Detection and Forensics against Stealthy Data Falsification in Smart Metering Infrastructure

Shameek Bhattacharjee* and Sajal K. Das†

* Dept. of Computer Science, Western Michigan University, Kalamazoo, MI, USA
† Dept. of Computer Science, Missouri University of Science and Technology, Rolla, MO, USA

Emails: shameek.bhattacharjee@wmich.edu; sdas@mst.edu

Abstract—False power consumption data injected from compromised smart meters in Advanced Metering Infrastructure (AMI) of smart grids is a threat that negatively affects both customers and utilities. In particular, organized and stealthy adversaries can launch various types of data falsification attacks from multiple meters using smart or persistent strategies. In this paper, we propose a real time, two tier attack detection scheme to detect orchestrated data falsification under a sophisticated threat model in decentralized micro-grids. The first detection tier monitors whether the Harmonic to Arithmetic Mean Ratio of aggregated daily power consumption data is outside a normal range known as safe margin. To confirm whether discrepancies in the first detection tier is indeed an attack, the second detection tier monitors the sum of the residuals (difference) between the proposed ratio metric and the safe margin over a frame of multiple days. If the sum of residuals is beyond a standard limit range, the presence of a data falsification attack is confirmed. Both the ‘safe margins’ and the ‘standard limits’ are designed through a ‘system identification phase’, where the signature of proposed metrics under normal conditions are studied using real AMI micro-grid data sets from two different countries over multiple years. Subsequently, we show how the proposed metrics trigger unique signatures under various attacks which aids in attack reconstruction and also limit the impact of persistent attacks. Unlike metrics such as CUSUM or EWMA, the stability of the proposed metrics under normal conditions allows successful real time detection of various stealthy attacks with ultra-low false alarms.

Index Terms—Statistical Anomaly Detection, Cyber-Physical Security, Smart Grid, False Data Injection, Data Falsification.

I. INTRODUCTION

Advanced Metering Infrastructure (AMI) is a key component of the smart grid technology, that measures data on loads and electricity (power) consumption of customers [19]. Such data are measured by smart meters installed at the customer site. The AMI data is expected to play a decisive role in the demand forecast, automated demand response, load adjustment, and management of daily and critical peak shifts [33]. Therefore, the integrity of AMI data is indispensable.

Several real incidents of isolated and organized data falsification and their losses to the utilities have been reported in [31], [32]. Existing research on defense against the falsification of power consumption data has been mostly restricted to electricity theft, [8], [12], [15], [27] where individual meters report lower than actual usage. Since smart meters belonging to rogue customers reduce the actually measured reading of power consumption, such an adversarial strategy is termed as a deductive mode of data falsification. However, recent works in this area [2], [20], and recent real case studies [28], have recognized the possibility of additive and camouflage modes of data falsification as well. In additive attacks, higher than actual power consumption can be sensed by a meter as a byproduct of static and dynamic load altering attack [20] or hardware tampering, affecting both customers and utilities. For camouflage attacks, the total margin of deductive attacks is balanced by the additive attacks, such that one set of customers benefit with lesser bills at other’s expense, while ensuring that the mean aggregate power consumption from a microgrid remains practically unchanged. Therefore, a defense is required against all three modes of attack.

Additionally, due to the economic and civilian impacts, the AMI could be a target of powerful stealthy adversaries, such as rival nations [9], utility insiders [32], organized crime [29] and business competitors [9], [35], who possess the ability to compromise several smart meters (as in [32]), alter smaller amounts of data per meter, to avoid easy detection (from proximity, consensus, or classification based detectors) while significantly impacting operations in the long-term [9], [12], [32]. Furthermore, stealthy or persistent adversaries may possess partial knowledge of usual security mechanisms or complete knowledge of the actual defense mechanism, thus enabling them to employ stealthier falsification strategies. Due to the aforementioned competence, the realistic strategy space for data falsification is much larger than what has been assumed by existing research in this field. Nonetheless, apart from cyber attacks, false data from smart meters can be easily launched through physical attacks (by physical/wireless tampering of hardware; e.g. optical ports, rewiring etc.,) on the smart meters (as committed in 2012 Puerto Rico attack [29], [32]) or through optical probe toolkit shown by analysts at InGuardians Inc. [29]. Hence, cryptography [26] or network based intrusion detection is not enough to counter this problem.

Contributions of our work: In this paper, first we discuss some possible stealthy strategies to launch additive, deductive and camouflage modes of data falsification. To detect the presence of organized data falsification attacks, we propose a two-tier, light weight, real time, statistical anomaly detection scheme that detects the presence and mode of various attacks. The first tier uses a Harmonic Mean to Arithmetic Mean Ratio metric of the aggregate power consumption data, to identify discrepancies in the time series behavior. The second tier uses the Sum of Residual under the Ratio Curve metric to confirm
whether the discrepancies in the first tier is indeed an attack or not. The second tier is particularly relevant when the effective margin of false data introduced by the adversary is very low and higher detection sensitivity needs to be achieved without degrading false alarm performance.

Subsequently, through a system identification phase, we first establish the ‘normal behavior’ of the two proposed metrics under no attacks. The normal limits of Harmonic to Arithmetic mean ratio metric is termed as a ‘Baseline Safe Margin’ and the limits of Sum of Residuals under the Ratio Curve is termed as ‘Baseline Standard Limit’. These metrics and their normal limits are carefully designed from real datasets, such that their values when observed under attacks, deviate from their normal limits with a high sensitivity. Then, we establish theoretical properties of the proposed metrics that trigger unique signatures under each mode of attack and type of adversarial strategy. Finally, we propose an attack reconstruction scheme using observed changes in the direction, sign and magnitude of the proposed metrics to associate the signatures with different attack types that could guide site security officers or demand control mechanisms for a suitable response. Thereafter, to mimic persistent adversaries with complete knowledge about our defense mechanism, we design attack strategies that will just ensure evasion. Against such persistent adversaries, we quantify our performance by the extent to which we limit the attacker’s impact of attack (e.g. revenue accrued per unit time), while preserving ultra-low false alarm rates while accounting for base rate fallacy.

We validated our approach through experiments on real data sets acquired from an actual AMI infrastructure from Texas, USA (800 houses for 3 years) and Dublin, Ireland (5000 houses for 535 days). Most of the results correspond to the Texas data due to longer duration, while Irish dataset is used to prove the generality, scalability and sensitivity trade-offs. Results show that our model is able to detect, decipher, and confirm various attacks launched by stealthy and persistent adversaries in real time across different datasets. We show that our model is particularly robust to different fractions of compromised meters and very low margins of false data which is typically a problem with existing mechanisms.

Both the proposed detection metrics are privacy non-intrusive by design, since they do not require profiling of fine grained customer specific data. Both metrics are more stable invariants of power consumption across multiple AMI datasets, which achieves high detection sensitivity and ultra-low false alarms simultaneously. Existing works in this area, do not implement novel falsification strategies, do not assume stealthier margins of false data (≤ 400W), do not assume wide variation in the fractions of compromised meters (5%-70%), do not assume persistent adversaries that have full knowledge of defense mechanism, or do not emphasize on real time detection. We quantified detection limits and impact resilience to justify the improvement.

The paper is organized as: Section II discusses limitation of related work, Section III discusses the threat model, Section IV describes the dataset and the system, Sections V and VI discusses the proposed method and its theoretical analysis respectively, followed by experimental results in Section VII.

II. SOME LIMITATIONS OF PREVIOUS WORKS

Existing research on data falsification from smart metering infrastructure can be roughly classified into Classification based detection, State Estimator based detection, and Consensus based Detection, which are confined to the study of electricity theft (deductive falsification).

Classification based approaches [7], [8] include multi-class Support Vector Machines (SVM), Neural Networks, Radial Basis Function based models. Classification based schemes are computationally intensive, require full and fine grained profiling of each smart meter, not scalable for real AMI sizes, and do not provide real time detection. A comparative study of all classification based schemes in [6] concludes that while these schemes require full and continuous profiling of every customers’ energy consumption, the detection rates of most of these schemes are approximately 60%-70%. Moreover, only two schemes provide a quantitative false alarm rate.

State based detection [5], [10], [13], need extra hardware deployed at different places in the AMI and the distribution grid for sanity checking [6]. Extra hardware required is costly to the extent that it “has been recognized as a practical deterrent for utility providers to use such solutions in scale as reported by [30]”. Some research proposes checking the non-technical losses (NTL) at transformer meters. However, [6] observes that NTL values vary due to a large number of factors other than attacks and hence also suffers from high false alarm rates. Also, NTL approach fails to detect camouflage attacks or additive attacks induced through load altering [2].

Consensus (Aggregate) based detection include mean aggregate outlier inspections, non-parametric measures such as Exponential Weighted Moving (EWMA) and CUSUM Control Chart (CUSUM) of the aggregates, and parametric measures such as Auto Regressive Moving Average (ARMA) models, to detect false data injection. The comparisons could be either instantaneous or historical consensus. If the difference between predicted consensus and the observed consensus (often called as residual) is above a threshold, then attack is inferred. Works such as [15], [25], [27], use mean or median power consumptions or compare proximity of meter readings to the mean and standard deviation of power consumption. ARMA based models [15] analyze each customer meter’s time series data separately to increase accuracy, using ARMA-GLR detector. However, in many practical cases, the consumption cannot be accurately modeled as an ARMA process as mentioned in [8], [15], resulting in success rates of 62% only. Consensus or Aggregate based approaches usually do not require separate profiling of every meter, or additional hardware, hence much less costly and feasible than the other two approaches.

Weaknesses of Consensus Methods under Stealthy Attacks:

Many aggregate or consensus based mechanisms, use non parametric statistics like EWMA and CUSUM [13], [15] and parametric statistics like ARMA [15] to ‘smoothen’ the mean power consumptions to get a stable estimated trend, which is compared with observed sample consensus or per meter mean sample measurements for attack detection. However, we observed some general difficulties with these approaches. Figure 1, shows arithmetic mean of the power...
consumption from our real data set [34] for the same time period for three different years. Note, that the mean power consumptions readily vary over time within the same year. Additionally, there is a large difference between the means on the same days in successive years. Such large variations in mean consumptions create two major roadblocks. First, false alarms or missed detections are more (depending upon the chosen weights of different statistics), due to the inherently unstable nature of the mean consumption. Secondly, the normal ‘residual’ difference between actually observed sample and the ‘smoothened’ mean consumption is large (above 130W from our studies), giving ample opportunities for the adversary to induce changes in such metrics that stay within the normal residual difference, while still accruing substantial attack impact. As an example, for a micro-grid of 200 meters, suppose 40% of them have been compromised. Then an average margin of false data of 325W from each compromised meter, introduces an average error residual of 130W in the mean. The calculation is shown in Appendix B. In other words, it is difficult to identify such an attack given the large legitimate variation of metrics derived from mean/median measures. However, at the same time the monetary (loss) impact of this attack is $RR = 75 \text{ dollars/day}$.

![Fig. 1. Instability of mean power consumption](image)

Furthermore, approaches such as [2], [8], [25], [27], that compare whether individual meter’s data is within the standard deviation fail, since an average falsification margin of 325W is much lesser than the typical standard deviation of the power consumption datasets (≥ 400W [34], [36]). Thus in most existing works, assumed margins of false data ranges between 400W upto 1500W and detection rate sharply degrades to zero for margins less than 400W. Powerful and stealthy adversaries having a higher initial attack budget can compromise a larger number of meters and/or inject lower margins of false data (below the standard deviation) per meter, to remain undetected while gaining incremental benefits in the long term. Hence, approaches based on proximity of data within estimated instantaneous or historical mean and/or standard deviation of aggregates fail under lower margins of false data. Furthermore, reports on 2012 Puerto Rico attack [29], [32], indicated that it is possible to compromise a large fraction of smart meters. Finally, due to the instantaneous smoothening of the consensus measures, the danger of the instantaneous residual difference being within the normal residual limit is increased when attackers increase their falsification margin incrementally over each time slot. For all the above reasons, solutions from traditional sensor networks or control systems cannot be borrowed. Given these limitations, there is a need for a real time, light weight, anomaly detection which is focused on detecting orchestrated attacks from multiple number of meters particularly with lower margins of false data. This motivates us to propose a novel scheme that provides security forensics to detect various falsification attacks in AMI with ultra low false alarms. Note that our approach while focusing on low margins of false data, also works for higher margins.

III. THREAT MODEL

False power consumption data could be achieved by manipulation of (i) inputs into the meters, (ii) data at rest in the meter, or (iii) in-flight from the meter. If an adversary is able to capture the data collectors/hops over the mesh network that connects AMI with the utility, it may be easier to intercept larger number of unique meter’s data without physically compromising many meters. Scope: We assume an orchestrated attack where an organized and powerful adversary launches data falsification from several compromised smart meters concurrently. The emphasis is especially on adversaries that are more inclined to keep the margin of false data smaller while having a higher number of compromised nodes $M$ to affect a long-term damage while avoiding easy detection. Since AMI is a new research area, large scale real world malicious samples are hard to find. Therefore, we generated attack samples over our real dataset, assuming an unbounded strategy space that does not favor our defense mechanism and shown the detection limits. Proposed method does not intend to detect isolated uncoordinated attacks or detect isolated rogue customers, since such a problem has been investigated by numerous previous works.

A. AMI Data Falsification Attack Modes

Let $P_i^t(\text{act})$, be the actual/authentic power consumption measurement from a meter $i$ at time slot $t$ that is unknown to the utility, while $P_i^t$ is the reported power consumption which the utility receives. Under no falsification, the reported consumption data $P_i^t = P_i^t(\text{act})$ is unbiased, while under attacks $P_i^t$ is biased by the following falsification modes:

**Deductive:** The adversary reports $P_i^t = P_i^t(\text{act}) - \delta_i$, where $\delta_{\text{min}} \leq \delta_i \leq \delta_{\text{max}}$ is a false bias, amounting to electricity theft.

**Additive:** The adversary reports $P_i^t = P_i^t(\text{act}) + \delta_i$. An additive attack could be launched by a competing utility on its rival company’s meters, inducing loss of business confidence by the customers of the victim utility, due to higher than actual bills. Furthermore, if a victim company participates in automated demand response, load may be changed by a Dynamic Load Altering attack [20], causing multiple smart meters to sense/record higher than actual consumption concurrently. The victim company ends up paying undue incentives for resultant peak shifts. A recent report in [31], showed utilities facing lawsuits by customers accusing unfairly high bills.

**Camouflage:** The adversary divides the compromised meters into two teams equal in number, which simultaneously adopt an additive and deductive modes of attack, respectively. This mode favors one set customers at the expense of others. It cannot be detected by simple mean based comparison approaches, because no suspicion is raised due to negligible change in the aggregate mean consumption.
B. Margin of False Data and Fraction of Compromised Meters

All $\delta_t$ are generated randomly within an interval $[\delta_{\min}, \delta_{\max}]$. Let $\delta_{avg}$ be the average value of $\delta_t$ for each compromised meter, termed as margin of false data, whose magnitude depends upon the time horizon of the intended attack and damage. Hence, $\delta_{avg}$ is part of the adversarial strategy. All units of $\delta_{avg}$ values are in Watts and lower values are more stealthy. For a smarter attack, the distribution of $\delta_t \in [\delta_{\min}, \delta_{\max}]$ should be a variant of some uniform distribution such that the resultant change in shape of the distribution of AMI consumption data is not very obvious [8]. Given the unimodal nature of power consumption in a residential AMI network [2], [8], any intelligent attacker would refrain from a strategy that would make the resultant distribution multi-modal or in general alter its shape. As an illustration, a comparison between normally distributed and uniformly distributed $\delta_t$ with the same $\delta_{avg}$, and its effect on the resultant shape of the distribution is depicted in the Appendix C. Note that our detection scheme works under both cases. However, our results consider variants of uniformly distributed case, being the smarter strategy.

We assume that organized adversaries compromise a certain number $M$ out of $N$ smart meters, depending on a total attack cost budget represented by $TC$. The fraction of compromised nodes is $\rho_{mal} = \frac{M}{N}$. Given a microgrid of fixed $N$ meters, we use different values of $M$ to study sensitivity to various attack budget. Separately, for an adversary with a given $M$, we vary $N$ to study scalability of the proposed solution.

C. Revenue (Impact) of an Undetected Attacker

The attacker’s revenue per day denoted by $RR$, defined in terms monetary worth of electricity power (in dollars), quantifies its benefit from an attack. Revenue $RR$ per day in dollars is defined as:

$$RR = \frac{\delta_{avg} \times M \times \eta \times E}{1000}$$  \hspace{1cm} (1)

where $\eta$ is the number of reports a day, and $E = $0.12 is the average per unit (KW-Hour) cost of electricity in USA. Additionally, we define breakeven time $T_{BE}$, as the time required for the total cumulative revenue accrued from attacks to match the initial total attack cost $TC$.

D. Stealthier Data Falsification Strategies

Unlike the existing works that assume a random noise as falsification strategy, we assume some stealthy strategies that may be inspired by partial knowledge of usual security mechanisms or complete knowledge about the actual defense mechanism.

Data Order Aware Attack: is motivated to minimize chances of detection against mechanisms utilizing proximity (distance) of reported data from meters with historical data or consensus. In Fig. 2, the green line corresponds to the actual power consumption from the compromised meters. The black and red lines correspond to falsified consumption data following a non-data order aware and a data order aware strategy under same $\delta_{avg}$ and $\rho_{mal}$. 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.

![Image](https://via.placeholder.com/150)

**Fig. 2. 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 $P_t^{(1)}(act) \leq \ldots \leq P_t^{(M)}(act)$. Then adversary generates $M$ random numbers for $\delta_t$, sorted as $\delta_{\min} \leq \ldots \leq \delta_{\max}$. For an additive attack, the lowest observed power consumption data is changed with the highest $\delta_{\max}$, while highest observed power consumption data is modified with lowest $\delta_{\min}$, and so on, such that $P_t^{(1)}(act) + \delta_{\max}, \ldots, P_t^{(M)}(act) + \delta_{\min}$. For a deductive attack, the highest observed power consumption data is changed with the highest $\delta_{\max}$, while the lowest observed power consumption data is changed with the lowest $\delta_{\min}$. Hence, $P_t^{(1)}(act) - \delta_{\min}, \ldots, P_t^{(M)}(act) - \delta_{\max}$. For a camouflage attack, the sorted $P_t^{(1)}(act), \ldots, P_t^{(M)}(act)$ 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 distribution.

Incremental Evolving Attack: Instead of immediately falsifying data with the intended $\delta_{avg}$, adversary incrementally increases margin of false data by $\Delta \delta$ watts on every time slot, leading up to the intended $\delta_{avg}$. A gradual change on every time slot causes the approaches based on instantaneous moving averages to fail, since the average readsjusts itself within the detection threshold continuously.

Persistent Attacks: These are strategy parameter pairs of ($\rho_{mal}, \delta_{avg}$) for which the adversary would persist in the system undetected for our proposed defense mechanism. When the adversary knows of the exact defense mechanism and the detection thresholds used, it can estimate ($\rho_{mal}, \delta_{avg}$) pairs that will just ensure evasion. Axelsson et.al. [1], in his seminal paper on intrusion detection termed this as ‘security of the detector’, as an evaluation property that quantifies the risk, should the detection mechanism be leaked. Since the adversary never gets detected in such cases, the security of our proposed approach is evaluated by quantifying the extent to which we limit the impact of an undetected attack or reduce the undetectable strategy space versus relative frequency of false alarms. We define breakdow point, as the $\delta_{avg}$ and $\rho_{mal}$ pair for which the proposed defense model just fails to detect an attack and examine the $RR$ and $T_{BE}$ at these breakdown points.
IV. DATA SET AND SYSTEM MODEL DESCRIPTION

We utilize real AMI datasets from two sources. The first dataset is collected from Pecan Street Project [34] consisting of 800 houses for a period of 3 years (2014, 2015, 2016) in Texas, USA. The second dataset is taken from Ireland Social Sciences Data Archives [36] consisting data from 5000 houses for a period of 535 days (around 2 years) (2009-2010) from 6 regions around Dublin, Ireland. All data belong to residential customers. Since the Irish dataset was for a limited time, a longitudinal study on it was not feasible. Hence, most of the learning methods in this paper utilizes the first dataset. In contrast, the second dataset is used to test the generalizable nature, sensitivity and scalability of our proposed detection mechanism to other AMI datasets that have a completely different population size, climate, and habits. Unless stated explicitly, all results by default correspond to the Texas data.

Data Distributions: Let $P_i$ denote the steady state power consumption distribution of any meter $i$. Investigations on real hourly power consumption data sets [34], from different microgrids, shows that each $P_i$ follows an approximate log normal distribution. Figs. 3(a) and 3(b), show the example of data distributions for a 200 meter population from Texas. It is observed that such log normal distributions are closely clustered to each other such that the variation among them is not arbitrarily large, as seen in Fig. 3(a). Given this observation, it can be argued that the combination of the individual log normals can be well approximated by a mixture distribution that is also log-normal [2], as seen in Fig. 3(b), which shows the trend for the mixture distribution ($P_{mix}$) of the power consumption of all houses. The shape indicates an approximate lognormal distribution.

To address concerns on the generality of these observations, we repeat the above for the Irish dataset. Fig. 4(a) refer to power consumption distribution of individual houses in the Ireland dataset, whose mixture distribution counterpart is shown in Fig. 4(b). The results clearly indicate that aggregate power consumption across multiple data sets from different regions follow a particular trend that generalizes. Hence, the validity of our experimental work is preserved regardless the dataset. Our findings are also corroborated by another work [21], with similar results obtained from a different data set. Hence, it can be safely concluded that the observed trend is generic rather than being specific to one data set.

Box-Cox Transformation: We converted $P_{mix}$ into an approximate Gaussian distribution using a box cox transformation technique. Box Cox transformation techniques are used for approximating non-normal data into a normal distribution. Hence, all power consumption data is transformed onto a natural logarithm scale to obtain an approximate Gaussian distributed random variable denoted as $P_{mix}$. Therefore, we denote $P_i = ln(P_{i}^2 + 2)$ as the effective power consumption data from each $i$ on an $ln$ scale at any time slot $t$ that is used by the detection scheme. All $t$’s are slotted hourly in our dataset which is usual practice as reported in several works [8], [15]. Additional results of this approximate normal mixture model for different months in the recent past and the QQ plot (for normality test) is depicted in the figures in Appendix A.

The box-cox transformation serves a dual purpose. The transformation to a lower dimensional real axis, increases relative sensitivity to the change in Harmonic Mean to Arithmetic mean ratios because some interesting statistical properties are more prominent in a lower dimensional real axis. Secondly, Gaussian approximation helps in the approximation of some mathematical bounds more tractable. For Texas and Irish data, 67.7% and 66.9% of the samples are within the first standard deviation, meaning the separation among data points are reduced. However, the transformed datasets remain unbalanced around the mean as 64% and 69% of the samples are on the left (i.e., lesser than) of the mean and 36% and 31% of the samples are the right (i.e., greater than) of the mean, for Texas and Irish datasets respectively. Due to this inherent asymmetry, more data points (even after transformation) are on the left side (i.e., less than) of the mean. This is a factor that triggers unique signatures under attacks (see Section VI-B).

Deployment Issues of proposed Detector: The exact topology of the AMI was not revealed by any of the projects from where the data was collected. Hence, it was infeasible to directly study implications of topological aspects over the proposed detection mechanism. Given an AMI system (see Fig 5), any proposed intrusion detection module can be implemented in a decentralized or centralized manner in an AMI network as reported in [3], and micro-grid size affects the sensitivity of the attack detection. In a decentralized implementation, a detection module is usually deployed in either FAN (field area network...
A. System Identification of Invariants

In statistical anomaly detection, identifying the normal behavior of the detection metric under no attacks is termed as system identification. A more stable detection metric (i.e. invariant) under ‘normal’ conditions is often more accurate.

We denote \( p_t = [p^1_t, \cdots , p^N_t] \), as the power consumption data on the \( ln \) scale obtained from \( N \) meters at any time slot \( t \). Let \( p \) denote the steady state mixture random variable, such that \( p_t \) is the realization of \( p \) at time \( t \). The harmonic mean and arithmetic mean of \( p_t \) on a particular time slot \( t \) is denoted by \( HM_t \) and \( AM_t \) and are defined as:

\[
HM_t = \frac{N}{\sum_{i=1}^{N} \frac{1}{p^i_t}} \quad AM_t = \frac{\sum_{i=1}^{N} p^i_t}{N} \tag{2}
\]

1) Harmonic to Arithmetic Mean Ratios: First, all \( HM_t \) and \( AM_t \) are calculated for each time slot \( t \), over a time window indexed by \( T \). Each \( T \) composed of 24 time slots, represent each day of a year, such that \( T \in \{1, \cdots, 365\} \). At the end of each window \( T \) (one day), the average Harmonic mean to Arithmetic mean ratio denoted by \( Q^r(T) \) is given by:

\[
Q^r(T) = \frac{\sum_{t=1}^{24} HM_t(T)}{\sum_{t=1}^{24} AM_t(T)} \quad \forall \ T \in \{1, \cdots, 365\} \tag{3}
\]

where \( 0 \leq Q^r(T) \leq 1 \), since \( \sum_{t=1}^{24} HM_t(T) \leq \sum_{t=1}^{24} AM_t(T) \), due to the well known Pythagorean mean inequality, \( HM_t \leq AM_t \).

V. PROPOSED METHODOLOGY

We propose the use of daily aggregate Harmonic Mean to Arithmetic Mean Ratios as a Tier 1 metric for anomaly detection. Using a two year long real dataset, we first establish that our Tier 1 metric is very stable under normal conditions, without requiring any smoothening procedure and is scalable for different datasets. Then, we define the normal limits of the proposed Ratio metric (termed as Safe Margins) by empirically studying the ratio distribution. Subsequently, we propose the sum of residual distances between the instantaneous ratio samples and the chosen safe margin (termed as Residual Under the Curve (RUC)) as a Tier 2 metric for the same two years. We derive the normal limits of the RUC metric (termed as Standard Limit) from the historical data which is simplified due to the inherent stability of these metrics. The study of normal behavior of proposed metrics, safe margins and standard limits under no attacks is termed as the ‘system identification’ phase. The safe margins and standard limits act as detection thresholds. Then, we propose a two tier attack detection scheme. The first tier checks whether the daily aggregate Harmonic Mean to Arithmetic Mean Ratios, are within the established safe margins. Once, there is an anomalous deviation in the first metric, the directional deviation of the second metric from the standard limit is used to confirm the presence of organized data falsification or rule it out as a legitimate change. The two tier approach is required to increase the detection sensitivity for low attack strengths without increasing the false alarms.

Figs. 6(a) and 6(b) shows the proposed ratio metric, \( Q^r(T) \), for two different meter population sizes 200 and 800 meters respectively; for years 2014 and 2015 from Texas, without using any moving average smoothening mechanism. This is in sharp contrast to the arithmetic mean trends shown earlier in Fig. (1), that fluctuates readily for the same dataset over the same time periods. To validate the generality of this observation, Fig. 7(a), shows the time series of \( Q^r(T) \) for a completely different (Irish) dataset with 5000 meters from six regions. Figs. 6(a), 6(b) and 7(a) indicate that \( Q^r(T) \) is a...
highly stable metric making it suitable for anomaly detection. Thus, it is established that legitimate changes that affect the mean power consumption, does not significantly affect the daily average ratio between $HM$ and $AM$.

The reasoning behind the stability observed in the ratio samples is related to the coarse grained weak positive correlation in the daily granularity time scale and we show it later in Section VI-A with real datasets. Another reasoning behind using this metric for attack detection is the asymmetric growth and decay rates of harmonic means compared to arithmetic means that are derived from the exclusive Schur-Concavity properties of harmonic means. The proof of this and resultant attack detection properties are derived and established in detail in Section VI-B.

Finally, it can be observed that the $Q'(T)$ time series show a weak cyclo-stationarity in the sense that the mean ratio is similar during the same window $T$ in successive years. Further time series smoothening may reduce jitters and false alarms, but invariably falls into the trap of incremental evolving attacks. Instead, we take a different approach to reduce false alarms without sacrificing missed detection using the proposed RUC metric as discussed later.

2) Safe Margin for Ratio Metric: The distribution of all $Q'(T)$ samples for meter populations of 200 and 800 respectively for the 2 years is Gaussian distributed (See Appendix F). The mean and standard deviation of the ratio distribution for 200 meter case is $\mu_r = 0.917$ and $s_r = 0.0085$, with 69.57% of the ratio samples within the first standard deviation from $\mu_r$. The percentage of ratio samples within the second and third standard deviations are 96.03% and 98.95%. The corresponding $(\mu_r, s_r)$ values for 800 meter population is $(0.906, 0.007)$, and $\mu_r \pm 2s_r$ contains 97% of the samples. Fig. 7(b), shows the ratio distribution for the 5000 houses irish dataset to be similar to ratio distribution of Texas data. A comparison between Figs. 6(a) and 6(b) show that ratio metric is less stable for lower sized meter populations. Intuitively, detecting low margin attack signature in 200 meter population will be more difficult. Hence, we show more results for 200 meters representing the worst case.

Defining a threshold parameter, $\kappa \in (0, 3s_r]$, we build a ‘safe margin’ as an interval around the observed instantaneous ratio samples $Q'(T)$, on every time window on the historical (training) datasets. Hence, the upper and lower limits of the safe margin boundary denoted by $\Gamma_{\text{high}}(T)$ and $\Gamma_{\text{low}}(T)$ are:

$$\Gamma_{\text{high}}(T) = Q'(T) + \kappa$$  \hfill (4) \hfill \\
$$\Gamma_{\text{low}}(T) = Q'(T) - \kappa$$  \hfill (5) \\

Larger $\kappa$ values produce wider safe margins. An obvious approach to minimize false alarms would have been to set the $\kappa$ to maximum or minimum observed values in $Q'(T)$ distribution. Such wide safe margin would in turn increase missed detection for stealthier attacks having lower $\delta_{\text{avg}}$ and/or $\rho_{\text{mat}}$. For example, a $\kappa = 2s_r$, instead of $\kappa = 3s_r$, will have more detection sensitivity but on average about 4.5% legitimate observed ratio samples would be outside this safe margin, which is still an unacceptable false alarm rate for good anomaly based intrusion detectors [1]. To bypass this problem (known as base rate fallacy), we now propose another metric called RUC, for the second tier which preserves increased detection sensitivity without sacrificing false alarms.

3) Sum of Residuals between Ratio and Safe Margin: We propose another metric that maintains, at each time window $T$, the sum of the residuals between the ratio curve and the chosen safe margin (denoted by $RUC(T)$) over a sliding frame of past $FS$ days. To calculate $RUC(T)$, we first calculate $\nabla(T)$ that denotes the ‘signed residual distance’ between the observed ratio and the safe margin by:

$$\nabla(T): \begin{cases} 
Q'(T) - \Gamma_{\text{high}}(T), & \text{if } Q'(T) > \Gamma_{\text{high}}(T); \\
Q'(T) - \Gamma_{\text{low}}(T), & \text{if } Q'(T) < \Gamma_{\text{low}}(T); \\
0, & \text{otherwise}; 
\end{cases}$$  \hfill (6) \\

The $\nabla(T)$ value could be positive or negative depending on whether the instantaneous ratio sample is above the upper safe margin $\Gamma_{\text{high}}(T)$, or below the lower safe margin $\Gamma_{\text{low}}(T)$. The $\nabla(T)$ is zero, if the value of the ratio observed is within $[\Gamma_{\text{low}}(T), \Gamma_{\text{high}}(T)]$. Given this, at any window $T$, we propose to keep record of the sum of the residual distances $RUC(T)$ over a sliding frame of the past $FS$ days, such that:

$$RUC(T) = \sum_{j=T-FS}^T \nabla(j) \quad (7)$$

Our second metric, $RUC(T)$ for the historical dataset (2014, 2015) is shown in Fig. 8 using a sliding frame of $FS = 7$ days. The sign of $\nabla(T)$ represents direction of the change and plays a key role in attack reconstruction as shown later. The rationale behind keeping such a signed metric is because for legitimate changes, even when the ratio goes out of the safe margins, it oscillates between the upper and lower safe margins, making the $RUC(T)$ often closer to zero.

![Fig. 8. Standard Limits of RUC with $\kappa = 2s_r$.](image)

4) Inferring Detection Thresholds for Testing Set: From system identification, we calculate a safe margin parameterized by $\kappa$ and a ‘standard limit’ (that depends on $\kappa$) acting as the thresholds of these two metrics. Since these thresholds are obtained from historical data, we denote them by a superscript ‘h’. For any time window $T^h$ in a testing set (current), let $Q'(T^h)$ denote the historical value of the ratio metric on the corresponding $T$-th day in the previous years. Let $Q'(T^h)$ be calculated as a weighted average of the ratios samples on the $T$-th day in the previous years. Unlike most other schemes, the choice of weights do not affect performance drastically since difference between ratios in successive years is minimal owing to their high stability. We capture both extreme choices
of weights in our study by giving 0.99 weight to 2015 data and 0.01 to 2014 for 200 meter case study; and we give equal weightage to 2014 and 2015 ratios for the 800 meter case study. We will verify in the results that performance is not drastically different in either choices of weights. Hence, to summarize the safe margin for the T-th day (of an year) in the testing set is $Q^*(T^h) \pm \kappa$. For instance, if $\kappa = 2\sigma_r$, then the upper and lower margins are $\Gamma_{\text{high}}(T^c) = Q^*(T^h) + 2\sigma_r$ and $\Gamma_{\text{low}}(T^c) = Q^*(T^h) - 2\sigma_r$.

Now, we learn a normal range for the RUC values known as ‘standard limit’ that is independent of $T$. Let $\{\text{RUC}(T)^y\}$ denote the set of the sum of residual distances observed in $y$-th (training) year, as shown in Fig. 8. Now, we need to choose an appropriate $\tau_{\text{max}}(h)$ and $\tau_{\text{min}}(h)$ as upper and lower thresholds from $\{\text{RUC}(T)^y \mid y \in \{1, 2\}\}$. Suppose $\tau$ represents the appropriate candidate thresholds. The appropriate thresholds $\tau_{\text{max}}(h)$ and $\tau_{\text{min}}(h)$ from the set of $\text{RUC}(T)^y$ is obtained by Algorithm 1 and marked in Fig 8 that avoids over-fitting. The $c_{\text{max/min}}$ and $p_{\text{max/min}}$ are the corresponding cost and penalty functions used for the $\tau_{\text{max}}$ (searching among all non-zero positive RUC($T$)) and $\tau_{\text{min}}$ (searching among all non-zero negative values in RUC($T$)). Notice, the design of RUC($T$) is such that the search space is very small, due to very less non-zero RUC($T$) values, which reduces the complexity of the search problem described via Algorithm 1.

**Algorithm 1 Calculate $\tau_{\text{max}}(h), \tau_{\text{min}}(h)$**

```plaintext
for $T, \tau, y$ do
  if ($\text{RUC}(T)^y < \tau$) then
    $c_{\text{max/min}} = \frac{|\tau - \text{RUC}(T)^y|}{2}$
  else
    $p_{\text{max/min}} = 2|\text{RUC}(T)^y - \tau|$ 
  end if
end for

$\tau_{\text{max/min}}(h) = \arg\min_{\tau} |c - p|$
```

**B. Tier One Detector**

While designing the safe margin criterion, we utilized an interval $\kappa$ from the historical ratio distribution. If the observed ratio is outside this safe margin, we declare a suspected anomaly. Let $c$ denote the current (testing) year (2016 in our case). Then, if the observed ratio sample of the current time window $Q^*(T^c)$ is within the interval $[Q^*(T^h) \pm \kappa]$, then the situation is considered normal. If not, then an anomaly is suspected.

$$Q^*(T^c) : \begin{cases} \in [Q^*(T^h) \pm \kappa] & \text{No Anomaly;} \\ \notin [Q^*(T^h) \pm \kappa] & \text{Anomaly Suspected;} \end{cases}$$

where $T^c$ is the current time window, $T^h$ is the corresponding time window of the prior year, $Q^*(T^h)$ is the corresponding historical ratio value in the previous years for the same window $T$, and $\Gamma_{\text{high}}(T^c)$ and $\Gamma_{\text{low}}(T^c)$ are the safe margins at $T^c$ day of the testing set.

**C. Tier Two Detector**

Given that the ratio metric was off safe limits and the RUC($T^c$) is outside the normal interval $[\tau_{\text{min}}(h), \tau_{\text{max}}(h)]$, it is confirmed, that there is an organized data falsification attack.

$$\text{RUC}(T^c) : \begin{cases} \in [\tau_{\text{min}}(h), \tau_{\text{max}}(h)] & \text{No Attack;} \\ \notin [\tau_{\text{min}}(h), \tau_{\text{max}}(h)] & \text{Attack Inferred;} \end{cases}$$

An illustration of the 2016 testing set with a camouflage attack, using a data order aware attack strategy with $p_{\text{mal}} = 40\%$ and $\delta_{\text{avg}} = 120W$ in Figs. 9(a) and 9(b). The blue lines (with markers) show the proposed metrics in the testing set under no attacks, and the dotted lines represent the safe margins. Observe the red lines representing metrics under attacks. The ratio signature is not evident, which demonstrates the need for the second tier.

![Fig. 9. Ultra Low Camouflage (a) Ratio (b) RUC](image-url)

**VI. THEORETICAL EXPLANATION AND PROPOSED METRIC PROPERTIES**

In the previous section, we just proposed the defense methodology, but did not understand why and how it detects data falsification. In this section, we formalize the different mathematical and security properties we discovered, that provides this understanding. This section first derives the generic security lessons/properties and then the implications of power consumption data falsification in AMI is concluded.

**A. Why is the ratio metric a stable invariant?**

We provide a mathematical explanation for the invariance observed in ratio metric across multiple data sets over various years. Most residential households in an area tend to have certain coarse grained shared behaviors during a typical day although individual differences exist. Therefore, power consumption of different households are not completely independent but intuitively should possess some weak positive correlation. Since the $Q^*(T)$ is a daily metric, the stability of proposed metrics will get captured (if its proven to be related to the correlation) although different datasets will have different mean values of $Q^*(T)$ samples. The difference in datasets only affects the mean value of the proposed $Q^*(T)$ but not its stability over time windows within that dataset (as seen in Figs 6 & 7). Now, suppose we denote the average difference between power consumption of any two houses in the series $p^1, \ldots, p^N$ over a time window $T$ (a day) as $\xi(T) = \text{Avg.}(\{p^{i+1} - p^i\})$. Since humans exhibit a shared routine of habits through a typical day, the average difference between any two meters averaged over $T$ (equaling a day) is unlikely to be arbitrarily different from each other. In other words, the distribution of $\xi(T)$ will be weakly stationary. This
claim can be verified by Fig. 10(a) showing the time series of $\xi(T)$ for the Irish dataset.

Now we establish an important link between the observed invariance of $\xi(T)$ and Tung’s Theorem and its corollaries [18], [16] (1975). Tung’s Theorem “proposed the theoretical upper and lower bounds on the absolute difference between Arithmetic and Geometric Mean in any series data.” An extension of this Theorem described ‘the upper and lower bounds on the absolute difference between Harmonic and Arithmetic Mean’ in a series data [16]. The theorem states:

Given a series $a \equiv \{1 \equiv a_1, \cdots, a_N \equiv B\}$, 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 $A_N$ is given by:

$$\frac{(B - 1)^2}{N(B + 1)} \leq |A_N - H_N| \leq (\sqrt{B} - \sqrt{1})^2$$ \hspace{1cm} (10)

If minimum and maximum values are $a_{\text{min}}$ and $a_{\text{max}}$ respectively, then Eqn. 10 can be rewritten as:

$$\frac{(a_{\text{max}} - a_{\text{min}})^2}{N(a_{\text{max}} + a_{\text{min}})} \leq |A_N - H_N| \leq (\sqrt{a_{\text{max}}} - \sqrt{a_{\text{min}}})^2$$ \hspace{1cm} (11)

where $a_{\text{max}} \approx a_{\text{min}} + (N - 1)\xi$. This means that the bounds of $|A_N - H_N|$ is dependent on $\xi$ and $a_{\text{min}}$ only. Since, it is evident from Figs. 10(a) and 10(b), that $\xi$ and $a_{\text{min}}$ are both stable, therefore $|A_N - H_N|$ should also exhibit high stability. This explains the invariance of harmonic to arithmetic mean ratios across multiple datasets and their subsets. We believe that this result has broader implications and not just restricted to smart metering infrastructure. Most cyber-physical systems having sensory redundancy exhibit positive correlation while sensing some physical phenomena, under appropriate temporal and spatial granularities. When these spatio-temporal granularity is appropriately designed, it will enable engineers to use our metric as an invariant for anomaly detection. The above explanation that is backed by results from real datasets prove that our approach is an alternative to deal with the common difficulty of handling shifting trends in various power consumption datasets.

B. Establishing Effect of Pythagorean Mean Properties under Data Falsification Attacks

The strictly Schur-Concavity property of Harmonic Means compared to the non-strict concavity of Arithmetic means can be exploited to derive unique security properties/lessons under different data falsification modes that facilitate deeper understanding of changes in the proposed metrics under attacks.

Fig. 11. Schur Concavity

Let $X = \{x^1, \cdots, x^N\}$, denote a data series from N sources (such as number of meters) each indexed as $i$, such that the mean and standard deviation of $X = (\mu, \sigma)$. The harmonic mean $HM(X)$ and arithmetic mean $AM(X)$ is defined as:

$$HM(X) = \frac{N}{\sum_{i=1}^{N} \frac{1}{x^i}}$$
$$AM(X) = \frac{\sum_{i=1}^{N} x^i}{N}$$ \hspace{1cm} (12)

Schur Concavity is described by the following criterion:

$$(x_{1j} - x_{1k}) \left(\frac{\partial y}{x_{1j}} - \frac{\partial y}{x_{1k}}\right) \leq 0 \hspace{1cm} \forall x \in \mathbb{R}^d$$ \hspace{1cm} (13)

where $x_{1j} \neq x_{1k}$. It can be verified from Fig. 11 that the y-axis represents the range of AM and HM functions for a simple two member data series $X = (x_1, 2)$. Where $x_1 \in \{0, \infty\}$ is the x-axis representing the domain of the AM and HM functions. Note, that while AM is both concave and convex, HM is strictly Schur Concave. Therefore, when a subset of $x^i$ values in $X$ is changed (visualize $x_1$ getting biased from original value say $x_1 = 1$), AM growth/decay rates are linear as well as symmetrical for additive and deductive biases. On the other hand, due to the exclusive Schur-Concavity property growth or decay rates in HM are asymmetric to the rate of change in AM. Therefore, this asymmetry causes a change in the ratio value of the observed HM and AM.

Additionally, we found that the increase or decrease in the ratio value is also dependent on the position of the datapoint (lesser or greater) being attacked w.r.t the actual mean, the margin of false data and the mode of data falsification. Table I, summarizes the various lessons/properties we derived for various attack types, and margins of false data, in terms of the effects of various attacks on the proposed ratio metric, for attacked datapoints on left (L) (lesser) or right (R) (greater) of the actual mean. These properties are generic and not specific to the dataset, but their implications on the power consumption dataset are explained separately later.

Now, we give a short explanation of Table I: Suppose, sub portions of a data set generated from multiple sources experience additive and deductive manipulation by an attack bias $\epsilon$. The rates of growth in HM and AM that occur under additive attacks (denoted by ‘+’) is represented by $|\Delta HM^+|$ and $|\Delta AM^+|$. Similarly, rates of decay in HM and AM occurring under deductive (denoted as ‘−’) and camouflage attacks (denoted as ‘−’) are represented by $|\Delta HM^-|$, $|\Delta AM^-|$, and $|\Delta AM^+|$, $|\Delta AM^+-|$. The L# denotes the observation/lesson/property number, Mode represents type of falsification, Position ‘L’ and ‘R’ denotes whether
the position of the attacked datapoint is on the left and right of the true mean. The necessary and sufficient conditions for experiencing each observation is also provided in the rightmost columns. ‘X’ denotes ‘do not care’ conditions. The statement of every property and lessons are provided in detail in Appendix D. We also provide a theoretical illustration of how Schur concavity creates these unique properties under various falsification attacks in Appendix D.

To conclude, the unique properties of Pythagorean Mean’s growth decay rates is the first reason to use Harmonic to Arithmetic Mean ratios for attack detection. The second reason to use HM to AM ratio is its high stability under no attacks as established in Figs. 6(a), 6(b) and 7(a). Third, the closed forms of Harmonic Means do not exist and only coarse approximations are possible. The approximation errors increase with increasing number of data points and number of attacked data points, and the entire underlying time series is not very stable. Thus, it makes it difficult if not impossible for an adversary to exactly back calculate a strategy to beat our method with 100% success rate.

Necessary and Sufficient Conditions for Ratio Change:
Below, we write the ‘approximation expressions’ for various necessary and sufficient conditions listed in Table I. We verify the accuracy of these approximation experimentally in the results section. Theoretical verification and approximation of bounds are provided in Appendix D and E respectively.

The approximate (average case) lower bounds are:
\[
k^-(\text{low}) = k^+(\text{low}) = \frac{\sqrt{N-M}}{M} + \frac{\sigma}{\sqrt{M}} \sqrt{\frac{N-M}{N-1}} + \sigma
\]
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:
\[
k^+(\text{high}) = k^-(\text{high}) = \max(\sigma^2, \frac{2\sigma}{M} + \frac{\sigma}{\sqrt{M}} \sqrt{\frac{N-M}{N-1}} + 2\sigma)
\]

The average conditions for deductive on the left and additive on the right side of the mean is:
\[
k^-(\text{high}) = k^+(\text{high}) = \sigma \sqrt{\frac{N}{N-1}}
\]

In the worst case \(k_{\text{low}} > \frac{\left(\sum_{i=1}^{N} - \mu\right)}{\sigma} + \frac{\sigma}{\sqrt{N-1}} + \mu\) - \(x_i\), and \(k(\text{high}) > |x_i - \mu| + \sigma^2\). The worst case expressions can be used for verification purposes, when the smallest datapoint in the series is additive attacked, while the highest datapoint is deductive attacked. The approximation of the above bounds are provided in Appendix E.

Implications of Lessons on Real Consumption Dataset:
Note that Table I lists ‘general’ attack signature properties of the ratio metric. When applied to real power consumption datasets and attacks, these lessons have to be surmised in the relevant context. Recall, that power consumption data set has 64% of the data points on the left side (<) of the mean and 36% are on the right (>) of the mean. Hence, on average, the probability of attacked datapoints will be more on the left of the true mean. From Table I, we know that attacked data-point’s authentic position relative to the true mean, affects growth and decay rates of HM and AM. Therefore, we predict that experimental results will confirm to L1, L3, L4, L7 given the focus of the paper is on lower margins of false data and most of the authentic data is on the left of the true mean. To conclude, HM to AM ratio should increase under additive attacks, and decrease under deductive and camouflage attacks, if the attacker does not know the defense mechanism and the margins of false data are lower than the standard deviation. For higher margins of false data, a decrease in proposed metrics is predicted for all modes of attack due to L2, L5, L6.

C. Parameters for Unbiased Security Performance Evaluation
Here, we define four metrics required for fair and unbiased security performance evaluation and analysis of our proposed anomaly detection technique. The following also explains why we do not use to the traditional ROC curves for evaluation.

Time to Detection: is the difference (in days) between the start of attack and the time it was detected. This metric is applicable for non-persistent and non-optimal attacks against our detector.

Expected Time between False Alarms \(E(T_{fa})\): For evaluating security of proposed approach against persistent attacks, the standard ROC curves for security evaluation are a biased measure for three reasons: (i) False alarm rates are misleading since it depends on the time duration of study; (ii) Low false alarms rates in ROC curves can be misleading due to base rate fallacy elucidated by Axelson et. al. in his seminal paper [1] and is not ideal for intrusion detection but component diagnostics; (iii) It is particularly difficult to get detection rate with ROC under persistent/undetected attacks, since the adversary never gets detected. To prevent such a bias, a recent work [23] showed that expected time between false alarms versus the impact of an undetected attack averaged over varying thresholds (\(\kappa\) in our case) is a more unbiased measure for performance evaluation for intrusion detection mechanisms under persistent attacks. If the impact of undetected attack

### Table I

<table>
<thead>
<tr>
<th>Lesson #</th>
<th>Mode</th>
<th>Position</th>
<th>Property</th>
<th>Ratio Effect</th>
<th>Necessary</th>
<th>Sufficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1</td>
<td>Additive (+)</td>
<td>L</td>
<td>(</td>
<td>\Delta HM</td>
<td>&gt;</td>
<td>\Delta AM</td>
</tr>
<tr>
<td>L2</td>
<td>Additive (+)</td>
<td>L</td>
<td>(</td>
<td>\Delta HM</td>
<td>&lt;</td>
<td>\Delta AM</td>
</tr>
<tr>
<td>L3</td>
<td>Deductive (-)</td>
<td>L</td>
<td>(</td>
<td>\Delta HM</td>
<td>&gt;</td>
<td>\Delta AM</td>
</tr>
<tr>
<td>L4</td>
<td>Deductive (+)</td>
<td>L</td>
<td>(</td>
<td>\Delta HM</td>
<td>&lt;</td>
<td>\Delta AM</td>
</tr>
<tr>
<td>L5</td>
<td>Deductive (+)</td>
<td>R</td>
<td>(</td>
<td>\Delta HM</td>
<td>&gt;</td>
<td>\Delta AM</td>
</tr>
<tr>
<td>L6</td>
<td>Additive (+)</td>
<td>R</td>
<td>(</td>
<td>\Delta HM</td>
<td>&lt;</td>
<td>\Delta AM</td>
</tr>
<tr>
<td>L7</td>
<td>Camouflage (+,-)</td>
<td>X</td>
<td>(</td>
<td>\Delta HM</td>
<td>&gt;</td>
<td>\Delta AM</td>
</tr>
</tbody>
</table>

does not arbitrarily increase for higher $E(T_{fa})$, then it is a good detection metric. Hence, we used this approach for security evaluation of our method rather than the standard ROC curves. However, our definition of $E(T_{fa}) = \sum_{\eta}^{\infty} \eta P_A$ is more unbiased than the one suggested in [23] which calculated $E(T_{fa})$ as an average of false alarms over the total time.

**Impact (of undetected attack) per Unit Time (I):** The $I = \frac{RR}{24}$ revenue damage per hour used as a measure to quantify impact of undetected attack, where $RR$ is attack revenue per day as defined earlier in Sec. III (Eqn. 1). We plot $E(T_{fa})$ versus $I$ for security evaluation of our method.

**Break Even Time ($T_{BE}$):** Assuming a fixed average cost of $\$500$ to compromise a meter (from reports in a real attack [29], [32]), let the total investment of attack is $TC = 500 * M$ in dollars. Then, $T_{BE} = \frac{TC}{RRCost}$ is the time (in number of years) for total revenue accrued to break-even $TC$ invested, where $RR$ is the revenue accrued per day for the undetected attack. Given a particular microgrid $N$, we vary $M$ to show how our methods prolongs $T_{BE}$ under persistent attacks.

### VII. Experimental Evaluation

A real dataset for the past 3 years (2014, 2015, 2016) was gathered from Pecan Street Project [34] for 200 and 800 houses. We used the 2014 and 2015 datasets for system identification, while 2016 data set was used as a testing set for evaluation. A validation data set [36] of 5000 houses from Ireland for 2 years (2009, 2010) was also used to prove generality, scalability, and sensitivity. We fed the real data into a virtual AMI, and generated various attack samples over data from the testing set. The experimental section is organized into (A) Attack Forensic Trends (B) Performance Evaluation (C) Comparison with existing works.

Subsection (A), shows the forensic signatures for each attack strategy, modes, and stealth levels in terms of $\delta_{avg}$, using a fixed threshold $\kappa = 2\sigma_r$ (that gives a desired baseline false alarm rate as shown later), and a fixed fraction of compromised nodes (40%), to prove that the properties predicted earlier match with the results. We consider the frame length of $FS = 7$ days over which the standard limit of $RUC(T)$ is calculated considering a corresponding $\kappa = 2\sigma_r$. We develop an attack reconstruction scheme that remaps the observed forensic signature to conclude presence and mode of attack.

Subsection (B), shows our performance by parameterizing all $\rho_{mal}$, $\delta_{avg}$, $\kappa$, and $N$, values to report the detection limits. We provide the bounds of $(\delta_{avg}, \rho_{mal})$ pairs where our scheme fails to detect, and quantify how our approach limits the impact of persistent attacks. We tested over multiple attack start points and randomized identities of compromised meters while reporting performance to avoid sampling biases.

Subsection (C), compares our work with some existing works under a comparable threat model with the same datasets.

#### A. Attack Forensics and Signatures

While we emphasize on results with $\delta_{avg} < 400W$, our method also works for $\delta_{avg} > 400$ as well. Within $\delta_{avg} < 400$, the comparatively higher $(\rho_{mal}, \delta_{avg})$ pairs produce a clear signature in the first tier itself. We term this class as Low

**Attack Strength.** If either $\delta_{avg}$ and/or $\rho_{mal}$ are lowered further, clear attack signatures are not obtained in the first tier, but the second tier reveals the attack. We term this attack class as Ultra Low Attack Strength. For all strategies with $\delta_{avg} \geq 400$, we term them as High Attack Strength.

1) **False Alarm Baseline:** Fig. 12(b), shows the number of false alarms per year for various values of $\kappa$ for our testing set of 800 houses (2016 data) under no attacks. It shows just one false alarm, while using a standard limit corresponding to $\kappa = 2\sigma_r$. For such an ultra-low false alarm rate (0.27%), this $\kappa$ is known as a baseline threshold where detection sensitivity is the worst. To represent the worst detection case, the attack forensic subsection results uses the baseline $\kappa = 2\sigma_r$. The performance evaluation subsection shows how detection is improved further by decreasing $\kappa$ from the baseline value if more false alarms are tolerable by an utility.

![False Alarm Performance](image-url)

**Fig. 12. False Alarm Performance**

(a) All $\kappa$ (b) $N = 800$ with $\kappa = 2\sigma_r$

2) **Ultra Low Attack Strength:** Figs. 13(a) and 13(b) exhibit the signatures of the ratio and RUC metric, under a deductive and data order aware attack with $\rho_{mal} = 40\%$, and margins as low as $\delta_{avg} = 50W$. The attack starts from the 41st day which is depicted by a bold black dotted line. As predicted, Fig. 13(a) shows a gradual decrease in the ratio, compared to the original testing set’s ratio from the attack start point. But owing to the low $\delta_{avg}$, the deviation in the ratio signature is not enough to confirm the attack as it closely follows the expected ratio trend. However, the corresponding RUC plot in Fig. 13(b) for the same scenario, reveals the confirmation of presence of an attack within 10 days from the start. Lack of clear signature in the first tier ratio confirms an attack of ultra low strength. In Fig. 13(b), the $RUC(T^c)$ on the 50-th day goes below the lower standard limit and is of negative sign, which rules out the options such as additive and one sided deductive and camouflage. Checking the directional change in the moving average of HM and AM from the 42nd day will show a gradual decrease in both the means. This confirms a ultra low strength deductive attack. We can see that even a margin of false data as low as 50W gets detected.

Figs. 14(a) and 14(b), exhibit the corresponding signatures for an additive attack with $\rho_{mal} = 40\%$ and $\delta_{avg} = 80$ with a data order aware strategy. The ratio signature follows lesson
(L1) and slightly increases, on account of being additive and negative. The \( \delta_{\text{avg}} \) being much lower than \( k_{\text{low}} = 458 \) (calculated by Eqn. 14) for this attack strength pair. But due to ultra low attack strength, the first tier is not enough to confirm an attack as seen in Fig. 14(a). However, Fig. 14(b), shows that second tier is able to detect the attack within 6 days.

The corresponding results for camouflage attacks had already been shown earlier in Figs. 9(a) and 9(b). Note that the signature for camouflage were more bursty due to the additive and deductive bias components. Note that the detection took about 5 days, which is earlier than additive owing to the higher \( \delta_{\text{avg}} \) of 125\( W \). This is to show that time to confirmation of attack is dependent upon the \( \delta_{\text{avg}} \) value as well.

3) Low Attack Strength: Fig. 15(a), shows signatures under additive attacks with \( \rho_{\text{mal}} = 40\% \) and \( \delta_{\text{avg}} = 200\). The \( \delta_{\text{avg}} \) is less than most existing works, but the ratio metric shows a clear indication of an attack. Figs. 15(b), shows the scenario for deductive and camouflage attack for the same \( \rho_{\text{mal}}, \delta_{\text{avg}} \) for a duration 42nd to 65th day. To show an additional perspective, the non data order aware result is also shown in Fig. 15(b), that proves the relative ease of detection for non-data order aware attacks.

Fig. 14. Ultra Low Additive (a) Ratio (b) RUC

Fig. 15. Low attack strength (a) Additive (b) All Mixed

4) Incremental Evolving Attacks: Figs. 16(a) and 16(b), show the tier one signature with incremental evolving attacks with 2 watts increment on every time slot up to \( \delta_{\text{avg}} = 200\% \). If we used a residual check at every time step, this strategy would not be captured easily due to readjusting.

5) Validation with Irish Data Set: To show that our signatures reported are consistent across other data sets, we demonstrate the deductive and additive attack signatures for the Irish Dataset. The Irish dataset did not have two full year’s data. Hence, we used safe margins widths from Texas dataset onto the ratio metric of the Irish dataset (since variance in ratio samples were similar). Figs. 17(a) and 17(b), show unique signatures for the Irish data set for deductive attacks \( (\rho_{\text{mal}} = 10\%, \delta_{\text{avg}} = 70\% \) and Figs. 18(a) and 18(b) show the same for additive attacks \( (\rho_{\text{mal}} = 20\%, \delta_{\text{avg}} = 100\% \), that are consistent with the observations from the Texas data set. This shows that the proposed metrics and inferred baseline thresholds obtained from Texas Data for the RUC generalizes well for the Irish Data set. Higher \( \rho_{\text{mal}} \) in this dataset get more easily detected hence lower \( \rho_{\text{mal}} \) is shown. From our experience, the Irish dataset showed more attack sensitivity.

Fig. 16. Incremental Evolving Attack: (a) Additive (b) Camouflage

Fig. 17. Irish Data Set: Deductive: (a) Ratio (b) RUC

Fig. 18. Irish Data Set: Additive: (a) Ratio (b) RUC

6) Validation with High Attack Strengths: Fig. 19, is an exhibit which proves that our proposed model works for higher attack strengths as well. In the following result the \( \rho_{\text{mal}} = 40\% \) and \( \delta_{\text{avg}} = 800\% \).

Fig. 19. High Attack Strengths: \( \Delta_{\text{avg}} = 800\% \)

<table>
<thead>
<tr>
<th>TABLE II</th>
<th>CONCLUDING SECURITY STATE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ratio</td>
<td>Limit</td>
</tr>
<tr>
<td>Up</td>
<td>Outside</td>
</tr>
<tr>
<td>Down</td>
<td>Outside</td>
</tr>
<tr>
<td>Down</td>
<td>Outside</td>
</tr>
<tr>
<td>Down</td>
<td>Inside</td>
</tr>
<tr>
<td>X</td>
<td>Outside</td>
</tr>
<tr>
<td>X</td>
<td>Inside</td>
</tr>
</tbody>
</table>

7) Attack Scene Reconstruction: Observing features from the signatures help us conclude the mode of falsification, type of strategy employed, estimated start time of attack, time of stoppage of attack (if any). The features are: (i) Direction (going up increasing), down (decreasing) or same) of the ratio, harmonic, and arithmetic mean (ii) Limits (outside or inside) of the RUC bound. (iii) Sign of RUC (positive or negative). For example, \( RUC(T) \) is positive and outside the \( T_{max} \)h, and
the observed AM and HM has shown an increase, a stealthy additive attack is confirmed. But if ratio and $RUC(T)$ is negative but observed HM and AM increases, then it is an additive attack with larger margins ($\delta_{avg} < k_{low}$). Intuitively, the rest of the conclusions can be made. S,N,X in the Table II denotes Stealthy, Non-Stealthy, and Don’t care.

### B. Performance Limits and Security Evaluation

For different candidate thresholds of $\kappa$, we show how our approach restricts the impact of optimal persistent attack for a given AMI network. Second, we study detection limits for different combinations of $\rho_{mal}$ and $\delta_{avg}$ at (known as breakdown point) for a given AMI network that just escapes detection. We show the adversary’s breakeven time $T_{BE}$ at these breakdown points is very high. Finally, we show detection sensitivity to different $\delta_{avg}$ with changing microgrid sizes for a given attack budget $M$ and a desired false alarm rate.

#### Limiting Impact of Persistent Attacks: Figs. 20(a) and 20(b), depict the limiting impact of undetected attacks if the attacker knew the threshold and detection method and remained undetected with an attack that just escapes detection. The Y-axis, denotes the impact per unit hour (I) by the adversary, while the X-axis denotes the expected time between false alarms. The largest point in X-axis correspond to the $E(T_{fa})$ values obtained for thresholds ('standard limit') corresponding to the baseline safe margin $\kappa = 2s_r$. As we reduce $\kappa$ from this baseline, the expected time between two false alarms decreases due to increasing false alarms. To show this, we decrease the $\kappa$ with a step-size of 0.02 and get a unique value of $E(T_{fa})$ and $I$ for every possible $\kappa$ value for both additive and deductive attacks. Figs. 20(a) and 20(b) (for deductive and additive attacks), clearly show that as the average time between two false alarms increases, the impact of undetected attack does not increase significantly. At $E(T_{fa}) = 134$, the impact is limited to $0.58$ for deductive attacks, and $0.62$ dollars for additive attacks, even when $40\%$ of the meters had been compromised. Similarly different lines in Figs. 20(a) and 20(b) correspond to $E(T_{fa})$ vs $I$ various $\rho_{mal}$.

#### Breakdown Points and Breakeven Time: For every value of $\rho_{mal}$, there is a corresponding $\delta_{avg}$ value on (or below) which our detection method starts to fail. We call this ($\rho_{mal}$, $\delta_{avg}$) pair as a breakdown point. Given a particular $\rho_{mal}$, and full knowledge of the defense mechanism, the attacker can use this corresponding $\delta_{avg}$ to cause maximum damage while remaining undetected. The impact of such a strategy is studied to quantify the time extent to which our method forces the adversary to breakeven its total cost.

From Table III(a) (left table), we observe that breakdowns occur at a comparatively larger $\delta_{avg}$ for smaller $\rho_{mal}$ and vice-versa. However, owing to small $\rho_{mal}$, the impact ($RR$ per day) is also the least. The attacker can increase the $\rho_{mal}$ (say 70%) and increase its daily revenue. However, at $500 as investment per meter (from [29], [32]), it will take 6.79 years to recover the money invested, because our method catches any $\delta_{avg}$ as low as 70W or above on average. A high $T_{BE}$ depicts a level of discouragement for adversary to continue an undetected attack that takes 6 years to breakeven since it is forced to attack at very small $\delta_{avg}$ values to ensure evasion. Table III(a), provides a comprehensive list of the breakdown points and their $T_{BE}$ for deductive attacks. Similarly, the detection limits for additive attacks is shown in Table III(b). Note that all these values are for thresholds that yielded one false alarm in a year. A slight increase in tolerable false alarm rate will further reduce the undetectable strategy space. Note that the largest $\delta_{avg}$ in the tables (120W) for any $\rho_{mal}$ is about four times less than breakdown points in existing works [2], [8] which fail to provide any detection when $\delta_{avg} < 400$ using the same datasets.

#### Scalability of N versus Sensitivity to $\delta_{avg}$: Fig. 21, shows a 3D view of sensitivity to different margins of false data under deductive attacks. (y-axis) for various deployments of microgrid sizes $N$ (x-axis), if the adversary has a given $M$ for the Irish dataset. The $\kappa$ threshold required for getting one false alarm per year changes for various values of $N$ as shown on the z-axis. The results show that the worst case where only 100 out 4000 meters ($\rho_{mal} = 2.5\%$) are compromised, the detection sensitivity limit is at $\delta_{avg} = 150W$. At the other extreme, when 300 out 4000 meters ($\rho_{mal} = 75\%$) are compromised, we detect the attack at $50W$. Both 150W and 50W are much lower than 400W which the existing works assume. Hence, it is a good decentralized solution for intrusion detection since microgrid sizes usually range from 100-2000.

**Table III**

<table>
<thead>
<tr>
<th>$\rho_{mal}$</th>
<th>$\delta_{avg}$</th>
<th>$RR$</th>
<th>$T_{BE}$</th>
<th>$\rho_{mal}$</th>
<th>$\delta_{avg}$</th>
<th>$RR$</th>
<th>$T_{BE}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>120</td>
<td>6.9</td>
<td>3.9</td>
<td>10</td>
<td>120</td>
<td>7.4</td>
<td>3.6</td>
</tr>
<tr>
<td>20</td>
<td>95</td>
<td>10.9</td>
<td>5.0</td>
<td>20</td>
<td>100</td>
<td>11.5</td>
<td>4.7</td>
</tr>
<tr>
<td>30</td>
<td>80</td>
<td>14.6</td>
<td>5.5</td>
<td>30</td>
<td>85</td>
<td>14.6</td>
<td>5.5</td>
</tr>
<tr>
<td>40</td>
<td>65</td>
<td>14.9</td>
<td>7.3</td>
<td>40</td>
<td>80</td>
<td>18.4</td>
<td>5.9</td>
</tr>
<tr>
<td>50</td>
<td>60</td>
<td>17.2</td>
<td>7.9</td>
<td>50</td>
<td>80</td>
<td>23.0</td>
<td>5.9</td>
</tr>
<tr>
<td>60</td>
<td>70</td>
<td>24.1</td>
<td>6.7</td>
<td>60</td>
<td>65</td>
<td>29.3</td>
<td>5.5</td>
</tr>
<tr>
<td>70</td>
<td>70</td>
<td>28.2</td>
<td>6.7</td>
<td>70</td>
<td>90</td>
<td>36.2</td>
<td>5.2</td>
</tr>
</tbody>
</table>

**Discussion on some Special Cases:** For additive attacks, we know that the ratio metric increases for lower attack strengths but decreases for higher attack strengths mainly due to right skewness of the AMI data. Naturally, an intermediate range
in $\delta_{\text{avg}}$ values exists where ratio will cease to increase and the start to decrease for additive attacks. We term this as a cross over point. The attacker may be tempted to target this as a potential strategy too by launching a strategy where $\delta_{\text{avg}} \sim k_{\text{low}}$, thinking that the ratio will stay same. However, we show that it is only possible to delay detection but not possible to avoid detection for all practical purposes. Plugging in the scenario of $\rho_{\text{mal}} = 40\%, N = 200, \sigma(2016) = 417$ in the Eqn 14, we calculate the necessary condition bound $k_{\text{low}} = 458W$. We attacked with $\delta_{\text{avg}} = 450$ and $\delta_{\text{avg}} = 490$ with several randomized combinations of $\rho_{\text{mal}}$ and plotted the worst case among all experimental rounds in Figs. 22(a) and 22(b). Fig. 22(b), shows that attack is detected within 45 days and 12 days after attack launch respectively. The average time to detection for these $\delta_{\text{avg}}$ over all experimental rounds was ‘much lower’ at 21 days. We conclude that for additive attacks it might be possible to delay but not escape detection by back calculating $k_{\text{low}}$ and $k_{\text{high}}$, due the highly irregular safe margin, non-existence Harmonic mean’s closed form.

![Fig. 22. Worst Case $\delta_{\text{avg}} \sim k_{\text{low}}$: (a) Ratio (b) RUC](image)

**C. Comparison with Existing Approaches:**

Now we compare our work with variations of EWMA and CUSUM techniques as discussed in the Section II.

**Comparison with EWMA:** The EMWA fails regardless the choice of weights and variation in thresholds of the safe margin. The average performance of an EWMA based metric is typically performed through ROC curves. The ROC for EWMA based approach under the given threat model is shown in Fig. 23(a). It clearly shows that as false alarm rates increase to up-to 25%-30%, the detection rate does not increase above 50%, for most stealthier ($\rho_{\text{mal}}; \delta_{\text{avg}}$) pairs. For instance, an attack of $\rho_{\text{mal}} = 50\%$ and $\delta_{\text{avg}} = 200$ is not detected by EWMA at a threshold which ‘gives an expected time between two false alarms of 13 days’. We detect this attack with a threshold (safe margin) that gives an expected time between two false alarms of 134 days.

![Fig. 23. Comparison: (a) EWMA (b) CUSUM](image)

**Comparison with CUSUM:** The $S_h$ and $S_l$ correspond to two CUSUM metrics that monitor increasing and decreasing trends in a process variable. We concluded that this not applicable under stealthy attacks for a process such as electric power consumption owing to its large variations. Therefore, the safe margin ends up being really wide due to the large variance in the CUSUM samples caused by drifts in mean consumption trends. Also these trends are inconsistent across years which rules a historical safe margin based approach. Therefore, CUSUM misses stealthier attacks altogether (as seen in Fig. 23(b)), although the false alarm is not poor. More insights and supporting results are provided in the Supplementary Results. The ROC curves of CUSUM will be absurd therefore and hence not included.

**VIII. Conclusions and Future Work**

In this work, we showed that various falsification modes such as deductive, camouflage and additive attacks launched by organized or persistent adversaries with stealthy strategies and ultra low attack strengths can be detected (or impacts limited) in real time in a light weight and non-private intrusive manner, while having ultra-low false alarms. We show that even if the attacker knows the mechanism and the thresholds and stays undetected, the method inherently forces the adversary to attack with strategies that limit the impact of undetected attacks, making such strategies less attractive to the adversary. In future, we will study on-off and data omission attacks which will require a little modification of the tier 2 detection level. We will propose a method for identifying the meters that are injecting false power consumption data that will be guided by the knowledge gained from this paper, such as mode of attack, stealth level, strategy used. Since the security properties are generic, the broader impact of the methodology and security properties for data falsification will be laid out.

**Acknowledgements** The authors would like to thank and acknowledge Mr. Aditya Thakur and Mr. Praveen Madhavarapu for assistance with the experiments. A major portion of the work was completed at Missouri University of Science and Technology, where Dr. Shameek Bhattacharjee was a post-doctoral research fellow during 2015-2018. The work has been supported by the following NSF grants: CNS-1818942, CNS-1545037, CNS-1545050, and DGE-1433659.

**REFERENCES**


[34] [Online] (2018, Dec) Pecan Street Project, Available: https://www.pecanstreet.org/

