This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TDSC.2018.2889729, IEEE Transactions on Dependable and Secure Computing

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 accuracy of critical tasks such as automated billing and pricing, demand forecast, automated demand response, load adjustments, 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 InGaurdians 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

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TDSC.2018.2889729, IEEE Transactions on Dependable and Secure Computing

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 ($\leq 400W$), 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

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TDSC.2018.2889729, IEEE Transactions on Dependable and Secure Computing

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 $RR = 75$ dollars/day.

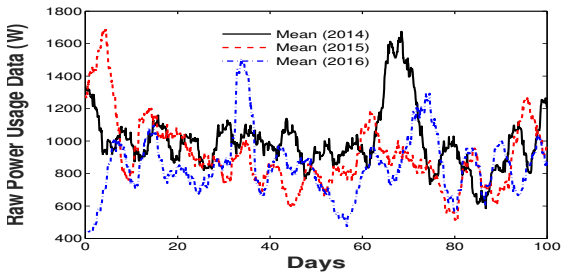


Fig. 1. Instability of mean power consumption

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

Deductive: The adversary reports $P_t^i = P_t^i(act) - \delta_t$, where $\delta_{min} \leq \delta_t \leq \delta_{max}$ is a false bias, amounting to electricity theft.

Additive: The adversary reports $P_t^i = P_t^i(act) + \delta_t$. 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.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TDSC.2018.2889729, IEEE Transactions on Dependable and Secure Computing

B. Margin of False Data and Fraction of Compromised Meters

All δ_t are generated randomly within an interval $[\delta_{min}, \delta_{max}]$. Let δ_{avg} be the average value of δ_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, δ_{avg} is part of the adversarial strategy. All units of δ_{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 δ_t with the same δ_{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:

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

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. 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* δ_{avg} *and* ρ_{mal}. Even as the same revenue is achieved with both strategies, the chances of detection (using proximity based mechanisms) are lesser in data order aware strategy due closer proximity to the actual data.

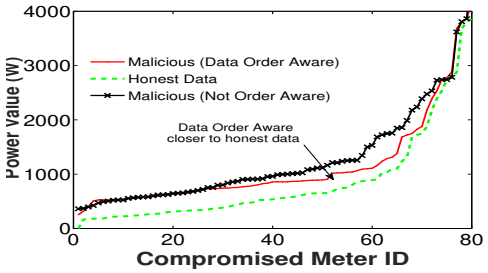


Fig. 2. Illustration: Benefit of Data Order Awareness

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 $P_t^{(1)}(act) \leq , \cdots, P_t^{(m)}(act), \leq P_t^{(M)}(act)$. Then adversary generates M random numbers for δ_t, sorted as $\delta_t^{min} \leq, \cdots, \leq \delta_t^{max}$. For an additive attack, the lowest observed power consumption data is changed with the highest δ_t^{max}, while highest observed power consumption data is modified with lowest δ_t^{min}, and so on, such that $P_t^{(1)}(act) + \delta_t^{max}, \cdots, P_t^{(M)}(act) + \delta_t^{min}$. For a deductive attack, the highest observed power consumption data is changed with the highest δ_t^{max}, while the lowest observed power consumption data is changed with the lowest δ_t^{min}. Hence, $P_t^{(1)}(act) - \delta_t^{min}, \cdots, P_t^{(M)}(act) - \delta_t^{max}$. For a camouflage attack, the sorted $P_t^{(1)}(act) \leq, \cdots, \leq P_t^{(M)}(act)$ 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 δ_{avg}, adversary incrementally increases margin of false data by Δq watts on every time slot, leading upto the intended δ_{avg}. A gradual change on every time slot causes the approaches based on instantaneous moving averages to fail, since the average readjusts 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 has knowledge 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 *breakdown point*, as the δ_{avg} and ρ_{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.

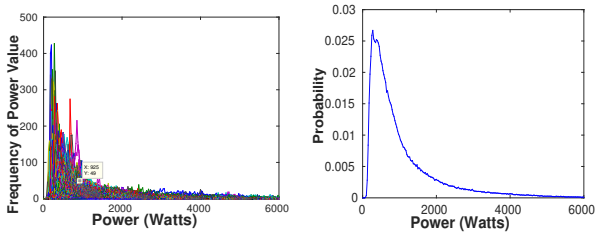
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TDSC.2018.2889729, IEEE Transactions on Dependable and Secure Computing

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.

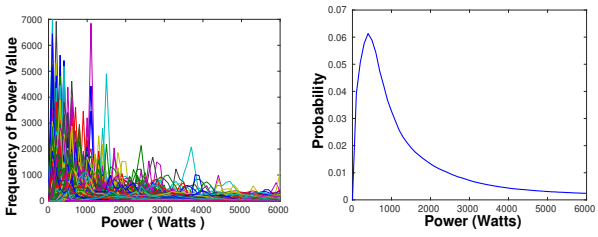
To address concerns on the generality of these observations, we repeat the above for the Irish dataset. Fig. 4(a) refer to power consumption distribution of individual houses in the Ireland dataset, whose mixture distribution counterpart is shown in Fig. 4(b). The results clearly indicate that aggregate 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.



(a) (b)

Fig. 3. Power Consumption (*Texas, US*): (a) 200 houses (b) Mixture P^{mix}

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 $p_t^i = ln(P_t^i + 2)$ 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.



(a) (b)

Fig. 4. Power Consumption (*Ireland*): (a) 5000 houses (b) Mixture P^{mix}

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 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).

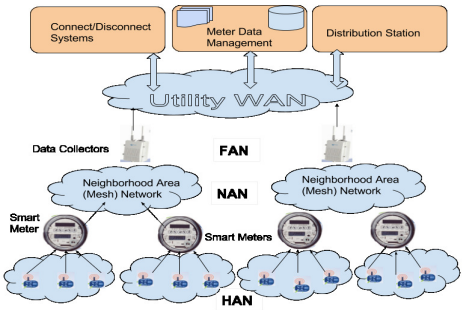


Fig. 5. AMI: Generic System View

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

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TDSC.2018.2889729, IEEE Transactions on Dependable and Secure Computing

if present) or NAN (neighborhood area network) gateway, that guards different smaller micro-grid subnets within the whole AMI network separately, while a centralized module is deployed in the Utility WAN gateway having a global view of a larger size of meters. Decentralized detectors have the advantage of more relative detection sensitivity and easier localization of the compromised parts of the AMI. To capture both implementation variations, we show results for both 200 and 800 houses from Texas that mimic different implementations respectively. For Irish datasets, various subsets sizes from 200 to 4000 were studied as decentralized implementation variants (the latest edge device DCN-3000 handles atmost 1000 houses [37]). We concluded that our detection method is best suited as a decentralized edge solution.

The technical contribution of this paper is divided into two sections: (a) *Proposed Methodology:* explaining the two-tier Pythagorean mean based detection scheme; (b) *Theoretical Explanation and Properties:* explaining the relationships between various Pythagorean Means (*Harmonic and Arithmetic Means*) that trigger deviations in the proposed metrics 'under various attack modes' and formalizes them as a set of properties/lessons. Using lessons/properties learned from such explanation as a basis, we justify how and why the proposed methodology is successful in detecting stealthy attacks and provide reasoning on the invariance of the proposed metrics over time and across datasets. Such separation of methodology from the theory has been done to elucidate a step by step guide when one intends to implement our approach.

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.

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.

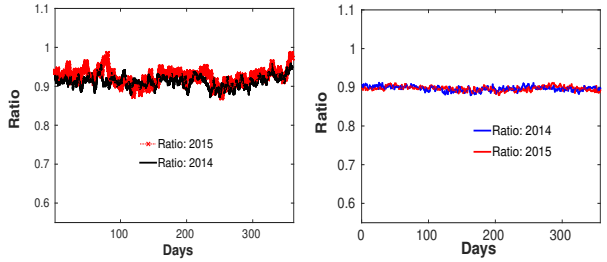
We denote $\boldsymbol{p}_t = [p_t^1, \cdots, p_t^N]$, as the power consumption data on the ln scale obtained from N meters at any time slot t. Let $\boldsymbol{p}$ denote the steady state mixture random variable, such that $\boldsymbol{p}_t$ is the realization of $\boldsymbol{p}$ at time t. The harmonic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

$$HM_t = \frac{N}{\sum_{i=1}^{N} \frac{1}{p_t^i}} \qquad AM_t = \frac{\sum_{i=1}^{N} p_t^i}{N} \tag{2}$$

*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:

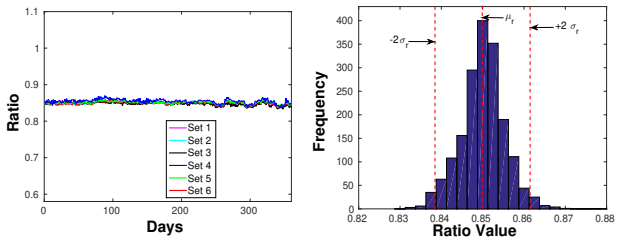
$$Q^r(T) = \frac{\sum_{t=1}^{24} HM_t(T)}{\sum_{t=1}^{24} AM_t(T)} \qquad \forall \quad T \in \{1, \cdots, 365\} \tag{3}$$

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, $HM_t \leq AM_t$.



(a) (b)

Fig. 6. Legitimate HM/AM Ratio: (a) 200 meters (b) 800 meters



(a) (b)

Fig. 7. Irish Data (5000 meters; 6 regions): (a) Ratios (b) Ratio Distribution

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

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TDSC.2018.2889729, IEEE Transactions on Dependable and Secure Computing

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 $Q^r(T)$ 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 $Q^r(T)$ 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 μ_r. The percentage of ratio samples within the second and third standard deviations are 96.03% and 98.95%. The corresponding (μ_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 $Q^r(T)$, on every time window on the historical (training) datasets. Hence, the upper and lower limits of the safe margin boundary denoted by $\Gamma_{high}(T)$ and $\Gamma_{low}(T)$ are:

$$\Gamma_{high}(T) = Q^r(T) + \kappa \tag{4}$$

$$\Gamma_{low}(T) = Q^r(T) - \kappa \tag{5}$$

Larger κ values produce wider safe margins. An obvious approach to minimize false alarms would have been to set the κ to maximum or minimum observed values in $Q^r(T)$ distribution. Such wide safe margin would in turn increase missed detection for stealthier attacks having lower δ_{avg} and/or ρ_{mal}. 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:

$$\nabla(T): \begin{cases} = Q^r(T) - \Gamma_{high}(T), & \text{if} \quad Q^r(T) > \Gamma_{high}(T); \\ = Q^r(T) - \Gamma_{low}(T), & \text{if} \quad Q^r(T) < \Gamma_{low}(T); \\ = 0, & \text{otherwise;} \end{cases} \tag{6}$$

The $\nabla(T)$ value could be positive or negative depending on whether the instantaneous ratio sample is above the upper safe margin $\Gamma_{high}(T)$, or below the lower safe margin $\Gamma_{low}(T)$. The $\nabla(T)$ is zero, if the value of the ratio observed is within $[\Gamma_{low}(T), \Gamma_{high}(T)]$. 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:

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

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.

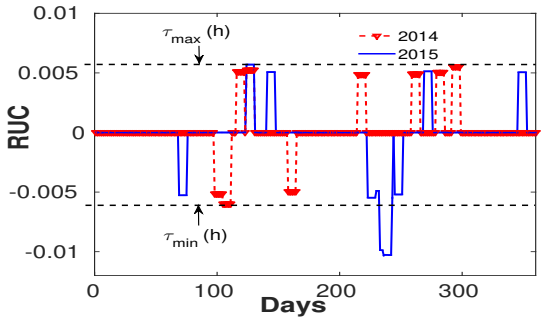


Fig. 8. Standard Limits of RUC with $\kappa = 2s_r$

*4) **Inferring Detection Thresholds for Testing Set**:* From system identification, we calculate a safe margin parameterized by κ and a 'standard limit' (that depends on κ) 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^c in a testing set (current), let $Q^r(T^h)$ denote the historical value of the ratio metric on the corresponding T-th day in the previous years. Let $Q^r(T^h)$ 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

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TDSC.2018.2889729, IEEE Transactions on Dependable and Secure Computing

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 $Q^r(T^h) \pm \kappa$. For instance, if $\kappa = 2s_r$, then the upper and lower margins are $\Gamma_{high}(T^c) = Q^r(T^h) + 2s_r$ and $\Gamma_{low}(T^c) = Q^r(T^h) - 2s_r$.

Now, we learn a normal range for the RUC values known as 'standard limit' that is independent of T. Let $\{RUC(T)^y\}$ 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 $\tau_{max}(h)$ and $\tau_{min}(h)$ as upper and lower thresholds from $\{RUC(T)^y \mid y \in 1, 2\}$. Suppose τ represents the appropriate candidate thresholds. The appropriate thresholds $\tau_{max}(h)$ and $\tau_{min}(h)$ from the set of $RUC(T)^y$ is obtained by Algorithm 1 and marked in Fig 8 that avoids over-fitting. The $c_{max/min}$ and $p_{max/min}$ are the corresponding cost and penalty functions used for the τ_{max} (searching among all non-zero positive $RUC(T)$) and τ_{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 $\tau_{max}(h), \tau_{min}(h)$

```
for T, τ, y do
  if (RUC(T)^y < τ) then
    c_{max/min} : |τ − RUC(T)| / 2
  else
    p_{max/min} = 2|RUC(T)^y − τ|
  end if
end for
τ_{max/min}(h) = arg min_τ |c − p|
```

B. Tier One Detector

While designing the safe margin criterion, we utilized an interval κ 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 $Q^r(T^c)$ is within the interval $[Q^r(T^h) \pm \kappa]$, then the situation is considered normal. If not, then an anomaly is suspected.

$$Q^r(T^c) : \begin{cases} \in [Q^r(T^h) \pm \kappa] & \text{No Anomaly;} \\ \notin [Q^r(T^h) \pm \kappa] & \text{Anomaly Suspected;} \end{cases} \quad (8)$$

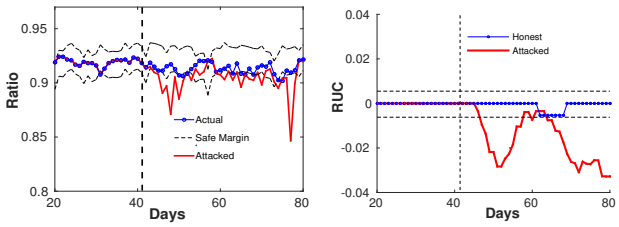
where T^c is the current time window, T^h is the corresponding time window of the prior year, $Q^r(T^h)$ is the corresponding historical ratio value in the previous years for the same window T, and $\Gamma_{high}(T^c)$ and $\Gamma_{low}(T^c)$ are the safe margins at T^c day of the testing set.

C. Tier Two Detector

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

$$RUC(T^c) : \begin{cases} \in [\tau_{min}(h), \tau_{max}(h)] & \text{No Attack;} \\ \notin [\tau_{min}(h), \tau_{max}(h)], & \text{Attack Inferred;} \end{cases} \quad (9)$$

An illustration of the 2016 testing set with a camouflage attack, using a data order aware attack strategy with $\rho_{mal} = 40\%$ and $\delta_{avg} = 120W$ 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.



(a) (b)

Fig. 9. Ultra Low Camouflage (a) Ratio (b) RUC

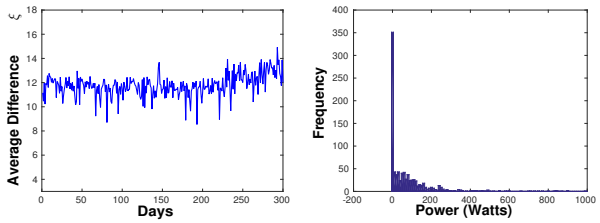
VI. Theoretical Explanation and Proposed Metric Properties

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 $Q^r(T)$ 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 $Q^r(T)$ samples. The difference in datasets only affects the mean value of the proposed $Q^r(T)$ 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 $p^1, \cdots, p^N$ over a time window T (a day) as $\xi(T) = Avg.\Big(|p^{i+1} - p^i|\Big)$. 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

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TDSC.2018.2889729, IEEE Transactions on Dependable and Secure Computing



(a) (b)

Fig. 10. Irish Dataset: (a) Average Difference $\xi(T)$ (b) a_{min} values

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:

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

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

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

where $a_{max} \sim a_{min} + (N-1)\xi$. This means that the bounds of $|A_N - H_N|$ is dependent on ξ and a_{min} only. Since, it is evident from Figs. 10(a) and 10(b), that ξ and a_{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.

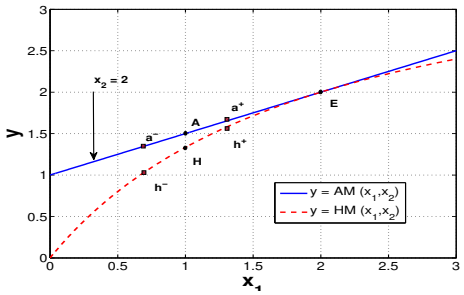


Fig. 11. Schur Concavity

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

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

Schur Concavity is described by the following criterion:

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

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 $\boldsymbol{X}$ 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 ϵ. 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 $|\Delta HM^{+}|, |\Delta HM^{+,-}|$, and $|\Delta AM^{+}|, |\Delta AM^{+,-}|$. The $L\#$ denotes the observation/lesson/property number, Mode represents type of falsification, Position 'L' and 'R' denotes whether

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TDSC.2018.2889729, IEEE Transactions on Dependable and Secure Computing

TABLE I
EFFECT OF DIFFERENT ATTACKS ON PYTHAGOREAN MEANS GROWTH DECAY RATES

Lesson #	Mode	Position	Property	Ratio Effect	Necessary	Sufficient
L1	Additive(+)	L	$\|\Delta HM^{+}\| > \|\Delta AM^{+}\|$	Increases	$\epsilon < k^{+}(llow)$	$\epsilon < k^{+}(lhigh)$
L2	Additive(+)	L	$\|\Delta HM^{+}\| < \|\Delta AM^{+}\|$	Decreases	$\epsilon > k^{+}(llow)$	$\epsilon > k^{+}(lhigh)$
L3	Deductive(-)	L	$\|\Delta HM^{-}\| > \|\Delta AM^{-}\|$	Decreases	$\epsilon > 0$	$\epsilon > k^{-}(lhigh)$
L4	Deductive(-)	R	$\|\Delta HM^{-}\| < \|\Delta AM^{-}\|$	Increases	$\epsilon < k^{-}(rlow)$	$\epsilon < k^{-}(rhigh)$
L5	Deductive(-)	R	$\|\Delta HM^{-}\| > \|\Delta AM^{-}\|$	Decreases	$\epsilon > k^{-}(rlow)$	$\epsilon > k^{-}(rhigh)$
L6	Additive(+)	R	$\|\Delta HM^{+}\| < \|\Delta AM^{+}\|$	Decreases	$\epsilon > 0$	$\epsilon > k^{+}(rhigh)$
L7	Camouflage(+,-)	X	$\|\Delta HM^{+,-}\| > \|\Delta AM^{+,-}\|$	Decreases	X	X

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: $k^{-}(rlow) = k^{+}(llow) =$

$$k_{low} = \frac{\sigma}{M} + \frac{\sigma}{\sqrt{M}}\sqrt{\frac{N-M}{N-1}} + \sigma \tag{14}$$

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: $k^{+}(lhigh) = k^{-}(rhigh) =$

$$k_{high} = max(\sigma^2, \frac{2\sigma}{M} + \frac{\sigma}{\sqrt{M}}\sqrt{\frac{N-M}{N-1}} + 2\sigma) \tag{15}$$

The average conditions for deductive on the left and additive on the right side of the mean is:

$$k^{-}(lhigh) = k^{+}(rhigh) = \sigma\sqrt{\frac{N}{N-1}} \tag{16}$$

In the worst case $k_{low} > \{(|x_i - \mu| + \sigma\sqrt{\frac{N}{N-1}}) + \mu)\} - x_i$, and $k(high) > |x_i - \mu| + \sigma^2$. 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 *Axelsson 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 (κ in our case) is a more unbiased measure for performance evaluation for intrusion detection mechanisms under persistent attacks. If the impact of undetected attack

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TDSC.2018.2889729, IEEE Transactions on Dependable and Secure Computing

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 $E(T_{fa}) = \frac{\sum_1^{\eta_{FA}} T_{BFA}}{\eta_{FA}}$ 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 $I = RR/24$ 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, $T_{BE} = \frac{TC}{RR*365}$ is the time (in number of years) for total revenue accrued to breakeven 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 δ_{avg}, using a fixed threshold $\kappa = 2s_r$ (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 $\kappa = 2s_r$. We develop an attack reconstruction scheme that remaps the observed forensic signature to conclude presence and mode of attack.

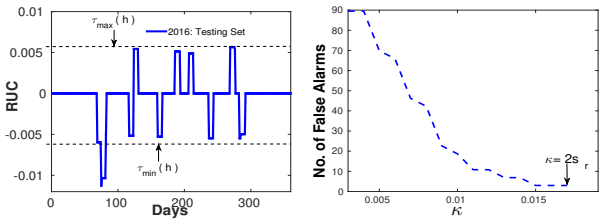
Subsection (B), shows our performance by *parameterizing all ρ_{mal}, δ_{avg}, κ, and N, values* to report the detection limits. We provide the bounds of $(\delta_{avg}, \rho_{mal})$ pairs where our scheme fails to detect, and quantify how our approach limits the impact of persistent attacks. We tested 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

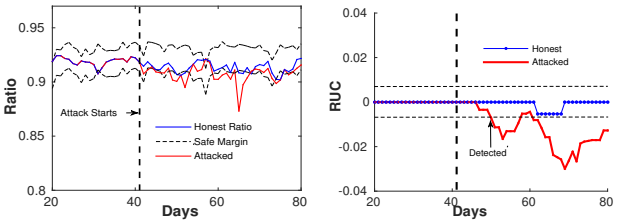
While we emphasize on results with $\delta_{avg} < 400W$, our method also works for $\delta_{avg} > 400$ as well. Within $\delta_{avg} < 400$, the comparatively higher $(\rho_{mal}, \delta_{avg})$ pairs produce a clear signature in the first tier itself. We term this class as *Low Attack Strength*. If either δ_{avg} and/or ρ_{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 κ 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 $\kappa = 2s_r$. For such an ultra-low false alarm rate (0.27%), this κ 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 $\kappa = 2s_r$. The performance evaluation subsection shows how detection is improved further by decreasing κ from the baseline value if more false alarms are tolerable by an utility.



(a) (b)

Fig. 12. False Alarm Performance(a) All κ (b) $N = 800$ with $\kappa = 2s_r$



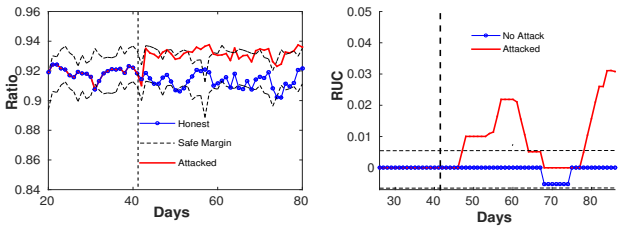
(a) (b)

Fig. 13. Ultra Low Deductive (a) Ratio (b) RUC

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 δ_{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

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TDSC.2018.2889729, IEEE Transactions on Dependable and Secure Computing



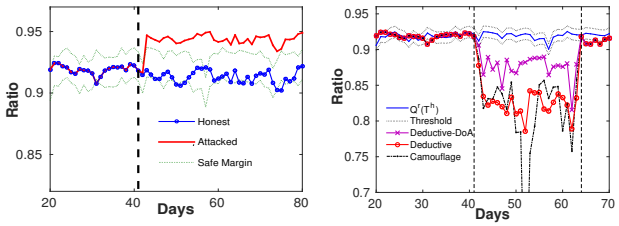
(a) (b)

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

(L1) and slightly increases, on account of being additive and the δ_{avg} being much lower than $k_{low} = 458$ (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.

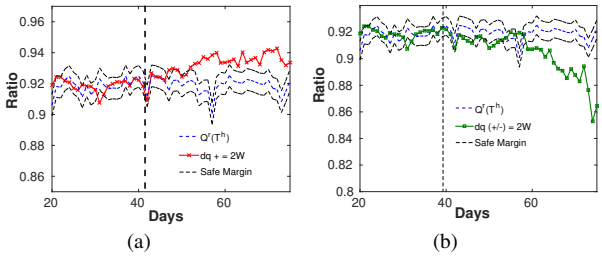
The corresponding results for camouflage attacks had already been shown earlier in Figs. 9(a) and 9(b). Note that the signature for camouflage were more bursty due to the additive and deductive bias components. Note that the detection took about 5 days, which is earlier than additive owing to the higher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

3) Low Attack Strength: Fig. 15(a), shows signatures under additive attacks with $\rho_{mal} = 40\%$ and $\delta_{avg} = 200$. The δ_{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 ρ_{mal}, δ_{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.



(a) (b)

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

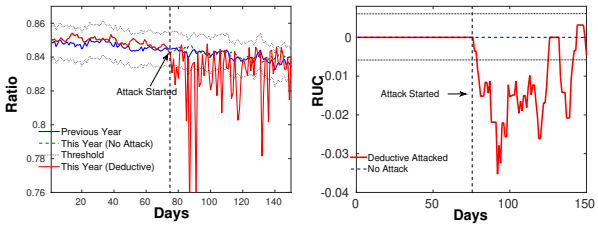


(a) (b)

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

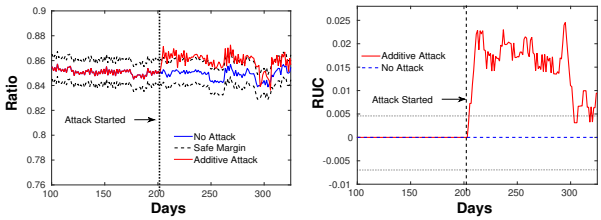
4) Incremental Evolving Attacks: Figs. 16(a) and 16(b), show the tier one signature with incremental evolving attacks with 2 watts increment on every time slot upto $\delta_{avg} = 200W$. If we used a residual check at every time step, this strategy would not be captured easily due to readjusting.

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 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 ($\rho_{mal} = 10\%$, $\delta_{avg} = 70W$) and Figs. 18(a) and 18(b) show the same for additive attacks ($\rho_{mal} = 20\%$, $\delta_{avg} = 100W$), 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 ρ_{mal} in this dataset get more easily detected hence lower ρ_{mal} is shown. From our experience, the irish dataset showed more attack sensitivity.



(a) (b)

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



(a) (b)

Fig. 18. Irish DataSet: 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 $\rho_{mal} = 40\%$ and $\delta_{avg} = 800$.

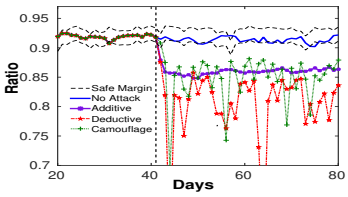


Fig. 19. High Attack Strengths: $\Delta_{avg} = 800$

TABLE II
CONCLUDING SECURITY STATE

Ratio	Limit	Sign	HM,AM	Conclusion
Up	Outside	Positive	Up, Up	Additive, S
Down	Outside	Negative	Up, Up	Additive, N
Down	Outside	Negative	Down, Down	Deductive, X
Down	Outside	Negative	Down, Same	Camouflage, X
X	Inside	X	X, X	No Attack, X

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 $\tau_{max}(h)$, and

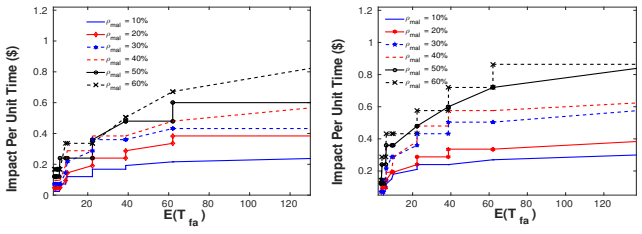
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TDSC.2018.2889729, IEEE Transactions on Dependable and Secure Computing

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 κ, 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 (ρ_{mal}, δ_{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 δ_{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 κ from this baseline, the expected time between two false alarms decreases due to increasing false alarms. To show this, we decrease the κ with a step-size of 0.02 and get a unique value of $E(T_{fa})$ and I for every possible κ 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 ρ_{mal}.



(a) (b)

Fig. 20. Limiting Impact of Persistent Attacks: (a) Deductive (b) Additive

Breakdown Points and Breakeven Time: For every value of ρ_{mal}, there is a corresponding δ_{avg} value on (or below) which our detection method starts to fail. We call this (ρ_{mal}, δ_{avg}) pair as a *breakdown point*. Given a particular ρ_{mal}, and full knowledge of the defense mechanism, the attacker can use this corresponding δ_{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 δ_{avg} for smaller ρ_{mal} and vice-versa. However, owing to small ρ_{mal}, the impact (RR per day) is also the least. The attacker can increase the ρ_{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 δ_{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 δ_{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 δ_{avg} in the tables ($120W$) for any ρ_{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.

TABLE III
BREAKDOWN POINTS

ρ_{mal}	δ_{avg}	RR	T_{BE}
10	120	6.9	3.9
20	95	10.9	5.0
30	80	14.6	5.5
40	65	14.9	7.3
50	60	17.2	7.9
60	70	24.1	6.7
70	70	28.2	6.7

ρ_{mal}	δ_{avg}	RR	T_{BE}
10	130	7.4	3.6
20	100	11.5	4.7
30	85	14.6	5.5
40	80	18.4	5.9
50	80	23.0	5.9
60	85	29.3	5.5
70	90	36.2	5.2

Scalability of N versus Sensitivity to δ_{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 κ 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 400 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.

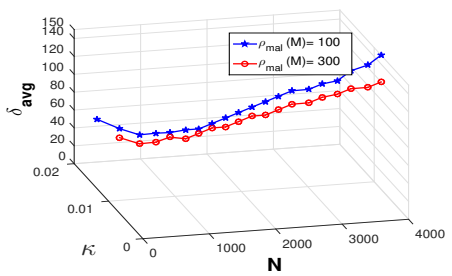
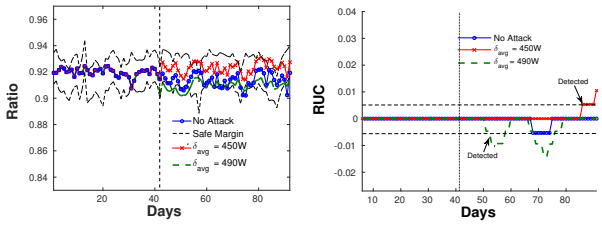


Fig. 21. Scalability and Sensitivity Performance: Irish Dataset

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

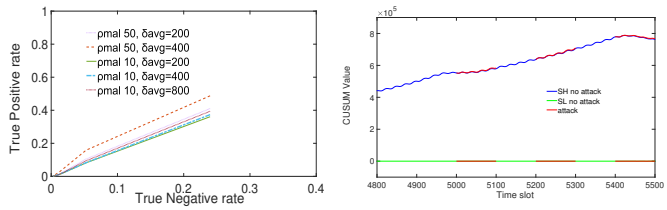
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TDSC.2018.2889729, IEEE Transactions on Dependable and Secure Computing

in δ_{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 $\delta_{avg} \sim k_{low}$ 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 $\rho_{mal} = 40\%, N = 200, \sigma_{(2016)} = 417$ in the Eqn 14, we calculate the necessary condition bound $k_{low} = 458$W. We attacked with $\delta_{avg} = 450$ and $\delta_{avg} = 490$ with several randomized combinations of ρ_{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 δ_{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_{low} and k_{high}, due the highly irregular safe margin, non-existence Harmonic mean's closed form.



(a) (b)

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



(a) (b)

Fig. 23. Comparison: (a) EWMA (b) CUSUM

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 $(\rho_{mal}, \delta_{avg})$ pairs. For instance, an attack of $\rho_{mal} = 50\%$ and $\delta_{avg} = 200$ 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.

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 this 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 privacy 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

[1] S. Axelsson, "The Base-rate Fallacy and the Difficulty of Intrusion Detection", *ACM Trans. Inf. Syst. Secur.*, Vol. 3(3), pp. 186-2015, Aug. 2000.

[2] 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.

[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] T. Chan, G.Golub, R. LeVeque "Algorithms for Computing the Sample Variance: Analysis and Recommendations", *The American Statistician*, Vol. 37(3), pp. 242-247, 1983.

[5] 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):2959-2966, Aug. 2013.

[6] 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):105-120, April 2014.

[7] A, Jindal, A. Dua, K. Kaur, M. Singh, N. Kumar, and S. Mishra, "Decision Tree and SVM-Based Data Analytics for Theft Detection in Smart Grid" *IEEE Trans. on Industrial Informatics*, Vol. 12(3), June 2016.

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

[9] T. Koppel, "Lights Out: A Cyberattack, A Nation Unprepared, Surviving the Aftermath", *Crown Publishers, New York*, 2015.
[10] 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.
[11] R. Lu, X. Liang, X. Li, X. Lin, and X. Shen, "EPPA: An efficient and privacy-preserving aggregation scheme for secure smart grid communications," *IEEE Trans. on Parallel and Distributed Systems* 23(9):1621-1631, Sept. 2012.
[12] S. McLaughlin, D. Podkuiko, and P. McDaniel, "Energy theft in the advanced metering infrastructure", *Proc. of Critical Information Infrastructures Security*, Springer-Verlag, pp. 176-187, Sept. 2009.
[13] S. McLaughlin, B. Holbert, S. Zonouz, and R. Berthier, "AMIDS: A multi-sensor energy theft detection framework for advanced metering infrastructures", *IEEE SmartGridComm*, pp. 354-359, Nov. 2012.
[14] J. Nagi, K. Yap, S. Tiong, S. Ahmed, M. Mohamad, "Nontechnical Loss Detection for Metered Customers in Power Utility Using Support Vector Machines", *IEEE Trans. On Power Delivery*, Vol. 25(2), pp. 1162-1172, 2010.
[15] D. Mashima and A. A Cardenas, "Evaluating electricity theft detectors in smart grid networks", *Springer Intl. Workshop on Recent Advances in Intrusion Detection*, pp. 210-229, Sept. 2012.
[16] B. Meyer, "Some Inequalities for Elementary Mean Values", *AMS Mathematics of Computation*, Vol. 42, No. 165, pp. 193-194, 1984.
[17] M. Tariq and H. V. Poor, "Electricity Theft Detection and Localization in Grid-tied Microgrids", *IEEE Transactions on Smart Grid*, Vol. no. 99, 2016.
[18] 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.
[19] R. Mohassel, A. Fung, F. Mohammadi, and K. Raahemifar, "A survey on advanced metering infrastructure", *Elsevier Journal of Electrical Power & Energy Systems*, 63:473-484, Dec. 2014.
[20] A. Rad and A.L. Garcia, "Distributed internet-based load altering attacks against smart power grids", *IEEE Trans. on Smart Grids*, 2(4):667-674, Dec. 2011.
[21] R. Sevlian and R. Rajagopal, "Value of aggregation in smart grids", *IEEE SmartGridComm*, pp. 714-719, Oct. 2013.
[22] Y.L. Sun, W. Yu, Z. Han, K.J. Ray Liu, "Information Theoretic Framework of Trust Model and Evaluation for Ad Hoc Networks",*IEEE Journal on Sel. Areas in Communications*, 24(2):305-317, Feb. 2006.
[23] D. Urbina, J. Giraldo, A. Cardenas, N. Tippenhauer, J. Valente, M. Faisal, J. Ruths, R. Candell and H. Sandberg, "Limiting the Impact of Stealthy Attacks on Industrial Control Systems", *ACM CCS*, pp. 1092-1105, 2016.
[24] W. Wang and Z. Lu, "Cyber security in smart grid: Survey and challenges", *Computer Networks*, 57(5):1344-1371, Apr. 2013.
[25] E. Werley, S. Angelos, O. Saavedra, O. Cortes, and A. Souza, "Detection and identification of abnormalities in customer consumptions in power distribution systems", *IEEE Trans. on Power Delivery*, 26(4):2436-2442, Oct. 2011.
[26] J. Xia and Y. Wang, "Secure key distribution for the smart grid", *IEEE Trans. on Smart Grid*, 3(3):1437-1443, Sept. 2012.
[27] W. Yu, D. Griffith, L. Ge, S. Bhattarai and N. Golmie, "An integrated detection system against false data injection attacks in the Smart Grid, *Security and Commun. Networks*, 8(2):91-109, Jan. 2015.
[28] [Online] (2017, Mar) Telegraph News, Available: http://www.telegraph.co.uk/news/2017/03/06/smart-energy-meters-giving-readings-seven-times-high-study-finds/
[29] [Online] (2012, Apr) Krebs on Security, Available: http://krebsonsecurity.com/2012/04/fbi-smart-meter-hacks-likely-to-spread/
[30] [Online] (2017, Jun) EPRI Product Abstract, Available: https://www.epri.com/#/pages/product/000000000001026553/
[31] [Online] (2009, Dec) NY Times, Available: http://www.nytimes.com/2009/12/14/us/14meters.html?ref=energy-environment&_r=0
[32] [Online] (2013, Feb) Maxim Integrated, Available: https://www.maximintegrated.com/content/dam/files/design/technical-documents/white-papers/smart-grid-security-recent-history-demonstrates.pdf
[33] [Online] (2010, July) Our Energy Policy, Available: https://www.smartgrid.gov/files/The_Smart_Grid_Promise_DemandSide_Management_201003.pdf
[34] [Online] (2018, Dec) Pecan Street Project, Available: https://www.pecanstreet.org/
[35] [Online] (2008, Dec) Energy.Gov, Available: http://energy.gov/sites/prod/files/oeprod/DocumentsandMedia/14-AMI_System_Security_Requirements_updated.pdf
[36] [Online] (2018, Dec) Irish Social Science Data Archives, Available: http://www.ucd.ie/issda/data/
[37] [Online] (2014, Dec) Echelon, Available: http://www.echelon.com/assets/blt21a75c85e69387d5/DCN-3000-Series-Distributed-Control-Node-datasheet.pdf

Shameek Bhattacharjee is an Assistant Professor at the Department of Computer Science at Western Michigan University, USA. He received his Ph.D. and M.S. from the University of Central Florida, Orlando in 2015 and 2011 respectively and his B.Tech from West Bengal University of Technology, India, 2009. Between 2015-2018, he worked as a post-doctoral researcher at the Center for Research in Wireless Mobility and Networking (CReWMaN) at Univ. of Missouri S & T at Rolla, MO, USA, where he is also affiliated as an Adjunct Faculty. His current research interests include information security in cyber-physical systems, wireless and social networks, particularly in topics such as anomaly detection, trust models, secure crowd-sensing, dependable decision theory.

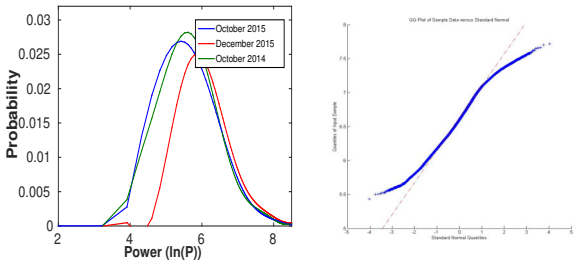
Sajal K. Das is a professor of Computer Science and the Daniel St. Clair Endowed Chair at the Missouri University of Science and Technology, where he was the Chair of Computer Science Department during 2013-2017. His research interests include cyber-physical security and trustworthiness, wireless sensor networks, mobile and pervasive computing, crowdsensing, cyber-physical systems and IoTs, smart environments (e.g., smart city, smart grid and smart health care), cloud computing, biological and social networks, and applied graph theory and game theory. He has published extensively in these areas with over 700 research articles in high quality journals and refereed conference proceedings. Dr. Das holds 5 US patents and coauthored 4 books Smart Environments: Technology, Protocols, and Applications (John Wiley, 2005), Handbook on Securing Cyber-Physical Critical Infrastructure: Foundations and Challenges (Morgan Kauffman, 2012), Mobile Agents in Distributed Computing and Networking (Wiley, 2012), and Principles of Cyber-Physical Systems: An Interdisciplinary Approach (Cambridge University Press, 2018). His h-index is 83 with more than 28,000 citations according to Google Scholar. He is a recipient of 10 Best Paper Awards at prestigious conferences like ACM MobiCom and IEEE PerCom, and numerous awards for teaching, mentoring and research including the IEEE Computer Societys Technical Achievement Award for pioneering contributions to sensor networks and mobile computing, and University of Missouri System Presidents Award for Sustained Career Excellence. He serves as the founding Editor-in-Chief of Elseviers Pervasive and Mobile Computing Journal, and as Associate Editor of several journals including the IEEE Transactions on Dependable and Secure Computing, IEEE Transactions on Mobile Computing, and ACM Transactions on Sensor Networks. Dr. Das is an IEEE Fellow.

Appendix for: Detection and Forensics against Stealthy Data Falsification in Smart Metering Infrastructure

Shameek Bhattacharjee and Sajal Das

I. Appendix A: Data Distribution Characteristics

Fig. 1(a), shows the sample mixture distributions on a transformed ln scale for different months. Note that the shape parameter remains similar but the location and scale parameters of p^{mix} keep changing. The extent of normality of the whole transformed power consumption data on the ln scale is shown in Fig. 1(b).



(a) (b)

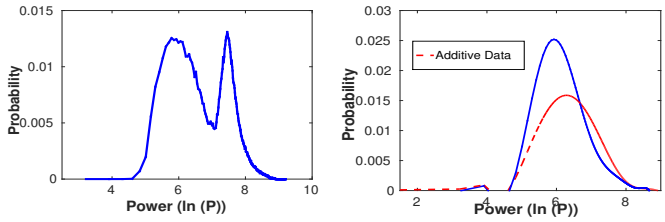
Fig. 1. Texas Data (a) p^{mix} for different months (b) The Q-Q Plot

II. Appendix B: Impact of Lower Margins of False Data

Consider AMI micro-grid of $N = 200$ smart meters, $M = 80$ implying $\rho_{mal} = 0.4$ or 40%. This scenario is totally realistic for a powerful and organized adversary deploying a decentralized detector. The actual aggregate power consumption distribution has a mean and a standard deviation of $\mu_A = 1200$ units and $\sigma_A = 400$ units, respectively. We studied from the real data sets that the average difference between the EWMA and the instantaneous mean is about $115W$. The average upper and lower bounds are 130 and 100. If the amount of additive error to be introduced in the final mean is say $\Lambda = 130$ units, the δ_{avg} for each malicious node is given by $\delta_{avg} = \frac{\Lambda * N}{M} = 325W$. This is an illustration that attacks with lower δ_{avg} will be rarely detected if measures of EWMA and ARMA of the mean consumption is used. However, their impact at the same time is quite large. Additionally, CUSUM of the mean does not work due high error residual due to the unstable mean power consumption.

III. Appendix C: Uniform vs Normally Distributed False Data

Figs. 2(a) and 2(b) show the difference between a normally distributed δ_{avg} versus a uniformly distributed δ_{avg}. Note that the change in the shape of the distribution is obvious in the first case, while for the uniformly distributed false data, there is no obvious change in the shape of the distribution. Hence, the second one might be a more stealthier choice for attack implementation.



(a) (b)

Fig. 2. Attack Distributions (a) Obvious Attack (b) Smarter Attack

IV. Appendix D: Understanding Security Properties and Signatures

We have identified the following security properties in terms of effects that various attack modes have over the Pythagorean means and consequently, the proposed metrics:

When sub-portions of a data set generated from multiple sources experience additive and deductive manipulation by a bias ϵ respectively, the growth and decay rates of HM are asymmetric when compared to the growth and decay of AM. Owing to this, the ratio of the biased HM and AM must differ compared to non-biased ones.

Property 1: Numbers on the left side of the true mean, when changed with additive bias of ϵ, the HM growth rate is faster than the corresponding AM growth rate, under the necessary condition that $\epsilon < k^{+}(llow)$, and the sufficient condition that $\epsilon < k^{+}lhigh$.

Lesson 1: Owing to property 1, the ratio of the biased HM and AM *increases* than the ratio of original data, (i.e., $Q^{+}(l) > Q$), if the $\epsilon < k^{+}(llow)$ holds.

Property 2: Numbers on the left side of the true mean, when changed with additive bias of ϵ, the HM growth rate is slower than the corresponding AM growth rate, under the necessary condition that $\epsilon > k^{+}_{llow}$ and sufficient condition $\epsilon > k^{+}(lhigh)$.

Lesson 2: Owing to property 2, the ratio of the biased HM and AM *decreases* than the ratio of the original data, (i.e. $Q^{+}(l) < Q$), if the $\epsilon > k^{+}(lhigh)$, holds.

Property 3: Numbers on the left side of the true mean, when changed with deductive bias ϵ, the HM decay rate is faster than the corresponding AM decay rate, for the necessary condition $\epsilon > 0$, and sufficient condition of $\epsilon > k^{-}(lhigh)$.

Lesson 3: Owing to property 3, the ratio of the biased HM and AM denoted by $Q^{-}(l)$ is lesser than the original datum ratio Q for the sufficient condition that $\epsilon > k^{-}(lhigh)$.

Property 4: Numbers on the right side of the true mean, when changed with deductive bias ϵ, the AM decay rate is faster than the corresponding HM decay rate rate, under

the necessary condition that $\epsilon < k^{-}(rlow)$ and sufficient condition $\epsilon < k^{-}(rhigh)$.

Lesson 4: Owing to property 4, the ratio of the biased HM and AM denoted by $Q^{-}(r)$ is greater than the original ratio Q, provided the $\epsilon < k^{-}(rlow)$ holds.

Property 5: Numbers on the right side of the true mean, when changed with deductive bias ϵ, the HM decay rate is faster than the corresponding AM decay rate, under the necessary condition that $\epsilon > k^{-}(rlow)$ and sufficient condition $\epsilon > k^{-}(rhigh)$.

Lesson 5: Owing to Property 5, the biased ratio $Q^{-}(r)$ is lesser than the original datum ratio Q, if the corresponding conditions hold.

Property 6: Numbers on the right side of the true mean, when changed with additive bias ϵ, the HM growth rate is slower than the corresponding AM growth rate, for necessary condition that $\epsilon > 0$ and sufficient condition is $\epsilon > k^{+}(rhigh)$.

Lesson 6: Owing to property 6, the ratio of the biased HM and AM denoted by $Q^{+}(r)$ is lesser than the original datum ratio Q under the sufficient condition of $\epsilon > k^{+}(rhigh) = \sigma\sqrt{\frac{N}{N-1}}$.

Property 7: The corresponding rates of growth and decay in the HM triggered by the additive and deductive biases of the same bias margin ϵ are unequal. The HM decays at a greater rate than it grows for the same ϵ. In contrast, the AM shows equal growth and decay rates for additive and deductive biases with same ϵ.

Lesson 7: Owing to property 7, the ratio of the biased HM and AM denoted by $Q^{+,-}(l)$ is lesser than the original datum ratio Q, since HM is effectively reduced, given the original data point being biased are on the same side of the true mean. This property helps in detecting camouflage attacks as explained later.

Illustration of Properties: Let us illustrate the properties using two numbers $X = (x_1, x_2)$. Let the standard deviation of X be σ. In Fig. 3, the x axis represents a variable say x_1. Let us fix another variable $x_2 = 2$, such that the actual ordered data set is $X = (x_1, 2)$. The y-axis represents the value of $AM(x_1, x_2)$ or $HM(x_1, x_2)$.

Hence, the AM function of $(x_{1j}, 2)\forall j = \{0, \infty\}$ is represented by the solid blue line showing linear growth with x_1, and is neither strictly concave or convex. On the other HM function of $(x_{1j}, 2)\forall j = \{0, \infty\}$ is represented by a dashed red line is a strictly Schur Concave Function.

For original data set X, the $AM = 1.5$ and $HM = 1.33$ represented by points A and H in Fig. 3. Hence, their ratio $Q = \frac{HM}{AM} = 0.88$. Now let us change one data point x_1, (which qualifies as a subset), with addition of a small amount of bias $\epsilon = 0.3$, mimicing an additive attack on numbers on the left side of the true mean.

Let us now compare the resultant growth rates of HM and AM in the biased data $X^{+} = (1+0.3, 2)$. The biased HM and AM are represented by points $h^{+} = 1.57$ and $a^{+} = 1.65$. The additive growth of HM is denoted by $\Delta HM^{+} = h^{+} - H = 0.23$. The growth of AM is denoted by $\Delta AM^{+} = a^{+} - A = 0.15$. Thus, $\Delta HM^{+} > \Delta AM^{+}$. Therefore growth rates of HM and AM are asymmetric and the numbers on the left of the mean with additive bias induces a faster rate of HM growth

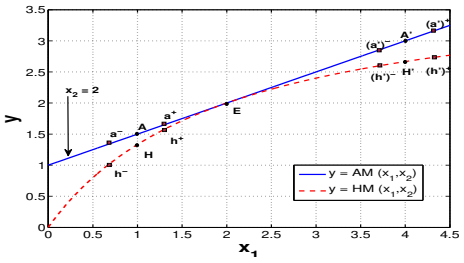


Fig. 3. Pythagorean Asymptotes

than the corresponding AM growth. This illustrates Property 1 and Property 2.

Furthermore, using Eqns. 14 and 15 (approximate average case bounds in the manuscript), we calculate $k^{+}(llow) = 2.12$ and $k^{+}(lhigh) = 3.535$. Note that, $\epsilon = 0.3 < k^{+}(llow)$, the biased ratio of HM and AM $Q^{+} = 0.95$ increases compared to the original $Q = 0.888$. Hence Lesson 1 is proven. If we consider an $\epsilon = 3.359 > k^{+}(lhigh)$. In such a case, $X^{+} = (1 + 3.359, 2)$, and $Q^{+} = 0.86282147031 < 0.888$. Hence, Lesson 2 is proven.

Similarly, suppose x_1 is biased with deduction of the same bias $\epsilon = 0.3$, such that $X^{-} = (0.7, 2)$. Points $a^{-} = 1.35$ and $h^{-} = 1.037$ correspond to the biased arithmetic and harmonic means respectively. Thus decay in HM and AM are given by $\Delta HM^{-} = h^{-} - H = -0.303$ and $\Delta AM^{-} = a^{-} - A = -0.15$. Ignoring the signs which denote decay, a comparison of corresponding rates of decay for the HM for additive and deductive attacks proves $|\Delta HM^{-}| > |\Delta AM^{-}|$. This proves property 3. Note that the biased ratio of HM and AM is $Q^{-} = 0.76 < Q$. This proves Lesson 3, where ratio drops if points already on the left of the true mean when *reduced* by a bias. Additionally, when we compare rates of HM bias of deductive with additive attack for the same $\epsilon = 0.3$, it can be verified that:

$$|\Delta HM^{-}| > |\Delta HM^{+}|$$

. However, $|\Delta AM^{+}| = |\Delta AM^{-}|$. Hence, HM decays at a faster rate than it grows for the same bias, while AM growth decay rates are same. This proves property 7 and Lesson 7. This is an intuition behind camouflage attacks detection signature of camouflage attacks discussed later.

Now let us see all the cases where, the number of the right side of the mean i.e. x_2 is changed with additive and deductive bias of $\epsilon = 0.3$. Now $X^{+} = (1, 2 + 0.3)$. The $HM^{+} = 1.39$ and $AM^{+} = 1.65$. The $\Delta HM = 0.05$ and $\Delta AM = 0.15$. Hence, HM mean grows at a slower rate than AM when numbers on right side are changed with additive bias. Hence, it proves property 6. Owing to the slower growth of HM, $Q^{+} < Q$, which proves Lesson 6.

Finally, x_2 is changed with a deductive bias of $\epsilon = 0.3$, making $X^{-}(r) = (1, 2 - 0.3)$. Hence, the $HM^{-} = 1.259$ and $AM^{-} = 1.35$. The $\Delta HM = 0.08$ and $\Delta AM = 0.15$. This illustrates that AM decays at a faster than HM when numbers on right side are changed with deductive bias given $\epsilon < k^{-}(llow) = 1$. Hence, property 4 is proven. As a consequence, $Q^{-} > Q$, owing to the slower growth of HM. This proves Lesson 4. To show lesson 5, x_2 needs to be

changed with a deductive bias of $\epsilon > k^{-}(rlow)$, but this is not a feasible strategy since $2 - k^{-}(rlow)$ is negative.

V. Appendix E: Approximation of the Bounds

Let us assume that actual data series from n sources is given as: $X(n) = \{x_1, \cdots, x_n\}$. For ease of analysis, let us assume that one data point say x_n is compromised, with a false data of δ_n added to it. Relative to the mean, the necessary condition that creates a difference in the ratios, by triggering an increase in variance will cause a drop in the ratio. Let the compromised data series be denoted by: $X'(n) = \{x_1, \cdots, (x_n + \delta_n)\}$. Now the sample mean of the actual data series is:

$$\overline{x} = \frac{1}{N} \sum_{i=1}^{N} x_i \tag{1}$$

The sample variance of the actual data series is:

$$\sigma^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left({x_i}^2 - 2\overline{x}x_i + \overline{x}^2 \right)$$

$$= \frac{1}{n-1} \left[\sum_{i=1}^{n} {x_i}^2 - N\overline{x}^2 \right]$$

$$\sigma^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left({x_i}^2 - \overline{x}^2 \right) \tag{2}$$

For easing the analysis, let us assume the mean $\overline{x} = 0$ such that,

$$\sigma^2 = \frac{1}{n-1} \sum_{i=1}^{n} {x_i}^2 \tag{3}$$

In contrast, assuming $\overline{x} = 0$, the sample mean of the compromised series is given by:

$$\overline{x''} = \frac{\sum_{i=1}^{n-1} x_i + (x_n + \delta_n)}{N} = \frac{\delta_n}{n} \tag{4}$$

Therefore the sample variance of the compromised series using Eqn.(2) is $\sigma_{''}^2 = \frac{1}{n} \sum_{i=1}^{n} \left({x_i''}^2 - \overline{x''}^2 \right)$

$\sigma_{''}^2 = \frac{n-1}{n}\sigma^2 + (1 - \frac{1}{n})\frac{\delta_n^2}{N} + 2x_n\delta_n$

The ratio is drops if $\sigma_{''}^2 > \sigma^2$. If we assume, x_n is equal to value of mean $(= 0)$, the following inequality (necessary) condition should be satisfied,

$$\delta_n > \sigma \sqrt{\left(\frac{n}{n-1} \right)} \tag{5}$$

Note that, Eqn. 5 holds only assuming x_n and $\overline{x}$ to be equal to zero. For $x_n \neq 0$ and $\overline{x} \neq 0$, it means that $x_n + \delta_n > \{|x_n - \overline{x}| + \sigma\sqrt{(\frac{n}{n-1})} + \overline{x}\}$ However, if data is normally distributed, the real margin of false data is closer to the necessary condition, because the difference between any data point and $\overline{x}$ is not arbitrarily high. Hence, the average case is given by,

$$\delta_{avg} > \sigma + \sigma \sqrt{\left(\frac{n}{n-1} \right)}$$

Generalizing, let M be the number of data points compromised without bias. Following the analysis in [4], it can be shown that on the average, the necessary condition is:

$$\delta_{avg} > \frac{\sigma}{M} + \frac{\sigma}{\sqrt{M}} \sqrt{\left(\frac{n-M}{n-1} \right)} + \sigma = k_{low}$$

The above bound is necessary but may not sufficient to effect a ratio drop additive attacks. But if δ_{avg} below k_{low} then the ratio cannot drop. This is infact established experimentally in the paper. The ratio narrowly rises and drops around k_{low}.

The sufficient condition is:

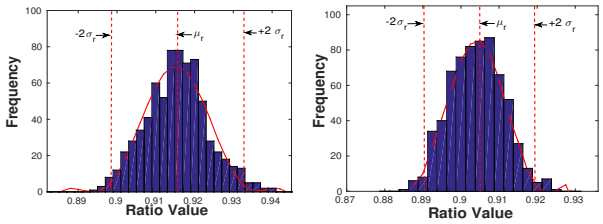
$$\delta_{avg} > max\{\frac{2\sigma}{M} + \frac{\sigma}{\sqrt{M}} \sqrt{\left(\frac{n-M}{n-1} \right)} + 2\sigma, \sigma^2\}$$

.

The theoretical sufficient condition is of little practical significance given the context of our dataset. It is more relevant for very small series data.

VI. Appendix F: Distribution of Ratio Metric Samples across Multiple Datasets

Fig. 4(a) and Fig. 4(b), shows similarities between the standard deviations in the ratio metric when compared to the Irish dataset, although the mean values are different when compared to the Irish dataset. This means that the safe margin design works across datasets due to the high stability of the ratio based metric.



(a) (b)

Fig. 4. Ratio Distribution (Texas Data): (a) 200 meters (b) 800 meters