Towards Fast and Semi-supervised Identification of Smart Meters Launching Data Falsification Attacks

Shameek Bhattacharjee
Missouri University of Science and Technology, Rolla, USA
shameek@mst.edu

Aditya Thakur
Missouri University of Science and Technology, Rolla, USA
astvd3@mst.edu

Sajal K. Das
Missouri University of Science and Technology, Rolla, USA
silvestri@cs.uky.edu

ABSTRACT

Compromised smart meters sending false power consumption data in Advanced Metering Infrastructure (AMI) may have drastic consequences on the smart grid's operation. Most existing defense models only deal with electricity theft from individual customers (*isolated attacks*) using supervised classification techniques that do not offer scalable or real time solutions. Furthermore, the cyber and interconnected nature of AMIs can also be exploited by organized adversaries who have the ability to *orchestrate* simultaneous data falsification attacks after compromising several meters, and also have more complex goals than just electricity theft. In this paper, we first propose a real time semi-supervised anomaly based consensus correction technique that detects the presence and type of smart meter data falsification, and then performs a consensus correction accordingly. Subsequently, we propose a semi-supervised consensus based trust scoring model, that is able to identify the smart meters injecting false data. The main contribution of the proposed approach is to provide a practical framework for compromised smart meter identification that (i) is not supervised (ii) enables quick identification (iii) scales classification error rates better for larger sized AMIs; (iv) counters threats from both isolated and orchestrated attacks; and (v) simultaneously works for a variety of data falsification types. Extensive experimental validation using two real datasets from USA and Ireland, demonstrates the ability of our proposed method to identify compromised meters in near real time across different datasets.

CCS CONCEPTS

• **Security and privacy → Trust frameworks**; **Intrusion detection systems**; • **Theory of computation → Semi-supervised learning**; • **Hardware → Smart grid**;

KEYWORDS

Data Falsification; Advanced Metering Infrastructure; Smart Grid Security; Anomaly Detection; False Data Injection; Cyber-Physical System Security

ASIA CCS '18, June 4–8, 2018, Incheon, Republic of Korea

ACM ISBN 978-1-4503-5576-6/18/06…$15.00
https://doi.org/10.1145/3196494.3196551

ACM Reference Format:
Shameek Bhattacharjee, Aditya Thakur, and Sajal K. Das. 2018. Towards Fast and Semi-supervised Identification of Smart Meters Launching Data Falsification Attacks. In *ASIA CCS '18: 2018 ACM Asia Conference on Computer and Communications Security, June 4–8, 2018, Incheon, Republic of Korea.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3196494.3196551

1 INTRODUCTION

Advanced Metering Infrastructure (AMI) is one of the building blocks of the smart grid technology, responsible for collecting data on loads and consumer's electricity consumption [16]. Such data are usually collected by smart meters installed on the customer site, and are expected to play a pivotal role in current and future smart grids. For example, AMI data will be at the basis of critical tasks such as automated billing and pricing, demand response, forecast, load adjustments [31], and management of daily and critical peak shifts. Hence, the integrity of AMI data is of utmost importance. However, data falsification attacks target the integrity of AMI data.

In the literature, defense against the falsification of electricity consumption data, has been mostly focused on electricity theft, [9, 13, 14, 23], where individual customers are the primary adversaries, who report lower than actual usage for lesser electricity bills. Since the actually measured reading of power consumption is reduced, such an adversarial strategy is a *deductive* mode of data falsification. Such attacks from individual rogue customers are usually uncoordinated and we term them as *isolated attacks*.

However, it is recognized that the cyber and interconnected nature of AMIs can be exploited by more organized adversaries, (e.g., organized criminals [30] and business rivals [10, 27]), who are more equipped to bypass cryptographic defense, compromise several smart meters, and alter a large or small amounts of data simultaneously, thereby significantly impacting the smart grid's operations [10, 13, 30]. We term such attacks as *orchestrated attacks*. Orchestrated physical attacks tampering the meter hardware to produce false data was reported in [29, 30]. Thus, cryptography or network intrusion alone cannot protect against this threat.

The goals of organized adversaries may not be restricted to monetary benefits on the customer billing side resulting from electricity theft. As an example, higher than actual power consumption can be reported by a meter as a byproduct of static and dynamic load altering attack [17] or hardware tampering affecting both customers and utilities. Such an attack is termed as an *additive* mode of data falsification. An additive attack launched by a utility on its rival company's meters, may induce loss of business confidence by the customers of the victim company, due to higher bills. The expected future use of AMI data for demand response, forecast and load planning may induce additive attacks to benefit customers by

drawing undue incentives during demand response [1]. Note that, an organized attacker may also perform a balancing additive and deductive attacks to evade detection methods that use mean aggregates, which are termed as *camouflage* mode of data falsification.

Previous works on orchestrated or isolated attacks have many disadvantages. Classification based techniques use computationally expensive Multi-Class Support Vector Machines (SVMs) [8, 9], Neural Networks [7] and only focus on retrospective identification (takes 6 months to 2 years), lacking the possibility of detecting the attacks or compromised meters before serious damages. Other works [1, 11] utilize a complete supervised approach by maintaining continuous fine grained meter specific historical evidence which is impractical and error prone for large scale AMI networks, and requires separate training for particular attack types. State based detector needs special hardware which is very costly as elaborated in Section 2. Some consensus based approaches use traditional measures of central tendency such as median and mean [14, 22, 23] or their variants. Such consensus measures get easily affected by larger margins of false data or larger fractions of compromised meters (when using instantaneous consensus metrics). They also fail under camouflage attacks (typical in orchestrated attacks), may lead to larger error rates (when using historical consensus metrics), since the mean aggregate of power consumption data readily changes (proved later by real datasets).

In this paper, we propose an anomaly based consensus correction scheme and a semi-supervised learning based trust scoring model, that detects occurrence as well as the specific type of falsification of power consumption data (referred to as *attack context*), and then identify the compromised meters injecting such false data in an AMI, regardless of isolated or orchestrated attacks. Specifically, we propose a novel metric based on *harmonic to arithmetic mean ratios of daily power consumption* to detect anomalies and infer the attack context. Based on the inferred attack context, we calculate a *resilient mixture mean and standard deviation* as an approximate consensus measures that weaken the alterations caused by the false data from orchestrated attacks. Subsequently, a set of discrete rating levels is associated to each meter over time using the proximity of its reported data to this resilient mixture mean. Then, a Folded Gaussian distribution based weighing procedure is used to assign weights to each of the discrete rating levels. Based on rating levels and weights observed over a time window, a trust value is calculated per meter that classifies compromised meters.

We validated our model through extensive experiments on real datasets acquired from two different AMI infrastructures with varying sizes and regions. Results show that our proposed method is able to detect and decipher additive, deductive, and camouflage attacks launched by organized adversaries in real time. We demonstrate that our method is robust to a high fraction of compromised meters (upto 75%), is able to identify compromised meters from non-compromised ones over margins of false data, thus making it scalable for large sized AMI. Additionally, our method identifies against isolated attacks from individual meters. We compare our results with three existing works to show improvement.

2 LIMITATIONS OF RELATED WORK

Existing works on AMI data falsification can be classified into Classification based, State Estimation based, and Consensus based methods. Classification based approaches [7–9] require extensive training phases and multi-class SVMs for each customer separately, in order to detect electricity thefts. They are computationally complex and only allow retrospective identification. A study comparing classification methods [7] concluded that the accuracy of most of these models are only 60% to 70%, although they suffer from privacy intrusion and complexity issues.

State based detection techniques [6, 12, 13] in contrast, require additional monitoring hardware deployed at various points across the AMI and distribution network for consistency checks. Additional hardware requirement is costly to the extent that it has been recognized as a practical deterrent for utility providers to use such solutions in scale. Some works monitor non-technical loss (NTL) at the transformer meter. However, in [7] it is observed that NTL could vary due to large number of factors other than attacks (e.g., legitimate changes due to unexpected weather) and hence suffers from high number of false alarms. Moreover, the NTL approach cannot detect for camouflage or load altering induced attacks.

Consensus based methods [14, 22, 23] use smoothened moving average of median or mean power consumption for detection, followed by information theory to identify meters. Most works except [14] assume isolated electricity theft from a small number of malicious meters that does not greatly bias the consensus. But this assumption on unbiased consensus may not hold for organized adversaries with higher attack budgets launching orchestrated attacks. Some works such as [22, 23], use historical mean/median power consumption for comparison of bad behavior. However, the mean power consumption varies readily due to contextual factors such as weather, customer habits etc. as shown later from studies with our real datasets. Some works [1, 11] use a supervised learning of historical proximity patterns of each meter with instantaneous consensus, but *fails for higher fractions of compromised meters (≥ 40%)*. Additionally, supervised approaches become cumbersome for large scale grids due to large training sets and require labels which may not be available or accurate. Another major limitation these methods, is that the assumed margins of false data per meter are usually fixed and are also typically very high (600W-1500W), which favors easier detection. As shown later, the mean consumption can easily get affected by both larger margins of false data or legitimate consumption changes (e.g., sudden weather changes), which increases errors. This is evident from [14, 22], where fine grained monitoring still yields accuracy of about 62%. Note that cryptography based approaches are not enough since physical attacks can also cause data falsification [29, 30].

3 SYSTEM ARCHITECTURE AND DATASET DESCRIPTION

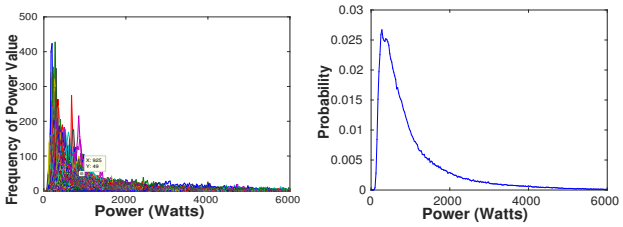
We consider a set of N smart meters reporting power consumption data to a data concentrator periodically. Let the i-th smart meter report a datum P_t^i at the end of time slot t. We model P_t^i as the realizations of a random variable (r.v.) P^i denoting the power consumption distribution of the i-th smart meter. A Neighborhood Area Network (NaN), formed by a collection of houses is governed by a NaN gateway node, that may act as data concentrators collecting data from multiple smart meters in an area. Multiple NAN gateways may be connected to form a Field Area Network (FAN),

governed by a FAN gateway, which in turn is connected to the Utility Wide Area Network (WAN).

Decentralized defense models are deployed at either NAN or FAN gateways while centralized detection frameworks are deployed at the WAN [3]. Since the datasets did not reveal the actual topology, we show results on smaller subsets of meters and as a whole, to mimic both deployment possibilities and understand performance scalability with varying micro-grid size N.

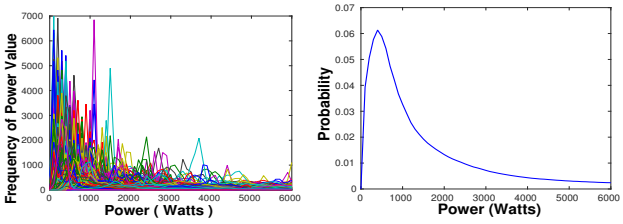
3.1 Dataset Description

To study the distribution of P^i, we investigated hourly (i.e., t slotted hourly) reported real power consumption datasets of 700 houses from Austin, Texas [25] and 5000 houses from Dublin, Ireland [26] that belong to residential customers. We observed that each P^i follows an approximate lognormal distribution in the Texas dataset. We also observed that all such log-normal distributions are clustered close to each other such that the variance between them is not arbitrarily large. The evidence is shown in Fig. 1(a). Given this observation, we claim that the combination of the individual lognormals can be well approximated by a mixture distribution which is also log-normal (as evident from Fig. 1(b)). We denote P^{mix} as the random variable with approximate lognormal mixture distribution considering all houses in the grid. The trends from Texas dataset, also matches with trends from the same experiments over the Irish dataset with similar results on consumption for a different population for different years as shown in Figs. 2 (a) and 2(b), proving generality of observations.



(a) (b)

Figure 1: Texas Dataset: (a) All Houses, (b) Mixture



(a) (b)

Figure 2: Irish Dataset: (a) All Houses, (b) Mixture

3.2 Gaussian Approximation of the Data

With an aim to ease mathematical tractability and exploit certain known properties of Gaussian distributions, we seek to convert the approximate lognormal distributions to an approximate Gaussian distributions. For this, we use a (NIST recommended) power transformation procedure [2] which is described by the following:

Given a data set $\boldsymbol{d} = \{d^1, d^2 \cdots, d^n\}$, where n denotes the total number of data points, the power transformation of $\boldsymbol{d}$ is given by $\boldsymbol{d(\lambda)} = \{d^1(\lambda), \cdots, d^n(\lambda)\}$, such that:

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

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

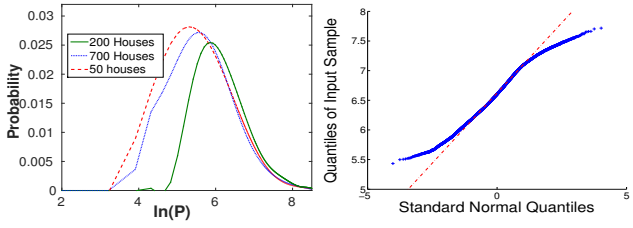
$$\lambda = \underset{\lambda \in \mathcal{R}}{\text{argmax}}\, f(\boldsymbol{d}, \lambda^*)$$

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

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

such that $\bar{d}(\lambda) = \frac{\sum_{i=1}^{n} d^i(\lambda)}{n}$. Using Eqn. 1, any P^i can be converted to obtain an *approximate* gaussian distributed r.v. denoted as p^i. For Texas dataset, we found $\lambda = -0.03$, which is closer to zero. Therefore, to simplify implementation, we used effective $\lambda = 0$ for the transformation of both datasets in this paper. Note that our proposed model does not require the data to be perfectly Gaussian.

We performed the above procedure for 3 subsets of smart meter population sizes (see Fig. 3(a)) to prove topology invariance on the Gaussianity. Hence, we validate the claim that, power consumption of meters in a micro-grid can be approximated by a Gaussian distribution. We also denote $p_t^i = ln(P_t^i + 2)$ as the effective power consumption report from each i on a power transformed scale (since $\lambda = 0$) at any time slot and p^{mix} denotes the aggregate mixture. The transformation is done to exploit certain statistical properties exhibited by the Gaussian distributions. The extent of Gaussianity is depicted through a Q-Q plot in Fig. 3(b). While the Gaussian approximation resulted in 67% and 69% of the datapoints to be within the first standard deviation (for texas and irish datasets), the distributions remain unbalanced around the mean with 64% of the total datapoints on the left and 36% on the right of the mean on average. This asymmetry contributes to unique observations under orchestrated attacks as discussed in Section 5.1.1.



(a) (b)

Figure 3: p^{mix} for Texas Data: (a) Various N, (b) QQ-Plot

3.3 Time Domain Granularities

While real data is collected hourly (known as time slots t), we calculate various consensus/anomaly detection metrics at the end of a finite 'time window' (denoted by T), that is a collection of l time slots. Here, l depends on how fine or coarse grained monitoring is desired. Let T_{re} be the average time taken by customers to react to the environmental factors that may trigger sudden legitimate changes in power consumption in houses. For example, in warmer sunny days during winter season, it is expected that most houses would reduce heaters within some time $T_{re} < l$. Sufficient research exists to show that using $l = 24$ hour window is reasonable [9].

Therefore, T has a daily time granularity. For confirming the presence of falsification, we calculate a cumulative average of anomaly detection metric over a 'sliding time frame' of F time windows.

4 ATTACK MODEL AND IMPLEMENTATION

In this section, we elaborate on four factors characterizing the adversarial strategy, namely, *fraction of compromised meters, attack types, falsification margins, and falsification distributions* that can be employed by organized or isolated adversaries. To the best of our knowledge, we have ensured that the numbers quantifying the adversarial strategy do not favor or suit the proposed defense mechanism. Since real malicious data samples are not available, we generated the malicious samples by applying the following four aspects of adversarial strategy over the real dataset:

Fraction of Compromised Meters: Power consumption data from smart meters can be sabotaged by an organized adversary or isolated rogue customers [13]. Let organized adversaries compromise M meters based on their attack budget and control a certain fraction $\frac{M}{N} = \rho_{mal} \in [0,1)$ of the N smart meters. For example, $\rho_{mal} = 0.50$, means 50% of the total number of meters is compromised. Note that ρ_{mal} could be very high in decentralized micro-grids, where N is typically smaller. Unlike most existing works, we explore the full possible spectrum of ρ_{mal} varying from 1% to 90% for varying network sizes N, while studying performance. Although the defense model is focused on an orchestrated attack with larger ρ_{mal}, we show in Section 6, that our trust model is also identifies small scale isolated attacks from lone 'rogue meters'.

Attack Types: The organized adversaries falsify data from multiple compromised meters simultaneously using one of falsification attacks *deductive, camouflage, additive*, based on its objective and intent. Falsification of data is achieved by accordingly changing the actual power consumption value P_t^i by some amount δ_t. For example, for deductive falsification, the actual power consumption data P_t^i from the i-th compromised meter at time t is modified as $P_t^i - \delta_t$. Similarly, for additive falsification, the modified attack sample is $P_t^i + \delta_t$ from a compromised meter. For camouflage falsification, half of the compromised meters launch additive falsification while the other half launches deductive falsification with the same average value of δ_t. The δ_t amount of false data is chosen randomly from within a margin $(\delta_{min}, \delta_{max})$ according to some falsification distribution with a strategic average falsification margin δ_{avg}.

Average Falsification Margin: We denote δ_{avg} as the *average margin of false data* for each compromised meter. The strategic value of δ_{avg} by a rational adversary is some value that ensures some minimum revenue but also prevents easy detection. Unlike existing works, which either do not clearly articulate the exact δ_{avg} (such as [14]) or achieve good performance for high δ_{avg} values such as 1200W-1500W [11], 600-900W [1], 400W-430W [9], we explore δ_{avg} values ranging from as low as 50W upto 2000W, to show the classification performance over a broad range of possible δ_{avg} values. The micro-grid sizes of 200, 800 and 5000 are used to show the scalability of performance error rates with relative sensitivity to δ_{avg}. We report improved performance on different parameters for $\delta_{avg} > 350W$ that is compared to existing works, and derive conclusions on δ_{avg} required to evade detection.

Falsification Distribution: Additionally, we argue that the distribution of δ_t within $(\delta_{min}, \delta_{max})$ should be some variant of uniform distribution such that the resultant shape of power consumption distribution remains unchanged, making it a smarter and less obvious attack. In contrast, the effect of normally distributed δ_t on the resultant shape is quite apparent. A comparison between normally and uniformly distributed δ_t is shown in Figs. 4(a) and 4(b). Note that, while our defense model works under both cases the results mostly consider variants of the uniformly distributed strategy. Apart from this (a) *Random strategy*, the following falsification distributions are also possible: (b) *Periodic:* Targeting the dynamic or time of use (TOU) electricity pricing [11], where attacks are launched on specific times when the price/demand of electricity is high. We implemented a periodic strategy where attacks happen on every 12 hours in a day,(c) *Incremental:* Instead of immediately attacking with the intended δ_{avg}, the adversary increases its average falsification margins by a minuscule amount $d\delta$ on every time slot till it reaches its intended δ_{avg}. We implemented an incremental strategy where $d\delta = 2W$ updated 4 hourly. (d) *Omission:* No data reaches the utility since communication/data is jammed/dropped. This is implemented by replacing data with null values from a subset of meters.

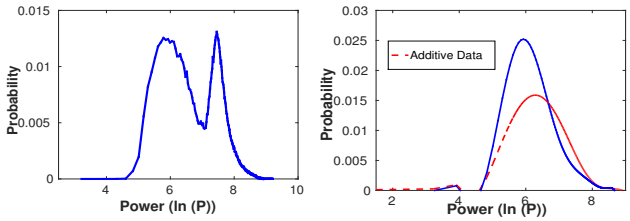


Figure 4: Attack Distribution (a) Obvious, (b) Smarter

5 PROPOSED FRAMEWORK

The proposed framework has two major parts: (a) Anomaly-driven consensus correction model, and (b) Trust scoring model. The consensus correction model provides robust consensus as inputs to the proposed trust scoring model, which improves the classification.

5.1 Anomaly Driven Consensus Correction

The objective of the anomaly based consensus correction module is to prevent the consensus measure (aggregate mean and standard deviations) from getting too biased due to orchestrated attacks. First, we show that the arithmetic mean is not a stable historical invariant for aggregate power consumption. Therefore, historical mean cannot be used as a consensus measure, and instantaneous mean cannot be used since distinguishing legitimate changes in the mean from malicious changes is difficult. The consensus correction module has four phases: (i) proposed detection metric training, (ii) inferring presence of organized falsification, (iii) detecting type of falsification, and (iv) calculation of resilient consensus (resilient mean and standard deviation (μ_{MR} and σ_{MR}) in a time window.

5.1.1 **Ratio of Harmonic to Arithmetic Mean.** Now, we show that the ratio of harmonic to arithmetic mean metric is better than other consensus/aggregate based measures for anomaly detection and consensus correction due to: (i) Higher invariance to legitimate changes in consumption, (ii) Pythagorean mean special asymmetry property facilitates attack detection.

Let $p_t^{mix} = \{p_t^1, \cdots, p_t^N\}$ denote the power consumption data series on an power transformed scale gathered from N smart meters at any time slot t (t is slotted hourly). The harmonic mean (HM_t) and arithmetic mean (AM_t) of aggregate power consumption in a time slot is defined as:

$$HM_t = \frac{N}{\sum_{i=1}^{N} \frac{1}{p_t^i}} \qquad AM_t = \frac{\sum_{i=1}^{N} p_t^i}{N}$$

All HM_t and AM_t over the time window T are recorded, such that the corresponding daily averages are given by $HM_{avg}(T) = \left(\sum_{t=1}^{24} HM_t\right)/24$ and $AM_{avg}(T) = \left(\sum_{t=1}^{24} AM_t\right)/24$ over this window T (T is a daily window consisting of 24 time slots). Similarly, let the average daily standard deviation be denoted as $SD_{avg}(T) = \left(\sum_{t=1}^{24} \sigma_t\right)/24$. Many prior works such as [14, 22, 24], propose the use of arithmetic mean or its derived smoothening statistics (such as Cumulative Sum/Moving Averages of $AM_{avg}(T)$) for sequential anomaly detection. Thereafter, they propose to use historical mean as a consensus in the event of an attack or the mean prior to the attack detection. However, Fig. 5(a) shows how actual arithmetic mean power consumption fluctuates for the same time windows for three years [25] in the Texas dataset without showing any repeating historical pattern or a stable time series. Due to high fluctuations of the instantaneous arithmetic mean, the error residual between the derived smoothening statistic and the actual mean is large. Thus, it will be difficult to identify legitimate changes from a malicious one by monitoring the time series. It will also cause large errors if historical arithmetic mean is used as a consensus. This is evident from the high rates of false alarms and missed detection reported in [7, 9]. To circumvent this problem, we propose to use the ratio of $HM_{avg}(T)$ and $AM_{avg}(T)$ as the detection metric by:

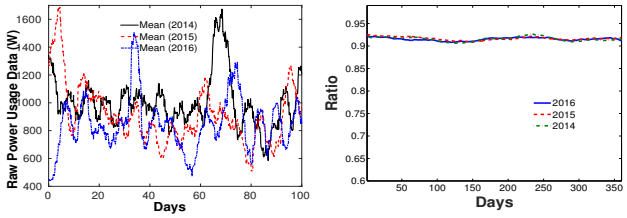
$$Q_{avg}^{ratio}(T) = \frac{HM_{avg}(T)}{AM_{avg}(T)} \tag{3}$$

We denote μ_{ratio} and σ_{ratio} as the mean and standard deviation of $Q_{avg}^{ratio}(T)$ observed in the dataset. Let us explain three reasons for choosing Eqn. 3, as a metric for detecting presence of attacks.

(i) High Invariance to Legitimate changes: From our experimental study, we observed that the time series of $Q_{avg}^{ratio}(T)$ samples over different years and across multiple datasets is highly stable over time in contrast to the time series of arithmetic mean of power consumption. Fig. 5(b), shows the daily $Q_{avg}^{ratio}(T)$ over three different years (2014, 2015, 2016) for the Texas dataset, while Fig. 6(a) shows the ratio $Q_{avg}^{ratio}(T)$ for six different meter populations for a completely different AMI data set in Dublin, Ireland during 2009-2010. Both, Figs. 5(b) and 6(a), prove that $Q_{avg}^{ratio}(T)$ is a highly stable invariant metric across different data sets, as compared to the aggregate arithmetic means. Note that, $Q_{avg}^{ratio}(T)$ cannot exceed 1, due to the $HM \leq AM$ property [18].

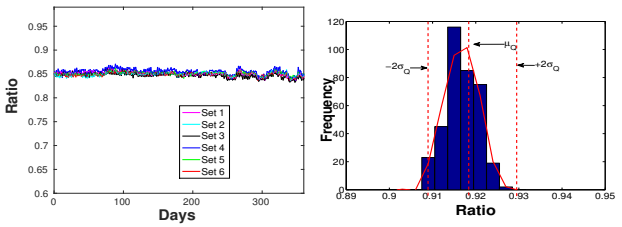
Apart from the stability over time, $Q_{avg}^{ratio}(T)$ also exhibits historical stability over different years, unlike arithmetic mean which shows large differences in the readings on the same day in successive years. In fact, without using moving averages, the standard deviation of the ratio samples σ_{ratio} of $Q_{avg}^{ratio}(T)$ is 0.017 and 0.012 for the Texas and Irish datasets respectively. Using a smoothening moving average will further lower the standard deviation and produce a more stable invariant under normal conditions.

Additionally, higher the variance in the power consumption dataset, the lesser is the mean of ratio sample distribution and vice-versa. Hence, we conclude that the $Q_{avg}^{ratio}(T)$ is a more robust metric for anomaly detection than other typical measures such as mean, mode, median due to its high invariance to legitimate changes in data over successive days across years.



(a) (b)

Figure 5: (a) $AM(T)$ unstable, (b) $Q_{avg}^{ratio}(T)$ of Texas Dataset



(a) (b)

Figure 6: (a) $Q_{avg}^{ratio}(T)$ of Irish Dataset, (b) Ratio Distribution

(ii) Special asymmetry property of Pythagorean Means: Asymmetric growth (or decay) rates of harmonic mean compared to the symmetric growth (or decay) of arithmetic mean under various attacks, is another reason which helps to infer the presence and type of falsification precisely, quickly, and with high sensitivity. With this attack context information, it is possible to estimate the true consensus accurately. When a *subset* of p^i values in p^{mix} are increased/decreased (with a false bias δ), the AM value grows or decays linearly. On the contrary, the behavior of HM is inherently different and can be summarized as follows:

Property (1) : We observe that HM *grows slower* and *decays faster* than corresponding AM, when sub-portions of a data set generated from multiple sources experience additive and deductive manipulation, respectively.

Property (2) : Growth and decay rates of HM under the same δ is unequal when used for additive and deductive attacks, unlike AM which show equal rates. In HM, decay rate is larger than its growth rate induced by the same δ.

Property (3) : We observed that growth and decay rates of HM compared to AM and the effects on the proposed ratio metric also depend on (a) whether the datapoint being biased are on the lesser (left) or greater (right) than the actual arithmetic mean, and (b) the magnitude of δ.

The above properties can be mathematically illustrated by the following: Consider a *sorted* series with two numbers $U = (u_1, u_2)$ such that its mean and standard deviation is (AM, σ). In Fig. 7, the x-axis represents the variable u_1. Let us fix the u_2 as constant such that $u_2 = \{2\}$ is a singleton set, while u_1 is a set such that $u_1 \in \mathbb{R}^+$.

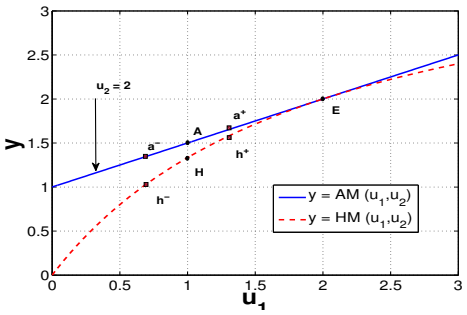


Figure 7: Growth Decay Rates of HM and AM

Hence, the cartesian product of u_1 and u_2 is the set $U = u_1 \times u_2$, whose elements are a two tuple dataset. In Fig. 7, let the y-axis represent the value of $AM(u_1,u_2)$ or $HM(u_1,u_2)$ for every possible element in the set U. Given any element say $U = (1,2)$, visualize increasing/decreasing value of u_1 as mimicking additive or deductive biases to $U = (1,2)$, that changes both the $AM(u_1,u_2)$ or $HM(u_1,u_2)$. In Fig. 7, AM function of $(u_1,2) \quad \forall \quad u1 \in (0,\infty]$ (represented by the solid blue line) shows a linear growth with increasing u_1, and is neither strictly concave or convex. On the other hand, the HM function of $(u_1,2) \quad \forall \quad u_1 \in (0,\infty]$ (represented by a dashed red line) is a strictly Schur-Concave Function [32]. This difference in concavity is the theoretical basis that trigger changes in the proposed ratio metric under various attack types and is illustrated below using the same example.

Illustration of Properties: For dataset $U = (1,2)$, let the $AM = 1.5$ and $HM = 1.33$ be represented by points A and H as marked in Fig. 7. Hence, their ratio value say $Q^{ratio} = \frac{HM}{AM} = 0.88$.

Suppose in $U = (1,2)$, $u_1 = 1$ is biased with deduction of $\delta = 0.3$, such that $U^- = (0.7,2)$. Points $a^- = 1.35$ and $h^- = 1.037$ correspond to the biased arithmetic and harmonic means respectively. Thus, decay in HM and AM are $\Delta HM^- = h^- - H = -0.293$ and $\Delta AM^- = a^- - A = -0.15$. Ignoring the signs which signify decay, $|\Delta HM^-| > |\Delta AM^-|$, proving that HM decays faster than AM. Note that the biased ratio of HM to AM in this case is $Q^- = 0.76 < Q^{ratio} = 0.88$. Suppose, the same $\delta = 0.3$ is instead added to u_1. However, with $\delta = 0.3$, added to u_1, the $Q^+ = 0.95 > 0.88 = Q^{ratio}$. Since $|0.76 - 0.88| > |0.95 - 0.88|$, it proves Property 2. While, we may be temped to believe that additive attacks increase ratio while deductive attacks decrease them, this is not true. Consider an additive bias value of $\delta = 3.5$ instead, that is added to data-point $u_1 = 1$. The resultant ratio in this case is $Q^+ = 0.82$, which is a decrease from the original ratio 0.88. Therefore, the fact that *magnitude of δ, plays a role in the observed rise or drop in the ratio is established.* Finally, suppose the $\delta = 0.3$ were added to u_2 instead of u_1. Note that in $U = (1,2)$, the data-point $u_2 = 2$, is on the greater than (right side) of the true mean of U ($= 1.5$). Now the $Q^+ = 0.80 < Q^{ratio} = 0.88$. *This proves that position of the data-point being biased by an δ, w.r.t to true AM also plays a role in the ratio change.* The above clearly explains property 1,2 and 3. The necessary and sufficient conditions of δ_{avg}, for observing a drop or rise in ratio metric under each attack type and the effect of biased datapoint's position relative to the true mean is detailed in Appendix A. The basic conclusion from the illustration in Appendix A, is that *the ratio metric decreases only when the biased datapoints move original datapoints away from the true arithmetic mean in a way that flattens the shape of the distribution (increasing the sample standard deviation), while, rise in ratio occurs if the biased datapoints move original datapoints closer to the true mean (increasing the sample standard deviation).*

(iii) Effect of Attacks on the Ratio Metric : Since we know that datapoints regardless of meter ids are more frequently on the lesser than of the mean (as shown in Section 3.2), it is a natural consequence that the more attacked data-points, will also be lesser than the true mean, regardless the attack type. Given this, a deductive attack, will cause more datapoints to move further away from the true mean. Therefore, we can conclude that the ratio is going to decrease for deductive attacks, regardless the δ_{avg}. For camouflage attacks, also since the HM has a higher decay rate than growth rate of AM for the same δ bias, the ratio is bound to decrease regardless the δ_{avg}. In contrast, the additive attacks with lower margins of false data will cause the final biased datapoints to be proximate to the true mean, thus reducing the sample standard deviation, and therefore increasing the ratio metric. However, for higher margins of δ_{avg}, the biased datapoints will end up being greater than true mean, and the ratio will show a decrease.

Above conclusions on increase and decrease of ratios have been experimentally verified in Fig 8, where $\rho_{mal} = 40\%$ was used for different attacks with varying δ_{avg}. The deductive and camouflage attacks correspond to a $\delta_{avg} = 600W$. The additive small and additive large denote the ratio lines under an attack of $\delta_{avg} = 200$W and $\delta_{avg} = 800$ respectively. The exceptions to the above observations may happen, if the attacker possess complete knowledge of system, defense mechanism and insider leaks; the details of which is discussed in Sec. 5.2.4.

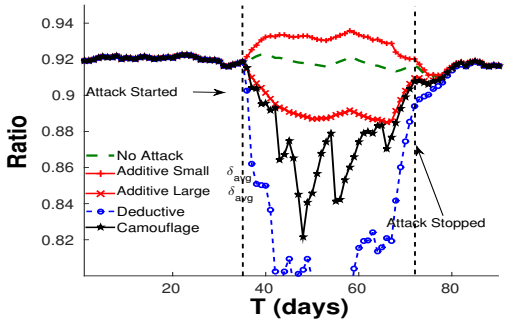


Figure 8: $Q^{ratio}_{avg}(F)$ under various attacks

5.1.2 **Inferring presence of Organized Falsification:** We know that asymmetric growth and decay rates of the $HM_{avg}(T)$ and $AM_{avg}(T)$ trigger a decrease or an increase in the $Q^{ratio}_{avg}(T)$, as false data is injected from a subset of numbers generating these means. Leveraging this knowledge, we propose the unsupervised and semi-supervised versions of the detection criterion that indicates the presence of an organized falsification, and thereby the need to invoke a suitable type of consensus correction.

A sustained drop or rise in ratio indicates malicious activity. Therefore, we need a collective anomaly detection (monitoring a subsequence of states) instead of point anomalies (monitoring each state independently). To capture collective anomalies, we define a sliding frame that contains the cumulative average of $Q^{ratio}_{avg}(T)$

samples over the last F days. If the cumulative average in the current frame F has deviated from the cumulative average in the previous frame $F-1$ by a threshold ϵ, then this forms a premise for a sustained change in the ratio metric. Formally, the unsupervised detection criterion is:

$$Q_{avg}^{ratio}(F): \begin{cases} \in Q_{avg}^{ratio}(F-1) \pm \epsilon & \text{No Anomaly;} \\ < Q_{avg}^{ratio}(F-1) - \epsilon & \text{Orchestrated Attack;} \\ > Q_{avg}^{ratio}(F-1) + \epsilon & \text{Low Additive Attack;} \end{cases} \quad (4)$$

where ϵ is a threshold parameter such that $\epsilon \in (0, 3\sigma_{ratio}]$. The choice of ϵ controls whether the consensus correction step will be invoked or not. The appropriate ϵ can be learned by studying the trade-off between ϵ and classification error rate, over various δ_{avg} and ρ_{mal} combinations as shown later in Figs. 11(a) and 11(b). Note that, the required sensitivity of ϵ to attacks need not to be very precise, since the purpose of Eqn. 4 is to only catch evidence of orchestrated attacks, that disturb the consensus significantly and therefore need the consensus correction. In contrast, isolated or smaller scale attacks (with low δ_{avg}/ρ_{mal}) do not drastically deviate the ratio metric and thus may not get detected under the given ϵ. However, at the same time such small scale attacks will also not affect the consensus in a way that causes large classification errors.

Therefore, observing $Q_{avg}^{ratio}(F)$ over time is enough to conclude whether an orchestrated falsification is happening. If $Q_{avg}^{ratio}(F)$ has decreased (or increased) more than ϵ, than the previous frame $Q_{avg}^{ratio}(F-1)$, it is an evidence of the start of an orchestrated falsification. Let this frame be marked as F_{trig}, such that $F_{trig}-1$ is the last frame with a normal ratio value.

Now, after a period of sustained drop (or rise) in the ratio metric outside the ϵ, an increasing (or decreasing) $Q_{avg}^{ratio}(F)$ may indicate that attacks are now ceasing. As seen in Fig. 8, the ratio increases (decreases) back to the normal range $Q^{ratio}(F_{trig}-1) \pm \epsilon$ value on the 72nd day, when our implemented attack was stopped on the 68-th day. Note that, isolated attacks from individual customers, may not have a drastic effect on the ratio margin ($\pm\epsilon$), and these attacks are countered by the *trust model* discussed later. This is a very simple but very powerful technique to differentiate between legitimate changes due to environment and false data injections.

Semi-Supervised Version of Detection Criterion: One disadvantage of the unsupervised detection criterion (Eqn. 4), is that it may miss incremental attacks where δ_{avg} slowly increases over time, such that the drop of ratio compared to the previous time window will be within the ϵ. However, if enough historical (attack free) data is available (e.g. the Texas Dataset), the historical normal range $Q_{avg}^{ratio}(F^{hist}) \pm \epsilon$ of the ratio can be learned easily given its stable nature. In such as case, even with incremental attacks, the ratio metric will eventually cross the learned stable historical range.

5.1.3 **Inferring Type of Data Falsification:** Once falsification is inferred at F_{trig}, observing the direction of $HM_{avg}(T)$ and $AM_{avg}(T)$ growth or decay, indicates the *type* of data falsification: additive, deductive, camouflage. For notational simplicity, henceforth we will refer to $HM_{avg}(T)$ and $AM_{avg}(T)$ as HM and AM respectively. The resilient mean ($\mu_{MR}(T)$) and standard deviation ($\sigma_{MR}(T)$) for window T, is referred as μ_{MR} and σ_{MR} respectively.

If both HM and AM values have increased compared to F_{trig}, then it is an additive attack. In deductive attack, both HM and AM decreases from the F_{trig}. In camouflage, the AM does not change much and HM decreases. The various possibilities are depicted in Table 1. A pictorial view of this is shown later in Fig. 12 and Fig. 13.

5.1.4 **Consensus Correction:** We calculate the resilient mean at each window T (μ_{MR}), as an *estimation* of the actual mean, given the information on presence and the type of attack. We exploit the robustness of HM and AM, under different attack types for estimation of the actual mean.

Table 1: Inferring Attack Types

Ratio	HM,AM	Inference	μ_{MR}
Down	Up,Up	Additive	HM-(AM-HM)
Down	Down,Down	Deductive	AM+(AM-HM)
Down	Down,Similar	Camouflage	HM
Similar	Up,Up	Legit Up	AM
Similar	Down,Down	Legit Down	AM

Mean Correction: The choice of μ_{MR} is guided by how the detected attack type biases the actual values of HM and AM. For additive attacks, the growth in HM is less than AM due to slower growth rate although both increases from actual AM. Hence, we deduct corrective factor $(AM-HM)$ from the observed HM to estimate the μ_{MR}. For deductive attacks, HM has a faster decay rate than AM. Since, $HM \leq AM$, for deductive attacks, HM is even lesser than the reduced AM. Hence, we add to the observed AM, the corrective factor $(AM-HM)$, to estimate the μ_{MR} such that $\mu_{MR} = AM + (AM-HM)$ is closer to the actual mean and far from deductive outliers at the same time. We choose to add $(AM-HM)$ to the AM because of larger HM drop for deductive attacks can cause $(AM-HM)$ value to be very high (when ρ_{mal} and/or δ_{avg} is high). Adding it to HM may be far less than the true mean. Hence, adding it to AM makes μ_{MR} closer to the actual AM.

For camouflage attacks, HM works as a good measure of μ_{MR} due to its stability to partial presence of false additive data. In fact, using the HM for camouflage helps distinguish meters launching additive falsification from meters launching deductive falsification, because HM is not symmetrically distant from the additive and deductive outliers, unlike AM. The deductive meters will have trust values lesser than honest meters but higher trust than additive meters. This is because HM will be closest to the data generated from honest meters followed by deductive and furthest from data generated from additive meters. Alternatively, if the separate identification of additive and deductive outlier meters are not desired AM may be used for μ_{MR}. In general, AM is more robust mean for camouflage attacks, when ρ_{mal} for camouflage attack is $\geq$ 50%. The extent of drop in the $Q_{avg}^{ratio}(T)$ is an indication of ρ_{mal} and δ_{avg}. The larger the drop in $Q_{avg}^{ratio}(T)$ larger is the ρ_{mal} and δ_{avg} and larger the bias in the observed mean. In case, no organized attack is detected from anomaly detection phase, μ_{MR} is equal to the observed AM. Table 1, summarizes the calculation of μ_{MR}.

Standard Deviation Correction: The σ_{MR} will increase regardless the type of data falsification attack (except for low additive attacks). Therefore, a directional correction is not possible like μ_{MR} based on the attack types. Using the measured σ_{MR} of the last time window, before detection of orchestrated attack, may not be wise since

there may be a longer delay between the launch and the actual detection of the falsification (such as in incremental attacks). Alternatively, one may be tempted to use the historical value of $SD_{avg}(T)$ on the corresponding T-th day in the previous years. However, this would add to the storage complexity. Moreover standard deviation on the same days on successive years are not necessarily same. We studied the distribution of $SD_{avg}(T)$ over the years and found that a distinct mode of $SD_{avg}(T)$ distribution occurs at $425-475$W range in the non-transformed scale. The probability of $SD_{avg}(T)$ being around $425-475$W is very high over 50% while all other ranges are less than 10%. As an approximation, we choose σ_{MR} as ln(450), whenever orchestrated attacks have been confirmed for using it in the subsequent trust model. The distribution of $SD_{avg}(T)$ is shown in Appendix B.

5.2 Consensus Aware Trust Scoring Model

The trust scoring model has three parts: discrete rating criterion, Folded Gaussian distribution based weights, inverse power law kernel based trust metric.

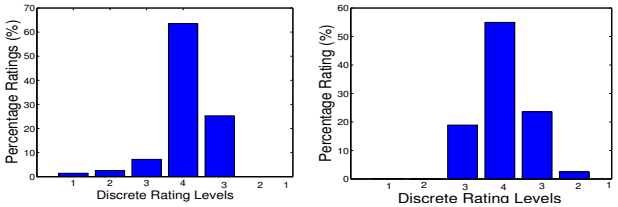
The discrete rating criterion assigns a rating level to each meter i, by comparing proximity of its reported data p_t^i with the resilient mean consensus μ_{MR}, over a time window of length T. Then, weights are assinged to these discrete rating levels according to both prior frequency of occurrence, density of each rating level in a Folded Gaussian distribution, and their proximity to μ_{MR}. This step finally yields an aggregate weight R^i for each i. Then, an inverse power law kernel is used to map the R^i weights into a trust value TR^i between 0 and 1, for linearly separable classification of compromised meters from honest meters.

5.2.1 Discrete Rating Levels:. We propose a criterion to assign a discrete rating level to the reported p_t^i based on its proximity to μ_{MR}. The σ_{MR} is the corrected standard deviation of all p_t^i from the calculated μ_{MR} in the window T. We define $\Delta_{abs} = \sigma_{MR}$. The absolute difference between the p_t^i for any meter i and the μ_{MR} is denoted by $\Theta^i_{diff} = |p_t^i - \mu_{MR}|$. Given this, the discretized rating levels denoted by l is given by Table 2, using the $68-95-99.7$ rule for Gaussian distributions to assign p_t^i as belonging to one of the 4 possible rating levels (bins) according to proximity to the μ_{MR}. The highest rating 4 is closest in terms of proximity to μ_{MR}, and similarly lower ratings are obtained if the meter's data is further from the μ_{MR}. Over a time window of say T hours, the ratings on each time slot t for meter i is collected to form a rating vector sequence r^i, which is sorted as $r^i_{sort} = r_0 \le r_1 \le \cdots \le r_{T-1}$.

Table 2: μ_{MR} based Discrete Rating Levels

Scenario	Discrete Rating Level(l)
$\Theta^i_{diff} \le \Delta_{abs}$	4
$\Delta_{abs} < \Theta^i_{diff} \le 2\Delta_{abs}$	3
$2\Delta_{abs} < \Theta^i_{diff} \le 3\Delta_{abs}$	2
$otherwise$	1

5.2.2 Folded Gaussian based Weights:. Now we find the corresponding (normalized) weights of each rating in the r^i_{sort} which is denoted as $W^i = w_0, \cdots, w_{T-1}$. Figs. 9(a) and 9(b), signify the approximate Gaussian nature of the rating distributions, under no attacks for two real meters from Texas dataset. It is clear that the most common and highest rating level is 4 followed by all others. This gaussian nature is known as *Folded Gaussian* where variables around the mean do not have different signs, since only the magnitude of the level is important. Intuitively, meters with more observed lower ratings should have lesser weights. The sorting makes it easier to give lower weights to smaller ratings through Eqn. (5) by dividing the rating space over the considered time window. Then via Eqn. 6, the distance between this weight x_t from the highest rating level (which is 4 known from no attacks) is determined. If the distance is larger, it assigns a non-linearly decreasing density value based on the shape of Gaussian distribution.



(a) (b)

Figure 9: Real Rating Distribution (a) Meter 1 (b) Meter 2

Additionally, higher percentage of lower ratings in a window, will give even lesser weights to those smaller ratings, than a scenario with lower percentage of low level ratings and vice versa achieved through Eqn. (6). We denote $\mu_{BR} = 4$ as the best or highest possible rating level, σ^i_{dr} denote the standard deviation of discrete ratings of each meter from $\mu_{BR} = 4$ in a window length T. The σ^i_{dr} for each meter will be different based on different observations compared to common mixture data, which captures certain individual differences in consumption. First, a weight parameter x_t distributed between 1 to 4 is calculated as:

$$x_t = 1 + \frac{(K-1)t}{(T-1)} \qquad \forall \quad t = 0, \cdots, T-1 \tag{5}$$

where $K = 4$, is the total number of discrete rating levels in the system, T is the window size. Therefore, the corresponding raw weight cw_t of the rating at time index t is:

$$cw^i_t = \frac{1}{\sigma^i_{dr}\sqrt{2\pi}} e^{-\frac{(x_t-\mu_{BR})^2}{2(\sigma^i_{dr})^2}} \tag{6}$$

The weights yielded from Eqn. (6), are normalized by $w^i_t = \frac{cw^i_t}{\sum_{t=0}^{T-1} cw^i_t}$. Let $I(l,t)$ be an indicator function which indicates whether a particular rating level l occurs in that time slot. All weights corresponding to each unique rating level l, such $l = \{1, \cdots, 4\}$ within T is added up, such that $WD(l) = \sum_{t=0}^{T-1} w_t I(l,t)$.

$$\text{where,} \quad I(l,t) = \begin{cases} 1, & \text{If } l \text{ occurred in time slot } t \\ 0, & \text{Otherwise} \end{cases} \tag{7}$$

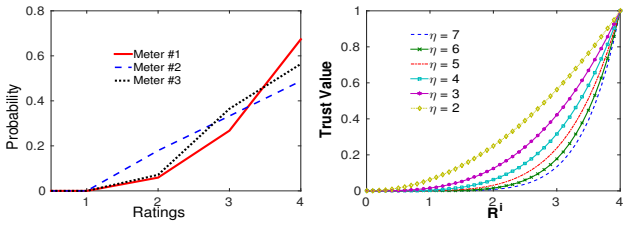
For example, sum of weights in W^i corresponding to each occurrence of rating level 2 is denoted by $WD(2)$. The aggregate weight rating R^i of the i-th meter is a continuous value between 1 and 4 and is given by:

$$R^i = \sum_{l=1}^{K} l \times WD(l), \quad R^i \in \{1,4\} \tag{8}$$

5.2.3 Inverse Power Law based Trust Value:. We know that the Θ^i_{diff} is an unsigned value which can be visualized as a folded Gaussian distribution, where ratings 3,2,1 regardless of whether they are on the right or left of the rating level 4 are treated as the same random variable. Therefore, the aggregate weight (R^i), when interpreted as a trust score must also follow a folded gaussian shape, meaning $R^i = 4$ represents the greatest trustworthiness followed by a exponential 'discounting' of trust, as R^i decreases. For this, we propose the inverse power law inspired kernel trick to transform the R^i into a final trust value, TR^i, between 0 and 1, by:

$$TR^i = \frac{1}{(K)^\eta}(R^i)^\eta, \quad TR^i \in \{0,1\} \tag{9}$$

where η is a scaling factor controlling the rate of discounting. The Eqn. (9), gives exponentially less trust to R^i as it decreases from the maximum value of 4, in adherence to the Folded Gaussian shape of the rating distribution of legitimate meters (shown in Figs. 10(a) and 10(b)). The scaling factor η depends on the skewness of folded gaussian in the benign data set. The Eqn. (9) produces trust values such that compromised and non-compromised meters have *linearly separable*, which enables to calculate an unsupervised threshold for classification. The trust maintenance over time uses a forgetting average [4] for periodic attacks.



(a) (b)

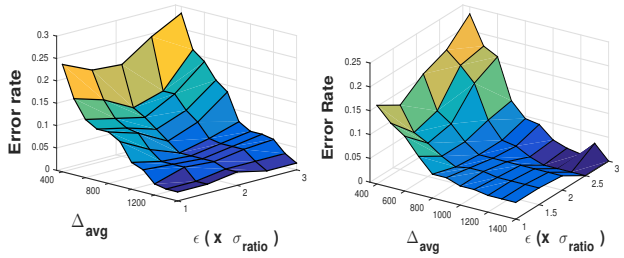
Figure 10: (a) Folded Gaussian (b) Inverse Power Law Kernel

5.2.4 Under Advanced Persistent Threats:. For advanced persistent adversaries, possessing full knowledge of our defense mechanism is 'not' enough to escape detection completely. The adversary has to ensure that deviation in the ratio metric *never* exceeds ϵ on every time window. To completely escape detection, it is mandatory for adversary to possess four additional knowledge: (i) Closed Form Expressions of Harmonic Means (ii) Exact (non-attacked) harmonic and arithmetic mean on each time slot (iii) ϵ value (iv) skewness of data distribution. The skewness knowledge is public (right skewed) and has no attack cost. Different microgrids will have completely different ϵ values, hence the attacker needs to know ϵ of each micro-grid. The ϵ value may be leaked by compromising an utility insider for each micro-grid, or the database storing all the ϵ values. Both possibilities increase the attack cost. Assuming, that attacker knows the ϵ, it further needs the exact knowledge of HM and AM in each time slot. This is rather implausible for adversary to know at runtime, unless it compromises 100% of the meters. This is because the means do not have any stable historical trend or time series, so attacker cannot reliably predict them. Unless actual HM and AM is known, one cannot ensure that the resultant $Q^{ratio}_{avg}(F)$ from the attack will have a ratio deviation that is lesser than the ϵ. Most importantly, the exact closed form expressions of harmonic means do not exist, and is an open problem in real analysis. Several approximations exist [18], but note that it does guarentee success everytime and the defender needs success only once to raise an alarm. Even if the attacker knew everything somehow, we show that the signatures are visually evident for δ_{avg} values as low as $50W$ (see Appendix C).

The sensitivity to successful identification of compromised meters is different from sensitivity to successful detection of presence of orchestrated attacks. Given our model, we report that if that attacker's $\delta_{avg} < 300W$, the missed detection rate drastically increases to 52%, since such δ_{avg} is much less than standard deviation which ends of with higher rating levels. For the attacker to keep atleast 50% of its compromised meters undetected, the highest possible $\delta_{avg} = 300W$ when $\rho_{mal} = 40\%$. However, at such low δ_{avg}, the impact per unit time is less (see Appendix D).

6 EXPERIMENTAL RESULTS

Data sets from 200, 700, and 5000 houses, were obtained from PeCanStreet Project [25] and Irish Social Science Data Archives [26], containing hourly power consumption data from Texas, Austin and Dublin, Ireland respectively. The different microgrid sizes mimic decentralized and centralized deployments of defense frameworks. We studied results of anomaly detection and trust model for all types of data falsification. Additionally, we studied the performance scalability of Irish data for 5000 houses. For anomaly detection results, a period of no attacks is followed by a period of attacks. Deductive attack results have lesser δ_{avg} than the others, since realistic values of power consumption are lower bounded by zero. For clarity of representation, we show anomaly detection results using cumulative moving averages over time frame (F) of length 7 days. Additionally, for easy depiction of meter classification results, all the compromised meters are assigned lower meter id than the honest ones. Finally, we show performance over all values of ρ_{mal} and δ_{avg} and compare with existing work.



(a) (b)

Figure 11: Error Rates: (a) $\rho_{mal} = 40\%$ (b) $\rho_{mal} = 10\%$

6.1 Choice of ϵ for consensus correction

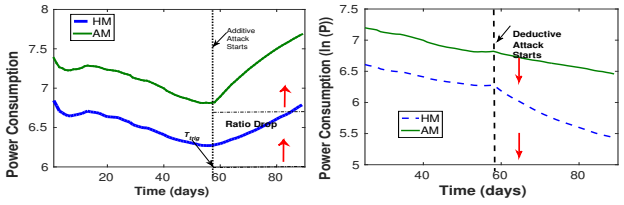
The choice of ϵ decides, whether the anomaly detection raises an alarm or not, which in turns decides whether or not the consensus correction is invoked that subsequently affects classification error rates. In the real-world, the drop/change in the ratio depends on the attacker's ρ_{mal} and δ_{avg} values which is always unavailable to the defense framework at run-time. To provide a suitable recommendation on the desired ϵ, we performed some experiments for an appropriate recommended value of ϵ.

It is important to understand that, smaller ρ_{mal} and δ_{avg} pairs will cause smaller drop/change in the ratios and therefore, ϵ will need to be small to capture them. However, this aspect is offset by

the fact that they also do not affect the consensus in a drastic way. In that case, an unwarranted consensus correction will cause the classification errors to increase. As a proof, it can be observed from Fig. 11(a) and 11(b), error rates are higher for very low values of ϵ. On the flip side, if the ϵ is too large, it will fail to raise alarm for many possible ρ_{mal} and δ_{avg} pairs. Therefore, the consensus correction will not be invoked, the trust scoring model will be executed with a biased consensus increasing the classification error rates. Again this could be verified from Fig. 11(a) and 11(b), where we see the error rates are very high when $\epsilon \sim 3\sigma_{ratio}$. Naturally, there will be an intermediate optimal region of ϵ, where the error rates will be minimized. From computational study, we observed that regardless the wide variation of δ_{avg} (from 350W to 1400W), the error rate minima is achieved at $\epsilon = [1.5\sigma_{ratio}, 2\sigma_{ratio}]$ of the ratio sample distribution, at $\rho_{mal} = 10\%$,$\rho_{mal} = 40\%$ respectively. Therefore, although ρ_{mal} and δ_{avg} may not be known apriori, the recommended ϵ value could be learned a priori, by the above manner. Following this, we have used $\epsilon = 2\sigma_{ratio}$, for all performance results, even for results parameterizing different ρ_{mal}.

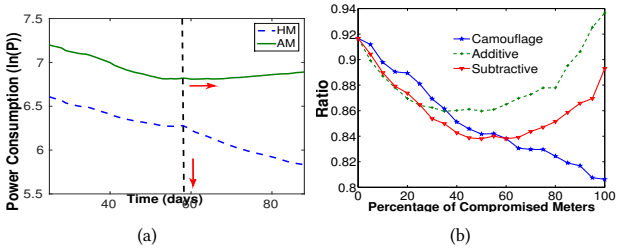
6.2 Inferring Presence and Type of Falsification

Figure 12(a), shows the directional changes in HM and AM and ratio drop to distinguish between legitimate changes and malicious attacks. In the first 57 days, HM and AM changed but their growth/decay had a symmetry, indicating legitimate changes in consumption through same $Q^{ratio}_{avg}(F) \sim 0.92$ value. However, when additive attack phase was launched at 58th day, the $Q^{ratio}_{avg}(F)$ started to decrease from within 2 days of attack, due to slower increase of HM compared to AM. The directional change of both HM and AM values from F_{trig} (arrow upwards) indicate the additive nature of attack. The inference of attack and its type is quick. The Fig. 12(b) and Fig. 13(a), show real time anomaly detection for deductive and camouflage attacks respectively.



(a) (b)

Figure 12: Ratio Change: (a) Additive (b) Deductive

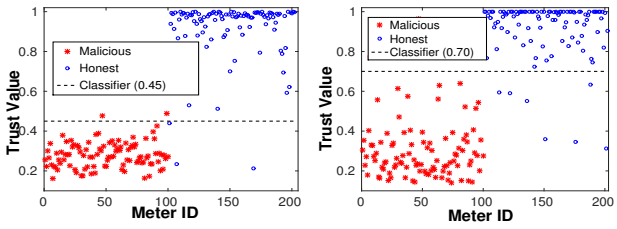


(a) (b)

Figure 13: Ratio Change: (a) Camouflage (b) Over all ρ_{mal}

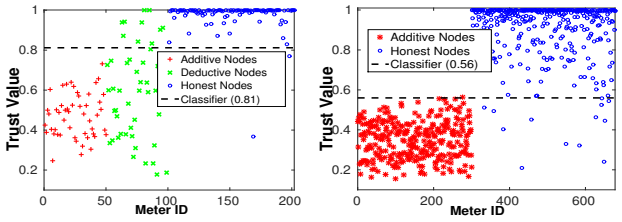
6.3 Robustness of Anomaly Detection over ρ_{mal}

$Q^{ratio}_{avg}(F)$ based detection is robust across larger fractions of compromised meters. Fig. 13(b), show that the drop in the $Q^{ratio}(F)$ is larger than the chosen ϵ (i.e. $Q^{ratio}(F)$ is decreased for $\rho_{mal} <$ 85% values compared to the ratio when ρ_{mal} is 0%). This is the reason for successful anomaly detection to higher fractions of compromised meters. We also observe, that the minimum ratio is achieved when $\rho_{mal} \sim 50\%$, for additive and deductive attacks. However, for camouflage attacks, the ratio always decreases with increasing ρ_{mal}, since the deductive portion of the camouflage attack only affects the Harmonic Mean while keeping the AM same. Hence, the resultant ratio decreases regardless the ρ_{mal}.



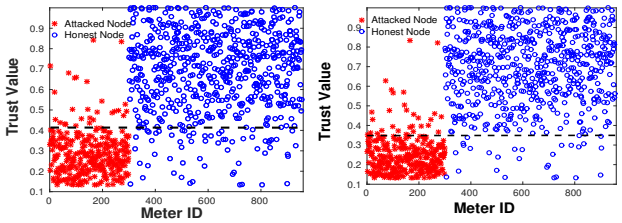
(a) (b)

Figure 14: Texas Dataset: (a) Additive (b) Deductive



(a) (b)

Figure 15: Texas Dataset (a) Camouflage (b) 700 houses



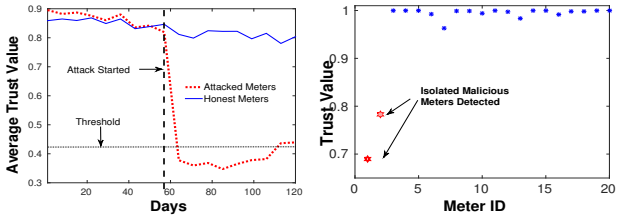
(a) (b)

Figure 16: Irish Dataset (a) $\delta_{avg} = 500$ (b) $\delta_{avg} = 600$

6.4 Compromised Meter Classification

Fig. 14(a) and Fig. 14(b), show the performance in terms of steady state trust values, under additive and deductive falsification when 50% meters are compromised for $\delta_{avg} = 900$ and $\delta_{avg} = 760$ for a small subset of 200 houses from Texas dataset. It is clear that compromised meter's trust values, marked in red (asterisk) are significantly less than non-compromised smart meters marked in blue (circles), such that they are linearly separable through a threshold. The threshold is obtained through a standard K-means unsupervised learning classifier. Similarly, Fig. 15(a) shows the steady state trust distribution under camouflage attacks with $\rho_{mal} = 50\%$, where additive and deductive meters are marked in green and red while non-compromised are marked in blue. Fig. 15(b) confirms the scalability results for 700 houses in Texas dataset under an additive attack. Note that, the false alarm (FA) rate scale well with 3% and 2.8% for 200 and 700 houses given same δ_{avg} as evident from Fig 14(a) and Fig. 15(b).

To prove that our work is valid, scalable and robust across different data sets, we show the performance of steady state trust values using a bigger subset of the Irish dataset with 1000 houses for deductive attacks with lower margins of false data. Figs. 16(a) and 16(b), shows trust value distribution and classification performance for $\delta_{avg} = 500$ and $\delta_{avg} = 600$. We see that even for a large and different region, with lower δ_{avg} than Texas data, the false alarm and missed detection rates are 9% and 8%. A performance accuracy for 5000 houses over all δ_{avg} and ρ_{mal} values and our comparison with existing works is shown later in Sec. 6.9.



(a) (b)

Figure 17: (a) Fast Time to Classification (b) Isolated Attacks

6.5 Time to Detection of Compromised Meters

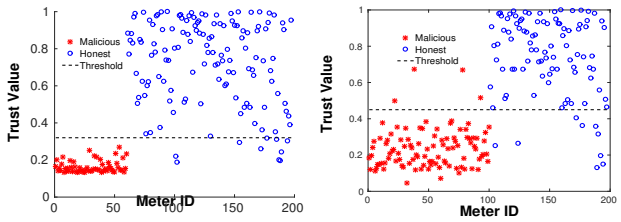
One key advantage of our work is that it allows for quick identification of compromised meters compared to most accurate classification based methods [8, 9]. Fig. 17(a), shows that difference in the evolution of trust values after the attack is launched. Since, it is not legible to show every single meter in one plot, we plot the average trust of the compromised set and the average trust of the honest set over time. Fig. 17(a) shows that the average trust of compromised set of meters falls below the threshold within 10 days from the start of attack. Therefore steady state as described in Figs. 14(a) and 14(b) is achieved within 10 days on average. Hence, rogue meters are identified before drastic damages have been inflicted.

6.6 Classification against Isolated Attacks

Although, our work is focussed on defending against orchestrated attacks, with large number of compromised meters, our work is still valid in identifying isolated malicious meters that may act alone. In such isolated attacks, the ratio drop may not be observably significant unless δ_{avg} is abnormally high. However, this is not necessary as such isolated attacks in such cases would not drastically affect the means and get revealed through the discrete rating criterion and the proposed trust model. As an evidence, the trust of two isolated meters launching additive and deductive attacks with $\delta_{avg} = 600$, (shown in Fig. 17(b)), is far less than other honest meters. Hence, this is a proof of validity for isolated attacks. Hence our defense model is sensitive to small scale low ρ_{mal} attacks.

6.7 Classification against Omission and Incremental Attacks

Fig. 18(a), shows the evidence that the proposed model can detect omission attacks, where 30% of the meters dropped the data. Fig. 18(b), is the performance against incremental attacks after 45 days of the initial attack was launched. The amount of false data was incremented by 2W per meter every 4 hours.

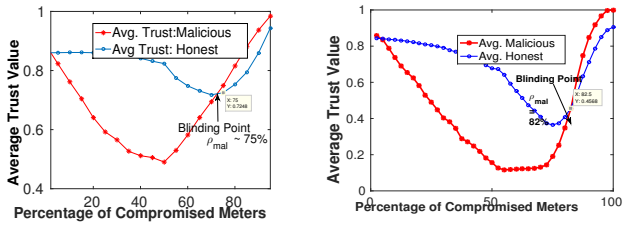


(a) (b)

Figure 18: (a) Omission (b) Incremental

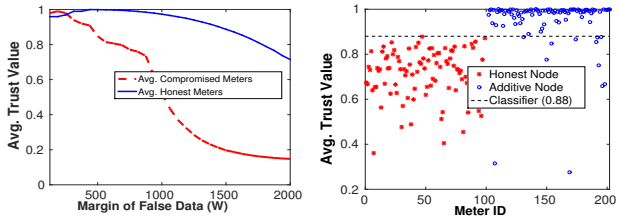
6.8 Avg. Performance over ρ_{mal} and δ_{avg}

One benefit of our work is the robustness to higher fractions of compromised meters compared to most consensus based methods that fail when ρ_{mal} and/or the δ_{avg} is high. Fig. 19(a) and Fig. 19(b), shows the average trust values of all compromised meters versus honest ones. Our technique completely fails at 75% percentage of compromised meters for a 200 house Texas dataset, which is resilient. Similarly, Fig. 20(a), shows the average trust difference among compromised and honest sets of meters over various δ_{avg} under additive attacks, when $\rho_{mal} = 50\%$. Fig. 20(b), shows the performance of our model, when compromised meters alternate true and false behavior periods based on pricing. We can observe, that although the difference between trusts are lesser than earlier case, we can still classify the compromised ones.



(a) (b)

Figure 19: Robustness over ρ_{mal} (a) Additive (b) Deductive



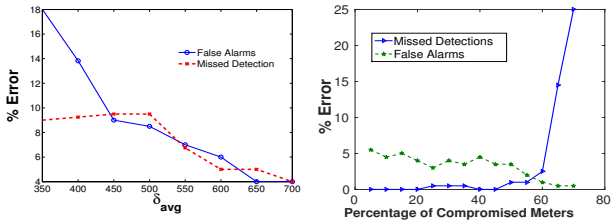
(a) (b)

Figure 20: (a) Avg. Trust vs. δ_{avg} (b) Periodic Attacks

6.9 Performance Accuracy & Comparison with Existing Works

Fig. 21(a), shows our classification error rate over different δ_{avg} for all 5000 houses in Irish dataset with $\rho_{mal} = 40\%$. To understand how it compares with existing works, Table 3 describes performance of other works in terms of various parameters such as False alarms(FA), Missed Detection(MD), learning type (S=supervised, SU=semi-supervised), and time to detection. Apart from this, we also qualitatively compare the level of privacy intrusion and complexity compared to other schemes.

Fig. 21(a) shows that the worst case false alarm (FA) and missed detection (MD) rate for $\delta_{avg} = 350$W is 18% and 9%. At $\delta_{avg} = 400$



(a) (b)

Figure 21: Classification Performance with Scalability (a) *All 5000 houses* over δ_{avg} (b) 200 houses over ρ_{mal}

(used in [9] over the same dataset,did not report ρ_{mal}), our work with $FA = 13.8\%$ and $MD = 9.3\%$ outperforms [9] at a high ρ_{mal} value. At $\delta_{avg} = 600$W, we have $FA = 6\%$ and $MD = 5\%$ outperforming [1] that reported FA=9% and MD=8% but over a much smaller set of 200 houses without scalability evidence and higher δ_{avg} of 700W-800W that favors good performance. False alarms increase with decreasing δ_{avg} and below 400 it increases more, because the standard deviation of the dataset usually range around $400W$. To our best knowledge, only [19] using a synthetic data reported lesser error rates than us. But this happens only if number of users are less than, 25 making this comparison unfair and therefore not scalable. Apart from this, [11] reports detection of around 90%, but for δ_{avg} as high as $1200W$ to $1800W$, which will facilitate easy detection in our case anyway.

Fig. 21(b), shows our classification error rate over different ρ_{mal} values for a smaller dataset of 200 houses in Texas. It can be seen that missed detection and false alarm rates are less than 2% and 5%, upto $\rho_{mal} = 0.60$. Above $\rho_{mal} > 65\%$, the performance degrades, but missed detection rates are still lower at 13% and 25%, where other works fail completely. A smaller population is chosen since realistic attack budgets may become a significant ρ_{mal} percentage for the smaller micro-grids sizes, making such a study practical.

Table 3: Comparison with Existing Work

Parameter	Proposed	CPBETD [9]	ARMA [14]	Entropy [1]
FA	13%	29%	33%	11%
MD	9%	24%	28%	8%
δ_{avg}	400W	400W	NA	800W
ρ_{mal}	$\geq$ 40%+	NA	NA	$\leq$ 40%
Size	5000	5000	200	200
Learning	SU	S	S	S
Detection Time	$\leq$ 10 days	1 yr	1 mo	1 mo

7 CONCLUSION AND FUTURE WORK

We conclude that Harmonic to Arithmetic Mean ratios is an effective light weight indicator of organized falsification over different types of falsification attack and robust under higher fractions of compromised meters while distinguishing legitimate changes in the data to malicious ones and helps in consensus correction. A semi-supervised folded gaussian trust model produces trust values, which identifies meters launching both organized or isolated attacks within a few days of attack, while preserving lower missed detections and false alarms rates, even when percentage of compromised meters are significantly higher. We showed that the method is generic and applicable across different real smart meter datasets. In future, we will extend our work to attacks where $\delta_{avg} < 350$ which may be realistic for advanced, persistent and long term adversaries who sacrifice immediate benefit for long term gains. We will discuss the theoretical details of the Pythagorean means for very low $\delta_{avg} < 350$ in the future work.

ACKNOWLEDGMENTS

This research is supported by the NSF grants CNS-1545037, CNS-1545050. Sajal K. Das is also a distinguished visiting professor at Zhejiang Gongshang University, Hangzhou, China.

REFERENCES

[1] S. Bhattacharjee, A. Thakur, S. Silvestri, S.K. Das, "Statistical Security Incident Forensics against Data Falsification in Smart Grid Advanced Metering Infrastructure", *ACM CODASPY*, pp. 35-45, 2017.
[2] P. Box, D. Cox, "An analysis of transformations", *Journal of the Royal Statistical Society*, Series B. Vol. 26 (2): pp. 211-252, 1964.
[3] 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.
[4] Y. Chae, L.C. DiPippo, Y. L. Sun, "Trust Management for Defending On-Off Attacks", *IEEE Trans. on Parallel and Distributed Sys.*, Vol. 26, No. 4, pp. 1178-1191, Apr. 2015.
[5] A.A. Hafez, Yue Xu, A. Josang, "A Normal Distribution based rating aggregation method for generating product reputations", *Web Intelligence*, Vol. 13(1), pp. 43-51, 2015.
[6] S.-C. Huang, Y.-L. Lo, and C.-N. Lu, "Non-technical loss detection using state estimation and analysis of variance", *IEEE Trans. on Power Systems*, 28(3), pp. 2959-2966, Aug. 2013.
[7] 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*, 19(2), pp. 105-120, April 2014.
[8] A, Jindal, A. Dua, K. Kaur, M. Singh, N. Kumar, S. Mishra, "Decision Tree and SVM-Based Data Analytics for Theft Detection in Smart Grid" *IEEE Trans. on Industrial Informatics*, Vol. 12(3), pp. 1005-1016, June 2016.
[9] P. Jokar, N. Arianpoo, V. Leung, "Electricity Theft Detection in AMI Using Customers' Consumption Patterns", *IEEE Trans. on Smart Grid*, Vol. 7(1), pp. 216-226, 2016.
[10] T. Koppel, "Lights Out: A Cyberattack, A Nation Unprepared, Surviving the Aftermath", *Crown Publishers, New York*, 2015.
[11] 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 (DSN)*, pp. 407-418, 2016.
[12] C.-H. Lo and N. Ansari, "CONSUMER: A novel hybrid intrusion detection system for distribution networks in smart grid", *IEEE Trans. on Emerging Topics in Computing*, 1(1):33-44, 2013.
[13] S. McLaughlin, D. Podkuiko, P. McDaniel, "Energy theft in the advanced metering infrastructure", *Critical information infrastructures security (CRITIS'09)*, Springer, pp. 176-187, 2009.
[14] D. Mashima, A. Alvaro, "Evaluating Electricity Theft Detectors in Smart Grid Networks", *Springer Heidelberg*, pp. 210-229, 2012.
[15] B. Meyer, "Some Inequalities for Elementary Mean Values", *AMS Mathematics of Computation*, Vol. 42, No. 165, pp. 193-194, 1984.
[16] R. Mohassel, A. Fung, F. Mohammadi, K. Raahemifar, "A survey on Advanced Metering Infrastructure", *Journal of Electrical Power & Energy Systems*, Vol. 63, pp. 473-484, Dec. 2014.
[17] A. Rad, 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.
[18] C.R. Rao, X. Shi, Y. Wu, "Approximation of the expected value of the harmonic mean and some applications", *National Academy of Sciences*, Vol. 111(44), pp. 15681-15686, 2014.
[19] S. Salinas, M. Li, and P. Li, "Privacy-preserving energy theft detection in smart grids: A P2P computing approach," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 9, pp. 257-267, Sep. 2013.
[20] R. Sevlian and R. Rajagopal, "Value of aggregation in smart grids", *IEEE SmartGridComm*, pp. 714-719, Oct. 2013.
[21] 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.
[22] E. Werley, S. Angelos, O. Saavedra, O. Cortes, 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, 2011.
[23] W. Yu, D. Griffith, L. Ge, S. Bhattarai, N. Golmie, "An integrated detection system against false data injection attacks in the Smart Grid, *Sec. and Commun. Networks*, Vol. 8(2), pp. 91-109, 2015.
[24] D. Urbina, J. Giraldo, A. Cardenas, N. Tippenhauer, J. Valente, M. Faisal, J. Ruths, R. Candell, Henrik Sandberg, "'Limiting the Impact of Stealthy Attacks on Industrial Control Systems", *ACM CCS*, pp. 1092-1105, 2016.
[25] www.smartgrid.gov/project/pecan_street_project_inc_energy_internet_demonstration.html
[26] Irish Social Science Data Archives, Available at: http://www.ucd.ie/issda/data/
[27] http://energy.gov/sites/prod/files/oeprod/DocumentsandMedia/14-AMI_System_Security_Requirements_updated.pdf
[28] https://skyvisionsolutions.files.wordpress.com/2014/08/utility-smart-meters-invade-privacy-22-aug-2014.pdf
[29] https://krebsonsecurity.com/2012/04/fbi-smart-meter-hacks-likely-to-spread/
[30] https://www.maximintegrated.com/content/dam/files/design/technical-documents/white-papers/smart-grid-security-recent-history-demonstrates.pdf
[31] https://www.smartgrid.gov/files/The_Smart_Grid_Promise_DemandSide_Management_201003.pdf
[32] W. Xia, Y. Chu, "The schur convexity of gini mean values in the sense of harmonic mean", *Mathematica Scientia*, Vol. 31(3), pp. 1103-1112, 2011.

Appendix

A NECESSARY AND SUFFICIENT CONDITIONS FOR RATIO DROP AND RISE

In this section, we provide a detailed explanation of the necessary and sufficient conditions for δ_{avg} in terms of $\rho_{mal} = M/N$, the mean μ and standard deviation σ of the power consumption dataset. Each occurrence of ratio rise or drop given an attack type, has an upper bound and lower bound on the value of δ that is dependent on the position of bias as well. For example, $k^-(rlow)$, denotes the lower bound for a deductive attack on the datapoints on the right side of the actual mean, and so on. The average case approximate lower bounds are: $k^-(rlow) = k^+(llow) =$

$$k_{low} = \frac{\sigma}{M} + \frac{\sigma}{\sqrt{M}}\sqrt{\frac{N-M}{N-1}} + \sigma \tag{10}$$

where + and − denote additive and deductive attacks and l and r denote whether the position of biased data-points are on the left or right side of the true mean. The average case approximate upper bounds are: $k^+(lhigh) = k^-(rhigh) =$

$$k_{high} = max(\sigma^2, \frac{2\sigma}{M} + \frac{\sigma}{\sqrt{M}}\sqrt{\frac{N-M}{N-1}} + 2\sigma) \tag{11}$$

The average conditions for deductive attacks on data-points the left and additive attacks on datapoints on right side of the true mean is:

$$k^-(lhigh) = k^+(rhigh) = \sigma\sqrt{\frac{N}{N-1}} \tag{12}$$

The worst case happens when additive attack changes only the minimum value (say x_i) of the dataseries, and deductive attack changes only the maximum value of the dataseries. The worst case expressions have little practical significance, but could be for purpose of verification. The worst case expressions for k_{low} and k_{high} are:

$k_{low} > \{(|x_i - \mu| + \sigma\sqrt{\frac{N}{N-1}}) + \mu)\} - x_i$, and

$k_{high} > |x_i - \mu| + \sigma^2$

B DAILY STANDARD DEVIATION DISTRIBUTION

The Fig. 22, shows the probability bar plot for $SD_{avg}(T)$ for the Texas dataset. We see that in most cases under no attacks, the probability of $SD_{avg}(T)$, being between 425W-475W, centered around 450W is 0.52. Probabilities of all other ranges are much lesser. Hence, the mode of the distribution is a reasonable approximation for σ_{MR}, under attacks being confirmed.

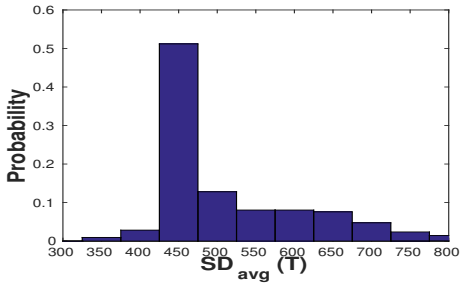


Figure 22: Historical $SD_{avg}(T)$:σ_{MR} approximation

C STEALTHY PERSISTENT ATTACKS

Fig. 23, shows an example that our ratio metric may work even under ultra-stealthy margins of false data. Here $\rho_{mal} = 40\%$ and δ_{avg} is as low as $50W$, where our attack was implemented from the 41st day. We estimated that this ρ_{mal}, δ_{avg} will cause deviation less than the ϵ with all the knowledge we possessed. Indeed, this works well for most of the time, but is not guaranteed to escape detection altogether, since all the meters whether compromised or not have erratically changing data that is difficult to predict beforehand. In future work, we will propose a mathematical detection criterion for such stealthy attacks.

Targeted one sided attacks, may happen theoretically, where attack only attacks datapoints greater than the mean, then the observations will reversed. Deductive attacks may show an increase in ratio, but in this case it will be restricted to only attacking 36% of the total datapoints which restricts the attack significantly.

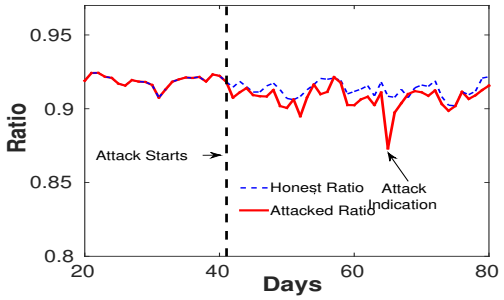


Figure 23: Deductive Attack with Full Knowledge

D COST BENEFIT UNDER MISSED DETECTION

The revenue of an adversary per day who has M undetected meters is given by: $RR = \frac{\delta_{avg} \times M \times \eta \times E}{1000}$, where η is the number of reports a day, and $E = \$0.12$ is the average per unit (KW-Hour) cost of electricity in USA. In [29, 30] mentioned that cost of compromising the smart meter is about \$500, in the puerto rico attack. Here, utility maintainance personnel asked for \$300 − \$700 from different customers and hacked their meters that reported lesser power consumption, and promised benefit over time. The optimal laser probes used for those attacks vary around 400. Therefore, for compromising 80 meters, the attack cost is about \$40,000. At $\delta_{avg} = 300$, we have missed detection of 42 meters, hence the revenue for attacker is \$36/day. At this rate, it would take about 3 years to recover the attack cost \$40,000. For $\delta_{avg} = 350$, only 7 meters remain undetected by our method, with an average revenue of \$6/day. At this rate, it will take 18 years to recover the attack cost. In future, we will explore how to identify meters reliably which have $\delta_{avg} < 300W$. That study will be useful if the attacker devises/offers novel cheaper ways of attack. Also, note that in terms of time to detection also our work is quicker (less than 10 days) compared to existing works. Hence, an attacker who does not intend to remain undetected, is not able to gain much attack revenue.