Attack Context Embedded Data Driven Trust Diagnostics in Smart Metering Infrastructure

SHAMEEK BHATTACHARJEE, Western Michigan University, USA
VENKATA PRAVEEN KUMAR MADHAVARAPU, Missouri Univ. of Science & Technology, USA
SIMONE SILVESTRI, University of Kentucky, USA
SAJAL K. DAS, Missouri University of Science and Technology, USA

Spurious power consumption data reported from compromised meters controlled by organized adversaries in the Advanced Metering Infrastructure (AMI), may have drastic consequences on a smart grid's operations. While existing research on data falsification in smart grids, mostly defend against isolated electricity theft, we introduce a taxonomy of various data falsification attack types, when smart meters are compromised by organized or strategic rivals. To counter these attacks, we first propose a *coarse grained and a fine grained anomaly based security event detection technique* that uses indicators such as deviation and directional change in the time series of the proposed anomaly detection metrics to indicate: (i) occurrence and (ii) type of attack, (iii) attack strategy used, *collectively known as 'attack context'*. Leveraging the attack context information, we propose three *attack response metrics* to the inferred attack context: (a) an unbiased mean indicating a robust location parameter; (b) a median absolute deviation indicating a robust scale parameter, (c) an attack probability time ratio metric indicating the active time horizon of attacks. Subsequently, we propose a trust scoring model based on *Kullback-Leibler (KL)* divergence, that embeds the appropriate unbiased mean, median absolute deviation and attack probability ratio metric at runtime to produce trust scores for each smart meter. These trust scores help classify compromised smart meters from the non-compromised ones. The embedding of the attack context, into the trust scoring model, facilitates accurate and rapid classification of compromised meters, even under large fractions of compromised meters, generalize across various attack strategies and margins of false data. Using real data sets collected from two different AMIs, experimental results show that our proposed framework has a high true positive detection rate, while the average false alarm and missed detection rates are much lesser than 10% for most attack combinations for two different real AMI micro-grid datasets. Finally, we also establish fundamental theoretical limits of the proposed method, which will help assess applicability of our method to other domains.

Additional Key Words and Phrases: Advanced Metering Infrastructure, Trust, Smart Grid Security, Smart Metering, Anomaly Detection, Data Integrity, Data Falsification Attacks, Artificial Intelligence based Security

ACM Reference Format:
Shameek Bhattacharjee, Venkata Praveen Kumar Madhavarapu, Simone Silvestri, and Sajal K. Das. 2020. Attack Context Embedded Data Driven Trust Diagnostics in Smart Metering Infrastructure. *ACM Transactions on Privacy and Security.* 1, 1, Article 1 (October 2020), 35 pages. https://doi.org/10.1145/3426739

Authors' addresses: Shameek Bhattacharjee, Western Michigan University, Rolla, USA, shameek.bhattacharjee@wmich.edu; Venkata Praveen Kumar Madhavarapu, Missouri Univ. of Science & Technology, Rolla, USA, vmcx3@mst.edu; Simone Silvestri, University of Kentucky, Lexington, USA, silvestri@cs.uky.edu; Sajal K. Das, Missouri University of Science and Technology, Rolla, USA, sdas@mst.edu.

2471-2566/2020/10-ART1 $15.00
https://doi.org/10.1145/3426739

1 INTRODUCTION

Advanced Metering Infrastructure (AMI) is one of the basic units of the smart grid technology. AMI collects data on loads and customer's power consumption [20], from Smart Meters installed on the customer site (see Fig. 1). Such data plays a pivotal role in several critical tasks such as automated billing, demand response, load forecast and management [9, 20, 23].

Apart from automated billing, strategic tasks are expected to be performed by future smart grids, based on the AMI power consumption data. For example, AMI will have implications on tasks such as daily and critical peak shifts [8, 14, 36]. When the consumption increases beyond a critical limit, emergency 'peaker plants' are currently used by most utilities for additional power generation to meet the demand. However, such peaker plants are extremely carbon as well as cost intensive. In the modern grid, the utility will also have the option for automated demand response where utilities pay customers to shut certain appliances temporarily (peak shifting) to obviate the need for additional generation [13, 35]. In general, an accurate short or long term data on loads and consumption will aid in accurate demand response, load forecast and planned generation in the future smart grid [1]. Therefore, the integrity of the AMI data is of utmost importance.

Defense against falsification of power consumption data from AMIs, has largely focused on *electricity theft* [12, 17, 19, 30], where individual customers are primary adversaries who report lower than actual usage for lesser bills. Since isolated smart meters belonging to rogue customers reduce the value of power consumption, we term such an adversarial attack as a *Deductive* mode of data falsification. However, it has been widely acknowledged that given the cyber and interconnected nature of AMI, it could potentially be the target of organized adversaries such as cyber criminals [34], utility insiders [38], or business competitors [15]. Organized adversaries can compromise *several smart meters* and then *spoof* false power consumption data [17] from smart meters. Organized adversaries are more equipped to crack/leak cryptographic secrets, have a higher attack budget, and possess the ability to simultaneously attack other elements of the grid (e.g., audit logs, transformers meters) in order to avoid easy consistency checks on false data. Existing research does not focus on defense against such adversaries and is only restricted to electricity theft from isolated customers as primary adversaries. Given that electricity theft is targeted at individual customer gain, the margin of false data is usually arbitrary [12] and typically high [16] such that there is a tangible benefit to each customer, thus facilitating easier detection.

Additionally, the goals of organized adversaries are not just restricted to monetary benefits on the customer side. As a recent example, in Netherlands [33] a manufacturer installed a large number of faulty smart meters (not proved whether it was erroneous or deliberate act) that reported 6 times higher than actual power consumption. We term such an attack as an *Additive* mode of data falsification. For example, an additive attack may be launched by a rival utility on its competing company's meters, that may induce loss of business confidence by the customers of the victim company, due to higher bills as reported in [32]. A class action lawsuit filed against a victim utility was reported in this case. If the utility participates in demand response, it may lose revenue from additive attacks for undue incentives payed to customers for induced peak shifts. Indirectly, additive attacks can be triggered by a load altering attack*(LAA)* [22], thus increasing the net consumption sensed by the smart meter. It may also be noted that adversaries may orchestrate large scale deductive attacks to cripple the utilities with huge revenue losses [38].

Apart from the additive and deductive attacks, we argue the possibility of more complex attack types in AMI. For example, a balancing additive and deductive attack with the same margin of falsification, could evade mean aggregate (or location parameter) based detection models. We term such a strategy as a *Camouflage* attack, which may be motivated for generating lesser bills to one set of customers at the expense of the other set. Such attacks may stay undetected, without raising

any suspicion because the total inflow and outflow of power measured (or mean predictions) at the transformer meters, and the total demand and reported usage remain relatively unchanged. The attacker in such a case does not need to attack other elements in the grid (for e.g. transformer meters) to prevent easy consistency checks. In general, random additive and deductive attacks may simultaneously coexist in the same AMI network, when launched by different adversaries with conflicting goals. We term such a scenario as a *Conflict attack*, that is a mixed attack type with unequal margins of falsification for each underlying simple attack type. Apart from the four attack types, there could be special attack strategies such as: (i) Data Omission attacks, where the data is prevented from reaching the utility (ii) On-Off attacks, where the adversary only attacks on specific hours of the day, and (iii) Data Distribution Order Aware Attacks, where the attackers ensure that the falsified data is more proximate to the original data distribution than usual assumed random strategies. While, some works have explored on-off strategies [12], data omission and order aware attacks have not gained attention. To summarize, existing defense frameworks cannot handle all of the above falsification types and strategies *simultaneously*.

Contributions: In this paper, we introduce a taxonomy of multiple possible data falsification attacks and strategies launched by organized adversaries. Then, we design two novel coarse grained and fine grained statistical invariants (based on Pythagorean Means) of aggregate power consumption data and learn their time series under no attacks. By exploiting knowledge of the impacts of each attack type on these invariants, *a coarse grained and a fine grained anomaly based security event detection criterion* is proposed that collectively indicates the *attack context that includes 'occurrence', 'type', and 'strategy' of different attacks.* Such detection criterion with attack context unlike existing works, better discriminate between attacks from legitimate changes and accordingly generates the following *attack responses: (i) replaces location and scale parameters with an robust mean and robust median absolute deviation respectively, (ii) calculates an attack probability time ratio metric.* Subsequently, we propose a Kullback-Leibler divergence based Relative Entropy Trust Model that embeds the attack responses from the attack context, in a way that identifies compromised meters with a high detection rate in a quicker time with less false alarms. Experimental results using real datasets from different AMIs, show that our detection technique is able to identify compromised meters with higher detection rates in quick time while incurring lower false positives, than recent works in the area, under various attack strategies employed by adversaries. Finally, we perform extensive formal security analysis to show the performance limits of our model.

Novelty: Our proposed work is the first effort to establish trustworthiness in AMI against multiple attack types and faults with coarse and fine grained attack strategies. Secondly, our focus is on orchestrated data falsification attacks devised by organized adversaries rather than just rogue customers. Our method works well for even higher fractions of compromised meters, unlike most statistics based methods due to the embedding of real time attack responses into the trust model. To demonstrate detection sensitivity in terms of margin of false data, we assume the full attack strategy space and show that detection rates are high across a wider threat landscape. Additionally, our method's time to detection of compromised meters is quick even under opportunistic attack strategies that are sporadic over time domain, via attack time probability ratio embedding. Our proposed method is light weight and gives better performance compared to the classical bad data detection mechanisms which use expensive multi-class SVM and neural network based training models. We also discuss about the limitations of our proposed framework *under adversary's knowledge of our defense mechanism*, which motivates the direction in which further research should be conducted.

The rest of the paper is organized as follows. Section 2 describes the system and threat models while Section 3 discusses the proposed framework with theoretical analysis. Section 4 includes

a special embedding method required to counter opportunistic attacks. Section 5 discusses the experimental results and Section 6 concludes the paper.

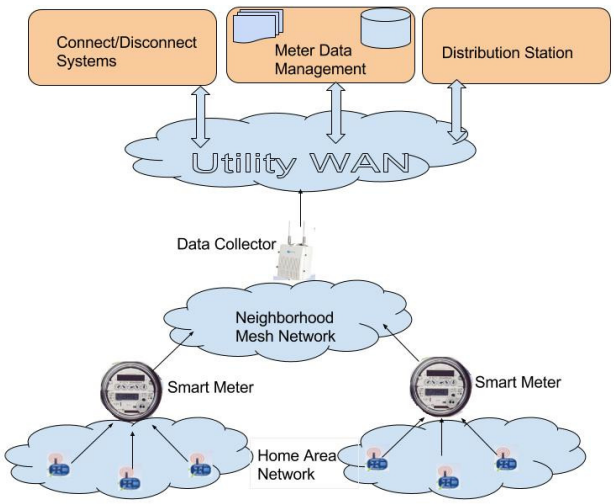


Fig. 1. Architecture of AMI [3]

2 SYSTEM AND THREAT MODELS

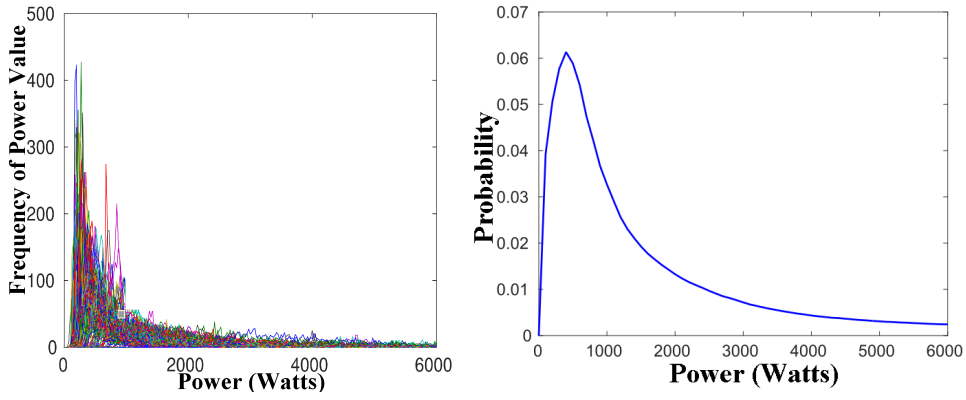
In this section, we discuss the network architecture of the AMI, characterize the distribution of two real datasets, and propose the threat model for organized data falsification in AMI.

2.1 Architecture

We consider a collection of N smart meters reporting power consumption data to a Neighborhood Area Network (NaN) Gateway (acts as an edge computing node) periodically and independently. The i-th smart meter, records an actual power consumption data, say $P^i_{act}(t)$ at the end of each time slot t (t is slotted *hourly*). The reported power consumption $P^i_{rep}(t)$ is equal to $P^i_{act}(t)$, if i is not compromised. However, $P^i_{rep}(t) \neq P^i_{act}(t)$, if i is compromised by an adversary. We model $P^i_{act}(t)$ as the realizations of a random variable P^i, that denotes power consumption from the i-th meter. The NaN gateway piggybacks data from each smart meter and sends it to the utility via a Wide Area Network (WaN) Gateway that collects data from multiple such NaN gateways. Occasionally, there is another network hierarchy known as the Field Area Network (FAN) gateway which connects NaN and WaN and may host edge computing services. Both FaN and WaN may host the security monitoring mechanisms. Deploying security mechanism at the FaN is a decentralized implementation, while deployments at the WaN is a centralized implementation. Our framework works regardless of the implementation. The current evaluation proposed mechanism assumes a decentralized implementation given the size of the microgrid datasets. Moreover, [6] has observed the benefits of decentralized security implementations over centralized ones.

2.2 Data Set Characterization and Transformations

To characterize the distribution of the random variable P^i from the i-th smart meter, we conducted preliminary investigations on real power consumption data sets with 800 [40] (Texas Dataset) and 5000 meters [42] (Irish Dataset) collected on an hourly basis. The Texas dataset contains data across the years 2014, 2015 and 2016. Throughout the paper, data from 2014 and 2015 are used as the historical training set, while 2016 serves as a testing set. The Irish dataset contains approximately 535 days of data from years 2009-2010, that we use to prove the generality of our results.



(a) (b)

Fig. 2. Power Consumption Distribution: (a) All Houses (b) Mixture

Each home consists of one smart meter in the datasets. We observed that for each meter, the power consumption can be approximated as a log normal distribution. We also observed that all such log normal distributions are *clustered close* to each other; that is, the variance between them is not arbitrarily large. Fig. 2(a) summarizes the results from all the houses in the Texas dataset. Thanks to this observation, we can approximate the aggregate of the individual log normals using a mixture distribution, which is also lognormal as evident from Fig. 2(b). Let P_{mix} denote the approximate lognormal mixture of all P^i.

Next we transform all P^i using a Box-Cox transformation technique [5] to obtain an approximate *normally distributed* r.v. denoted as $\hat{P^i}$. Let $\hat{P}_{mix}$ denote the mixture of all the $\hat{P^i}$. Results of $\hat{P_{mix}}$, for different months is depicted in Fig. 3(a). The box-cox transformation serves a dual purpose. First, it maps the data points to a lower portion real axis. Some interesting statistical properties of proposed Pythagorean Mean based invariants are more prominent in this lower-dimensional real axis which increases the relative sensitivity of Harmonic Mean to Arithmetic mean differences and their ratios (used for detecting anomaly) under false data injections. Below, we describe the box-cox transformation technique and how we apply it in our context.

2.2.1 Box-Cox Transformation:. The transformation of non-normal data into approximate normal distribution can be achieved using the following method. Given any set of data-points $\boldsymbol{D} = \{D^{(1)}, \cdots, D^{(k)}, \cdots, D^{(n)}\}$, where n denotes the total number of data points in $\boldsymbol{D}$, the box cox transformation of $\boldsymbol{D}$ is given by $\hat{\boldsymbol{d}} = \{d^{(1)}(\lambda), \cdots, d^{(k)}(\lambda), \cdots, d^{(n)}(\lambda)\}$:

$$d^{(k)}(\lambda) = \begin{cases} \frac{(D^{(k)})^{\lambda}-1}{\lambda} & \text{if} \quad \lambda \neq 0; \\ ln(D^{(k)}) & \text{if} \quad \lambda = 0 \end{cases} \tag{1}$$

where λ is an appropriate transformation parameter chosen from a possible set $\lambda^* \subseteq \mathcal{R}$, such that

$$\lambda = \underset{\lambda \in \mathcal{R}}{\operatorname{argmax}} f(\boldsymbol{D}, \lambda^*)$$

where $f(\boldsymbol{D}, \lambda^*)$ is the logarithm of the likelihood function given by:

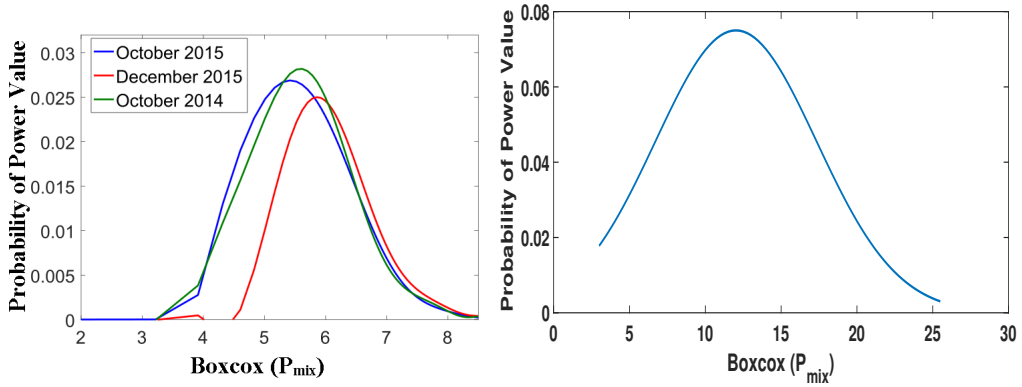
$$f(D, \lambda) = -\frac{n}{2} ln \left[\sum_{i=1}^{n} \frac{[d^{(k)}(\lambda) - \bar{d}(\lambda)]^2}{n} \right] + (\lambda - 1) \sum_{i=1}^{n} ln(d^{(k)}) \tag{2}$$

such that $\bar{d}(\lambda) = \frac{\sum_{i=1}^{n} d^{(k)}(\lambda)}{n}$ is the arithmetic mean of the transformed data.

2.2.2 Applying Transformation to the Datasets: The data from each smart meter i (analogous to D) is transformed onto the box cox transformed scale by using the above procedure. Thereafter, we build the time series of the whole dataset in the box cox transformed scale as:

$$\hat{p}(t) = \{\hat{p}^1(t), \cdots, \hat{p}^i(t), \cdots, \hat{p}^N(t)\}$$

where $\hat{p}(t)$ denotes the reported time series over all smart meters $i \in \{1, N\}$ at each time slot. The appropriate λ is learned from the historical training set (2014, 2015), and the same is applied to the testing set (2016) and Irish Dataset (2010). To prove the generality of this method, we repeated the experiments for the Irish data set [12], and reported similar results which are included in the preliminary version of our work [2]. The distribution for Irish dataset after box-cox transformation After the transformation, 67% and 68% of data points fall within the first standard deviation for the Texas and Irish data sets respectively. However, the resultant distributions as a whole are not symmetric about the mean and 64% and 69% of the data are lesser than the mean and 36% and 31% of the data are greater than of the mean. This asymmetry is another factor that affects the observations in the anomaly detection phase.



(a) (b)

Fig. 3. After BoxCox: (a) Monthly Texas (b) Yearly Irish

2.3 Time and Space Domain Granularities of Detection

In this paper, for various time series and trust score calculations, we use different time domain granularities. An hourly time granularity is referred to as a time 'slot' denoted by t. A daily time granularity consisting of 24 time slots is referred to as a time 'window' denoted by T. A collection of time windows is referred to as a time 'frame' denoted by F (e.g. 30 days). In this paper, the anomaly detection has two versions. The coarse-grained version runs on a daily basis (with T) while the fine-grained version runs on an hourly basis (with t). In this paper, the Kullback-Leibler (KL) divergence based trust model assigns trust of each smart meter at the end of a frame (say every (a) 10 days (b) 30 days) based on evidence and observations that are collected on an hourly basis. We assume that the whole defense framework is running in a decentralized manner, that is deployed or hosted in an edge device such as a NaN gateway in the AMI network. In such cases, the usual micro-grid size under observation varies between 100-1000 houses. For a larger grid, our trust model will also work for centralized implementations seamlessly, however, the anomaly based security event detection will need to be decentralized as pointed out by several prior works [6].

2.4 Threat Model

Now let us introduce a detailed threat model for our framework.

2.4.1 ***Adversary Types and Scope:*** In this paper, we keep the scope of threats to be a little less specific, since, in the real world, the defender has no control over what kind of adversary will attack its infrastructure. A good defense model is one can capture a wider range of attacks from various adversary types who can have various creative strategies to launch their intended attack objectives. In this light, we divide the threat model specification into four features: data falsification attack types, fraction of compromised meters, margins of false data, and attack strategies that specify a wide threat landscape. The attack strategies are further divided into continuous and opportunistic strategies. The paper's core contribution (Section 3) by default considers the continuous strategies. Thereafter, it explains the modifications required to the core contribution separately in Section 4 for opportunistic attack strategies.

We assume that the organized adversary belongs to either business competitors or organized cyber-criminals, who possess the ability to compromise several smart meters by bypassing cryptography or manipulating its sensory inputs, or utility personnel who might manipulate several smart meters physically [38] (e.g., via optical probes [34]). False power consumption data from a meter can be achieved in the following ways: (a) manipulation of inputs to the meter [22], (b) manipulating data content in the meter [34], and (c) in-flight from the meter [18] to NaN gateway. A meter is compromised if either the input, content, or output is modified from the actual value. The adversary launches data falsification from multiple such compromised smart meters concurrently. In another variation, the attacker could take control of NaN gateway where it could intercept data from multiple smart meters at once and launch smarter attack strategies as discussed later.

We assume attackers who may have a long or short term damage objective. *Long term damage* requires evading detection easily, while still benefiting from attacks. The adversary may accept some initial loss in the hope of avoiding easy detection and accruing incremental benefits over time. Examples of long term adversarial objectives include monetary gains in terms of electricity pricing and belief manipulation of learning based demand forecast models. In this case, the false data margins are typically smaller. A *short term damage*, on the other hand, requires inflicting the maximum damage in a short time, before getting detected. Examples of short term objectives include an attacker aiming to, gain quick revenue or masquerade an illegitimate demand response event. Due to the contrasting requirements on these two objectives, important adversarial parameters such as the fraction of compromised nodes ρ_{mal} and the margin of false data δ_{avg} can readily vary depending on the nature of time deadlines associated with such objectives. In the real world, we can encounter attackers with rational or irrational objectives, and hence it is important to explore all combinations of ρ_{mal} and δ_{avg}.

2.4.2 ***Data Falsification Attack Types in AMI:***. We define the manner in which the actual power consumption data $P^i_{act}(t)$ of each meter i is modified as the *mode* of data falsification. We identify the following modes:

Additive: The adversary reports $P^i_{rep}(t) = P^i_{act}(t) + \delta(t)$, where $\delta_{min} \leq \delta(t) \leq \delta_{max}$. This mode can lead to loss of business confidence from customers due to higher bills and masquerade a demand peak leading to remote disconnect of customer appliances, thereby inducing utilities to pay undue incentives.

Deductive: The adversary reports $P^i_{rep}(t) = P^i_{act}(t) - \delta(t)$, where $\delta_{min} \leq \delta(t) \leq \delta_{max}$. This mode can lead to loss of revenue for power utility companies.

Camouflage: The adversary divides the compromised meters into two teams equal in number, which simultaneously adopt an additive and deductive mode, respectively with an equal $\delta(t)$. This mode can favor a smart meter of one power utility at the expense of others and has less impact on the strategic decisions in the grid. It cannot be detected by mean (parametric) based anomaly detectors, because no suspicion is raised due to negligible change in the mean power consumption.

Conflict: It is a scenario where additive and deductive attacks coexist simultaneously but are not balanced (i.e. uncoordinated). Such a scenario represents random attacks possible if there are more than one uncoordinated adversarial teams or multiple dishonest customers acting randomly.

2.4.3 ***Average Margin of False Data*** δ_{avg}. The value $\delta(t)$ is generated randomly within an interval $[\delta_{min}, \delta_{max}]$, for $\delta_{min}, \delta_{max} > 0$, and accordingly added to or deducted from the actual power consumption. Note that, arbitrarily high δ_{max} may facilitate intuitively easy detection, while very low δ_{max} hardly accrues any revenue. The average of $\delta(t)$ is a strategic value, denoted by δ_{avg} and referred to as the *margin of false data*. The units of δ_{avg} values in this paper is in Watts.

Apart from the type of attack, the attacker chooses a strategic value of δ_{avg} in the interval $[\delta_{min}, \delta_{max}]$ as part of its strategic objective. The inflicted δ_{avg} may be high or low depending on the amount of damage it wants to inflict, and the short or long time horizon of the attack. However, the attack margin $\delta(t)$, maybe more in peak periods than non-peak periods, to exploit the time-dependent pricing of electricity.

2.4.4 ***Attacker Budgets and Fraction of Compromised Meters***. We assume that organized adversaries compromise a certain number M of N smart meters based on *attack budget*. The fraction of compromised nodes is $\rho_{mal} = \frac{M}{N}$, which can be high for smaller microgrids with a small N.

With higher ρ_{mal}, the adversary can afford to decrease the margin of false data (per meter) to avoid getting intuitively and easily detected. Although the attack cost increases in these cases, the adversary may reduce the chance of detection, and look to recover the initial cost in the long term. This is however not an option for adversaries with short term objectives. A concrete mathematical example of this aspect can be found in the preliminary version of this work [2].

2.4.5 ***Attack Strategies***. Now we represent different falsification strategies of additive, deductive, camouflage, and conflict attacks for any δ_{avg} and ρ_{mal}. The attack strategy describes how the $\delta(t)$ bias values are distributed over the interval $[\delta_{min}, \delta_{max}]$ as well as over the time domain. In this paper, we study attack strategies such as (*uniform random bias, data distribution order aware*) that are 'continuous' over time, while (*on-off, data omission*) are *opportunistic strategies* that are discontinuous and fine grained over the time domain due to their opportunistic nature.

Uniform Random Bias: Since the distribution of power consumption is unimodal, the attacker refrains from any strategy that would make the resultant distribution multi-modal. In that sense, a uniformly distributed random $\delta(t)$ injected into the actual smart meter data does not change the overall shape of the distribution but only affects its location and or scale parameters as proved in our early work [2]. Such variants of a uniform distribution over time have been adopted from [12].

Data Distribution Order Aware: We use this strategy where the adversary matches the $\delta(t)$ values with the actual consumption data recorded from its compromised set of meters. The adversary sorts the data from its controlled meters and the false bias $\delta(t)$ values from smallest to largest on every time slot. For an additive attack, the smallest recorded value is biased with the highest value in the set of $\delta(t)$, and so on. For deductive attack, the smallest data from the meter set is biased with smallest value in the set of $\delta(t)$. For camouflage attack, since compromised meter population is divided into 2 parts, the additive and deductive order aware strategy are implemented similarly. Such an attack can prevent obvious outliers, which is much better than a simple uniform random strategy, and therefore more proximate to the actual data distribution, thus making it harder to detect. Detailed implementation of the attack and proof that it follows the original data distribution more closely is provided in Appendix A. We use this specifically for deductive attacks in the paper, since $P^i_{rep}(t)$ are lower bounded by 0. This does not allow realizing the full δ_{avg} margin in the attack and too many zero values that can raise easy suspicion. Such a strategy is possible if the adversary controls some NaN gateway.

On-Off: In this case, the adversary's orchestrated attacks are distributed between OFF periods (no falsification is launched) and ON periods (falsification is launched) over time. Such an attack is inspired by dynamic pricing nature of the electricity. For example, if malicious purpose is electricity theft, the attackers launch attacks during time slots when the electricity prices are the highest.
Data Omission: The data may be prevented from reaching the NaN/FaN gateways, either (i) intentionally (omission attack) or (ii) due to accidental network failures (omission failure). An example of an omission attack is the jamming of the wireless channels that carry the data over the mesh network to the edge gateway. Such omission attacks cripple data availability which affects decision and analytics in the smart grid. Alternatively, some hops of the mesh network may fail or channels may occasionally drop some packets due to network collisions. Hence, no data reaches from particular meters on the concerned time slots, which is termed as an omission failure.

We use the uncleaned dataset (which contains missing data on certain times slots) to show that we can detect such omission failures. In contrast, we simulate omission attacks by dropping the data from ρ_{mal} fraction of compromised meters. We believe that omission failures tend to be very random and infrequent, while omission attacks are likely to be more frequent. Additionally, omission attacks depend on the opportunity to thwart the communication resources, and therefore the time between successive attacks is not necessarily periodic like an ON-OFF strategy.

2.5 Overview of the Proposed Framework

The proposed framework has three phases: (a) Anomaly-based Security Event Detection (b) Security Event-based Attack Context and Response Generation (c) Attack Context Embedded Trust Scoring Model. The anomaly detection phase indicates the nature of the security event in terms of the information such as the presence, type, strategy, and strength of the concerned data falsification attack. Such information extracted from the security event detection aids in the calculation of certain attack response metrics such as an unbiased robust mean, a median absolute deviation, and an attack probability time ratio by the attack context generation phase. Such attack response metrics are supplied to the trust scoring model phase that calculates a linearly separable score for each meter and uses it to identify the compromised meters launching data falsification attacks. The embedding of the attack context based response metrics improve the accuracy of compromised meter classification and the classification convergence times regardless of the attack types, margins, and strategies inflicted.

The anomaly detection phase is further divided into two parts: (i) coarse grained anomaly detection for attacks for all strategies except on-off and omission strategies; (ii) fine grained anomaly detection for on-off and omission strategies. Note that, the coarse and fine grained anomaly detectors run simultaneously in the framework since any attack strategy is possible in reality. While both anomaly detection variants help to calculate the robust mean and median absolute deviation, the attack probability time ratio is relevant only for the fine grained anomaly detector. The trust model is further divided into three parts: (a) estimating parameters of true proximity distributions (b) estimating parameters of observed proximity distribution with appropriate attack context embedding (c) The Kullback-Leibler Divergence calculation.

3 PROPOSED FRAMEWORK FOR DATA DRIVEN TRUST DIAGNOSTICS

In this section, we propose the coarse grained and fine grained anomaly based security event detection scheme. The proposed event detection scheme leverage the properties of how different data falsification types change the Pythagorean Means (such as Harmonic, Geometric, Arithmetic Means) of an attacked time series. We propose an invariant for both coarse and fine grained anomaly detection schemes, that is stable under no attacks. The evidence of invariant stability

is proved through two real datasets gathered from 200 from a solar village in Texas, USA [40], and 5000 smart meters in Dublin, Ireland. Then, we show how these invariants exhibit visibly evident changes under various attacks, which forms the premise for inferring the presence of attack, type of data falsification, and the strategy used by the adversary that collectively reconstructs the security event. Based on the nature of the security event, an attack context is generated (in the form of robust mean, median absolute deviation, attack probability time ratio). The attack context information is forwarded to the trust based scoring model which enables accurate identification of the compromised smart meters.

3.1 Anomaly based Security Event Detection

First, we propose the detection metric (or invariant). Second, we explain the reasoning behind the design of the proposed invariant. Third, we establish the normal range of the under no attacks. Fourth, we propose the detection criterion to detect the occurrence of an orchestrated attack that needs a consensus (location and scale) correction. Fifth, we show how the attack type could be determined given the incidence of attack. Finally, we show how the knowledge of the incidence of attack and its corresponding type is used to estimate an approximate robust mean and median absolute deviation (collectively called as robust consensus measures). Information on the robust consensus measures is supplied to the entropy based trust model for improved classification that maximizes detection sensitivity for a wide range of δ_{avg} and ρ_{mal} values while minimizing the incidence of false alarms.

3.1.1 Pythagorean Means. The various Pythagorean means (arithmetic, geometric and harmonic means) in a particular time slot t is given by:

$$AM(t) = \frac{\sum_{i=1}^{N} \hat{p}^i(t)}{N} \quad , \qquad GM(t) = \Big(\prod_{i=1}^{N} \hat{p}^i(t) \Big)^{\frac{1}{N}} \quad , \qquad HM(t) = \frac{N}{\sum_{i=1}^{N} \frac{1}{\hat{p}^i(t)}}$$

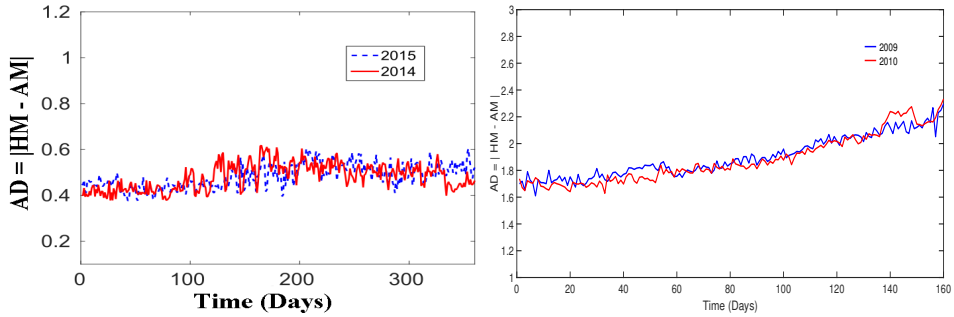
The average of all these hourly means $AM(t), GM(t)$ and $HM(t)$ over a particular day ($t \in [1, 24]$) is represented by $\overline{AM}(T)$, $\overline{GM}(T)$, and $\overline{HM}(T)$ respectively where $T \in [1, 365]$. For example, $\overline{AM(T)} = \sum_{t=1}^{24} AM(t)/24$) and so on. Due to the well known Pythagorean mean inequality, $\overline{HM}(T) \leq \overline{GM}(T) \leq \overline{AM}(T)$ holds.

3.1.2 Proposed Coarse Grained Invariant ($AD(T)$). From our statistical studies over two big datasets, we discovered that the time series of the absolute difference between average daily harmonic and arithmetic mean power consumption is an effective invariant across datasets. Theoretical reasoning behind the stability of the harmonic mean and arithmetic combination has been extensively discussed and presented in our previous work [3]. Formally, the coarse grained invariant is quantified by $AD(T)$ and is defined as:

$$AD(T) = \Big| \quad \overline{AM}(T) - \overline{HM}(T) \quad \Big| \tag{3}$$

Eqn. 3, is designed as an anomaly detection metric for two main advantages: First, the time series of $AD(T)$ is a highly stable invariant of the aggregate power consumption, compared to other parametric and non-parametric measures that are functions of the instantaneous or historical arithmetic mean power consumption as proved in our previous work [3]. Furthermore, our previous work in the context of smart transportation systems [26, 28] showed that this observation of stationarity in harmonic and arithmetic mean generalizes across application domains under careful spatial and temporal considerations. High invariance over time or a given context is one of the desired properties of anomaly detectors [7].

Fig. 4(a) shows the instantaneous values of $AD(T)$ for two different years (2014 and 2015). It can be verified that under no attacks, the average value of $AD(T)$ is about 0.49 and the values are relatively stable over time across both years. Similarly, Fig. 4(b), shows the time series of $AD(T)$ for the portion of the Irish dataset that has a historical overlap between two years 2009-2010. The $AD(T)$ of the Irish dataset is stable over history, since AD(T) on the T-th day is one year is not arbitrarily different from the AD(T) of the corresponding T-th day in the previous year. Both Figs. 4(a) and 4(b) is in complete contrast to the Fig. 5 that shows the average arithmetic mean $\overline{AM}(T)$ for the Texas dataset, can be seen as neither stable over time or over history. As it is well known that anomaly detection metrics ideally need high invariance under normal operations, we, therefore, conclude that $AD(T)$ a better invariant compared to any derivative of the popular arithmetic mean and standard deviation. Additionally, since the values are not arbitrarily different, the variance in the $AD(T)$ samples is also lesser compared to the variance in $\overline{AM}(T)$ samples. An elaborate theoretical explanation on the generality of the observed stability of the absolute difference between $\overline{HM}(T)$ and $\overline{AM}(T)$ across datasets is elaborated in detail in Appendix B.



(a) (b)

Fig. 4. Time Series of proposed $AD(T)$: (a) Texas Dataset (b) Irish Dataset

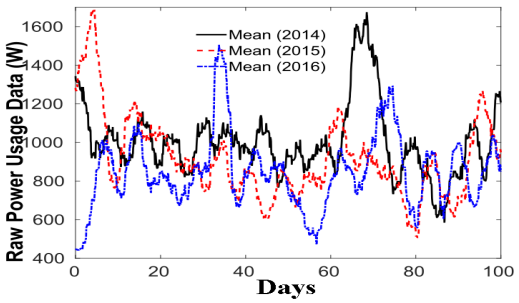
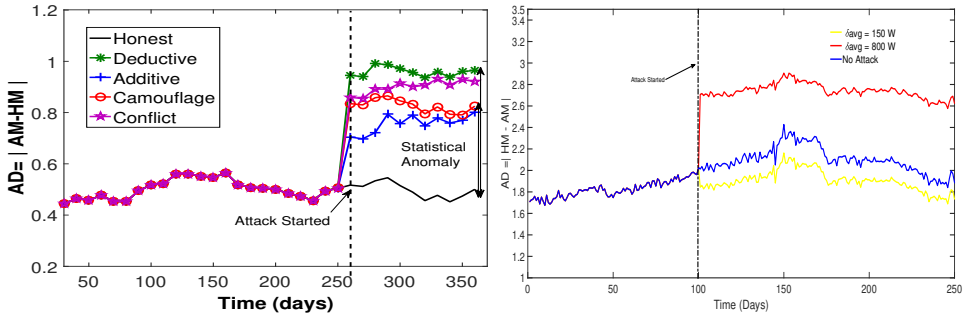


Fig. 5. Unstable $AM(T)$ for Texas Dataset

3.1.3 Summary of Security Properties of Proposed $AD(T)$. The second advantage is that harmonic, geometric and arithmetic mean possesses certain special mathematical properties that produce unique changes in the time series of $AD(T)$, whenever data falsification occurs from a subset of data sources, that otherwise produced a stationary $AD(T)$.

While harmonic mean and geometric mean is a strictly Schur-Concave function [29], the arithmetic mean is both Schur Concave and a Schur Convex function of its arguments (the numbers involved in the calculation of the means). Such a difference in the strictness of Schur-Concavity property produces six unique novel properties in the context of data falsification that we had identified. The direction of deviation depends on the skewness in the datasets, but the fact there



(a) (b)

Fig. 6. $AD(T)$ deviation under attacks (a) Texas Dataset (b) Irish Dataset

will be deviation is generic and independent of the datasets. These six properties are divided into two-sub groups based on the direction of change in the $AD(T)$. The direction of change in $AD(T)$ is dependent on whether the δ_{avg} is greater or lesser than a certain threshold Γ, given a particular ρ_{mal}, attack type, and the skewness in the data distribution. The theoretical and experimental proof of the properties has been established in our earlier work [3]. For the sake of completeness, we now provide a summary of these properties in harmonic and arithmetic means, that cause the deviation in $AD(T)$ under attacks.

Case 1: For all attacks with $\delta_{avg} > \Gamma$, the following hold true:

Property 1: Under additive attacks, the harmonic mean grows slower compared to the arithmetic mean, thus $AD(T)$ will increase.
Property 2: Under deductive attacks, the harmonic mean decays faster compared to the arithmetic mean decay rate, thus $AD(T)$ will increase.
Property 3: Given the same δ_{avg} and the same set of arguments, the decay in harmonic mean is larger for deductive attacks compared to growth in harmonic mean for additive attacks. Therefore, in a camouflage attack with the same δ_{avg}, the resultant harmonic mean will be lesser than the original harmonic mean, while the arithmetic mean will not change. Thus $AD(T)$ will increase.
Effect on properties 1,2, and 3 on $AD(T)$: It is easy to conclude that all the above three properties will cause the $AD(T)$ *to increase after attacks* compared to before attacks because the gap between $\overline{HM}$ and $\overline{AM}$ widens, so does its absolute value represented by $AD(T)$. This is experimentally verified in Fig. 6(a), where attack injected after the 250-th day shows a sharp increase in the $AD(T)$ for various attack types.

Case 2: For all attacks with $\delta_{avg} < \Gamma$, the above three properties are reversed and the following hold true:
Property 4: Additive attacks will show larger growth in harmonic mean compared to arithmetic mean growth, thus $AD(T)$ will experience a decrease.
Property 5: Deductive attacks will show smaller decay compared to arithmetic mean decay, thus $AD(T)$ will experience a decrease.
Property 6: $AD(T)$ will decrease if actual number of data points attacked with additive are smaller than the actual mean. This is typically true for power consumption datasets that are right skewed, hence the mean is shifted towards the right tail of the distribution. If such data is attacked, on

average more number of datapoints being modified will be smaller than the actual arithmetic mean. *Effect on properties 3,4, and 5 on $AD(T)$:* It is easy to conclude that all the above three properties will cause the $AD(T)$ *to decrease after attacks* compared to before attacks because the gap between $\overline{HM}$ and $\overline{AM}$ narrows, so does its absolute value represented by $AD(T)$. This is experimentally verified in Fig. 6(b), where attack injected after the 100-th day shows a decrease in the $AD(T)$.

Approximation of crossover Γ: For directional switching of the proposed $AD(T)$, the approximate bounds on the average value of Γ and its details have been published in our earlier work [3]. The closed form of Γ is not possible due to non-existence of the closed form. However, approximation of the lower and upper bounds are given by the following The approximate (average case) lower bounds are: $\Gamma^{-}(rlow) = \Gamma^{+}(llow) =$

$$\Gamma_{low} = \frac{\sigma}{M} + \frac{\sigma}{\sqrt{M}}\sqrt{\frac{N-M}{N-1}} + \sigma \tag{4}$$

where + and − superscripts denote additive and deductive manipulation and l and r denote whether the bias points are on the left or right of the actual mean. The approximate upper bounds are: $\Gamma^{+}(lhigh) = \Gamma^{-}(rhigh) =$

$$\Gamma_{high} = max(\sigma^2, \frac{2\sigma}{M} + \frac{\sigma}{\sqrt{M}}\sqrt{\frac{N-M}{N-1}} + 2\sigma) \tag{5}$$

3.1.4 Identifying Normal Range of $AD(T)$. Let the standard deviation of the $AD(T)$ samples in the training set be denoted as $\sigma_{AD(T)}$. Given that the $AD(T)$ metric is stable over history as evident from earlier results, the normal range can be a residual margin around the historical values. The margins can be a parameterized by a scalar factor $\gamma \in (0, 3]$ of the standard deviation of the $AD(T)$ samples, such that the upper threshold for $AD(T)$ at the T-th window in the testing set i given by:

$$AD^{test}_{max}(T) = AD_{hist}(T) + \gamma\sigma_{AD(T)}$$

and the corresponding lower threshold is:

$$AD^{test}_{min}(T) = AD_{hist}(T) - \gamma\sigma_{AD(T)}$$

Please note that, it is possible that smaller δ_{avg} (stealthy) or smaller ρ_{mal} values (isolated or small scale adversaries) will not create enough deviation for the $AD(T)$ to fall outside the $AD^{norm} \in [AD^{test}_{min}, AD^{test}_{max}]]$ range. However, such smaller attacks will also not drastically affect the consensus measures (mean and standard deviation). As mentioned earlier, the one of the purpose of the anomaly detection phase in our framework is to provide an unbiased instantaneous mean and median absolute deviation to the trust model across either a high ρ_{mal} or δ_{avg} values. Therefore, successful detection of incidence and type of attack, is only required when attacks are strong enough to influence the consensus significantly. To this end, the simple definition of $AD^{norm} \in [AD^{test}_{min}, AD^{test}_{max}]]$ is sufficient. If attacks are not detected, however, at the same time they do not affect the consensus in a significant way. In such cases, the trust scoring model proposed later will be still successful in detecting the compromised meters regardless.

3.1.5 Coarse Grained Detection Criterion for Presence of Organized Data Falsification. From Fig. 6(a), it is easy to conclude that for all attack types, the AD^{obs} is larger than the AD^{norm} learned from the training phase. The AD^{norm} act as a safe margin for the invariant, and anything outside of it is inferred as an orchestrated attack that needs a location and scale correction as a response. As long as the attack continues the $AD(T)$ remains higher than the normal values.

$$AD^{obs}(T) : \begin{cases} \in AD^{norm} & \text{No Organized Falsification ;} \\ > AD^{norm} & \text{Organized Falsification Occurred;} \end{cases} \tag{6}$$

3.1.6 Determining the Type of Data Falsification Attack. From the above, we conclude that an authentic change in the observed distribution may cause the mean consumption to increase or decrease but AD^{obs} remains the same as compared to the historical range of values $AD^{norm} = [AD^{min}, AD^{max}]$. An additive attack causes both HM and AM of consumption to increase but also causes AD^{obs} to increase compared to its normal range. This way a legitimate versus a malicious increase in the data can be distinguished. A deductive attack causes the HM and AM of mean consumption to decrease and causes AD^{obs} to increase from the historical range. Similarly, camouflage and conflict attacks do not have much change in the AM of the consumption but triggers a large increase in the $AD^{obs}(T)$. In this way, it is possible to infer which type of data falsification has been launched. A summary of the above discussion to determine the presence and type of attack is given in Table 1.

Table 1. Concluding Security Events

AD	AM	HM	GM	Conclusion
Increased	Increased	Increased	Increased	Additive
Increased	Decreased	Decreased	Decreased	Deductive
Increased	Same	Decreased	Decreased	Camouflage
Decreased	Increased	Increased	Increased	Additive Low
Increased	Any	Any	Any	Conflict
Same	Don't Care	Don't Care	Don't Care	No Attack

3.2 Attack Context Response Metrics

Given that an attack has been inferred that bias the instantaneous (hourly) consensus measures, we need a consensus correction scheme. The knowledge of the attack type could be leveraged to *unbias* the consensus measures. This is because, the manner and extent to which different instantaneous means such as, $HM(t), GM(t)$ and $AM(t)$ and corresponding standard deviations get biased by different attack types, is unique (from Property 1,2,3 and their corollary). Alternatively, one may be tempted to use the historical values of mean and standard deviation on the corresponding hours of the T-th day in the previous years. However, as already shown in Fig. 5, the mean values on the same days on successive years vary greatly and hence historical values are not reliable. Therefore, it is required that for a successful statistical detection, a robust mean (location consensus) and a robust measure of dispersion is calculated.

3.2.1 Estimation of Robust Mean as a Response:. For the calculation of robust mean, we need to reconstruct the actual mean from the observed mean using knowledge of how each attack type changes these means. Additionally, the extent of change triggered in the $AD(T)$ metric also depends on δ_{avg} and/or ρ_{mal}. Hence, an adjusted robust mean helps to estimate an approximate value closer to the original mean. Note that, the highest possible δ_{avg} is lesser in deductive attacks than additive ones because the feasible margin of deductive false data is bounded by zero. As the margins of false data or compromised fraction increases, the observed consensus gets more and more biased. To prevent this, we ought to have a consensus correction step. Otherwise statistics based trust models will not be able to identify the compromised meters.

From the statistical observations, we see that the $HM(t)$ is more proximate to the actual AM than the observed $AM(t)$, under the effect of additive attacks, due to a slower increase in HM as opposed to AM. However, the $HM(t)$ itself is not a robust mean consensus, if either δ_{avg} or ρ_{mal} is large. Therefore, we propose to use $\mu_R(t) = HM(t) - AD(t)$ as the estimated robust mean aggregate under additive attacks, which is closer to the original instantaneous arithmetic mean. Therefore, we deduct the $AD(t)$ since it is the index of the extra deviation caused by the attacks. As an example, in Table 3, for additive attack $HM - AD = 7.92 - 0.76 = 7.16$, which is very close the actual AM value of 7.053.

In contrast, for deductive attacks, due to $HM \leq GM \leq AM$ property, the $HM(t)$ is even lesser than the already biased $AM(t)$. But, $GM(t) + AD(t)$ is more robust than AM for deductive attacks, and results show that it is a good approximation to the actual mean. From the example in Table 3, it can be verified that for deductive attack, the robust mean $\mu_R = 6.29 + 0.79 = 7.08$ is closer to the actual mean 7.05. For camouflage attacks, AM is the most robust and hence μ_R is set as the AM. For conflict attacks, GM is an intermediate robust choice as it shows relative stability to both partially positive and negative outliers. The recommended mean correction for each attack type is tabulated in Table 2.

Security Incident	Choice of $\mu_R(t)$
Additive	HM-AD
Deductive	GM+AD
Camouflage	AM
Conflict	GM
No Attack	AM

Table 2. Robust Mean Responses

Parameter	Actual	Add	Deduct	Camo	Conf
AM	7.053	8.68	6.67	7.04	7.26
GM	6.860	8.35	6.29	6.65	6.89
HM	6.680	7.92	5.88	6.02	6.11
AD	0.373	0.76	0.79	1.02	1.15

Table 3. Attacks on Various Means in Texas Dataset

3.2.2 Estimating a Median Absolute Deviation as a response:. If the presence of an attack is discovered from the anomaly detector, then we know that the instantaneous standard deviation of the observed data is biased. The $\sigma(t)$ in the testing set under attacks will increase regardless of the type of data falsification attack (except for low additive attacks). Therefore, a directional correction of the standard deviation is not possible like $\mu_R(t)$ based on the attack types. While standard deviation is a very popular measure of dispersion (scale parameter) to build proximity distributions, we argue the use of a less common statistical measure of dispersion known as *'Median Absolute Deviation' (MAD)*, which is defined as follows:

For a univariate power consumption data at any time t, $\hat{p}(t) = \{p^1(t), \cdots, p^i(t), \cdots, p^N(t)\}$, the data's median is defined as $\tilde{p}(t) = Median(\hat{p}(t))$. The median absolute deviation is defined as: $MAD(t) = Median(|p^i(t) - \tilde{p}(t)|)$.

The MAD is a much more robust measure of dispersion (or more robust scale parameter) compared to the traditional standard deviation because MAD is more robust and remains less affected due to outliers (reducing false alarms under no attacks) and extreme values (under stronger margin attacks) compared to standard deviation. This is because measures such as standard deviation are derived from variance which uses squares of the difference between those outlying datapoints and the true mean. Squares produce very high values when datapoints are greater than 1, thus causing an unwarranted increase in the standard deviation. This is the cause of increased missed detection under attacks and increased false alarms under no attacks. Therefore, we depart from the traditional use standard deviation for characterizing the probability distribution of the proximity of individual smart meters data with the consensus.

The measured $MAD(t)$ of the historical time slots, before the inference of orchestrated attack is therefore embedded as the robust measure of dispersion or the robust scale parameter in the trust model in the event of an attack indication from the anomaly detector. As shown later, the mean correction, robust scale parameter as median absolute deviation and attack probability time ratio embedding facilitates quick detection, this approximation works well.

Both robust mean and median absolute deviation bias correction improves results significantly compared to the preliminary version of this work in [2]. The failure points for higher δ_{avg} values completely disappear. While, the above adjustment of mean location parameter and median absolute deviation may not always be perfectly close to the actual mean and median deviation, our results show that classification performance is much better under these approximate bias corrections rather than just using the exact harmonic mean and standard deviation as the location and scale parameters as done in our preliminary work [2].

3.3 Attack Context Embedded Relative Entropy based Trust Model

We pursue a light weight learning approach for identifying compromised smart meters that launch data falsification. The prior historical data set is considered as the authentic distribution of power consumption. From the historical data set, a *true proximity distribution* denoted as $\boldsymbol{X}^i$ for each smart meter is generated based on its reported consumption's proximity to the arithmetic mean of the authentic data set. Since the authentic historical data set is attack-free, the measure of consensus is arithmetic mean (AM), denoted by $\mu(t)$ and the standard deviation is $\sigma(t)$.

In the observed data set under test, we define $\mu_R(t)$ and $MAD_R(t)$ as the robust mean and median absolute deviation of the observed distribution based on the inferred security incident. In the testing set, a *current proximity distribution*, denoted by $\boldsymbol{Y}^i$, for each smart meter i is calculated based on the proximity of its reported consumption data $p^i_{rep}(t)$ to $\mu_R(t)$. In the absence of a detected security incident, the robust mean and median absolute deviation equals $\mu(t)$ and $\sigma(t)$ (like in the historical set). However, when an attack is present, the $\mu_R(t)$ is set according to Table 2 based on the inflicted attack type and strategy. This way the attack context is embedded via the appropriate robust mean as a response to the detected attack context. Similarly, the $MAD_R(t)$ is set to the historical median absolute deviation if there is an indication of an attack.

If the true distribution is very different from the current distribution, it is an indication that this meter's data is *unusually* different and this difference in the probability space is measured as *Kullback-Leibler divergence* (also called KL Distance) which measures the *relative entropy* between the two distributions. The higher the divergence between the two distributions, the more the indication of anomalous behavior. The trust of a meter is calculated at the end of the frame F (in days). The total number of observations over the time frame is given by TS. For the relative entropy trust model, we had time frames of length $F = 10$ days and $F = 30$ days. Therefore, the number of time slots monitored in the frame of observation is $TS = F * 24$.

3.3.1 True and Current Proximity Distributions as Meter Evidence:. We introduce a binary random variable $X^i = \{0, 1\}$ for each meter i, for $i = 1, \ldots, N$, which acts as a historical reference distribution. If the historical data reported $\hat{p}^i_{rep}(t)$ at time t from meter i falls within one standard deviation of $\mu(t)$, then $X^i = 1$, else 0. Formally,

$$X^i(t) = \begin{cases} 1 & \text{if } \hat{p}^i_{rep}(t) \in \{\mu(t) \pm \sigma(t)\}; \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

where $X^i(t)$ follows a Bernoulli distribution with parameter r, that is the probability of $X^i = 1$ is r, and the probability of $X^i = 0$ is $1 - r$.

Suppose, $S(X)$ be the variable that denotes the number of successes, that is $S(X^i) = \sum_{t=1}^{TS} X^i(t)$. Let $S(X) = k$ be the observed value of the variable for any meter i, such that number of success in the true distribution is $S(X^i) = \sum_{t=1}^{TS} X^i(t) = k$.

Similarly, we have a binary random variable Y^i for the current distribution of each smart meter, such that the probability of $Y = 1$ is q and the probability of $Y = 0$ is $1 - q$. In this case, the number of successes is denoted by a variable $R(Y^i) = \sum_{t=1}^{TS} Y^i(t)$. Let $R(Y) = j$ denote the number of successes for any such meter i such that number of successes in the current distribution is $R(Y^i) = \sum_{t=1}^{TS} Y^i(t) = j$. If an attack has been detected through the anomaly detection phase, then the robust mean $\mu_R(t)$ and the robust standard deviation $\sigma_R(t)$ is calculated, and the Y^i is calculated based on them. In this way attack context is embedded such that Y^i remains unbiased from the effects of orchestrated attacks. However, in the absence of any detected attacks, $\mu_R(t) = \mu(t)$. Formally, the current proximity distribution is given by:

$$Y^i(t) = \begin{cases} 1 & \text{if } p^i_{rep}(t) \in \{\mu_R(t) \pm MAD_R(t)\}; \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

Intuitively, in absence of attacks, the distribution of Y should be very close to X. On the contrary, the two distributions should show a difference when an attack is present.

3.3.2 Estimating Parameters of True and Current Proximity Distributions:. Next, we need to estimate the parameters r and q for corresponding distributions X^i and Y^i. An obvious estimate is the minimum variance unbiased estimate (frequentist), which is the sum of all successes divided by the total number of observations TS. However, this approach may cause $r = 0, q = 0$, or $r = 1, q = 1$, for which the relative entropy (see Eqn. 15) is undefined. Moreover, frequentist probability unbiased estimator makes sense only if there is a large set of observations [24]. However, since our trust model works on a shorter horizon of time (typically on a few days or monthly basis), such approaches are improper. Hence, we need to accommodate a Bayesian approach for estimation of r and q, so it is theoretically sound and mathematically tractable. Since the following is true for all meter's i, we drop the suffix i from the notational simplicity.

First, we estimate the parameter of r. We prove that the estimated probability $r = \frac{k+1}{TS+2}$, where k is the realization of the total number of successes observed. Thus $S(X) = k$ follows a binomial distribution with parameter r.

Hence, the probability of observing exactly k successes out TS times, given the probability of success of each trial was r, is given by,

$$P(S(X) = k|r) = \binom{TS}{k} r^k (1-r)^{TS-k} \tag{9}$$

The Bayesian posterior estimate of r, based on prior TS observations by Bayes theorem, is given as:

$$P(X(TS+1) = 1|S(X) = k) = \frac{P(X(TS+1)), S(X) = k)}{P(S(X) = k)} \tag{10}$$

The denominator is the marginal probability of $P(S(X) = k)$ marginalized over all possible outcomes of r. Hence,

$$P(S(X)) = \int_0^1 \binom{TS}{k} r^k (1-r)^{TS-k} f(r) dr \tag{11}$$

Assuming conditional independence between $S(X)$, r and $X_i(t+1)$ of the prior and likelihood can be solved as:

$$P(X(TS+1)), S(X) = k) \Rightarrow \; = \int_0^1 P(X(TS+1) = 1|r) P(S(X) = k|r) dr \tag{12}$$

Since there is no prior information on r, we assume a non-informative prior such that $f(r) = 1$, for Eqn. (11) and Eqn. (12). Plugging in Eqn. (11) and Eqn. (12) into Eqn. (10), it can be shown that:

$$P(X(TS+1) = 1|S(X) = k) = \frac{k+1}{TS+2} = r \tag{13}$$

Similarly,

$$q = \frac{j+1}{TS+2} \tag{14}$$

It can be verified that $r, q \neq 0, 1$. Hence, the logarithms of distributions X^i and Y^i for the i-th smart meter, (described in terms of probability parameters $r^{(i)} = \frac{k^{(i)}+1}{TS+2}$ and $q^{(i)} = \frac{j^{(i)}+1}{TS+2}$), in Eqn (15) is always defined and exist even as $k^{(i)} = 0$ or $j^{(i)} = 0$.

3.3.3 Kullback-Leibler Divergence based Scoring and Classification:. We adopt the *Kullback Leibler divergence* to measure the difference between the historical distribution X^i and the observed distribution Y^i for a smart meter. Note that X^i and Y^i are not consumption patterns but a trend on proximity to the middle quartile. Subsequently, the KL distance is transformed into a trust value between 0 and 1 by passing it through an inverse square root function that produces linearly separable trust values between compromised and honest meters via a single threshold.

The KL distance between two distributions X and Y for a smart meter i, is given by:

$$D_i(X^i||Y^i) = (1 - r^{(i)}) \times ln\Big(\frac{1 - r^{(i)}}{1 - q^{(i)}}\Big) + r^{(i)} \times ln\Big(\frac{r^{(i)}}{q^{(i)}}\Big) \tag{15}$$

The $D_i(X^i||Y^i)$ is a positive real value that indicates the divergence between the observed and the historical proximity distribution. Hence, the smaller the value of $D_i(X^i||Y^i)$ the better it is in terms of being trustworthy and the larger it becomes the less trustworthy it becomes since a larger divergence indicates a mismatch between the true and observed proximity distributions. Given this, the final trust value Q^i of a smart meter i, is given by:

$$Q^i = \frac{1}{1 + \sqrt{D_i(X^i||Y^i)}} \qquad 0 \leq Q^i \leq 1 \tag{16}$$

The rationale of Eqn. 16, is a scaling function that scales the lowest value in $D_i(X^i||Y^i)$ a trust score that is closest to 1 while the highest value in $D_i(X^i||Y^i)$ gets the exponentially lower trust score with increasing $D_i(X^i||Y^i)$. The exponential nature ensures a risk aversion towards progressively increasing distance in the probability space.

Limitation of Coarse Grained Anomaly Detection based Trust Model: Since the coarse grained anomaly detection has an observation granularity of 24 hours, it is not suitable for detection of opportunistic omission and on-off strategies that are discontinuous and sparsely distributed over the time domain. Therefore, an anomaly monitoring metric with a daily time granularity such as $AD(T)$ will not be sensitive and fail to provide the early indication of the attack's presence that is necessary to embed in the attack context.

Apart from failing to identify the incidence, type, and robust consensus, there will be another hurdle for the subsequent pipelined trust model. Since in most of the time slots, there are no attacks from the meters, the evidence against each meter will have reduced sensitivity when observed over a time frame. This is because the probabilities (modeled by evidence) in information theoretic measures (such as KL Divergence) are steady state long term measures. When observed over the time frame, the detection of meters will be delayed due to a lesser change in evidence counts. However, if the trust model is made aware of the *incidence of such non-continuous strategies* and the *approximate start and stop times of such attacks*, the evidence against meters collected on those specific slots may be weighted as more important (while others as less important). This would facilitate quicker classification of such meters while running the trust scoring model through information theoretic measures. This is achieved by calculating the fraction of the time frame that a meter was under such attack strategies (*defined later as attack probability time ratio*). This motivates the need for a fine grained anomaly detection phase that runs in parallel with coarse grained anomaly detection metric and associated attack context embedding.

4 FORMAL SECURITY ANALYSIS

We do the theoretical analysis in terms of attack parameters to formally specify the impact of attacks on the effectiveness of the defense method. Specifically, we assess the security level of our mechanism by taking into account what an intelligent adversary might do to bypass the invariant

based anomaly detection and the compromised meter detection trust model. Here, we also show closed form theoretical expressions of our observations that will help generalize our framework.

Theoretical Analysis of Deviation in $AD(T)$ under attacks: As a part of the theoretical security analysis of the anomaly detection phase, we provide the closed form approximate estimated deviation in the anomaly detection metric $AD(T)$. This can be estimated by calculating the expected harmonic mean and arithmetic mean, given an attack type, ρ_{mal} and δ_{avg}. Below we provide an estimation of the harmonic mean followed by the arithmetic mean. Finally, we show how closely the theoretical result from the closed form expression matches the experimental result to prove accuracy of analysis. We also show that change in $AD(T)$ observed experimentally also matches the theoretical analysis. Because our detector uses the values in a box-cox transformation domain, we have carefully estimated it for real data values and found their box cox equivalents on the transformed scale.

$$Nor(AM^{ba}(t)) = \frac{\sum_{i=1}^{N} P^i_{rep}(t)}{N}; \quad Nor(AM^{ba}(T) = \frac{\sum_{t=1}^{24} Nor(AM_{ba}(t))}{24} \tag{17}$$

Similarly, $Nor(HM_{ba}(T))$ and $Nor(GM_{ba}(T))$ can be calculated. For brevity, we drop the T from the following analysis. Since the closed form expression of the harmonic mean does not exist, we first estimate the new geometric mean $Nor(GM^{esaa})$ after the attack. Then, we harness the following Pythagorean equation that calculates the estimated harmonic mean from the estimated geometric and arithmetic means:

$$Nor(HM^{esaa}) \approx \frac{(Nor(GM^{esaa}))^2}{Nor(AM^{esaa})} \tag{18}$$

where $Nor(HM^{esaa})$ and $Nor(AM^{esaa})$ denote the estimated HM and AM values after an attack.

Estimation of the Geometric Mean after attack: Let $Nor(GM^{ba})$ denote the geometric mean of a power consumption data before attack in the original data domain and is defined by:

$$Nor(GM^{ba}) = \left(\prod_{i=1}^{N} P^i_{rep}\right)^{\frac{1}{N}} = \sqrt[N]{(P^1 \times P^2 \cdots P^M \times P^{M+1} \cdots \times P^N)} \tag{19}$$

Similarly, let the estimated geometric mean after additive attack from $\rho_{mal} = M/N$ meter and δ_{avg} in the original data domain be denoted as $Nor(GM^{esaa})$ such that:

$$Nor(GM^{esaa}) = \sqrt[N]{(P^1 + \delta_{avg}) \times (P^2 + \delta_{avg}) \cdots \times (P^M + \delta_{avg}) \times (P^{M+1}) \times \cdots \times (P^N)} \tag{20}$$

Now we need to convert each $P^i + \delta_{avg}$ term into a multiplier of P^i. Let the ratio between the δ_{avg} and the actual data from the $i-th$ meter before attack P^i be given by a dummy variable:

$$\alpha^i = \frac{\delta_{avg}}{P^i} \tag{21}$$

Since P^i is a completely random physical quantity, we will need to characterize the α variable as a property that is shared across datapoints under an attack.

From the studies, we know that for the power consumption distribution, most of the data points are within the first standard deviations from the mean (say $\overline{P}$). For the Irish and Texas dataset, more percentage of datapoints (70%) are lesser than the mean $\overline{P}$ compared to percentage of data points greater than the mean (30%) on average. While the 30% values are lesser and greater than the mean cancel the effect of each other on the estimation of P^i, the remaining fraction of samples represents an imbalance factor (say $\nabla = 0.40$). Since these fraction of samples are lesser than the arithmetic

mean $\overline{P}$, a corrective factor of $\nabla * \sigma$ should be deducted from the $\overline{P}$ to adjust for the approximate estimated value of a P^i. Therefore, we can re-write the α^i as:

$$\alpha = \frac{\delta_{avg}}{\overline{P} - \nabla\sigma}$$

$$Nor(GM^{esaa}) \approx \sqrt[N]{(P^1 + \alpha.P^1) \times (P^2 + \alpha.P^2) \cdots \times (P^M + \alpha.P^M) \times (P^{M+1}) \times \cdots \times (P^N)}$$
$$Nor(GM^{esaa}) \approx \sqrt[N]{(1+\alpha)P^1 \times (1+\alpha)P^2 \cdots \times (1+\alpha)P^M \times (P^{M+1}) \times \cdots \times (P^N)}$$
$$Nor(GM^{esaa}) \approx \sqrt[N]{(1+\alpha)^M \times P^1 \times P^2 \cdots \times P^M \times (P^{M+1}) \times \cdots \times (P^N)}$$
$$Nor(GM^{esaa}) \approx \sqrt[N]{(1+\alpha)^{\rho_{mal}*N} \times P^1 \times P^2 \cdots \times P^{\rho_{mal}*n} \times (P^{\rho_{mal}*n+1}) \times \cdots \times (P^N)}$$
$$Nor(GM^{esaa}) \approx \sqrt[N]{(1+\alpha)^{\rho_{mal}*N} \times P^1 \times P^2 \cdots \times P^{\rho_{mal}*n} \times (P^{\rho_{mal}*N+1}) \times \cdots \times (P^N)}$$
$$Nor(GM^{esaa}) \approx (1+\alpha)^{\rho_{mal}} \sqrt[N]{P^1 \times P^2 \cdots \times P^M \times (P^{M+1}) \times \cdots \times (P^N)}$$

From Eqn. 19, the above reduces to the following:

$$Nor(GM^{esaa}) \approx (1+\alpha)^{\rho_{mal}} Nor(GM^{ba})$$
$$Nor(GM^{esaa}) \approx \left(1 + \frac{\delta_{avg}}{\overline{P} - \nabla\sigma}\right)^{\rho_{mal}} Nor(GM^{ba}) \tag{22}$$

Plugging in the real values of σ, ∇, ρ_{mal}, δ_{avg} and $Nor(GM^{ba})$, we obtain the estimated theoretical geometric mean after the attack as $Nor(GM^{esaa}) = 410$, while the actual measured geometric mean after the attack was recorded as $Nor(GM^{expaa}) = 390$. This indicates that this is a reasonably close approximation. Now the next step is to calculate $Nor(AM^{esaa})$ to plug it in Eqn. 18 for estimation of the new harmonic mean $Nor(HM^{esaa})$.

Estimation of AM after attack: Let the $Nor(AM^{esaa})$ denote the arithmetic mean attack after attack. For the following estimation, assume the attack to be additive. Similarly, this method could be used to estimate other attack types. Given the $\rho_{mal} = M/N$ is the fraction of compromised meters and δ_{avg} is the average falsification margin per meter, then the estimated attacked arithmetic mean is under additive attack is:

$$Nor(AM^{esaa}) = Nor(AM^{ba}) + \left(\rho_{mal} * \delta_{avg}\right) \tag{23}$$

Estimation of HM after attack: The Eqn. 22 and Eqn. 23 can be plugged in the following:

$$Nor(HM^{esaa}) \approx \frac{(Nor(GM^{esaa}))^2}{Nor(AM^{esaa})} \tag{24}$$

Estimation of Box-Cox Equivalents of Means: Let the $Box(Nor(AM^{ba}(T), \lambda)$ denote the box cox equivalent of the mean before attack in the normal scale such that:

$$Box(Nor(AM^{ba}(T), \lambda) = \frac{(Nor(AM^{ba}(T))^\lambda - 1}{\lambda} \tag{25}$$

Similarly, $Box(Nor(HM^{ba}(T), \lambda)$, and $Box(Nor(GM^{ba}(T), \lambda)$ are corresponding box-cox equivalent values of harmonic and arithmetic means before the attack. Similarly, the box-cox equivalent values of them after attack $Box(Nor(AM^{esaa}), \lambda)$, $Box(Nor(GM^{esaa}), \lambda)$, $Box(Nor(HM^{esaa}), \lambda)$, can be easily estimated.

Final Estimation of $AD(T)$ after attack: Note that the box-cox equivalent of the arithmetic mean gives a slightly different answer compared to the arithmetic mean of data in a power transformation scale (the experimental result). Let the difference be $\kappa = |Box(Nor(AM^{esaa}), \lambda) - Box(Nor(AM^{ba}), \lambda)|$. The estimated arithmetic, geometric, and harmonic means calculated over

box-cox transformed arguments (what our method actually implements), after the additive attack is given by the following:

$$\overline{AM}^{esaa} = \overline{AM}^{ba} + \kappa; \qquad \overline{GM}^{esaa} = Box(Nor(GM^{esaa}); \qquad \overline{HM}^{esaa} = Box(Nor(HM^{esaa}) \quad (26)$$

For the estimation of arithmetic mean, the estimation of change (κ) will result in a closer approximation compared to direct box-cox calculation for a given ρ_{mal} and δ_{avg}. Let be the value of the $AD(T)$ metric after the attack be $AD^{esaa}(T) = |\overline{HM}^{esaa} - \overline{AM}^{esaa}|$. Thus, The expected deviation in the $AD(T)$ metric after an attack of ρ_{mal} and δ_{avg} for additive attacks is given by:

$$E(\Delta AD(T)) = |\overline{HM}^{ba} - \overline{AM}^{ba}| - |\overline{HM}^{esaa} - \overline{AM}^{esaa}| \quad (27)$$

The theoretical deviation in the $AD(T)$ metric for a $\rho_{mal} = 0.40$ and $\delta_{avg} = 800$W is 0.553. For the same attack the experimental result shows the change of $AD(T)$ to be 0.712. This indicates a reasonable approximation as well as the positive magnitude of change. Additionally, the theoretical value shows a increase in the $AD(T)$ which is also seen in the experimental result.

Table 4. Estimation Accuracy of Invariants with Irish Dataset

Parameter	Experimental	Theoretical
$\overline{AM}^{esaa}$	14.5245	14.257
$\overline{HM}^{esaa}$	11.8113	11.703
$AD^{esaa}(T)$	2.713	2.554
$E(\Delta AD(T))$	+0.712	+0.553

Optimal Evasion δ_{avg} against Anomaly Detection Invariants: For an optimal evasion of our anomaly detection step, the adversary would want to use the maximum δ_{avg}, which creates a deviation in the invariants, that is just within the designed safe margin. In practice, since the adversary does not know the current $AD(T)$ value (since he cannot possibly control 100%) of the meters, he relies on the historical $AD(T)$, which can be possibly known the adversary through a database hack. Therefore, the adversary would ensure that given its attack type, and the fraction of compromised meters, the δ_{avg}, should be such that the following condition satisfies:

$$|AD^{esaa}(T) - AD_{hist}(T)| < 0.75 * \sigma_{AD(T)} \quad (28)$$

Specifically, expanding the Eqn. 26, we get the theoretical expected change in the statistical invariants as a function of the ρ_{mal} and δ_{avg} (the two key variables apart from the attack type that changes the in-variants). Thus, the estimated optimal evasion δ_{avg} can be found by the adversary solving the following optimization problem:

$$\delta_{avg}^{evasion} = \arg\max_{\delta_{avg}} f(\delta_{avg}) \quad (29)$$

$$\text{s.t.} \qquad f(\delta_{avg}) < 0.75 \times \sigma_{AD(T)}$$

$$\text{where } f(\delta_{avg}) = |AD^{esaa}(T) - AD_{hist}(T)| \quad = \quad |(|\overline{HM}^{esaa} - \overline{AM}^{esaa}|) - AD_{hist}(T)|$$

Note that $\overline{HM}^{esaa}$ and $\overline{AM}^{esaa}$ are given by the following as a function of the attack:

$$\overline{HM}^{esaa} = Box\left(\frac{\left(\left(1 + \frac{\delta_{avg}}{\overline{P} - \nabla\sigma}\right)^{\rho_{mal}} Nor(GM^{ba})\right)^2}{Nor(AM^{ba}) + \left(\rho_{mal} * \delta_{avg}\right)}\right) \quad (30)$$

$$\overline{AM}^{esaa} = (\overline{AM}^{ba} + |Box\Big(Nor(AM^{ba}) + (\rho_{mal} * \delta_{avg})\Big) - Box\Big(Nor(AM^{ba})\Big)|) \quad (31)$$

We can see that the above equations are a function of the ρ_{mal} and δ_{avg}, which formally analyses the effect of any attack on the statistical invariants. We have proven the approximation accuracy

of our expression in Table 5 by showing how theoretical values approximate to experimental observations.

ρ_{mal}	Exp. $\delta_{avg}^{evasion}$	Theo. $\delta_{avg}^{evasion}$
20	400	380
30	370	360
40	350	330
50	330	320
60	320	300

Table 5. Evasion δ_{avg}: Experiment vs Theory

ρ_{mal}	Theo. Evasion δ_{avg}	$MAD^{evasion}(t)$	Current Mean
20	380	347	652
30	360	356	684
40	330	352	708
50	320	357	736
60	300	361	756

Table 6. Inferred MAD at Invariant Evasion Points

Formal Estimation of Robust Mean under Attacks: For robust mean closed form derivation, we just plug in the values of $Nor(AM^{esaa})$, $Nor(GM^{esaa}$, $Nor(HM^{esaa})$ or their box-cox transformed equivalents, (expressions derived previously) and plug into the Table 2 to find the theoretical value as shown below:

$$\mu_R^{Additive}(t) = Box^{-1}(\overline{HM}^{esaa} - AD(T)), \qquad \mu_R^{Deductive}(t) = Box^{-1}(\overline{GM}^{esaa} + AD(T)) \tag{32}$$

$$\mu_R^{Camouflage}(t) = Box^{-1}(\overline{AM}^{esaa}), \qquad \mu_R^{Conflict}(t) = Box^{-1}(\overline{GM}^{esaa}) \tag{33}$$

where $\overline{GM}^{esaa} = Box\left(\left(1 + \frac{\delta_{avg}}{\overline{P}-\nabla\sigma}\right)^{\rho_{mal}} Nor(GM^{ba})\right)$, and $\overline{HM}^{esaa} = Box\left(\frac{Nor(GM^{esaa})^2}{Nor(AM^{esaa})}\right)$

The $Box^{-1}(.)$ is defined as: $Box^{-1}(x) = (x * \lambda + 1)^{1/\lambda}$ where x is the value in box cox scale being remapped and λ is the box cox transformation parameter. The $\overline{AM}^{esaa}$ under camouflage is the same as the observed arithmetic mean, since it balances out the mean by virtue of its attack type.

Condition for Successfully Evading of Meter Detection: Note that, we already proved that as ρ_{mal} increases, our invariant criterion forces the δ_{avg} to be smaller. Hence, the attacker cannot unilaterally increase one attack parameter to arbitrarily change the median absolute deviation. Therefore, at the theoretical evasion δ_{avg}, we first present, the current median absolute deviation (under attacks) by varying from the ρ_{mal} from 20% to 60%, as listed in Table 6.

The trust score depends on the divergence between proximity distributions X_i and Y_i. The adversary has to bypass the invariant based anomaly detection to ensure that the mean and median absolute deviation correction does not take place. Furthermore, the adversary has to make sure that the majority of it's compromised meter readings are within the observed (biased) mean and the median absolute deviation (MAD) range. However, on average we say that to bypassing meter detection reliably the following condition needs to be satisfied for a given ρ_{mal}.

$$\delta_{avg}^{bypass} \leq min(\delta_{avg}^{evasion}, MAD^{evasion}(t))) \tag{34}$$

Let us look at a specific example from Table 6. For $\rho_{mal} = 40\%$, the $\delta_{avg}^{evasion}$ is 330 and the MAD at that evasion δ_{avg} based attack is 352. The min(330,352) is 330, which is the theoretical value to bypass the trust model. In our experiments, for $\delta_{avg} > 330$ (Fig 17), the missed detection rate is lower than 10%, however at when $\delta_{avg} < 330$, it starts missing meters and missed detection becomes about 30%. This is also repeated in the Texas dataset results in Fig.14 and 15, where below 330, the missed detection becomes between 30%-40% proving correctness.

5 SPECIAL CASE STUDY ON FINE GRAINED ANOMALY BASED TRUST MODEL

Now we propose the customized version of our trust model that can *run in parallel* for effective identification under on-off or omission attack strategies. It is important to note that the fine grained anomaly based detection will produce different responses than the coarse grained one, and therefore will invoke an augmented and modified version of the proposed trust model in Section 3.3 with novel embeddings of responses produced by the fine grained anomaly based security event detector.

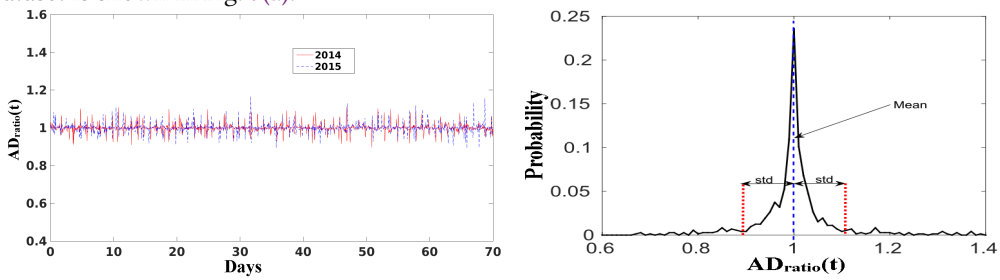
5.1 Fine Grained Anomaly based Security Event Detection

In this subsection, we will introduce the invariant (metric) for fine grained anomaly detection, justify the choice of invariant, establish a detection criterion for fine grained attacks, determine attack type, strategy, start and stop times, and calculate the attack probability time ratio.

5.1.1 Proposed Invariant. We propose a more fine-grained detection metric denoted by $AD_{ratio}(t)$ that is computed hourly, in contrast to $AD(T)$ that is computed daily. The $AD_{ratio}(t)$ is the ratio of the absolute difference between 'hourly' arithmetic and harmonic means between the previous $t-1$ and current time slot t. At any time slot t, the metric is defined as:

$$AD_{ratio}(t) = \frac{AD(t-1)}{AD(t)} \tag{35}$$

where $AD(t) = |HM(t) - AM(t)|$. The time series of the proposed metric AD_{ratio} for the Texas Dataset is shown in Fig. 7(a).



(a) (b)

Fig. 7. (a) Time Series of $AD_{ratio}(t)$ (b) Distribution of $AD_{ratio}(t)$

5.1.2 Identifying Normal Range of $AD_{ratio}(t)$. Fig. 7(b) shows the distribution of the proposed $AD_{ratio}(t)$ for the historical training dataset (2014 and 2015). It can be seen that the distribution of $AD_{ratio}(t)$ has a mean value of 0.998 with a standard deviation of 0.1. Very few sample $AD_{ratio}(t)$ values lie beyond the second standard deviation. Let $AD^{norm}_{ratio} \in [AD^{min}_{ratio}, AD^{max}_{ratio}]$ denote the normal range of this fine grained $AD_{ratio}(t)$ metric.

5.1.3 Investigating Effect of Various Attacks on $AD_{ratio}(t)$. For deductive attacks, we had mentioned that the decay rate of Harmonic Mean is larger compared to the decay in Arithmetic mean given the dataset. Therefore,

$$HM(t) - HM(t-1) > AM(t) - AM(t-1)$$

Solving the above, we get,

$$\frac{HM(t-1) - AM(t-1)}{HM(t) - AM(t)} < 1$$

$$\implies AD(t-1)/AD(t) < 1 \implies AD_{ratio}(t) < 1.$$

From the above, it is clear that a deductive or omission (which is a virtual deductive attack) attack when initiated, will cause a *sharp drop* in the proposed $AD_{ratio}(t)$ metric. When the attack stops,

there will be a *sharp rise* in the AD_{ratio} metric, since the harmonic mean has to increase more than the arithmetic mean to restore the original ratio that is very stable and $AD_{ratio}(t) \rightarrow 1$. Therefore, the difference between $HM(t) - AM(t)$, will be much lesser compared to $HM(t-1) - AM(t-1)$. Since the denominator decreases when the attack stops, the $AD_{ratio}(t)$, experiences a sharp rise. Experimental verification of this is provided in Fig. 8.

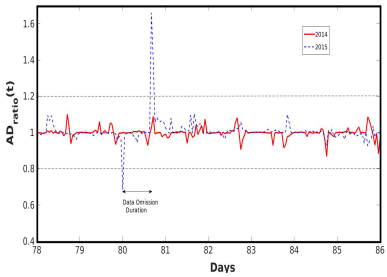


Fig. 8. Omission Attack Example

$\mathrm{AD_{ratio}(2c-1)}$	|FGAT|	$\mathrm{t^{oe}}$	**Conclusion**
$> AD^{max}_{ratio}$	High	Constant	Additive ON-OFF
$< AD^{min}_{ratio}$	High	Constant	Ded/Camo ON-OFF
$< AD^{min}_{ratio}$	High	Varying	Omission Attack
$< AD^{min}_{ratio}$	Sparse	Don't Care	Omission Failure

Table 7. Concluding Fine Grained Security Events

Similarly, for additive attacks, harmonic means have a slower growth rate compared to the arithmetic mean. Therefore,

$$HM(t) - HM(t-1) < AM(t) - AM(t-1)$$

Solving the above, we get

$$\frac{HM(t-1) - AM(t-1)}{HM(t) - AM(t)} > 1$$
$$\implies AD(t-1)/AD(t) > 1 \implies AD_{ratio}(t) > 1$$

From the above, it is clear that for additive attacks the $AD_{ratio}(t)$ must increase when attacks start, while for deductive and camouflage attacks the $AD_{ratio}(t)$ must decrease.

5.1.4 Detecting Incidence of Fine Grained Attacks. The following equation is similar to the coarse grained logic for confirming presence of opportunistic fine grained attacks.

$$AD_{ratio}(t) : \begin{cases} \in \{AD^{norm}_{ratio}(t)\} & \text{No Attack;} \\ \notin \{AD^{norm}_{ratio}(t)\} & \text{Fine Grained Attack} \end{cases} \tag{36}$$

5.1.5 Determining Fine Grained Attack Types and Strategies. To reconstruct the security events under fine grained attack strategies, we first need to record the sequence of time slots where the event $AD_{ratio}(t) \notin \{AD^{norm}_{ratio}(t)\}$ occurred over the observed time duration, into a vector $FGAT = \{t(1), t(2), \cdots, t(c), \cdots, t(C)\}$, where $c \in \mathbb{N}$ is the set of first C natural numbers. The odd and even entries of the set FGAT are represented by $t(2c-1)$ and $t(2c)$ respectively and $|FGAT|$ is the cardinality of this set over the time frame under observation. Additionally, let the time difference between the pairs of odd entries and even entries be $t^{oe} = |t(2c-1) - t(2c)|$.

There are three important facets to monitor. First, the set of t^{oe} values help distinguish between deductive ON-OFF and omission attacks having similar signatures. Second, the cardinality of $|FGAT|$ is important to *distinguish between the possibility of omission attack versus omission failures.* Third, whether $AD_{ratio}(t)$ corresponding to the odd entries$t = 2c-1$ in FGAT are greater than AD^{max}_{ratio} or smaller than AD^{min}_{ratio}, help differentiate between additive, deductive, and camouflage data falsification types.

If the t^{oe} is constant for all odd values of c, then there is an on-off attack. Given that t^{oe} is constant, if $AD_{ratio}(2c-1) > AD^{max}_{ratio}$, it is an additive on-off attack, while an $AD_{ratio}(2c-1) < AD^{min}_{ratio}(t)$, it is an deductive on-off attack. Therefore, odd entries $AD_{ratio}(2c-1)$ helps to distinguish between

additive, deductive or camouflage attacks. Since attacks are launched and stopped at periodic intervals, the $|FGAT|$ will not be singleton or sparse.

If the set of $t^{oe} = |t(2c-1) - t(2c)|$ consists of variable values, the $|FGAT|$ is not singleton or sparse, and $AD_{ratio}(2c-1) < AD^{min}_{ratio}(t)$, it is an omission attack (deliberate). On the other hand, if $t(2c-1) - t(2c)$ consists of variable values, $|FGAT|$ is singleton or sparse, the $AD_{ratio}(2c-1) < AD^{min}_{ratio}(t)$ is a omission failure due to non-adversarial reasons.

The missing data from a subset of houses at any time slot t is perceived as a deductive attack where actual power consumption values are replaced by null values which are lesser than actual data. This causes the harmonic mean to decay at a rate greater than compared to the decay in the arithmetic mean. Therefore, the difference between arithmetic mean and harmonic mean at time slot t increases compared to the previous time slot $(t-1)$ with no data omission. Therefore, the $AD_{ratio}(t)$ value between time slots t and $t-1$ experiences a sharp decrease. As long as the degree of omission stays same the AD_{ratio} is restored to normal value. When omission stops there will be another drastic change, where the harmonic mean will grow faster than the AM, such that the $AD(t)$ decreases compared to the $AD_{ratio}(t-1)$ calculated with missing data. Hence, there is a sharp drop in the proposed AD_{ratio}. This can be verified from Fig. 8.

5.2 Estimation of Attack Probability Time Ratio as a Response:

Apart from the robust consensus measures, which are required for fine grained attack strategies, we also need another additional response that needs to be embedded into the subsequent trust modeling step. This response is known as the *attack probability time ratio.*

The attack probability time ratio P_{attack} is an indicator of the fraction of time slots that the system was under attack over an observed time frame. For example, for an on-off attack having an ON period of 6 hours of attack in a day, $P_{attack} = 1/4$. Therefore, the fraction of time slots with no attack is $(1 - P_{attack})$, will be automatically considered as successes even when this meter is launching data falsification attacks. Therefore, in the probability space, these meters will not be further apart when there are on-off attacks versus no attacks. Hence, the time to detection of such meters will be significantly larger. To reduce this, we need to keep track of the P_{attack}, and embed this information in the trust model. Such P_{attack} can be estimated from our designed FGAT vector, by the

$$P_{attack} = \frac{\sum_{c=1}^{C} |t(2c) - t(2c-1)|}{TS}$$

5.3 Trust Scoring Model with Attack Probability Time Ratio Embedding

Since on-off and omission strategies are discontinuous over time, the number of failures will not be as high compared to the case of continuous attacks in an observed time frame. This will produce $q^{(i)}$ values of compromised meters which are still high and therefore proximate to the parameter r in the true distribution. Hence, the time to detection convergence of meters with missing data (omission) or discontinuous falsification of data (on-off) will be time consuming, due to lack of evident separation in the probability space, which leads to classification errors as well.

Since the fine grained anomaly detector gives an early indication on the time slots when such on and off attack happened (from FGAT vector), a lesser weight can be given to the number of successes observed by weighing it with the fraction of duration the system is not under attack ((i.e., $1 - P_{attack}$)) in the observed frame F. In this manner, the time to detection of these meters could be improved. Under these opportunistic attack strategies, which are captured in the fine grained anomaly detector, the Eqn. 14 in the trust model is modified by weight to the number of successes j. This weight is $(1 - P_{attack})$, which prevents the value of q to be very high even when the number of OFF periods is large compared to the ON period of attacks over the observed time frame

containing TS windows. Hence, due to the attack context awareness, the observed distribution q under evidence of on-off and omission attacks (from the fine grained detector) for each meter is modified as:

$$q^{(i)} = \frac{(1 - P_{attack})j^{(i)} + 1}{TS + 2} \tag{37}$$

Eqn. 37, can be explained by the following: Note that the q is the probability that $Y^i(t) = 1$, meaning the meter $i's$ reading is falling within the robust mean and median absolute deviation. However, in an on-off attack, there are off periods, where this compromised meter's data is likely to achieve a value of 1. Hence, the probability of q over a given time frame TS is not remarkably different from r. Since the probability of q is specified by the number of successes j, a discounting factor of $1 - P_{attack}$ is required, since these $1 - P_{attack}$ time was not under attacks was a part of the OFF period. that be counted on as the FGAT vector shows evidence of orchestrated data falsification on selective ON periods (e.g., when prices are high/demand is high, etc.,).

The value of q is lesser compared to a value that contributes the entire observed j towards the probability of success. This ensures a larger difference between q and r in the probability space, which facilitates quick classification that is apparent even when the attacker acts honestly in majority of the time slots. The modification by Eqn. 37 is termed as *attack probability ratio time embedding* that customizes the trust model for better and quick classification of the compromised meters.

Some Limitations of our Approach: The relative entropy based trust model detect compromised meters only if the δ_{avg} is greater than the median absolute deviation of the datasets. From the Eqn. 22 and Eqn., 23, it is clear that if the δ_{avg} is lesser than the MAD, in most time instances, the Y_i of the attacked meters will be within that deviation and therefore be labeled as one instead of zero more frequently. Thus, there will not be a significant change in the probability of $p(Y_i = 1) = q$ in the attacked set. Therefore the deviation between X_i and Y_i in the probability space, will not be evident to produce a divergence that could clearly classify the malicious meters from the honest ones. Our studies from real datasets indicate that the MAD ranges between $290W - 350W$. Therefore, in our approach the missed detection errors increase $\delta_{avg} < 300$. However, the error rates are better than existing works across datasets as shown in the comparison in Section 5.6.

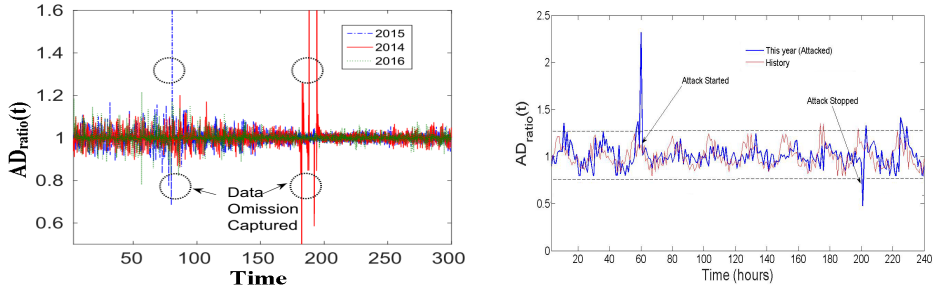
Intuitively, one solution to this limitation is to introduce multinomial evidence labels for each meter instead of binary labels (0,1), and then calculate the distances between the distributions in the probability space with a similar entropy measure. However, our experience showed that this is not enough to improve classification accuracy. This motivates the need for an alternative approach, that complements the relative entropy approach, when $\delta_{avg} < MAD$.

6 PERFORMANCE EVALUATION

We utilized two big datasets for the performance evaluation of our proposed method. The first dataset is an hourly power consumption dataset from PeCan Street Project [40], containing 200 and 800 houses from a solar village near Austin, Texas for years 2014, 2015, 2016. The 2014 and 2015 dataset is used for learning (training), while 2016 is used a testing set. Two 90 day periods representing two seasons in 2016 were used as a scenario under attacks to generate the malicious dataset. The malicious data sets were generated from the real data samples that were fed with our threat model with various ρ_{mal} and δ_{avg}. The second dataset is a power consumption dataset from 5000 houses from six micro-grid regions in Dublin, Ireland [42], which was utilized to prove the scalability and generality of our proposed approach. The datasets are publicly accessible.

The experimental section is divided into four parts: (i) First, we show some results related to the fine grained anomaly detection; (ii) Second, we show supervised classification results for 200 houses for all attack types over various δ_{avg} value (iii) Third, we show unsupervised classification (using

K-means) for 200 houses. (iv) Fourth, we show a performance evaluation in terms of classification error rates for both 800 houses and 5000 houses using unsupervised classification, to prove that error rates scale well for larger micro-grids and works across different combinations of ρ_{mal} and δ_{avg} for various datasets, (v) we show real time nature of detection of smart meters, (vi) a few comparisons of our performance with existing works.



(a) (b)

Fig. 9. (a) Data Omission Captured in Uncleaned Data (b) Deliberate Additive On-Off Attack Forensic

6.1 Fine Grained Anomaly Detection Forensics

Here we show some results on how the fine grained anomaly detection metric can detect opportunistic strategies such as Omission and On-Off.

Data Omission Strategy: Fig. 9(a), shows a result on the uncleaned real dataset with missing data. We do not know whether this was due to an attack or a network failure. Nonetheless, this is analogous to data omission, and our proposed fine grained anomaly detection metric $AD_{ratio}(t)$, can capture such events. Since the metric $FGAT$ contains only two entries for the whole year, it is evident that this is particular data omission is likely an isolated failure, rather than an attack. A magnified version was shown previously in Fig. 8, to prove that the $AD_{ratio}(t)$ first decreases (when omission starts) and then increases (when omission stops).

On-Off Strategy: We study a small timeline of say 10 days, and start additive attacks (ON) and then stop it (OFF), it is possible to detect the ON period of attacks with the proposed $AD_{ratio}(t)$ metric. As an example, Fig. 9(b), shows an additive attack with $\delta_{avg} = 600$, which was launched from the 60-th hour to the 200-th hour of this time-line. Note that, in additive attacks the harmonic mean grows at a much slower rate compared to the growth in arithmetic mean (given a sufficiently high δ_{avg}). Hence, at the 60-th slot the difference between the arithmetic mean and harmonic mean is larger than the previous time slots. There the ratio $AD_{ratio}(t)$, shows a sharp increase.

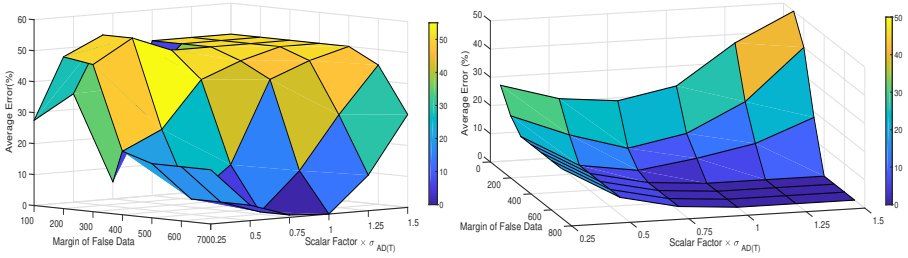
6.2 Effectiveness of the Anomaly based Attack Context Generation

The effectiveness of the anomaly detection step is directly related to the embedding of attack context in the proposed trust model which in turn preserves the classification accuracy, lowers false alarm rates and, improves time to accurate classification of the compromised meters. Therefore, the effectiveness of the anomaly detector is demonstrated through the minimization of classification error rates (defined as the average of missed detection and false alarm rates).

The effectiveness of the anomaly detector is also directly dependent on the value of threshold ($\pm\gamma\sigma_{AD(T)}$) around the historical $AD(T)$ value. Recall, that γ is the scalar factor that parameterizes the threshold variation. Therefore, to demonstrate the effectiveness of anomaly detector we show the error rates (average of missed detection and false alarms) as a function of the varying margins of false data and variable candidate thresholds in the anomaly detector. Through this, we also

demonstrate the optimal threshold range that the anomaly detectors should use to minimize the error rate in classification.

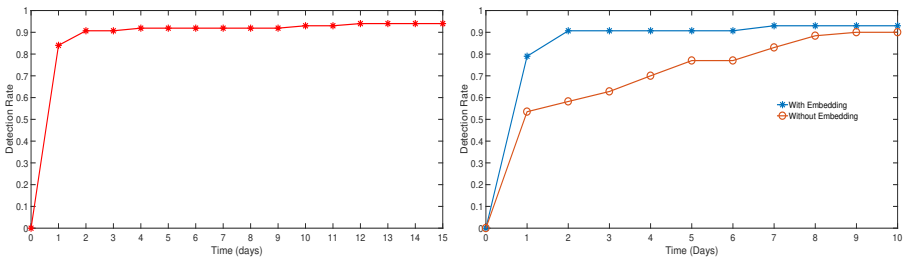
Effectiveness of Error Rate Minimization: We report a $0.75\sigma_{AD(T)}$ as a threshold that produces minimal error rates across extreme values of ρ_{mal} and over all trained values δ_{avg}. This study is done because the defender has no control on the actual ρ_{mal} and δ_{avg} values that will manifest. Fig. 10(a) clearly shows that a global minima for classification error rate exists for a threshold of $0.75\sigma_{AD(T)}$, which produces minimal error rates regardless of δ_{avg} among all candidate thresholds for the Irish dataset for $\rho_{mal} = 15\%$ under additive attacks. Fig. 10(b) shows that the minimal error rate is achieved for the same $0.75\sigma_{AD(T)}$ across all δ_{avg} for different $\rho_{mal} = 50\%$ under a deductive attack.



(a) (b)

Fig. 10. Error Rate Minimization (a) Low $\rho_{mal} = 15\%$; (b) High $\rho_{mal} = 50\%$

Effectiveness of Time to Detection (TTD): Figs. 11(a) is a CDF that is a testimony of the convergence times to the detection rate for an additive attack with $\rho_{mal} = 20\%$ and $\delta_{avg} = 600$ and a data-order aware strategy. The classification of compromised meters is not only accurate but also happens in a very quick time. The steady state detection rate as observed from Figs. 11(a) is achieved within 2 days. Additionally, Fig. 11(b), shows the effectiveness of the probability of attack time ratio embedding (as a result of the fine grained anomaly detector) into the trust model, and proves that it improves the time to detection of compromised meters significantly. The Fig. 11(b), shows the comparison between the CDF of detections with and without embedding under an on-off strategy with an on-to-off ratio of $1:3$. We can observe that the circled line corresponding to detection rate without the P_{attack} embedding approaches its steady state after atleast 10 days compared to the blue line with the probability of attack time ratio embedding that approaches the steady state detection rate of 90% within just 2 days.



(a) (b)

Fig. 11. (a) TTD of Compromised Meters (b) Comparative Effectiveness of P_{attack} embedding

6.3 Supervised Classification

In this case, the threshold is obtained from a small set of training meters from the training dataset which is then applied to the testing set with the full set of meters in test set. Later, we show how our proposed approach performs in an unsupervised mode as well.

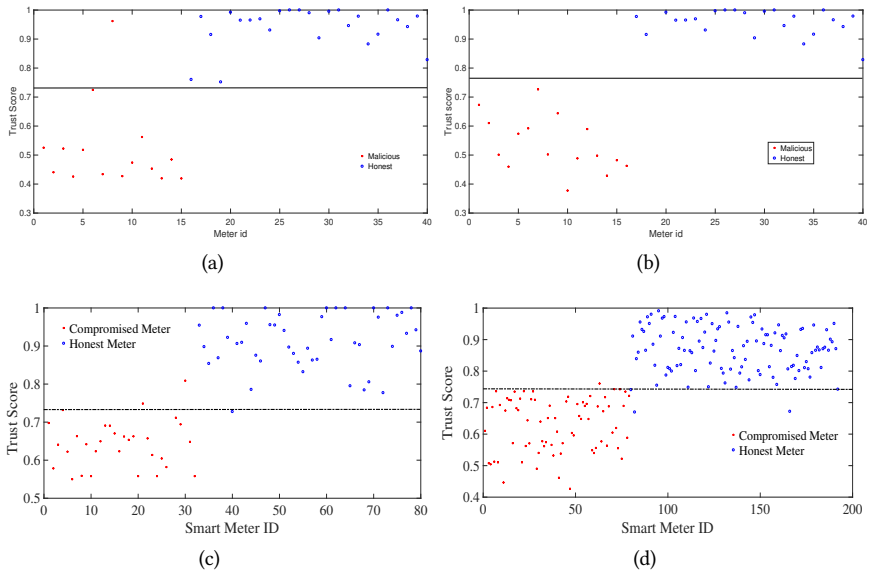


Fig. 12. Training Set: (a) Additive; (b) Deductive (c) Effect of Meter Sizes (d) Effect of Different Season

6.3.1 Training Set:. First, we use a training data set from 40 houses and use power consumption reported in 2014 for a month. In each training case, we labeled 40% meters as compromised ($\rho_{mal} = 0.4$) and alter their reported values with $\delta_{avg} = 500W$ and then plotted, the corresponding trust values. We chose intermediate values of ρ_{mal} and δ_{avg} to prevent overfitting or underfitting. We use the trust scores of these labels, to calculate a threshold that can linearly separate between compromised and non-compromised nodes. We use a decision tree based classifier called CART (Classification and Regression Trees) to find the supervised thresholds. The results of training for additive and deductive attacks are shown in Figs. 12(a), 12(b). Then we studied, the effect of meter training size by repeating this with 80 meters (See Fig. 12(c)) as well as the effect of the training time period (seasonal change) on all meters (See Fig. 12(d)) to test the sensitivity of training for supervised classification. The conclusion is that all thresholds are close.

6.3.2 Classification with Testing Set:. For testing illustration, we use 2016 dataset from Texas and the attack launching period is one month. We set $\rho_{mal} = 0.4$ and $\delta_{avg} = 600W$. *More results over completely different combinations of ρ_{mal} and δ_{avg} are presented later to prove the robustness performance.* Results for additive and deductive attacks shown in Figs. 13(a) and 13(b), exhibit a clear separation between honest and compromised nodes with a false alarm rate of 1.5% in both the cases. The missed detection rate is 5% and 8% for additive and deductive attacks, respectively.

6.4 Classification Performance Evaluation

Fig. 14(a), shows the classification error rates for a larger dataset of 800 houses in terms of missed detections and false alarms under additive attack for the unsupervised classification approach over all possible values of δ_{avg}, given a $\rho_{mal} = 0.50$. From the figure, we can conclude that the relative entropy approach works well for most values of δ_{avg} even when 50% of the nodes are compromised. Particularly, the missed detection is higher than false alarms, which means detection rate is more of a concern for additive attacks particularly, when $\delta_{avg} < 400$. We report 22% missed detection and 2% false alarm at $\delta_{avg} = 400$. At $\delta_{avg} = 300$, the missed detection rate increases to 39%. Therefore,

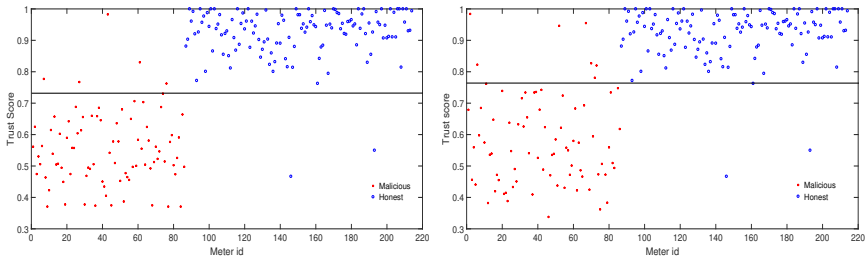


Fig. 13. Testing Sets: (a) Additive; (b) Deductive

we experimentally verify that is methodology is not well suited for the margin of false data lesser than the median absolute deviation of the dataset.

Fig. 14(b) shows the classification error rates in terms of missed detections and false alarms for unsupervised classification approach over all possible values of δ_{avg}, given $\rho_{mal} = 0.50$ under a deductive attack for 800 houses. This indicates the robustness of our solution across all margins of false data under deductive attacks. The missed detection rate does not have an upper evasion point compared to our preliminary work [2] and other information theoretic approaches.

Fig. 15(a) and 15(b), shows the classification error rates in terms of missed detections and false alarms for the unsupervised classification approach over all possible values of δ_{avg}, given a $\rho_{mal} = 0.20$ under a camouflage and conflict attack.

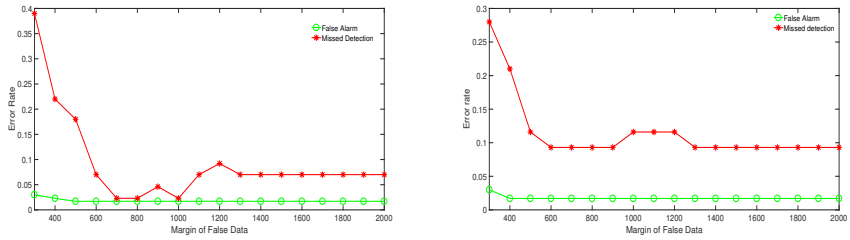


Fig. 14. Error Sensitivity Analysis over δ_{avg} (Texas): (a) Additive (b) Deductive

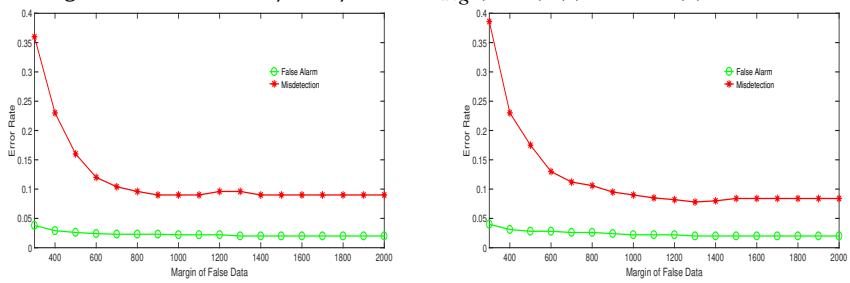
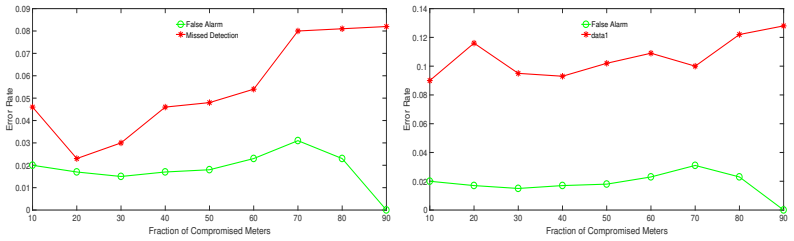


Fig. 15. Error Sensitivity Analysis over δ_{avg} (Texas): (a) Camouflage (b) Conflict

Figure 16(a) confirms that the error rate is within 10% for all possible fractions of compromised nodes as high as 90%, for the additive attack. This indicates the robustness of our solution to higher fractions of compromised nodes for additive attacks. Additionally, Figure 16(b), indicates the robustness of our solution to various margins of false data under deductive attacks. The missed detection rate does not have an upper evasion point in terms of ρ_{mal}.



(a) (b)

Fig. 16. Error Sensitivity Analysis over ρ_{mal} (Texas): (a) Additive (b) Deductive

6.5 Comparisons with Existing Work and Scalability of Error Rates

Fig. 17 shows that the false alarm rate for the Irish dataset across 5000 houses is less than 2%. Additionally, the missed detection rate is below 20% for any $\delta_{avg} \geq 350$W. Second, the Fig. 17, compares our performance for deductive attacks with existing works in terms of missed detection (MD) and false alarm (FA) rates, that use techniques such as One class SVM [12], multi-class SVM [12], F-Deta (Information Theory based) [16], folded Gaussian trust [4]. The proposed approach's performance in terms of FA and MD is shown in solid lines with season wide cross validation. From the figure, it is evident that across various margins of false data, our FA and MD rates are lowest compared to the other approaches. Additionally, across the same chosen δ_{avg}, our work remains resilient under high fractions of compromised meters compared to previous works.

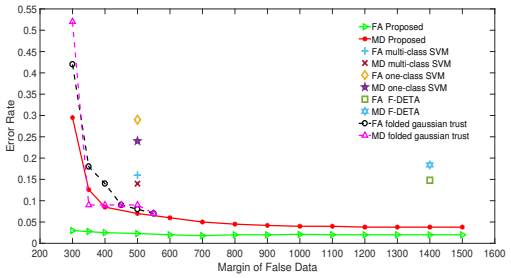


Fig. 17. Error Rate Comparison with Existing Works: Irish Dataset

Table 8, also quantifies the advantages and benefits of our framework in comparison to some of the recent works in this area, in terms of 'other aspects' that are not directly comparable with previous works. These aspects include ranges of studied margin of false data δ_{avg} and ρ_{mal}, detection rate convergence times, applicability to multiple attack types, and both coarse and fine grained opportunistic attack strategies. While our framework applies to all attack types, other works focus on deductive attacks except our previous work. Therefore, the numbers for our framework in Table 8 are for deductive attacks only for a fair comparison. However, our work is much broader compared to existing works since it addresses an umbrella of various threats simultaneously. Some entries in the table marked NA when a concerned parameter that is not reported explicitly. Moreover, our work shows error sensitivity performance over both datasets.

Our framework has a much better performance over a wide attack strategy space with ρ_{mal} ranging from 1% to 90% and δ_{avg} ranging from $300W$-$2000W$ compared to the existing works that assume a narrower or fixed attack strategy space in terms of ρ_{mal} and δ_{avg}. Works such as [16] reasonable missed detection rates, but assume a very high δ_{avg} of above 1000W which facilitates easier classification. The false alarm rate at only select δ_{avg} is provided and the detection time is not clear. At this assumption, our missed detection rate is less than 6% and false alarm rates are 8% for a larger dataset of 800 meters. The work in [12] has a small ρ_{mal} of 0.72%, but at their assumed

$\delta_{avg} = 400W$, our MD and FA rates are better for both additive and deductive attacks across lower and higher ρ_{mal} values while needing the same number observations per day. Our work can also perform classification in an unsupervised mode compared to the supervised approach with a high training time as reported in [12]. The upper evasion limit of high δ_{avg} and ρ_{mal} vanishes, compared to our preliminary work [2], due to the robust mean and median absolute deviation correction and convergence times are preserved under omission and on-off attacks. Our recent work [10] also showed that harmonic and arithmetic mean calculations are compatible with fully homomorphic encryption schemes enabling privacy preserving security computations in AMI. Therefore, our method unlike others will be compatible with AMI privacy requirements [31].

Table 8. Comparison with Existing Work

Parameter	Proposed	CPBETD [12]	ARMA [19]	Prior [2]	F-Deta [16]
False Alarm	1.5%-4%	29%	33%	11%	NA
Missed Detection	30%-0%	24%	28%	8%	10%-36%
δ_{avg}	300-3000W	400W	NA	700-800W	1000W-2000W
ρ_{mal}	$1\% - 90\%$	1%	NA	$\leq 40\%$	55%
Attack Type	All	Deductive	Deductive	All	Deductive
Detection Time	2-3 days	77days	30 days	30 days	NA
Opportunistic strategies	Yes	Yes	No	No	No

7 CONCLUSIONS AND FUTURE WORK

In this paper, we proposed coarse and fine grained anomaly based security event detection technique that serves as an early indicator of the presence of organized data falsification attack, infers the attack type, and strategy inflicted, which helps to reconstruct an attack context that includes a response metrics such as robust mean, standard deviation, attack probability time ratio, which depend on what kind of threat has been inflicted. Based on this attack context, the relative entropy trust model adapts itself dynamically in runtime, to produce linearly separable trust scores that can identify the compromised meters injecting false data with higher accuracy and in near real time. In all, we showed that our framework applies regardless of the high fraction of compromised nodes, and across various margins of false data in an unsupervised classification mode as well with very low time to detection of compromised meters.

In the future we will address the problem of anomaly detection and meter identification when the margin of false data, the upper and lower interval of false data is much smaller than the established bypass margin of false data. Such stealthy attacks are possible since every unit of electricity has a value, which the evidential model using robust mean and median absolute deviation will not be able to detect.

Acknowledgements: The work is supported by National Science Foundation grants under award numbers SATC-2030611, SATC 2030624, OAC-2017289, CNS-1818942, CNS-1545037, CNS-1545050, CPS 1943035, ECCS 1936131, NIFA - 2017-67008-26145. We thank the reviewers and associate editor for suggestions. We thank Mr Aditya Thakur for assistance with some of the experimental plots.

REFERENCES

[1] V. Agate, A. Khamesi, S. Gaglio, S. Silvestri, "Enabling peer-to-peer User-Preference-Aware Energy Sharing Through Reinforcement Learning", *IEEE International Conference on Communications (ICC)*, 2020.
[2] S. Bhattacharjee, A. Thakur, S. Silvestri, S.K. Das, "Statistical Security Incident Forensics against Data Falsification in Smart Grid Advanced Metering Infrastructure", *ACM Conference on Data and Application Security (ACM CODASPY)*, pp. 35-45, Mar. 2017.
[3] S. Bhattacharjee and S.K. Das, "Detection and Forensics under Stealthy Data Falsification in Smart Metering Infrastructure", *IEEE Trans. on Dependable and Secure Computing*, Vol. 16, Dec. 2018.
[4] S. Bhattacharjee, A. Thakur, S.K. Das, "Towards Fast and Semi-Supervised Identification of Smart Meters launching Data Falsification Attacks", *ACM Asia Conference on Computer and Communications Security (ACM ASIACCS)*, pp. 173-185, 2018.
[5] P. Box, D. Cox, "An analysis of transformations", *Journal of the Royal Statistical Society*, Series B. Vol. 26 (2): pp. 211-252, 1964.
[6] A. Cardenas, R. Berthier, R. Bobba, J. Huh, J. Jetcheva, D. Grochocki, and W. Sanders, "A Framework for Evaluating Intrusion Detection Architectures in Advanced Metering Infrastructures", *IEEE Trans. On Smart Grid*, Vol. 5(2), pp. 906-915, Mar. 2014.
[7] V. Chandola, A. Banerjee, V. Kumar, "Anomaly detection: A survey", *ACM Computing Surveys*, Vol. 41, Issue, 15, pp. 15-58, 2009.

[8] S. Ciavarella, J. Y. Joo and S. Silvestri, "Managing Contingencies in Smart Grids via the Internet of Things," in IEEE Trans. on Smart Grid, vol. 7, no. 4, pp. 2134-2141, July 2016.

[9] V. Dolce, C. Jackson, S. Silvestri, D. Baker, A. De Paola, "Social-Behavioral Aware Optimization of Energy Consumption in Smart Homes" *IEEE International Conf. on Distributed Computing in Sensor Systems (DCOSS)*, 2018.

[10] Y. Ishimaki, S. Bhattacharjee, H. Yamana, S.K. Das, "Towards Privacy-preserving Anomaly-based Attack Detection against Data Falsification in Smart Grid", *IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SMARTGRIDCOMM)*, Nov. 2020.

[11] R. Jiang , R. Lu, Y. Wang, J. Luo, C. Shen, and X. Shen, "Energy-Theft detection issues for advanced metering infrastructure in smart grids", *Tsinghua Science and Technology*, Vol. 19(2), pp. 105-120, Apr. 2014.

[12] P. Jokar, N. Arianpoo, and V. Leung, "Electricity theft detection in AMI using customers' consumption patterns", *IEEE Trans. on Smart Grid*, Vol. 7(1), pp. 216-226, Jan. 2016.

[13] A. Khamesi, S. Silvestri, D. Baker, A. De Paola, "Perceived-Value Driven Optimization of Energy Consumption in Smart Homes" *ACM Transactions on Internet of Things*, Vol. 1, Issue 2, 2020.

[14] A. R. Khamesi, S. Silvestri, "Reverse Auction-based Demand Response Program: A Truthful Mutually Beneficial Mechanism", *IEEE Mobile Ad Hoc Sensor and Smart Systems (IEEE MASS)*, 2020.

[15] T. Koppel, "Lights Out: A Cyberattack, A Nation Unprepared, Surviving the Aftermath", *Crown Publishers, New York*, 2015.

[16] V. B. Krishna, K. Lee, G. A. Weaver, R. K. Iyer and W. H. Sanders, "F-DETA: A Framework for Detecting Electricity Theft Attacks in Smart Grids," *IEEE/IFIP on Dependable Systems and Networks (IEEE DSN)*, 2016, pp. 407-418.

[17] S. McLaughlin, D. Podkuiko, and P. McDaniel, "Energy theft in the advanced metering infrastructure", *Proc. of Critical Information Infrastructures Security*, Springer-Verlag, pp. 176-187, Sept. 2009.

[18] S. McLaughlin, B. Holbert, S. Zonouz, and R. Berthier, "AMIDS: A multi-sensor energy theft detection framework for advanced metering infrastructures", *IEEE Conf. on Communications, Control, and Computing Technologies for Smart Grid Communications (SMARTGRIDCOMM)*, pp. 354-359, Nov. 2012.

[19] D. Mashima and A. Alvaro, "Evaluating electricity theft detectors in smart grid networks", *Springer Intl. Workshop on Recent Advances in Intrusion Detection*, pp. 210-229, Sept. 2012.

[20] R. Mohassel, A. Fung, F. Mohammadi, and K. Raahemifar, "A survey on advanced metering infrastructure", *Elsevier Journal of Electrical Power & Energy Systems*, 63:473-484, Dec. 2014.

[21] B. Meyer, "Some Inequalities for Elementary Mean Values", *AMS Mathematics of Computation*, Vol. 42, No. 165, pp. 193-194, 1984.

[22] A. Rad and A.L. Garcia, "Distributed internet-based load altering attacks against smart power grids", *IEEE Trans. on Smart Grids*, Vol. 2(4), pp. 667-674, Dec. 2011.

[23] E. Shin, A. R. Khamesi, Z. Bahr, S. Silvestri, D. A. Baker, "A User-Centered Active Learning Approach for Appliance Recognition" *IEEE International Conference on Smart Computing (SMARTCOMP)*, 2020

[24] Y.L. Sun, W. Yu, Z. Han, K.J. Ray Liu, "Information Theoretic Framework of Trust Model and Evaluation for Ad Hoc Networks",*IEEE Journal on Sel. Areas in Communications*, Vol. 24(2), pp. 305-317, Feb. 2006.

[25] S. H. Tung, "On Lower and Upper Bounds of the Difference Between the Arithmetic and the Geometric Mean", *AMS Mathematics of Computation*, Vol. 29, No. 131, pp. 834-836, 1975.

[26] J. P. Talusan, F. Tiasus, K. Yasumoto, M. Wilbur, A. Dubey, S. Bhattacharjee, "Smart Transportation Delay and Resiliency Testbed Based on Information Flow of Things Middleware," *IEEE International Conference on Smart Computing (SMARTCOMP)*, USA, 2019.

[27] E. Werley, S. Angelos, O. Saavedra, O. Cortes, and A. Souza, "Detection and identification of abnormalities in customer consumptions in power distribution systems", *IEEE Trans. on Power Delivery*, Vol. 26(4): pp. 2436-2442, Oct. 2011.

[28] M. Wilbur, A. Dubey, B. Leao, S. Bhattacharjee, "A decentralized approach for real time anomaly detection in transportation networks", *IEEE Conference on Smart Computing*, 2020.

[29] W. Xia, Y. Chu, "The schur convexity of gini mean values in the sense of harmonic mean", *Mathematica Scientia*, Vol. Vol. 31(3), pp. 1103-1112, 2011.

[30] W. Yu, D. Griffith, L. Ge, S. Bhattarai and N. Golmie, "An integrated detection system against false data injection attacks in the Smart Grid, *Security and Commun. Networks*, Vol. 8(2), pp. 91-109, Jan. 2015.

[31] [Online] Available at: https://skyvisionsolutions.files.wordpress.com/2014/08/utility-smart-meters-invade-privacy-22-aug-2014.pdf

[32] *NY Times*, Last Accessed Oct. 2020, [Online] Available at: http://www.nytimes.com/2009/12/14/us/14meters.html?ref=energy-environment&_r=0

[33] [Online] Last Accessed Oct. 2020, Available at: http://www.telegraph.co.uk/news/2017/03/06/smart-energy-meters-giving-readings-seven-times-high-study-finds/

[34] [Online] Last Accessed Oct. 2020, Available at: https://www.maximintegrated.com/content/dam/files/design/technical-documents/white-papers/smart-grid-security-recent-history-demonstrates.pdf

[35] [Online] Last Accessed Oct. 2020, Available at: https://energy-solution.com/2015/01/29/enabling-automated-demand-response-pge-dras/

[36] [Online] Last Accessed Oct. 2020, Available at: https://www.smartgrid.gov/files/The_Smart_Grid_Promise_DemandSide_Management_201003.pdf

[37] [Online] Last Accessed Oct. 2020, Available at: http://blog.comverge.com/intelligent-energy-management/does-ami-have-what-it-takes-for-demand-response/

[38] [Online] Last Accessed Oct. 2020, Available at: http://krebsonsecurity.com/2012/04/fbi-smart-meter-hacks-likely-to-spread/

[39] [Online] Last Accessed Oct. 2020, Available at: https://www.whitehouse.gov/sites/default/files/microsites/ostp/nstc-smart-grid-june2011.pdf

[40] [Online] Last Accessed Oct. 2020, Available at: https://www.smartgrid.gov/project/pecan_street_project_inc_energy_internet _demonstration.html

[41] [Online] Last Accessed Oct. 2020, Available at: http://energy.gov/sites/prod/files/oeprod/DocumentsandMedia/14-AMI_System_Security_Requirements_updated.pdf

[42] [Online] Last Accessed Oct. 2020, Available at: Irish Social Science Data Archives, Available at: http://www.ucd.ie/issda/data/

[43] Link: https://drive.google.com/file/d/1rWZ6O68Hm55Qf403RQp2M2o0HuD7hqam/view

Appendix A. Implementation of Data Order Aware Attacks:

In Fig. 18, the blue line corresponds to the actual power consumption. The red and yellow lines correspond to deductive attacked consumption data following a non-data order aware and a data order aware strategy respectively *under same* δ_{avg} *and* ρ_{mal}. Even as the same revenue is achieved with both strategies, the chances of detection (using proximity based mechanisms) are lesser in data order aware strategy due closer proximity to the actual data.

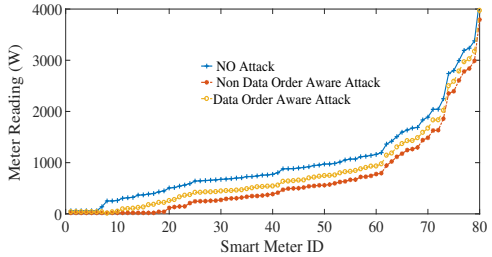


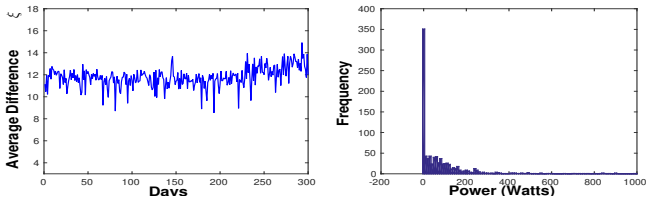
Fig. 18. Illustration: Benefit of Data Order Awareness

This strategy is implemented in the following manner: The adversary sorts the actual power consumptions observed from its set of M compromised meters such that $P_{act}^{(1)}(t) \leq, \cdots, P_{act}^{(m)}(t), \leq P_{act}^{(M)}(t)$. Then adversary generates M random numbers for $\delta(t)$, sorted as $\delta_t^{min} \leq, \cdots, \leq \delta_t^{max}$. For a deductive attack, the highest observed power consumption data is changed with the highest δ_t^{max}, while the lowest observed power consumption data is changed with the lowest δ_t^{min}. Hence, $P_{act}^{(1)}(t) - \delta_t^{min}, \cdots, P_{act}^{(M)}(t) - \delta_t^{max}$. For an additive attack, the lowest observed power consumption data is changed with the highest δ_t^{max}, while highest observed power consumption data is modified with lowest δ_t^{min}, and so on, such that $P_{act}^{(1)}(t) + \delta_t^{max}, \cdots, P_{act}^{(M)}(t) + \delta_t^{min}$. For a camouflage attack, the sorted $P_{act}^{(1)}(t) \leq, \cdots, \leq P_{act}^{(M)}(t)$ is divided into two parts, and corresponding portions are changed accordingly. This kind of attack therefore, is more aware of the current consumption trends as seen by the meters under adversarial control and minimizes the chances of the final reported value to be obvious outliers and more close to the actual power consumption distributi10n.

Appendix B. Explanation of Stability of Ratio Metric:

We provide a short theoretical and mathematical reasoning behind the observed stability of absolute difference between harmonic mean to arithmetic mean for AMI power consumption data across various data sets. Note that, the invariant $AD(T)$ is a daily metric and most residential households share certain coarse grained common behavioral routines, although individual fine grained differences exist. Hence, the power consumption of different households are not completely independent but exhibit some weak positive correlation (i.e. the power consumption of houses tend to increase or decrease together due to commonality of habits).

For example, most houses are tend to use more electricity on very cold days. Obviously, the strength of this positive correlation varies may vary from region to region, but in a particular Neighborhood Area it produces some common correlation. Since humans have common behavioral habits during a typical day, then intuitively the daily pattern of average difference in power consumption values between any two pair of houses averaged over T is not going to be arbitrarily different from each other. Let the average difference between power consumption of any two houses in the sorted series $p^1(T) \cdots, p^N(T)$ over T be denoted as $\xi(T) = \frac{1}{N}\sum_{i=1}^{N} |p^{i+1}(T) - p^i(T)|$. The distribution of $\xi(T)$ for the Irish dataset as shown in Fig. 19(a), follows a stable trend. Additionally, the variance in $\xi(T)$ is also less. The most important thing is that the $\xi(T)$ is stationary in the mean in the wide piecewise sense even for 5000 houses.



(a) (b)

Fig. 19. Irish Dataset: Distribution of (a) $\xi(T)$; (b) a_{min}

Now, we show an important relationship between the stability of $\xi(T)$ and Tung's Theorem [25], which proposed the theoretical upper and lower bounds on the absolute difference between Arithmetic and Geometric Means in any series data. The corollary for this Theorem [21] describes the upper and lower bounds on the absolute difference between Harmonic and Arithmetic Mean in a series data.

- *Tung's Theorem Corollary:* Given a sorted series $a = 1 \equiv a_1, \cdots, a_n \equiv B$, where 1 and B denote the minimum and maximum values of the series of n numbers. Let H_n and A_n denote the harmonic and arithmetic means respectively. Then the bounds on the absolute difference between H_n and A_n:

$$\frac{(B-1)^2}{N(B+1)} \leq |A_n - H_n| \leq (\sqrt{(B)} - \sqrt{(1)})^2 \tag{38}$$

For minimum (a_{min}) and maximum (a_{max}) values, Eqn. 38 can be rewritten as:

$$\frac{(a_{max} - a_{min})^2}{n(a_{max} + a_{min})} \leq |A_n - H_n| \leq (\sqrt{(a_{max})} - \sqrt{(a_{min})})^2 \tag{39}$$

where $a_{max} \sim a_{min} + (n-1)\xi$. Therefore, the bounds on $|A_n - H_n|$ is only a function of ξ and a_{min}.From Figs. 19(a) and 19(b), we know that both ξ and a_{min} are mostly stable; hence $|A_n - H_n|$ is also highly stable. This is one explanation on the stability of harmonic to arithmetic mean ratios across multiple datasets and subsets.

Received xxxxx; revised xxxx; accepted xxxx