contribute to the truthfulness score. The problem is modeled similar to a weighted regression approach where probability masses are explanatory variables and the expected truthfulness is a response variable. We apply Richard's generalized curve and Kohlrausch relaxation functions to model w_b and w_u. The expected truthfulness for any published event E_k (denoted as k for ease of representation) will be given as:

$$\tau_k = (w_b).b + (w_u).u \tag{3}$$

where, $0 < \{w_b, w_u\} < 1$. Hence, $0 < \tau_k < 1$.

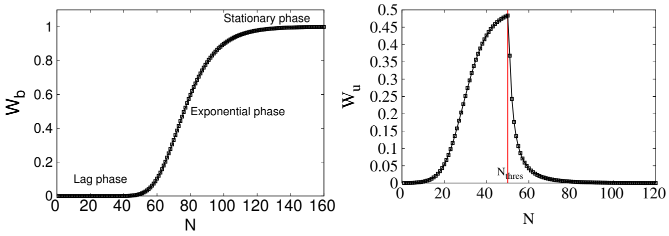
1) Design of Belief Coefficient (w_b): We mentioned that expected truthfulness should also consider the volume of the feedbacks, i.e., how many feedbacks have been received for an event apart from the belief mass b. Intuitively, lesser N (total number of feedbacks/ratings) should have low w_b, which in turn, contributes to a lesser expected truthfulness. However, w_b should gradually increase with N. Thus, to model this nature of w_b, we use a Generalized Richard's Equation normalized between 0 and 1 as:

$$w_b = \frac{1}{(1 + A_b e^{-B_b N})^{1/\nu}} \tag{4}$$

where $A_b > 0$, $A_b \neq \infty$ is the initial value of the coefficient, and B_b is the rate of growth, $\nu \neq 0$ is the parameter controlling the point where the curve enters into exponential growth.

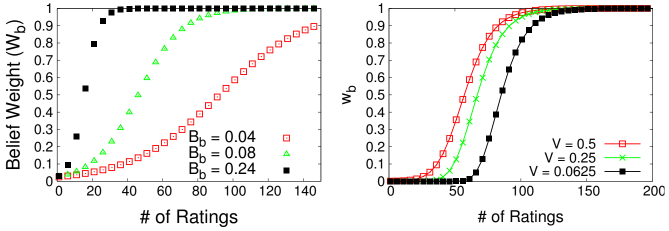
The nature of w_b is given by the sigmoid function (refer to Fig. 3a). More conservative (or high risk applications) systems will have less A_b and B_b to control rapid growth of w_b. This provision enables the CS administrator to nullify the effects of ballot stuffing. QoI based on only belief masses usually get influenced in the favor of the adversaries, if there are limited number raters in the system. Thus, under fewer number of feedbacks, w_b will be low, and it will progressively increase (in an exponential manner) if more feedbacks are available, eventually saturating after sufficient number of ratings are received. Presence of larger crowd of independent raters makes it impossible to sabotage the significant proportions of positive feedback in the adversary's favor.

Physical Significance of Choosing Richard's Curve: Using Richard's generalized curve as w_b is motivated from deductive reasoning and developmental learning studies in cognitive psychology. Intelligent humans are subconsciously rational enough to know that possibility of a bias negatively affecting a belief inference is more, if less number of people say the same thing, as opposed to the same thing endorsed by more number of people. Hence, an inference backed by more people/feedbacks is more trusted than the same inference backed by less people. The increasing confidence of belief with higher support can be modeled through *incremental change processes* [8], [18]. Mathematically modeling such incremental change processes is essential for most studies in belief learning problems and is done by exponential growth functions [18]. These models are characterized by a slower initial phase followed by an *inflection point* where the learning rate exponentially peaks in the face of increasing evidences and finally saturates into a stationary phase where the learning rate approaches an upper asymptote. Since, we want to award the confidence of the rating community to the belief mass, we model weight w_b through Richard's generalized curve from the family of exponential growth models. The advantage of Richard's curve over unlike existing models is that, it provides mathematical provisions to control point of inflection ν, rate of change to maximum value after the inflection B_b, and controlling the initial lower asymptote A_b.



(a) Belief: w_b (b) Uncertainty: w_u

Figure 3: Illustrations of Coefficients



(a) Growth Rate: B_b (b) Effect of Various ν

Figure 4: Parameters of Richard's Equation

Choice of A_b***,*** B_b ***and*** ν: The initial asymptote is the base value of the weight given to belief mass when no rating is received. If the system is not restrictive, then a higher initial weight w_b is required, and hence a lower value of A_b is recommended. In contrast, for a conservative system, the initial weight of w_b should be very low, to ensure that it should acquire a sufficient number of ratings before attaining a substantial weight.

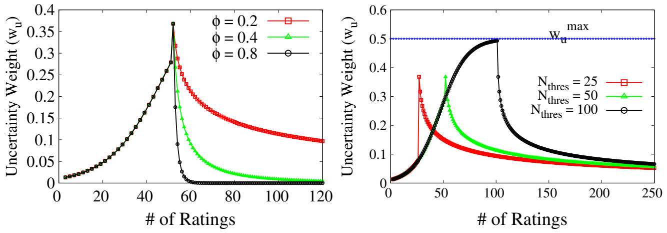
Fig. 4a shows the effect of B_b that controls the number of ratings N required to attain the maximum possible value of w_b once it enters the exponential phase. For example, if the concerned area is inherently crowded and higher N is expected, then B_b should be kept low such that full weight to w_b is awarded only after a sufficient number of ratings are received. If the system is less restrictive, it can lower the value of B_b.

The ν controls the value of N at which the curve first enters into the exponential growth phase. A lower value of ν is preferred if a CS administrator expects receipt of false ratings, or if the location historically receives lower number of ratings. Fig. 4b shows the different values of ν.

2) *Design of the Uncertainty Coefficient (*w_u*):* In Eqn. (3), w_u controls the contribution of uncertainty mass to the effective truthfulness. Intuitively, uncertainty is high if an incident has just occurred, and the majority of users are uninformed. However, it gets reduced as more feedbacks are received. Thus, for smaller values of N, we should have an increasing function for w_u and as this is also similar to growth curve, we model by a Richard's function bounded at w_u^{max}. However, once N attains a threshold value, say $N = N_{thres}$, the coefficient should start to decrease. The value of N_{thres} and w_u^{max} depends on the empirical data of relevant application scenario and risk attitude.

Kohlsrausch relaxation function [1] is used to model property of a system that evolves towards equilibrium after sudden perturbation or a trigger. After trigger point $N = N_{thres}$, we use this function to model the discounting effect of uncertain ratings on trustworthiness thus ensuring null invariance. Its parameter $0 < \phi < 1$ controls the rate of discounting of w_u over N. The larger the value of ϕ, the quicker is the decrease (refer to Fig. 5a). The following equation gives the variation of w_u with respect to the number of received feedbacks:

$$w_u = \begin{cases} \frac{w_u^{max}}{(1+A_u e^{-B_u N})^{1/\nu}}, & \text{if } N < N_{thres} \\ e^{-(N-N_{thres})^{\phi}}, & \text{if } N \geq N_{thres} \end{cases} \tag{5}$$



(a) Kohlrausch Factor ϕ (b) N_{thres}

Figure 5: Impact of Parameter Choices on w_u

where A_u and B_u are respectively the corresponding asymptote and growth parameters (as discussed in Eqn. (4)). $0 < w_u^{max} < 1$ is a fixed parameter controlling maximum allowable benefit of doubt for an event. Choice of w_u^{max} may be guided by risk attitude or availability of trusted agents [17].

Physical Significance of Choosing Kohlrausch Function: Earlier we mentioned that the interpretation of uncertainty in the feedbacks is different from that of evidence without uncertainty. Especially for decisions that are objective, people tend to give a benefit of doubt while trusting something if uncertainty is reported from a small number of people. But if there is high uncertainty even as more people have participated, the effect of this uncertainty does not contribute to the increase of trust, since the risk perception is magnified [8]. The uncertainty involves a trigger point or a *knot point*, about which there is a relatively brisk reorientation of the existing state of benefit of doubt into a qualitatively different state of discounting the benefit of doubt. Such phenomena in belief learning and developmental theory is known as *transformational change processes* [18]. They fit into a family of spline curves and such phase transitions are modeled by multiple equations around the knot point [7]. The nature of w_u mimics such effects on the interpretation of uncertainty.

Choice of ϕ***,*** N_{thres} ***and*** w_u^{max}: Kohlrausch factor ϕ determines how quickly w_u's discounting effect reaches its minimum after N_{thres} is reached. Fig. 5a shows the effect of various choices of ϕ. A CS administrator chooses a higher value of ϕ if proportion of uncertainty needs to be immediately discounted or vice-versa. Effects of A_u and B_u to w_u are similar to that of A_b and B_b to w_b.

A small N_{thres} would prevent w_u to reach its maximum value, before the uncertainty discounting starts. This is true for more conservative systems and is evident from Fig. 5b. A low w_u^{max} may be required when administrator comes to know about the ground truth (from other sources such as mobile trusted participants [17]), and does not want uncertainty mass to obtain higher weights.

C. QoI of Published Event

τ_k is the expectation that the published event E_k has actually happened. Now, the system needs to perform a regression to determine the odds of E_k being true or false which we model as the QoI. We have used the generalized linear models (GLM) for this purpose. When response/predictor variables is categorical (true/false, yes/no, etc.) with non-normal error distribution, we need a *link* function to provide the relationship between the predictor variable (linear) and the mean of the distribution (explanatory) defining the QoI. Thus, if Q_k is the response and τ_k is the mean, the link between them is established by the following *logit* function:

$$Q_k = \ln\left(\frac{\tau_k}{1-\tau_k}\right) \tag{6}$$

Q_k is the QoI of the event E_k which has value in the interval $[-\infty, +\infty]$. The logit function gives monotonically decreasing weights to all $\tau_k < 0.5$, and monotonically increasing weights for $\tau_k > 0.5$. Finally $Q_k = 0$, if $\tau_k = 0.5$.

We compare QoI score generated by QnQ with Jøsang's expected truthfulness (E^J), which is equivalent to τ_k in our approach (because the scale of both metrics have to be between 0 and 1 for fair comparison). From the earlier Table. I, E1, E2,E3,E4 end up with 0.04, 0.46 and 0.16 and 0.26. This solves the limitations/bias of Jøsang Scores.

D. QoI-based User Reputation Score

For any reporter i, we match the reports he had generated with the estimated QoI value of the corresponding events. We add Q_k for every unique reported event by a reporter i, to calculate the aggregate reputation score S_i.

$$S_i = \sum_{k=1}^{K} Q_k I(k,i) \tag{7}$$

$$\text{where,} \quad I(k,i) = \begin{cases} 1, & \text{If } i \text{ reported event } k \\ 0, & \text{Otherwise} \end{cases} \tag{8}$$

E. Normalized User Reputation Score

The aggregated reputation score S_i obtained from Eqn. (7) is a real number in the interval $[-\infty, +\infty]$. In order to make it intuitive and consistent with the definition of trust metrics, we use the logistic distribution function to map its values in the interval [-1, +1]. Therefore, the final reputation score (R_i) of a reporter i is given as:

$$R_i = \begin{cases} +\left(\frac{1}{1+e^{-\frac{S_i-\mu_s^+}{C^+}}}\right), & \text{if } S_i > 0 \\ -\left(\frac{1}{1+e^{-\frac{|S_i|-|\mu_s^-|}{C^-}}}\right), & \text{if } S_i < 0 \\ 0, & \text{if } S_i = 0 \end{cases} \tag{9}$$

where, μ_s^+ and μ_s^- are the mean reputation score for reporters with positive and negative S_i respectively. Similarly, $C^+ = \frac{\sqrt{3}\sigma_{s^+}}{\pi}$ and $C^- = \frac{\sqrt{3}\sigma_{s^-}}{\pi}$ where σ_{s^+} and σ_{s^-} are the standard deviations for reporters with positive and negative S_i respectively. Note if a user generates reports for the "same" event multiple times, it is considered as one reported event.

V. Performance Study

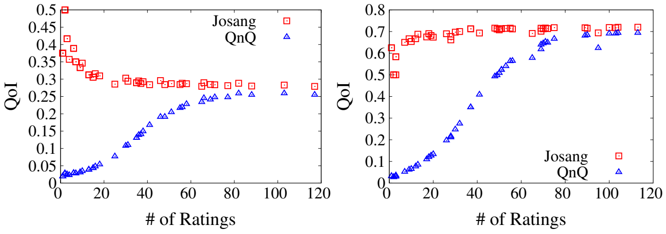
We present the simulation settings, comparisons with Jøsang's belief and D-S reputation model and show the efficiency of QnQ in terms of reputation scores, and incentives.

1) Simulation Settings: We developed a realistic environment of a vehicular participatory sensing system by choosing the simulation parameters from a Waze data set [3]. We simulate a city area of 200 × 200 sq.units as the region of interest. Out of $U = 2400$ number of active users, $U_{rp} = 800$ are reporters and $U_{rt} = 1600$ are raters. The city area is equally partitioned into four sub-regions; each initialized with $\frac{U_{rp}}{4}$ reporters and $\frac{U_{rt}}{4}$ raters at the start time. We consider a fixed attack budget of compromising 520 fake/rogue devices, out of which 120 are used for generating false reports and 400 are used for false ratings. These devices have been distributed uniformly in the simulated city area. The total simulation time is uniformly divided into $D = 240$ number of epochs, and each epoch is of a duration of $T = 30$ minutes. Each epoch has a predefined probability of occurrence of both true and false events. Every event has a event radius (50 sq. units) within which all reporters and raters are liable to report or rate. Each event has a tunable lifetime within which reports and feedbacks are accepted. For e.g., if an event E_i has occurred in epoch i and the duration of its lifetime is two epochs, then E_i can be reported and rated until epoch $i + 2$.

We have considered random paths along which a user moves with an average speed of 5 units/epoch. We have refrained from using any particular mobility traces since variations in mobility patterns can bias the number of raters and reporters and hence the results. Rather, our simulation is done by parameterizing the number of raters and ratings to account for all possible realistic combinations.

For the reporters, we emulate honest, selfish, and malicious behaviors in the following ways. 20% of the reporters are programmed as selfish while 10% act as malicious and the rest act honestly. Given that an incident has occurred, an honest reporter reports 99% of the time and has a minuscule probability of false report (simulating occasional wrong perception). Malicious reporters within a randomly generated location (chosen for false event) collude to generate fake reports of a fictitious incident with high probability $\approx 100\%$. Finally, selfish reporters intermittently report both true and false events. One class of selfish users reports more true events (about 60%) than false events, while the other class reports lesser true events (about 40%) than false events.

For the raters, we simulate both honest and compromised (malicious) raters. The compromised raters give positive ratings



(a) False Event (b) True Event

Figure 6: QoI Score Comparison

to false events and negative ratings to true events, while honest raters provide genuine ratings with 5% uncertain legitimately. The user ID that reported a particular event is prevented from rating it for obvious reasons. The percentages of compromised raters corresponding to an event varies with the variation in the population size. Since 400 out of 1600 raters are compromised), *on average* the fake rating percentage for true and false events is about 25%. We have discussed the effect of this in our scalability analysis (see Section V-6). Unless otherwise stated, the values of different parameters considered for experiments are: $A_b = A_u = 20$, $B_b = B_u = 0.08$, $\phi = 0.2$, $w_u^{max} = 0.5$, $N_{thres} = 40$.

2) *QoI of Events:* Fig. 6a shows a comparison between the QoI score achieved by QnQ vs. Jøsang's belief model for a false event. We observe that QnQ refrains from giving an undue high QoI score, unlike Jøsang's model for lower number of ratings. As higher number of ratings are received, the confidence of the crowd and the uncertainty discounting is taken into account to converge to the true value. Hence, malicious raters are unable to harness an advantage. This is however not true for Jøsang's model, and false events may end up getting higher scores if number of ratings were lesser. In contrast, Fig. 6b shows QoI score comparison for a true event. For QnQ, QoI converges to the true value only after sufficient amount of ratings are received, while for Jøsang's model this aspect does not matter. This is essential to prevent potential sabotaging by an organized minority of rogue raters. QnQ will always assign low QoI to events receiving low feedbacks. When number of ratings are limited there could be two possible options: (i) the published event may not be significant enough and does not draw attention of majority of raters, resulting in low QoI, and (ii) the place has an inherently low population and so N is not very high. The parameters A_b, A_u, B_b, B_u and ν (explained in Section IV-B) could be tuned to achieve higher QoI score at comparatively lower number of ratings to adapt to contextual requirements.

3) *User Reputation Scores: Quality and Quantity:* Fig. 7 shows how QnQ is able to reflect both quantity (i.e., total number of events participated) and quality (i.e., the number of events found to be true) of participation in the resultant user reputation score. The first observation is that three distinct groups of users emerge. The lowest group corresponds to malicious, the middle group to selfish and top group to honest users. Another key observation is that selfish and malicious users cannot increase their reputation even with boosting up their participation. This is a contrast to the Waze example presented in Section II-B. Since selfish users intermittently contribute true and false events, their scores are higher than malicious but lesser than the honest users. The selfish group has two tiers as explained below.

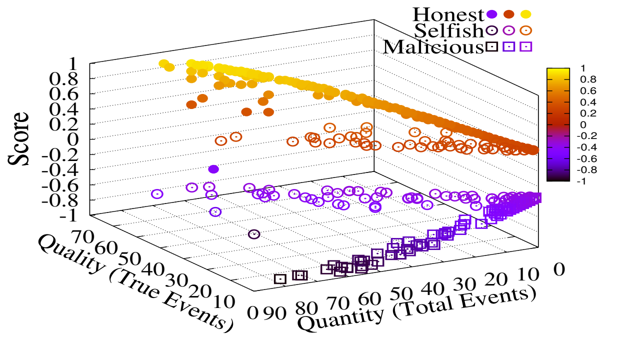
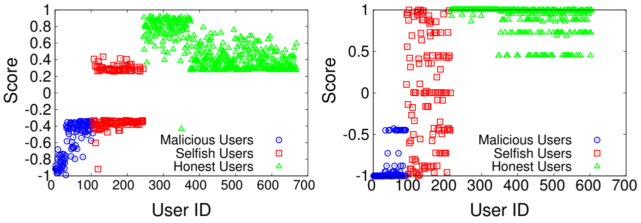


Figure 7: User Reputation Scores vs Quality and Quantity



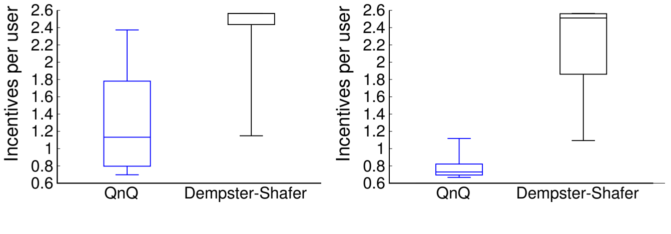
(a) QnQ (b) D-S

Figure 8: Classification of User Behavior

4) *Classification of Users with Fairness:* We considered 2 different types of selfish users : (i) reports more true events than false events, and (ii) reports more false events than true events. To be fair, the selfish users with higher number of genuine contributions should have higher scores than others from same class. This aspect can be verified from Fig. 8a and Table III. Here n_i is the total event participation of user i. Some honest users with lower scores are at par with some selfish users. These honest users have *a very low* event participation (honest #2 in Table III). In contrast, selfish #1, contributed about 25 good events along with 11 bad events. The system awards him a score marginally above 0.38 due to his 25 good events as opposed to 3 contributed by honest #2. However, his net score increase is not substantial due to 11 bad events. Both have much lower score than honest #1, which has one inadvertent bad event and 57 good events. Our model is better in accuracy and fairness than D-S based reputation score as shown in Fig. 8b, where many selfish users end up with very high scores.

Table III: Comparative Reputation Scores

Type	n_i	True #	False #	Score
Honest #1	58	57	1	0.959792
Honest #2	3	3	0	0.316258
Malicious #1	57	3	54	-0.926592
Malicious #2	8	0	8	-0.409386
Selfish #1	36	25	11	0.380653
Selfish #2	39	17	22	-0.473364

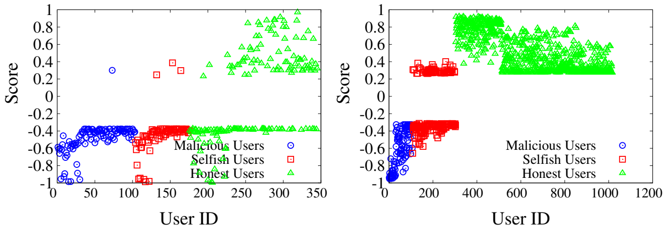


(a) Honest (b) Selfish

Figure 9: Incentive Disbursed: Securing Undue Leakage

5) Reducing Incentive Losses: We considered a reputation score-based incentive mechanism presented in [17] and compared the rewards disbursed. Fig. 9a shows that QnQ offers a larger variation of incentives disbursed to honest users according to the variations in quality and quantity. However, D-S gives higher incentives since it only awards quality but not quantity. Hence, users with lower participation also end up with a high score and hence a higher incentive. In contrast, Fig. 9b shows that mean QnQ-based incentives for selfish users is 50% that of honest ones and is three times less than that yielded by D-S based reputation model. Unlike D-S based model, QnQ can distinguish between honest and selfish behaviors and penalize the latter with low rewards thus preventing loss of revenue due to false contributions.

6) Scalability and Robustness Limits: Unlike all prior plots with population size of 2400, now we consider populations of 1200 and 3600 (fake devices inclusive), under the same attack budget of 520 devices. Fig. 10a show that even with low population of 1200, such that malicious raters form about 47% of the total raters, we still succeed to keep all of the malicious and selfish users in lower reputation tier, but misclassification rate of honest users increases. This is because the smaller crowds are unable to generate required confidence since bad mouthing causes many honest user's events to be rated low. Because our system follows a protective approach, this sacrifice is made when crowd is low. However, when crowd increases, reputation of all honest users are improved. This situation is evident from Fig. 10b.



(a) $U = 1200$ (b) $U = 3600$

Figure 10: Reputation Scalability

VI. Conclusion

In this work, we addressed the reputation scoring of users (under malicious and selfish behaviors) in the premise of crowd sourcing. We proposed a regression-based reputation model, QnQ, which is resilient to rogue contributions and null invariance and unifies quality and quantity of information contributed. The resultant reputation score provides a clear segregation among honest, selfish and malicious users, and implicitly guarantees fairness within each segregated group without sacrificing either participation or quality. In terms of rewarding the users, QnQ performs better than the D-S reputation-based incentive mechanism. Finally, recommendations on decision parameters help to adapt the model under varying conditions of risk and uncertainty. As a part of the future work, we intend to use this reputation score to solve an optimization problem with two objectives: (i) maximize truthful user participation, and (ii) minimize disbursement of undue incentives.

Acknowledgment

The work is partially supported by NSF grants under award numbers CNS-1545037, CNS-1545050, DGE-1433659 and IIS-1404673. Dr. Das is also a distinguished visiting professor of Zhejiang Gongshang University, Hangzhou, China.

References

[1] R. S. Anderssen, S. A. Husain, and R. Loy, "The Kohlrausch Function: Properties and Applications", *Anziam Journal*, vol. 45, pp. 800-816, 2004.

[2] H. Amintoosi, and S. S. Kanhere, "A Reputation Framework for Social Participatory Sensing Systems", *Springer MNA*, vol. 19, pp. 88-100, 2014.

[3] R. P. Barnwal, N. Ghosh, S. K. Ghosh, and S. K. Das, "Enhancing Reliability of Vehicular Participatory Sensing Network: A Bayesian Approach", *IEEE SMARTCOMP*, pp. 1-8, 2016.

[4] S. Bhattacharjee and M. Chatterjee, "Trust based Channel Preference in Cognitive Radio Networks under Collaborative Selfish Attacks", *IEEE PIMRC*, pp. 1502-1507, 2014.

[5] J. A. Burke, D. Estrin, M. Hansen, A. Parker, N. Ramanathan, S. Reddy, and M. B. Srivastava, "Participatory Sensing", *CENS*, 2006.

[6] Z. Feng, Y. Zhu, Q. Zhang, L. Ni, and A. Vasilakos, "TRAC: Truthful Auction for Location-aware Collaborative Sensing in Mobile Crowdsourcing", *IEEE INFOCOM*, pp. 1231-1239, 2014.

[7] R. Cudeck and K.J. Klebe, "Multiphase mixed-effects models for repeated measures data" *Psychological Methods*, vol 7, no. 1, pp. 41-63, 2002.

[8] J.R. Eiser and M.P. White, "A Psychological Approach to Understanding how Trust is Built and Lost in the Context of Risk", *Conference on Social Contexts and Responses to Risk*, 2005.

[9] K. L. Huang, S. S. Kanhere, and W. Hu,"Are You Contributing Trustworthy Data?:The Case for a Reputation System in Participatory Sensing", *ACM MSWiM*, pp. 14-22, 2010.

[10] C. Huang and D. Wang, "Topic-Aware Social Sensing with Arbitrary Source Dependency Graphs", *ACM/IEEE IPSN*, pp. 1-12, 2016.

[11] A. Jøsang,"An Algebra for Assessing Trust in Certification Chains", *NDSS*, 1999.

[12] A. Jøsang and R. Ismail, "The Beta Reputation System", *Bled eConference*, pp. 41-55, 2002.

[13] I. Koutsopoulos, "Optimal Incentive-driven Design of Participatory Sensing Systems", *IEEE INFOCOM*, pp. 1402-1410, 2013.

[14] P. Massa and P. Avesani, "Trust-Aware Bootstrapping of Recommender Systems", *ECAI Workshop on Recommender Systems*, pp. 29-33, 2006.

[15] P. Mohan, V. N. Padmanabhan, and R. Ramjee, "Nericell: Rich Monitoring of Road and Traffic Conditions using Mobile Smartphones", *ACM SenSys*, 2008.

[16] H. Mousa, S. Mokhtar, O. Hasan, O. Younes, M. Hadoud, and L. Brunie "Trust Management and Reputation Systems in Mobile Participatory Sensing Applications: A survey" *Computer Networks*, vol. 90, pp. 49-73, 2015.

[17] F. Restuccia and S. K. Das, "FIDES: A Trust-based Framework for Secure User Incentivization in Participatory Sensing", *IEEE WoWMoM*, pp. 1-10, 2014.

[18] N. Ram, K. Grimm, "Handbook on Child Psychology, Developement Science and Methods: Growth Curve Modeling and Longitudinal Factor Analysis", *Wiley*, pp. 758-785, 2015.

[19] M. B. Sinai, N. Partush, S. Yadid, and E. Yahav, "Exploiting Social Navigation", *arXiv preprint arXiv:1410.0151*, 2014.

[20] D. Wang, L. Kaplan, H. Le, and T. Abdelzaher, "On Truth Discovery in Social Sensing: A Maximum Likelihood Estimation Approach", *ACM IPSN*, 2012.

[21] G. Wang, B. Wang, T. Wang, A. Nika, H. Zheng, and B. Y. Zhao, "Defending against Sybil Devices in Crowdsourced Mapping Services", *ACM MobiSys*, 2016.

[22] B. Yu and M. P. Singh. "An Evidential Model of Distributed Reputation Management", *ACM AAMAS*, pp. 294-301, 2002.

[23] X. Zhang, Z. Yang, Z. Zhou, H. Cai, L. Chen, and X. Li, "Free Market of Crowdsourcing: Incentive Mechanism Design for Mobile Sensing". *IEEE TPDS*, vol. 25, no. 12, pp. 3190-3200, 2014.