Resilience Against Bad Mouthing Attacks in Mobile Crowdsensing Systems via Cyber Deception

[1]Prithwiraj Roy*, [2]Shameek Bhattacharjee*, [2]Hussein Alsheakh, and [1]Sajal K. Das

[1]Department of Computer Science, Missouri University of Science and Technology, Rolla, MO, USA
[2]Department of Computer Science, Western Michigan University, Kalamazoo, MI, USA
Emails: {przhr, sdas}@mst.edu, {shameek.bhattacharjee, hussein.s.alsheakh}@wmich.edu

***Abstract*—Mobile Crowdsensing System (MCS) applications deploy rating feedback mechanisms to help quantify the trustworthiness of published events which over time improve decision accuracy and establish user reputation. In this paper, we first show that factors such as sparseness, inherent error probabilities of rating feedback labelers, and prior knowledge of the event trust scoring models, can be used by strategic adversaries to hijack the feedback labeling mechanism itself with bad mouthing attacks. Then, we propose a randomized rating sub-sampling technique inspired from moving target defense and cyber deception to mitigate the degradation in the resulting event trust scores of truthful events. We offer a game theoretic strategy under various knowledge levels of an adversary and the MCS in regards to picking an optimal sub-sample size for bad mouthing attacks and event trust calculations respectively, by using a vehicular crowdsensing as a proof-of-concept.**

***Index Terms*—Mobile Crowdsensing Security, Moving Target Defense, Trust, Cyber Deception, Security of AI**

I. Introduction

Widespread availability of internet of things (IoT) and handheld mobile devices (e.g., smartphones, tablets, smartwatches, smart vehicles, roadside units) and rapid advances in pervasive sensing technologies have fueled the development of *Mobile Crowdsensing Systems* (MCS). A typical MCS involves a server that accumulates voluntary contributions given by either autonomous sensing agents (e.g., IoT devices) or by humans (via an App).

Based on the nature of the MCS applications, either individual contributions or a summary statistic of such contributions (called *events*) are published by the MCS service provider. Such published events guide human choices or automated decisions that improve the quality of life and civic-well being in smart cities. The real benefit of using an MCS paradigm is that precise and fine-grained information collection is possible without maintaining expensive dedicated infrastructure.

Real life examples of MCS include Google Waze [15], where contributions are in the form of 'reports' from humans indicating the presence of special road events (viz., jam, accident, weather hazard, crime scene, speed trap, gasoline prices). Based on the reports, Waze publishes events on its App. that allows intelligent traffic route selection. Most mobile apps and socio-economic networks, such as Yelp, run similarly.

However, a real challenge in MCS is the 'trust loop problem' [7]. Arguably, the collection of evidence to quantify trust is the most challenging aspect. To be precise, even if we obtain some evidence indicating reliability/trust, how can one be sure that this very evidence is itself reliable [7]? MCS also suffers from a cold start problem: When an MCS is launched it has no idea of prior trustworthiness of users or prior probabilities of event types. Therefore, mechanisms depending on prior trust of users to quantify event trust, do not work. Another class of methods propose the identification of rogue users by comparing the deviation of each user's report with a truth discovery mechanism. However, truth discovery mechanisms do not take into account orchestrated attacks, and assign reputation based on the truth discovery output.

To address the above issues, some existing methods offer a feedback monitoring mechanism that allows other users/agents known as labelers/raters, who provide positive, negative, or uncertain rating labels to the reports/events. This allows an initial truthfulness perception of the events, based on which the user reputation is built. With time, the MCS can remove low reputation users to improve reliability. However, this rating feedback motivates the possibility of orchestrated false ratings/feedbacks/labels (known as *feedback weaponizing attacks*) from a well-organized malicious adversary that biases the event trust scoring mechanism. This negatively affects reputation systems, incentive assignment, and publishing decisions. The three established feedback weaponizing attacks are ballot stuffing, bad mouthing, and obfuscation stuffing [3]. Ballot stuffing and obfuscation stuffing attack's goal is to make spam events look true, while bad mouthing attack try to suppress important events that really occurred.

Now, most of the existing methods for event trust scoring use variants of Josang's Belief Model [10], Beta Distribution [9], or Dempster Shafer Belief [11]. While some works [3] provide active resilience to ballot stuffing and obfuscation stuffing attacks, none of them *actively* mitigate bad mouthing attacks.

*Co-primary authors Equal Contribution

978-1-6654-2263-5/21/$31.00

A. Motivation

To motivate this paper, we present three important challenges not adequately addressed in previous works.

First challenge is the sparsity in the rating feedback of the MCS. Prior event trust models assume that deployment of an MCS automatically implies the availability of a substantial rating population that keeps the relative proportion of compromised feedbacks to the total number of feedbacks low enough; for event trust scores to remain unbiased [3]. Nonetheless, this assumption may not always be practical. On one hand, newly launched MCS may have a lower customer base, and hence the rating feedback labelers are lower to begin with. A rival business with a small attack budget may poison the event trust score and prevent good reporters from gaining reputation, thus forcing them to get less incentives. On the other hand, even when the MCS may have a high user base, the geographical spread of this user base may not be spatio-temporally uniform. For example, a downtown area has less crowd during nights; similarly some parts of a city may inherently be sparsely populated than other areas. In such instances, an adversary can poison event trust scores with a small attack budget.

The second challenge is error probabilities in the feedback apparatus. Existing event trust models do not mathematically incorporate, the error probability of rating feedback from an honest rater *combined* along with the possibility of feedback weaponizing attacks. When the error probability is combined with the presence of an attack, we observe that existing trust scoring models lead to biased results on true events even when a minority of the rating population is compromised, as elaborated in Section IV-C1.

The third challenge is about targeted feedback weaponizing attacks by adversaries with prior knowledge of our proposed defense model. In our paper, our defense needs to take into account according to Kirchoff's principle [16] that any defense model should not assume secrecy of the model.

B. Contributions of this Paper

In this paper, we establish the use of a moving target cyber deception based sub-sampling technique as a method to ensure active resilience against bad mouthing attacks when the rating sample sizes are smaller and the honest rating labelers have errors in their judgment, in presence of uncertainty and adversaries having knowledge of event trust scoring models. Specifically, we first establish some conditions under which linear models (e.g., Josang model and Beta trust model) as well as nonlinear models (e.g., QnQ [3]) fail. Then, we describe the sub-population sampling method under various adversarial scenarios, and available information present to the MCS provider. We show that when the rating consensus is lacking and the attack scale is unknown, a sub-sampled strategy for quantifying event trust (1) decreases the probability of evasion of bad mouthing attacks; (2) improves the estimation of event trust accurately, regardless of whether the adversary controls the majority or minority of the rating feedback population. We also provide a game theoretic formulation where we analyze strategic behaviors of MCS and adversaries by taking into consideration economics of security attack and defense, thereby showing improved resilience to bad mouthing attacks and boost in event trust scores, as compared to the existing methods.

The paper is organized as follows. Section II introduces preliminaries and different rating feedback systems. Section III describes the system architecture and threat model. Section IV presents the proposed approach while Section V reports experimental results. The final section concludes the paper.

II. Preliminaries

Typical event trust in MCS include three major phases: (1) accumulate ratings (feedbacks) on the event that serve as 'evidence'; (2) quantify an event's trust score by using the available ratings; and (3) classification decision on whether the event is truthful or not by comparing the event trust score with a hard or soft threshold.

A. Rating Feedback Systems and Event Trust

Rating feedback mechanisms specify a discrete state space of choices that any rater has about the authenticity of an event. The raters could be a human, mobile trusted agents, or watchdog module running an anomaly detector, which provide a rating conceptually belonging to three categories: positive feedback (α), negative feedback (β), and uncertain feedback (μ). Let the number of ratings per category be denoted as η_α, η_β, and η_μ and the event's evidence is denoted as $E : \langle N, \eta_\alpha, \eta_\beta, \eta_\mu \rangle$ where N is total population of ratings, $N = \eta_\alpha + \eta_\beta + \eta_\mu$. The event trust is quantified by the following known techniques.

B. Beta Trust Model

Beta trust treats the state space as binary while Beta reputation [9] can be applied to both binary and ternary evidence state spaces. The event trust (truthfulness) is as quantified as:

$$QoI_{beta} = \frac{\eta_\alpha + 1}{N + 2}, \qquad \text{where} \qquad 0 < QoI_{beta} < 1 \tag{1}$$

where η_α is the total number of positive feedbacks and N is the total number of feedbacks received.

C. Josang's Belief Model

Josang's Belief Model [10] explicitly handles uncertainty in the evidence, by specifying expected truthfulness (E) by the following linear score:

$$QoI_{jo} = b + (a).u, \quad \text{where} \quad b = \Big(\frac{\eta_\alpha + 1}{N + 3}\Big); u = \Big(\frac{\eta_\mu + 1}{N + 3}\Big) \tag{2}$$

Here b and u are the degrees of belief and uncertainty respectively, and a is the relative atomicity parameter that decides the extent to which the uncertainty should contribute to the event truthfulness. Note that $a = 0.5$ if there is no prior information available, such that $0 < QoI_{jo} < 1$ acts as a linear predictor of the estimated truthfulness of this event.

D. QnQ Belief Model

In *quality vs. quantity* (QnQ) model [3], the event trust is:

$$QoI_{QnQ} = w_b.b + w_u.u \tag{3}$$

where b and u are the same as Josang's Belief model as above, but w_b and w_u are Richard's and Kohlsrausch relaxation [6] functions. (Refer to [2], [3] for mathematical details.) We found that while ballot stuffing and obfuscation attacks are actively prevented by QnQ, bad mouthing attacks are passively avoided by virtue of the assumption of large crowd. In short, QnQ works well under sparse samples but only for ballot stuffing and obfuscation stuffing attacks but for not bad mouthing attacks.

E. Relationship between Voting Systems and Event Trust

Majority voting is central to dependable decision making in cooperative distributed systems. Here, we show that majority voting is implicitly related to how event trust scores are interpreted for event inference.

Binary rating can be viewed as voting for an event. Let the evidence be a tuple $\langle \eta_\alpha, \eta_\beta \rangle$ indicating the number of positive and negative ratings received. Suppose the MCS receives two events with evidences: $E1$: $\langle 20, 30 \rangle$ and $E2$: $\langle 21, 20 \rangle$. A decision rule by majority voting would indicate that the event is false for $E1$, while it is true for $E2$. However, the majority voting is a hard decision rule and lacks intelligence in the sense that it cannot quantify confidence on the event's likelihood of being actually true or false. For example, if $E3$: $\langle 98, 2 \rangle$, the answer will still be a true event and there will be no difference between $E2$ and $E3$ although the likelihood of $E3$ of being actually true is higher.

In contrast, models such as beta trust [9], have roots in AI and enable similar outcomes but via a soft decision rule that allows embedding the notion of confidence into event decisions. By taking the same example: $E1$'s beta trust score is 0.4038, while $E2$'s is 0.5116. To classify between a true versus a false event (a binary classification problem), one needs a logit link function to evaluate the trust score. A negative score probabilistically indicates the event is likely false while a positive score means the event is likely true [3], [4]. This is because the mid-point of 0.5 is assumed as a neutral decision boundary between true versus false event inferences. Therefore, for $E1$, $E2$, and $E3$, we get $\log(0.40/(1-0.40)) = -0.16$ and 0.020, and 1.5 respectively. Therefore, $E1$ and $E2$ will be inferred as a false and true event, respectively. We can observe that the final inference is the same in both voting and trust models. However, unlike voting, the trust scoring with AI approach allows to distinguish between two events that are positive in their score ($E2$ and $E3$), but for event with a greater score the decision maker has more confidence that it is indeed true. This aspect helps *in assigning proportional* incentives based on the trust unlike voting. This makes the AI based approach to event trust scoring more intelligent than majority voting, although they are similar in terms of decision output for the event.

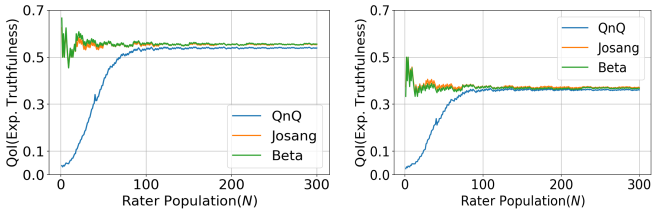


Figure 1: Trust Score under Bad Mouthing: (a) Majority honest, (b) Majority compromised

Ternary rating contains an additional category of neutral/uncertain unlike binary ratings. The authors in [4] showed that we can accommodate uncertainty by partial splitting of uncertain ratings in the ratio of the observed positive and negative ratings; and then adding it to the positive feedback. This way it can approximate the Beta Model by providing modified effective positive and negative ratings that can be applied to Beta Trust [4], if probability of all components getting compromised is uniform. For example, for 10 positive, 8 negative and 4 uncertain ratings, the effective positive posterior belief are obtained by partially splitting 4 uncertain ratings in the ratio of the observed η_α and η_β such that the modified positive votes are:

$$\eta''_\alpha = \lceil \eta_\alpha + \left(\frac{\eta_\alpha}{\eta_\alpha + \eta_\beta}\right).\eta_\mu \rceil; \quad \eta''_\beta = N - \eta''_\alpha; \quad QoI = \frac{\eta''_\alpha + 1}{\eta''_\beta + \eta''_\alpha + 2} \tag{4}$$

Thus, if the majority of the rating population is honestly giving correct feedback, the Beta and Josang belief models deliver an event trust score more than 0.5, as depicted in Fig. 1a where 60% raters are honest. This relationship is used to simplify some analysis in our paper.

III. System and Threat Models

In this section, we present the abstraction of the MCS, the threat of bad mouthing attacks, and novel considerations that have not been incorporated in previous works.

A. System Model

We assume a network of U users/devices subscribed to a crowdsensing application. Users have three types of roles for a given event E: (a) reporter (b) rater (c) passive consumers who do not participate in either reporting/rating.

Reporters: are the set of users who report to the MCS server indicating an 'event' of interest. The MCS server publicly publishes, either the individual reports or an aggregate decision from all reports, as an 'event' of interest, to all other users of that application during the cold start phase.

Raters: are the set of users (humans) or machines (drones) who provide a rating on the published events which indicate the crowd-sensed perception of the relative goodness of the

published events. We denote the total number of ratings received for a given event as N, and the number of positive, negative, and uncertain ratings are η_α, η_β, η_μ respectively.

Events: Each event E has a geo-boundary and a time period of viability. Users within the boundary and time period are liable to rate it. Reporters cannot rate the event reported by themselves, while a rater can rate a single event only once.

Rating Probabilities of Honest Raters: Let p_b be the probability that an honest rater accurately rates a legitimate event as true. Any rater when not controlled by an adversary has an error probability of p_e with which it mistakenly rates a legitimate event as false. For example, extreme weather conditions with low visibility can produce errors in judgment by drones. The probability that an honest rater rates a true event as uncertain is p_u. So from the above, for an honest rater $p_b + p_u + p_e = 1$. Therefore, the probability that an honest rater does not rate a true event with positive feedback is $1 - p_b$. This is a robustness issue rather than a security issue, but it still affects the event trust score outcome negatively.

B. Threat Model

Let us discuss several aspects that specify the threat model.

Bad Mouthing Attacks: The rogue raters may be controlled by an organized adversary launching Bad Mouthing Attacks that give false ratings to true events, causing legitimate events to get low event trust scores. This not only prevents the published legitimate event to be deleted but also degrades reputation of the honest reporters who reported that event. In future, when those reporters report again, their reports will be ignored due to their unfairly low reputation caused by bad mouthing attacks [3]. Such event suppression can cause serious impacts on a smart city environment. For example, actions required for serious events may not be taken, and can lead to further worsening of an emergency/disaster event.

Attack Budget and Adversary Capabilities: The adversary uses knowledge of defense mechanism and MCS to decide whether to apply his full budget or not. If purpose is served without using its full budget, then it is a rational choice for the adversary. The 'attack scale' is the fraction of compromised ratings to the total number of ratings (from defender's perspective). Among N raters that have provided ratings to a particular truthful event j, suppose K raters are not compromised by the malicious adversary, while rest $N - K$ raters are controlled by the adversary. In MCS, the rating feedback mass N for an event/item is often sparse for three reasons: (1) when a application is launched initially, the user base is not very high, reducing the chances of getting a high feedback mass; (2) lack of motivation or incentives to provide ratings; (3) due to mobility patterns of users a part of the city may have a sparse user footprint on specific times of a day.

When N is small, the adversary's attack budget $N - K$ can dominate the proportion of the rating sample N leading to biased event trust scores. The adversary can either compromise or simply recruit extra $N - K$ malicious raters to the MCS network. The MCS defender does not know which raters are compromised although the MCS can assume reasonably that the adversary would try to dominate the rating population to ensure event trust score is degraded (i.e., preferably below 0.5).

Adversary assumptions: (1) The adversary knows that the defender may use a traditional trust scoring method or our subsampling based method. This is following Kirchoff's principle in security, that secrecy of a defense mechanism cannot be assumed. (2) We assume a rational adversary who knows that there is a true event and therefore a compromised rater's p_e does not affect the input from a compromised rater. Rationality also means that if a ghost spam event is generated (by say selfish users [2]), our adversary refrains from participating in rating, because it goes against the rationale for bad mouthing. (3) Adversary has enough budget such that he can manage majority of the rating population if required. This makes sense due to the sparse sampling problem in MCS. However, once it has compromised a rater, it exhausts a resource and it cannot change it. (4) Adversary possess knowledge about the average rater population size N at given time/place and the probability of accuracy, p_b, of the honest rating users. (5) An intelligent rational adversary would always try to maximize his gain and will not have any strategy that causes a net loss. (6) Our approach makes sense only when no clear consensus is available in the rating population. For that we have chosen the situations where there is split vote of at least 30% or more for either true or false event. If more than 70% raters are manipulated then it is a dogmatic system and a mitigation strategy does not make sense. On the other hand, if more than 70% raters are honest then the full sampling strategy can infer the correct state and as we show in Fig. 3a, subsample size from our method converges to the full sample.

IV. Proposed Approach

The proposed approach is divided into four parts: First, we motivate the MTD and cyber deception approach to our solution. Second, we show that the strategic situation between the MCS and the adversary can be modeled into a two player game theoretic formulation with numerous strategies. Third, we provide a mathematical analysis of the randomized subsampling approach and the adversary's cost benefit analysis that helps to prune the strategy space of the attacker-MCS game. Fourth, we solve for best rational equilibrium strategies of the reduced game for both MCS and the adversary.

A. Randomized Rater Sub-Sampling as MTD

The philosophy of moving target defense (MTD) [1] mandates the creation of constantly shifting environments by a defender to introduce asymmetric uncertainty between defenders and attackers. The Department of Homeland Security, USA defines MTD as diverse strategies that continually shift and change over time to limit opportunities for attack, increase the cost of attacks and/or increase system resiliency. The guidelines in [1] specify dynamic system randomization at run time as

one overarching option for a defender to constantly change the effective attack surface.

Classical defense methods focus on host/resource/user level detection models. But they suffer from scalability and agility concerns in MCS because: (i) network sizes always vary rapidly as a function of time due to mobility, (ii) new users quickly join or leave, and (iii) behavior of users keep evolving. Therefore, in MCS, it is rather prudent to think about mitigation of or resilience against attacks rather than focusing on detection of attacks and identification of compromised users. This is complicated by the ease of biased ratings that change the event inference completely, under existing methods.

Motivated from the above, we propose a way to bring ideas of moving target defense (MTD) and cyber deception into event trust scoring for a more resilient MCS.

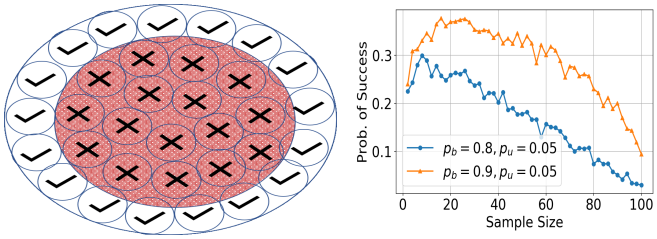
DEFINITION 1: Randomized Rater Subsampling: *Our main idea of dynamic system randomization is we hypothesize that instead of using all the ratings received for calculating a trust score, the defender uses a subsample of a strategic size from the set of ratings received, and then calculates the event trust from rating counts obtained from the subsample.*

Our hypothesis is given credence by the following illustration: Fig. 2a, offers a visual intuition of why randomized subsampling can mitigate the effects of feedback weaponizing attacks. We have already established in Sec II-E that if the majority of the ratings involved are not compromised, the event trust score progressively moves towards the correct event status (because of being above 0.5). There are three key considerations from Fig. 2a. (1) The big outer transparent circle represents the full sample N (total number of ratings received); (ii) the medium inner red shaded circle represents a set of compromised ratings is a subset of N but greater than 50% of N. (iii) the smallest circles with a tick or a cross represent a random subsample of a certain size (say n) drawn from the set N. The tick corresponds to samples where majority of the ratings in that n sized sample correspond to non-compromised rating set, while crosses correspond to the opposite.

We can conclude visually that the red circle size is more than half the outer circle. In this case, a full sample majority voting will always result in an outcome that corresponds to what the inner red circle indicates, which will be misleading and therefore be considered a failure for the defender MCS. In this case, the defender has zero chances of success in predicting an event trust score that is greater than 0.5 (for a true event).

However, if the defender picks a randomized subsample of a certain size n (represented by the smallest circles in Fig. 2a from set N; there are $\binom{N}{n}$ such possibilities). Without knowing which ones are compromised, we can see that success probability is non-zero. and thus better than using the whole sample. Regardless of the number of selections with ticks, it is better than zero chance of success using the full sample.

To justify this hypothesis, we conducted a numerical simulation to mimic the following scenario: A true event receives $N = 100$ ratings with 48 of them controlled by adversary giving a negative rating and $p_b = 0.9, p_u = 0.05$ for the remaining 52 honest raters. We then perform an exhaustive search over the subsample search space to figure the fraction of times (out of 1000 rounds of iteration) a subsample of a given size contains a majority of positive ratings (which is the accurate rating for the concerned event) in that selection. *This fraction is termed as the probability of success*, because a majority of true ratings in a subsample guarantees $QoI > 0.5$ that is required for inferring the event correctly.



(a) (b)

Figure 2: Subsampling: (a) Intuition (b) Optimal Sample Sizes

Fig. 2b, shows the y-axis as the probability of having more than 50% of positive ratings versus different possible sample sizes. We can see that the subsample size 16 maximizes this probability as the above mentioned settings. For another instance with $N = 100, K = 60, p_b = 0.8, p_u = 0.05$ the optimal sample size is 8. It also shows that even if the adversary compromised just short of majority of the rating sample, the classical use of using the full rating sample for trust calculation causes a probability of success = 0 (the rightmost point on the x-axis). In this case, 30% of the time, the MCS was successful in obtaining a trust score above 0.5.

B. Attacker Defender Game

Here we demonstrate that the strategic situation above can be modeled via game theory.

Defender's Perspective (MCS): In Section IV-A, we showed that using a subsample (1 to $N - 1$) of a certain size as a strategy by a defender will provide better outcomes if the majority of ratings are compromised. However, the defender knows that the adversary possesses the knowledge of its randomized subsampling approach. Based on this, defender expects that adversary would re-adjust its strategy to instead compromise a minority of ratings. Thus, there is a possibility that adversary might compromise an effective minority of the population if the defender chooses a sub-sampling approach. Hence, choosing the full sample, N, could also be a viable strategy in such a case. Hence, the feasible strategy could be any number between 1 and N.

Adversary's Perspective (MCS): The adversary expects the defender to play a sub-sampling strategy if it compromises a majority of the ratings. However, it also knows that the $1 - p_b$ is a feature that helps its cause. Thus, to overturn an event decision, the actual number required to compromise will be lesser than 51% of the total ratings received.

The adversary also knows that subsampling will provide better outcomes for the defender, it might contemplate switching to a strategy with a minority of ratings compromised. Therefore, the adversary 'theoretically' has an option to compromise 1 to N raters unless (i) subject to an upper bound on its attack budget; (ii) any strategy where his gain is lesser than its investment.

To be rational, the defender should select an optimal subsample size if any, that will maximize the probability of success out of all possible subsample or full sample sizes that could be potential strategies. The adversary will try to maximize its net benefit and would avoid any strategy that leads to a loss. The above presents a huge search space for finding equilibrium strategies in the attacker defender game. Therefore, our next effort is to prune the search space of strategies for MCS and adversaries before we solve the game formulation.

C. Pruning the Defender MCS Strategy Space

Now we provide an analysis of success probabilities with the randomized subsampling technique with the use of hypergeometric distribution [8]. The security status of the rater can be classified into one of the two mutually exclusive categories; compromised or non-compromised. Each rater is counted as a success if not compromised (or failure if compromised). Let there be K number of non-compromised (honest) and $N-K$ number of compromised raters in the received rating sample of size N. Let X be a random variable, denoting the number of non-compromised ratings in a chosen subsample from the population. Given that a subsample of size $n \leq N$ is drawn from such a population of size N, the probability of observing exactly k number of non-compromised raters from a population that originally contains exactly K non-compromised raters, can be specified by the pmf of a Hyper-Geometric distribution:

$$P(X=k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}} \tag{5}$$

The Eqn. 5 assumes that each non-compromised rater is never wrong in their judgment and produces a correct rating at all times. However, in the MCS model, each honest rater is either a human or sensor/drones that are prone to errors (p_e) and uncertainty (p_u). Hence, we need to modify the hypergeometric pmf, to find the effective true success probability of observing exactly k number of successes. Note that for MCS, success indicates those ratings that contribute to trust, and not the security status of the raters.

1) Embedding Error of Rating Labelers: To simplify, first we assume a binary rating system ($p_u = 0$), and then extend the analysis to ternary rating space in Sec IV-C4. Given exactly k honest raters are picked in a sub-sample, the chances that there will be at-least l or more successes (positive ratings) out of the k, can be modeled by a binomial distribution:

$$p(Y > l | X = k) = \sum_{j=l}^{k} \binom{k}{j} (p_a)^j (1-p_a)^{k-j} \tag{6}$$

where l is the lower bound required for successes under the normal fusion rule. For example, in case of majority voting $l = \lceil \frac{n+1}{2} \rceil$ (if n is even), and p_a is the chance that an honest rater provides a rating that eventually contributes to the increase in trust. The interpretation of how p_a is calculated changes with whether a Beta or Ternary Approximation to Beta Trust is used for event trust scoring and will be elaborated separately in Section IV-C4.

As l is the lower bound required for successes, based on whether subsample size is even or odd (n), l needs to be changed. If sample size is even, $l = \lceil \frac{n+1}{2} \rceil$ to ensure number of successes are majority in the selected subsample. In contrast, if we take odd samples then $l = \lceil \frac{n}{2} \rceil$ to obtain the majority of selected subsample. In this paper, we choose a discussion for even samples only, but it works seamlessly for odd samples with the only change in equations being $l = \lceil \frac{n}{2} \rceil$.

2) PMF of Success under Attacks: Here we derive the pmf for achieving a success where dominant rating attacks and errors coexist. Intuitively, the probability mass function will have a positive value, if there are at least $\lceil \frac{n+1}{2} \rceil$ honest raters are in the sample of size n. However, in theory, if the number of compromised raters is very low, then the minimum of honest raters in a larger sample size can be greater than $\lceil \frac{n+1}{2} \rceil$. The following is an illustrative example that explains this:
Lower Bound on Honest Raters: Considering two scenarios: Scenario I: $N = 100, K = 40$ and Scenario II: $N = 100, K = 90$ and a candidate subsample size $n = 70$. For scenario II, the $\lceil (n+1)/2 \rceil = 36$. However, given that there are 90 honest raters, even in the worst case, if all dishonest raters fall into subsample of size $n = 70$, the minimum possible honest users in this selected sub-sample is $n - (N-K) = 60$. Since 60 is greater than 36, the lower bound of Scenario II is $n-(N-K)$ while For scenario I, the lower bound is $\lceil (n+1)/2 \rceil$. Combining them, lower bound of the number of honest users in a sample of size n required for a success is:

$$k_{min} = max(\lceil \frac{n+1}{2} \rceil, n-(N-K)) \tag{7}$$

Upper Bound on the Number of Honest Raters: The maximum possible number of honest raters that can be picked depends on whether the total number of successes K is larger or smaller than the candidate sample size. Hence, the maximum possible upper bound for the number of honest raters:

$$k_{max} = min(n, K) \tag{8}$$

Probability of Success under attacks: The pmf of the resultant success in having a majority of the rating labels chosen as authentic labels is given by the following:

$$P(n) = \frac{\sum_{i=k_{min}}^{k_{max}} \binom{K}{i}\binom{N-K}{n-i} \sum_{j=\lceil (n+1)/2 \rceil}^{i} \binom{i}{j} (p_a)^j (1-p_a)^{(i-j)}}{\binom{N}{n}} \tag{9}$$

3) Optimal Sampling Size for Subsampling: Intuitively, the MCS' rational strategy should be to pick that sub-sample of

size n_{opt} that maximizes $P(n)$. Mathematically, we can write the optimal sample size as:

$$n_{opt} = \arg\max_{n}(P(n)) \tag{10}$$

Eqn. 10 gives the optimal sub-sample size n_{opt} for a given $N-K$ (no. of compromised raters) such that the probability of success is maximized. Hence, among all possible subsample strategies, the defender will pick n_{opt} since it maximizes its probability of success. The optimal probability of success achieved under the n_{opt} is:

$$P_{opt} = P(n = n_{opt})$$

Since we have already shown that the majority of ratings belonging to the true class is directly related to getting the desired event trust scores for accurate event truth inference, we would like to analyze how Beta and Josang's Models perform under the evidence extracted from the optimal subsample size.

4) Interpretation of p_a in Ternary State Space: Both positive and a portion of uncertain ratings (the benefit of doubt) contribute to event trust score. We envision the positive and the portion of the uncertain ratings that contributes to the event trust score as a 'success' since it contributes to an increase in the score when it is a true event (from Section II-E). Hence, p_a needs to be expressed in terms of p_b and p_u. Note that, the compromised $N-K$ raters always provide negative feedback. Therefore, the probability of having an honest user giving a rating that contributes to the event's trust score is $p_a^{beta} = p_b + p_u\frac{K}{N}$. Therefore, the *theoretical result* on the probability of success is given by the following:

$$P^{beta}(n) = \frac{\sum_{i=k_{min}}^{k_{max}} \binom{K}{i}\binom{N-K}{n-i} \sum_{j=\lceil (n+1)/2 \rceil}^{i} \binom{i}{j}(p_a^{beta})^j(1-p_a^{beta})^{(i-j)}}{\binom{N}{n}} \tag{11}$$

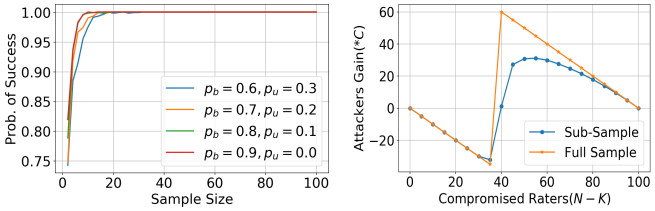
$$n_{opt}^{beta} = \arg\max_{n}(P^{beta}(n)) \qquad P_{opt} = P^{beta}(n = n_{opt}) \tag{12}$$

where P_{opt} *is the optimal probability of success* and n_{opt} *is the optimal sample size* that produces P_{opt}. In the experiment section, we will compare the theoretical result with the experiment to prove the correctness of the equation. A similar derivation can be done under Josang's Belief Model.

D. Pruning of Adversary Strategy Space

In the real world, the adversary may choose not to attack at all. However, the p_e may still cause false ratings and degrade event trust score. To examine whether this limits the achievable success, we assessed the effect of optimal subsampling under no attacks, using a numerical simulation (See Fig. 3a). We observed that optimal sample size converges to N (using Eqn. 11), if no attack is present in the system (regardless the p_b).

If adversary chooses to attack, we can prune its strategy space: Let the adversary incur a uniform cost, C, for each compromised rater. Then adversary's net payoff G is defined as follows: With just $(N-K)$ compromised raters, the adversary can overturn a decision worth the same as if it compromised



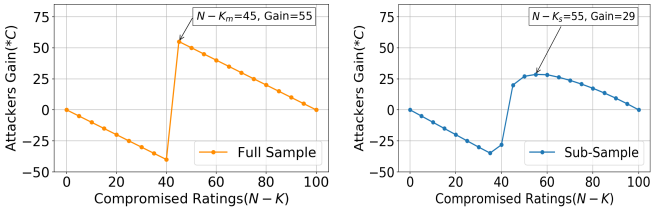
(a) (b)

Figure 3: Pruning:(a) No attack as a strategy (b) Attacker's Gain ($p_b = 0.7$ *and* $p_u = 0.25$).

all raters N. Therefore, investing $(N-K)C$, it gains a return equivalent of NC with a probability $(1-P_{opt})$, under defender's optimal sub-sample size. Thus, G is computed as the difference between return and investment as:

DEFINITION 2:The net payoff of the adversary $G = (1-P_{opt})NC - (N-K)C$.

Fig. 3b shows G for each possible value of compromised raters, when the defender plays its optimal sub sample size for a given $N-K$ (blue line) and full sample size (orange line). From Fig. 3b, we make a key observation: when $(N-K)$ is lesser, the attacker's net payoff under sub-sampling is larger than a full sampling. However, the G in all such cases is negative, which indicates a net loss for the adversary. Thus, we conclude that a rational adversary will at least want a net payoff to be non negative, and hence the minimum value of compromised raters for which G becomes positive is $N-K_m$. Therefore, we can prune all $N-K$ less than $N-K_m$ from the adversary's strategy space regardless of defender's actions.



(a) (b)

Figure 4: Attacker Payoff: (a) Full Sample (b) Sub sample

When defender uses a full sampling strategy: Fig. 4a (orange line) also verifies that choosing $N-K_m$ (that gives a positive G) is also the maximum payoff achievable by the adversary for any candidate $(N-K) > N-K_m$. This proves that adversary's best strategy is to compromise $N-K_m$ (*Strategy 1*) if the defender plays a full sample.

When the defender chooses the sub-sampling approach: Fig. 4b (blue line) shows that there is a particular compromised $N-K_s$ which maximizes the G. This proves that adversary's best strategy is to compromise $N-K_s$ (*Strategy 2*), if defender chooses to play a sub-sampling strategy.

Note that these choices of $N-K_m$ and $N-K_s$ is dependent on p_a, because G is related to $1-P_{opt}$, and P_{opt} is directly related to p_a. Fig. 4a and 4b shows how attacker's payoff changes with its number of compromised raters, given a p_b and p_u. Thus the attacker can compute the number of raters it needs to compromise that would maximize his gain, by taking the

maxima of the blue and orange lines corresponding to $N-K_s = \arg\max_{N-K}(G(n))$ and $N-K_m = \arg\max_{N-K}(G(N))$ It is evident from the figures that $N - K_s > N - K_m$ or $K_m > K_s$.

We have not included the case of no attacks as there is no loss for the attacker, even in presence of ratings caused due to errors on the part of honest users (p_e), and the optimal subsample size from Eqn. 12 still converges to N as shown in Fig. 3a.

E. Game Theoretic Formulation under Pruned Strategy space

Here we present a game theoretic approach for selecting a best rational strategy under a complete information zero sum game between the attacker and defender where both are aware of p_a and N and each other's possible set of strategies.

The adversary knows that if defender chooses the subsampling strategy it would do so with a subsampling size that gives P_{opt}. The MCS knows that the adversary has two options for calculating optimal compromised number of raters that would maximize the adversary's gain, for each of its optimal subsampling or full sampling strategy.

Formally, adversary has two strategies: Strategy 1: Compromise $N - K_m$ raters which maximizes G under full sampling strategy by defender; Strategy 2: Compromise $N - K_s$ raters, which maximizes G, if defender incorporates the sub-sampling strategy. The corresponding payoffs for each strategy are calculated in terms of adversary's gain/loss. The payoffs of the players are derived as follows:

1) Payoff Calculations for Strategies: We first calculate the adversary's pay-offs, given the defender's strategy.

If the defender plays a full sampling strategy, then the attacker is able to compute the minimum number of manipulation to overturn the decision of the defender. To perform this, adversary must ensure that $K.p_a \leq \frac{N}{2}$. Hence, $N-K_m = N(1-\frac{1}{2p_a})$ and the payoff with strategy 1 (i.e. $N - K = N - K_m$) is given by: $G1(full) = (1-0)NC - (N-K_m)C = K_mC$.

Since $N - K_m < N - K_s$, if the P_{opt} drops to zero at $N - K_m$, it will remain zero under $N - K_s$. Therefore, the payoff under Strategy 2 is given by $G2(full) = (1-0)NC-(N-K_s)C = K_sC$.

We now calculate the two pay-offs for the adversary, given the defender plays an optimal subsampling strategy. Let $P_1(n_1)$ and $P_2(n_2)$ be the maximum probability of success with subsampling method (calculated from Eqn. 12) under adversarial Strategy 1 and 2, respectively; where n_1 and n_2 are the 'corresponding' optimal sampling size selections. For notational simplicity, we denote $P_1(n_1), P_2(n_2)$ as just P_1, P_2.

So the attacker's payoffs with strategy 1 and strategy 2 is $G1(sub) = (1 - P_1)NC - (N - K_m)C = (K_m - P_1N)C$ and $G2(sub) = (1-P_2)NC-(N-K_s)C = (K_s-P_2N)C$ respectively. As cost C is a common scaling factor in all the payoffs, the final payoff matrix of the game is given in Table I after removing C from all the individual payoffs. In Table I the defender's (D), has two strategies: SS stands for subsampling technique and FS denotes a full sampling approach.

2) Game Solution: We solve the proposed two player zero-sum game to find the Nash equilibrium. Note, $p_a \leq 1$ (because $p_b, p_u \leq 1$), and both $0 \leq P_1, P_2 \leq 1$. As $K_m > K_s$, for the defender there exists a strictly dominating strategy but the attacker does not have any strictly dominating strategy. So the defender should always go for sub-sampling method to minimize his loss/attacker's gain. Thus, depending on the values of K_m, K_s, P_1, P_2, a Nash equilibrium can be obtained.

Table I: Complete Information Game.

		Adversary	
		$N - K_m$	$N - K_s$
MCS	FS	$-K_m, K_m$	$-K_s, K_s$
	SS	$-(K_m - P_1N), K_m - P_1N$	$-(K_s - P_2N), K_s - P_2N$

Let us illustrate this with an example: Let $p_b = 0.8$, $p_u = 0.15$. As depicted in the Fig. 4, the $N - K_m$ and $N - K_s$ for the adversary is 40 and 55 respectively. After calculating the payoffs as described in Table. I, we get the payoff matrix as shown in Table. II. Clearly, the attacker does not have a strictly dominating attack strategy but the defender's SS is a dominating strategy. Since the attacker's payoff from strategy 2 is higher than that of strategy 1, he will select for strategy 2, and defender will choose the sub-sampling method as the defense mechanism with the optimal size corresponding to Strategy 2, to minimize attackers gain and that will be our Nash equilibrium.

Table II: Payoff Matrix for $p_b = 0.8$

		Adversary	
		Strategy 1	Strategy 2
MCS	FS	-55, 55	-45, 45
	SS	-19, 19	-29, 29

3) Event Trust Scores under Optimal Sample Size: For beta distribution the trust is calculated based on the

$$QoI^{beta}_{proposed} = \frac{\eta_\alpha(n^{beta}_{opt}) + \left(\frac{\eta_\alpha(n_{opt})}{\eta_\alpha(n_{opt})+\eta_\beta(n_{opt})}\right).\eta_\mu(n_{opt})}{\eta_{opt}+2} \tag{13}$$

V. Experimental Results

In this section, we discuss the simulation, followed by numerical and simulation results.

A. Simulation Settings

We consider a vehicular crowd-sensing application as a proof of concept by using SUMO (Simulation for Urban Mobility [14] as a simulation environment. We have extracted the Open Street Map (OSM) for a small part of Manhattan city and created the network by considering the following scenarios: 1) Individual vehicle trips are created randomly in the selected network with a minimum trip length. 2) Accidents/traffic congestion is created by forcing a vehicle to stop at certain location and time. As every event has a certain time window of relevance, we collect the vehicle information in the vicinity of the reported event. We consider these users as potential raters who are liable to rate. A snapshot of SUMO simulation is shown in Fig. 5 where a traffic congestion is manually created and all the vehicles stranded are considered as

raters. We have considered different rating population sizes but maintaining $N = 100$ keeping in mind that the adversary has a limited budget and an intelligent adversary would attempt to attack the system only when the N is comparatively low which would keep his cost low. In each case, we have considered the compromised rater percentages to be between 30% to 70% for understanding the effect of varying attack scale. The rationale behind this assumption is based on the factor that if less than 30% raters are compromised then the gain of the adversary is minuscule as discussed in the threat model in section III-B. The p_b is varied between 0.7 to 0.95, while the p_u, is varied from 0.25 to 0. The compromised raters would always submit false ratings every time as they are controlled by the adversary. As per the attack strategy, these users are divided into honest and compromised groups and bad mouthing is performed accordingly. Simulation parameters description is given in Table. III.

Parameter	Symbol	Value
Total Ratings	N	100-300
Honest Raters	K	30%-70%
Prob. true rating (honest raters)	p_b	0.7-0.95
Prob. uncertain rating (honest raters)	p_u	0.25-0
Prob. effective true rating (honest raters)	p_a	-
Degree of belief	b	-
Degree of uncertainty	u	-
Relative atomicity	a	0.5
Optimal Prob. of success	P_{opt}	-
Optimal sample size	n_{opt}	-
Cost of compromising 1 Rater	C	-
Gain of adversary	G	-

Table III: Parameter Description and Value

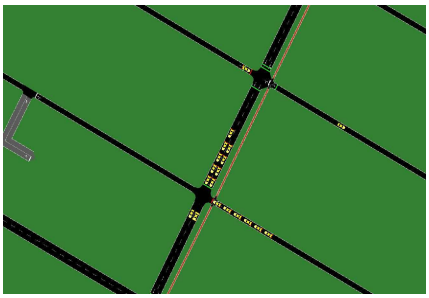
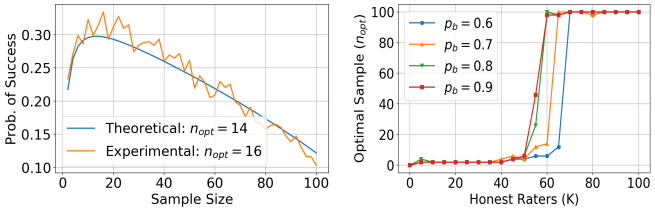


Figure 5: SUMO snapshot.

B. Implementation and Metrics of Performance

We evaluate performance over 1000 iterations using the data collected from the SUMO tool. In each iteration, different samples are picked to provide an average case performance that reduces bias. We show results not only for the game (which contains a strategic N-K) but also all possible N-K values to give a sense of how our method will perform under adversaries that may be non-rational or non-strategic.

Finally, for an optimal subsample size for the optimal attack strategy, raters of that size are randomly selected from the population over multiple events. In each event iteration, we calculate the Beta trust score using Eqn. 13. We record the number of iterations where the event trust score under our proposed method was above 0.5, *which counts as an event that our method successfully evaded an bad mouthing attack.* This is repeated under all possible (N-K) values.



(a) (b)

Figure 6: (a)Theoretical and simulation results: optimal Sample for $p_b = 0.75$ (b) Effect of p_b over optimal sample size

Metrics: We use the two metrics for performance evaluation: (i)Probability of Evasion: is the probability that the event trust score obtained by using the beta trust under our proposed method (MTD aware approach) gives a score of 0.5 or more (calculated over 1000 iterations in our simulation), to give the probability of evading a bad mouthing attack under our proposed approach. (ii)Boost in event trust score: This is the raw boost in trust score when our method is used under bad mouthing attacks, under all possible values of $N-k$, p_b, p_u.

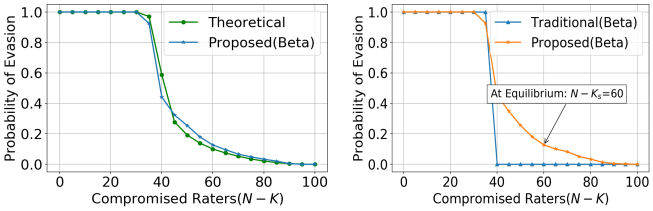
C. Optimal Sample Sizes

Fig. 6a shows a comparison between theoretical (from Eqn. 11) and experimental values of probability of success ($P(n)$) when total raters $N = 100$ and honest raters $K = 60$, with $p_b = 0.75$ and $p_u = 0.05$. We can conclude from Fig. 6a that simulation result closely follows the theoretical result, the optimal sample sizes n_{opt} from theoretical and experimental results are similar. In this situation although the honest raters is 60%, still we see that traditional full sample strategy fails with 0 probability of success (the right most point on the x-axis), because of p_e of the honest raters contribute to the adversarial objective. Thus it allows the adversary to have majority of the population giving negative feedback by controlling only 40% of the raters.

The change in the resulting optimal sample sizes for various K value (which indirectly depends on $N-K$ by the adversary) over different p_b values is shown in Fig. 6b for $N = 100$ and $p_e = 0.1$. It is evident that optimal sample size changes with different K and p_b. As $K.p_b$ increasingly dominates the population, the optimal sample sizes tend to increase and eventually reaches N.

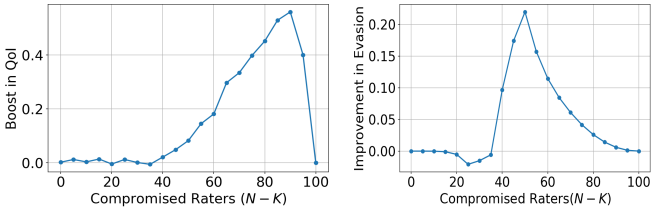
D. Performance Evaluation

We divide performance evaluation into two parts: (i) Illustrative Performance and (ii) Average Case Performance. The illustrative result is for a specific parameter setting while average case performance evaluation is result averaged over all possible combinations of parameters involved.



(a) (b)

Figure 7: Probability of Evasion ($p_b = 0.7$ and $p_u = 0.25$): (a) Theoretical vs Experimental (b) Comparison between Proposed and Traditional trust scoring



(a) (b)

Figure 8: Improvement over Traditional Method (a) Trust Score Increase (b) Probability of Evasion Increase

1) Illustrative Performance: We show theoretical versus experimental results of performance as well as comparison of our method versus traditional trust score with Beta distribution for setting $N = 100$, $p_b = 0.7$ and $p_u = 0.25$ for all $N-K$. The performance at equilibrium strategies are also presented.

Theoretical Versus Experimental Performance: The comparison of the probability of evading a bad mouthing attack successfully between theoretical and experimental results is shown in Fig. 7a, which verifies the accuracy of the model.

Improvement from Traditional Beta Trust: An illustrative result is shown in Fig. 7b the benefit of our method as opposed to traditional method in terms of probability of evasion. The percentage chances of getting event trust above 0.5 under the traditional beta trust which does not implement our MTD is compared to the same under proposed beta trust with the MTD approach. We observe that our proposed method has either equal or a better chance of evasion of bad mouthing attack regardless of the $N-K$ inflicted.

Note that where the performance between traditional and proposed is equal, those are the $N-K$ which are not part of the rational strategy space of the adversary. From the game solution, we have the Nash equilibrium where adversary compromises $N-K_s = 60$ as marked in the plot and in that equilibrium, probability of evasion by using the proposed model is better than the traditional model.

2) Average Case Performance: We provide an average case performance improvement in terms of probability of evasion and boost in trust scores, by averaging them over various p_b, p_u values for each $N-K$. Similarly, the average boost in event trust for using the proposed beta model is shown in Fig. 8a. The overall improvement in probability of evasion of the proposed model over traditional model in depicted in Fig. 8b. Please note that low boost or low improvement in the probability of evasion is seen for low $N-K$ because, the MCS gets an advantage regardless of whether MTD is used or not, and is not an indication of limiting performance of our method.

VI. Conclusion

In this work, we have presented a randomized sub-sampling method to improve resilience against bad mouthing attacks even when a large fraction of the rater population (especially during cold start phase) is compromised by a strategic adversary. We showed that there exists an optimal sample size that produces an increase in event trust for each potential value of total compromised raters. Finally, we modeled the problem as a two player zero sum game to conclude that there exists a pure strategy Nash equilibrium. We also showed improvement in terms of evading the effect of bad mouthing attack and the boost in event trust under such attacks. In future, we will study whether the approach applies to a setting where ballot stuffing and bad mouthing both are equally likely under a sparse sample setting where adversary can dominate the feedback apparatus.

Acknowledgments: This work is funded by NSF grants SATC-2030611, SATC-2030624, CNS-1818942, CNS-1545037, OAC-2017289 and DGE-1433659.

References

[1] S. Jajodia, A. K. Ghosh, V. Swarup, C. Wang, X. Sean Wang, "Moving Target Defense: Creating Asymmetric Uncertainty for Cyber Threats" *Springer*, 2011.
[2] S. Bhattacharjee, N. Ghosh, V. K. Shah and S. K. Das, "QnQ: A reputation model to secure mobile crowdsourcing applications from incentive losses" *IEEE Conference on Communications and Network Security (CNS)*, pp. 1-9, 2017.
[3] S. Bhattacharjee, N. Ghosh, V. K. Shah and S. K. Das, "QnQ: Quality and Quantity Based Unified Approach for Secure and Trustworthy Mobile Crowdsensing," *IEEE Transactions on Mobile Computing, vol. 19, no. 1*, pp. 200-216, 1 Jan. 2020.
[4] S. Bhattacharjee, S. Debroy, M. Chatterjee, "Quantifying Trust for Robust Fusion while Spectrum Sharing in distributed DSA networks , IEEE Transactions on Cognitive Communications and Networking, Vol. 3(2), pp. 138-154, Jun. 2017.
[5] H. Amintoosi, S. S. Kanhere, "A Reputation Framework for Social Participatory Sensing Systems" *Mobile Netw Appl 19*, pp. 88–100, 2014.
[6] R. Kohlrausch, "Theorie des elektrischen Rückstandes in der Leidner Flasche" Annalen der Physik und Chemie. Vol. 91 (1): pp. 179–213, 1854.
[7] F. Restuccia, N. Ghosh, S. Bhattacharjee, S. K. Das, T. Melodia, "Quality of Information in Mobile Crowdsensing: Survey and Research Challenges" *ACM Trans. Sen. Netw.* Vol. 13(4), Article 34, Dec. 2017.
[8] L. Wang, S. Ren, B. Korel, K. A. Kwiat and E. Salerno, "Improving System Reliability Against Rational Attacks Under Given Resources," *IEEE Trans. on Systems, Man, and Cybernetics: Systems*, Vol. 44(4), pp. 446-456, April 2014.
[9] A. Josang, I. Roslan, "The beta reputation system." *In Proceedings of the 15th bled electronic commerce conference*, Vol 5., pp. 2502-2511. 2002.
[10] A. Josang, "An algebra for assessing trust in certificate chains" *ISORC on Network and Distributed System Security (NDSS)*, Feb. 1999.
[11] B. Yu, M. P. Singh. "An evidential model of distributed reputation management." *In Proceedings of the first international joint conference on Autonomous Agents and Multiagent Systems: Part 1*, pp. 294-301. 2002.
[12] K. Murphy, "Machine Learning: A Probabilistic Perspective (Adaptive Computation and Machine Learning series)", *MIT Press*, 2012.
[13] A. Josang and J. Haller, "Dirichlet Reputation Systems," The Second International Conference on Availability, Reliability and Security (ARES'07), pp. 112-119, 2007.
[14] P. Lopez, M. Behrisch, L. Bieker-Walz, J. Erdmann, Y. P. Flötteröd, R. Hilbrich, L. Lücken, J. Rummel, P. Wagner and E. WieBner, IEEE International Conference on Intelligent Transportation Systems, pp. 2575-2582, 2018.
[15] [Online] www.waze.com
[16] https://en.wikipedia.org/wiki/Kerckhoffs%27s_principle