A Diversity Index based Scoring Framework for Identifying Smart Meters Launching Stealthy Data Falsification Attacks

Shameek Bhattacharjee
Western Michigan University
Kalamazoo, MI, USA
shameek.bhattacharjee@wmich.edu

Praveen Madhavarapu
Missouri Univ. of Sc. & Tech.
Rolla, MO, USA
vmcx3@umsystem.edu

Sajal K. Das
Missouri Univ. of Sc. & Tech.
Rolla, MO, USA
sdas@mst.edu

ABSTRACT

A challenging problem in Advanced Metering Infrastructure (AMI) of smart grids is the identification of smart meters under the control of a stealthy adversary, that inject very low margins of stealthy data falsification. The problem is challenging due to wide legitimate variation in both individual and aggregate trends in real world power consumption data, making such stealthy attacks unrecognizable by existing approaches. In this paper, via proposed modified diversity index scoring metric, we propose a novel information-theory inspired data driven device anomaly classification framework to identify compromised meters launching low margins of stealthy data falsification attacks. Specifically, we draw a parallelism between the effects of data falsification attacks and ecological balance disruptions and identify required mathematical modifications in existing Renyi Entropy and Hill's Diversity Entropy measures. These modifications such as expected self-similarity with weighted abundance shifts across various temporal scales, and diversity order are appropriately embedded in our resulting framework. The resulting diversity index score is used to classify smart meters launching additive, deductive, and alternating switching attack types with high sensitivity (as low as 100W) compared to the existing works that perform poorly at margins of false data below 400W. Our proposed theory is validated with two different real smart meter datasets from USA and Ireland. Experimental results demonstrate successful detection sensitivity from very low to high margins of false data, thus reducing undetectable strategy space of attacks in AMI for an adversary having complete knowledge of our method.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; **Artificial intelligence**; • **Security and privacy** → **Trust frameworks**; **Intrusion detection systems**; • **Hardware** → **Smart grid**;

KEYWORDS

Anomaly Detection; Interpretable ML based Security; Explainable AI based Security; Information Theory; Data Falsification Attacks; Smart Meters; IoT security; Bio-Inspired ML Approaches

ASIA CCS '21, June 7–11, 2021, Hong Kong, Hong Kong

ACM ISBN 978-1-4503-8287-8/21/06…$15.00
https://doi.org/10.1145/3433210.3437527

ACM Reference Format:
Shameek Bhattacharjee, Praveen Madhavarapu, and Sajal K. Das. 2021. A Diversity Index based Scoring Framework for Identifying Smart Meters Launching Stealthy Data Falsification Attacks. In *Proceedings of the 2021 ACM Asia Conference on Computer and Communications Security (ASIA CCS '21), June 7–11, 2021, Hong Kong, Hong Kong*. ACM, , USA, 14 pages. https://doi.org/10.1145/3433210.3437527

1 INTRODUCTION

As illustrated in Figure 1, the Advanced Metering Infrastructure (AMI) in a smart grid is composed of a collection of *smart meters* (end-point IoT devices) that collect power/energy consumption data from customers [19]. The smart meters periodically send data via a hierarchical communication network to various data management servers belonging to the utility company. These servers process data from smart meters for critical operations such as automated billing, load forecast, daily and critical peak shifts, and automated demand response [32]. Therefore, the integrity of the data from individual smart meters is pivotal to the success of the smart grid.

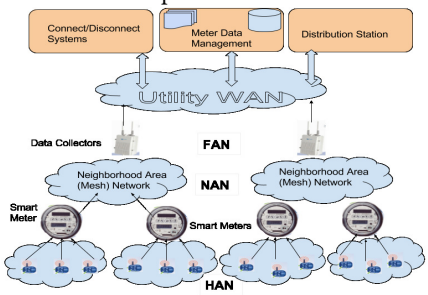


Figure 1: AMI Network Architecture

However, orchestrated cyber or physical attacks on smart meter data are getting increasingly likely as the smart meters are being connected to web-based Energy Management portals [32] and smart home networks at the customer's end [13]. An orchestrated cyber attack can compromise several smart meters connected to the same feeder and then spoof false power consumption readings. Real evidence of orchestrated large scale physical attacks on smart meters was reported in Puerto Rico [33, 34] where several hundreds of smart meters were tampered via an optical probe toolkit by collusion of customers and utility insiders, thus false power consumption data inflicted huge losses to the concerned utility.

1.1 Motivation and Key Challenges

Identification of those smart meters involved in the injection of false power consumption data is a key security challenge. The parameter quantifying the extent of falsification from original data

per meter device is termed as *'average margin of attack strength'*. Lower margins of attack strength are stealthier and hence harder to detect. Additionally, the parameter that quantifies the total percentage of such compromised smart meters in a micro-grid is termed as *'attack scale'*. Orchestrated and coordinated attacks often have larger attack scales compared to the isolated attacks. Moreover, a smart adversary can find a cheaper exploit to compromise smart meters, thereby allowing even a lower margin attack to have a significant impact on the utility when compared to adversary's total attack cost. Orchestrated attacks are usually launched by organized and stealthy adversaries (business competitors, organized cyber criminals). They expect to lower the margins of data falsification per meter by hiding behind the randomness of smart meter data, such that meters are not easily caught. Rival nation states may be motivated to launch organized attacks, since meter data dictate the generation and distribution of electricity to critical infrastructures.

Our analysis of existing works in smart meter data falsification found that most methods fail to identify meters when the attack margins are below 400W, regardless of the attack types and strategies (elaborated in the threat model section). Additionally, the possibility of physical attacks causing data falsification (optical laser attacks [34] and acoustic transduction attacks [5, 6]) render cryptography based and network traffic based intrusion detection methods on IoT devices inadequate. Cybersecurity practices such as static analysis, and signed software updates do not protect against a sensor recording false data since the physical attacks influence the output of sensor hardware that is trusted by software/firmware [6, 23]. Furthermore, several studies [31, 41] have noted that embedded/hardware/in-situ security of smart meters that provide some protection against physical attacks, is not cost effective due to the large scale nature of meter deployment and variety in commodity hardware. Therefore, providing a device level data driven behavioral anomaly scoring technique is necessary not only as an extra level of security, but as a principle approach for trusting the data from a distributed set of IoT devices (e.g. smart meters), which motivates our approach.

However, non-typical but benign conditions can cause changes in data due to weather conditions, and seasons; and low margin data falsification attacks can easily hide behind such randomness. Our anomaly based detection approach should distinguish between such events and changes that are caused due to attacks. Finally, the method has to generalize across various attacks and datasets.

1.2 Contributions of This Work

In this paper, we propose a novel information-theoretic anomaly scoring framework, called *Modified Diversity Index Scoring*, that captures smart meters launching additive, deductive, and alternating switching attack types across a wide range of very low to very high margins of attack strengths and attack scales, while also lowering false alarms and missed detection, compared to existing approaches, for various stealthy attack strategies.

Specifically, we first establish an analogy between the intelligent data falsification attacks in smart meters and the monitoring of ecological balance of species distributions in a geographical region. Next, we show that information-theoretic approaches, such as Renyi and Tsallis Entropies (popular in ecology), Shannon's Entropy and Kullback-Leibler Divergence common in computer security; are not sufficient to address this problem. Thereafter, by studying the effects of various attack types on the probability of relative abundance of each discretized space of the random variable of power consumption, we identify the need for modifications to the existing information theoretic measures. To this end, we introduce modifications to the concept of Renyi Entropy and Hill's Diversity Entropy by embedding a notion of a weighted expected self-similarity mapping of a smart meter IoT device across multiple temporal scales. Next, we embed an appropriate order of the entropy and a weighted relative abundance vector to capture subtle drifts in the horizontal, vertical and incline directions in the probability space, thereby resulting in a diversity index score. The higher the diversity index score, the more likely is the meter launching data falsification attacks. Thereafter, we offer a supervised approach to the learn the parameters of our proposed model that maximizes the separation of diversity index scores between the set of labeled compromised and honest meters, accompanied by cross-validation.

We validate the proposed framework with multiple *full year* real datasets, demonstrating its generalization across a wide range of attack strengths, scales, types, and strategies, across seasons. Experimental results show that our method exhibits lower false alarms and missed detection even when the average attack strengths *per meter* lower than $400W$ (which causes evasion in previous defenses) for both Texas Dataset (200 meters) and Irish Dataset (1300 meters). Specifically, we show that model generalizes to successfully detect deductive, alternating switching attacks and strategies that were not used to train the model. A comparison with existing works exhibits improved performance in terms of reduced undetectable attack strategy space when the attacker has knowledge of our method. We also provide a tradeoff between impact of missed attacks versus cost of base rate false alarms (when there no attacks in the test set).

The paper is organized as follows: Sec. 3 and Sec. 4 describe the system and threat models along with the datasets used. Sec. 5 presents the proposed method, while Sec. 6 discusses the optimal choice of parameters and threshold choice. Section 7 describes experimental results and the final section offers conclusions.

2 RELATED WORKS

Existing approaches for detecting smart meters launching data falsification attack can be broadly divided into three categories, (i) classical machine learning, (ii) information-theoretic, (iii) consensus-based statistical approaches.

The classical machine learning methods use SVMs [10], decision/regression trees (DRT) [9], and neural networks [4, 12]. In [10], the problem was investigated using SVM with a radial basis network when $\delta_{avg} = 450W$, but the percentage of compromised meters assumed was just 1%. In the DRT method [9], the approach does not parameterize the attack strategy space fully, and the attack strength and scales are unclear. Surveyed by[8], the false alarm rates reported by most neural network based methods are much higher even as they do not generalize for an unbounded attack strategy spaces and low profile attacks.

Information-theoretic approaches proposed in [1, 14] use Kullback-Leibler divergence to classify compromised meters, and is a competing approach to our model. Hence, we will show how our attacks perform under these defenses and our solution's detection performance compare with these approaches.

Consensus-based approaches used classical statistics [25, 26], time series [18], robust statistics and density-based learning [3] to identify such smart meters. We chose to compare with [3], since it outperforms the others, and parameterizes the attack strategy space with attack scales and strengths. State estimation based methods are not used since they depend on putting extra monitoring hardware in the higher layers of a smart grid. Note that [2] applies to stealthy attacks at a micro-grid level and but not at the meter level, thus it does not feature in our comparisons. The comparison of our method with existing research is provided in Sec 7.4.

3 SYSTEM MODEL AND DATASETS

Here we present the AMI architecture as a proof of concept for an IoT network, the real AMI datasets used, and the rationale behind their choice for validation of our proposed framework.

AMI Architecture: Let us consider a typical AMI micro-grid, where each house is equipped with a smart meter that records aggregate power consumption data (from all appliances inside the home) and periodically (e,g., hourly) sends them to the utility company over the AMI communication network. The AMI network typically consists of a Neighborhood Area Network (NaN) Gateway that aggregates data from multiple smart meters. Data from multiple NaN gateways are aggregated by a Field Area Network (FAN) gateway and sent to the utility wide area network. The FAN gateway may also host edge computing capabilities.

Let N be the total number of smart meters in a micro-grid, such that $i \in \{1, N\}$ uniquely identifies a smart meter ID. Formally, the actual power consumption of a i-th house at the t-th 'time slot' is denoted by $P^i_{act}(t)$, where t is slotted 'hourly'. If a meter is not compromised, then the actual power consumption is equal to the advertised power consumption P^i_t sent to the utility.

Choice of Datasets: We have used two real AMI datasets to validate the proposed framework. The first dataset is Ireland Social Sciences Data Archives [30] containing 5000 meters from six regions in the city of Dublin, Ireland, collected between 2009-2010. Three out of these six regions, have more than 1000 smart meters. The rationale for choosing this dataset is to investigate the scalability of our framework for large micro-grids. The second dataset is Pecan Street Project [29] containing hourly power consumption data from 215 houses from a Solar village in Texas, USA, collected between 2014-2016. Hence, we have chosen two datasets that are inherently different in terms of their geography, climate, randomness, and extreme difference in sizes.

4 THREAT MODEL

We assume an organized adversary that orchestrates data falsification attacks from multiple smart meters via cyber or physical exploit [2]. Smart meters receive power consumption from various appliances via the Home Area Network (HAN) and sends it to the utility side via the Neighbourhood Area Network. Either the (i) input to the smart meter, (ii) the power consumption data at rest inside the smart meter, (iii) or data in flight may be falsified. An example of falsifying power consumption data at rest is the Puerto-Rico Grid Attack of 2012, where hundreds of smart meter's optical ports were manipulated using laser probes by utility insiders [33, 34], causing the smart meters to record lower than actual power consumption. Similarly, load altering attacks reported in [20], have shown the possibility to change the inputs from appliance loads to the smart meter. Similarly, the data in flight from multiple smart meters to the NaN gateway may be falsified by a traditional man-in-the middle attack. Finally, another possibility is an organized adversary that controls a set of smart meters like a Botnet, collect data from intercepted smart meters, and inject advanced data falsification strategies, that we discuss under the stealthy attack strategies.

Our approach is agnostic of the exploit used to falsify the data. Of course, depending on the exploit the attack scales, strengths, and strategies will vary. Our intention is to capture various kinds of data falsification attack realizations instead of a specific one, since exploits tend to evolve over time and just because an attack has not been realized before, does not mean they will not be experienced in future. We capture this generic data falsification attack landscape by parameterizing the attack strategy space; taking into account the full range attack scales, strengths, strategy combinations in this section. The following features characterize our threat model:

Attack Scale: The fraction of compromised meters, $\rho_{mal} = \frac{M}{N}$, is the *attack scale*, where M is the number of unique smart meters compromised by an organized adversary in a given network. Traditional use of Kullback-Leibler Divergence (KLD) model with statistical aggregates work well, if $\rho_{mal}\%$ [1, 21] are smaller. However, resilience against higher ρ_{mal} has been reported only when associated margins of false data per meter is too high (which facilitates easier detection). However, if the attack budget is high, or a creative adversary finds a cheaper exploit to compromise a meter, or the network size is smaller, then the attack budget constraint does not automatically imply a lower fraction of compromised meters [1]. This is because, in reality, the value of M depends also on the creativity of its exploit, and the micro-grid size N. Given large values of ρ_{mal} are possible in the real world, we take into account a wide variation of ρ_{mal} between 0.10 to 0.90.

Average Margin of Attack Strength: Average margin of false data is the average extent of falsification introduced per meter. We observed that in most previous works, the average margin of false data is not parameterized as a variable except in two recent works [1, 3], which report that these methods completely fail to detect meters when their average margin of false data is $\delta_{avg} < 400$. This happens because the standard deviation of data streams are high (430W-480W in AMI applications) due to randomness of human activity, making it difficult for previous methods to achieve success. We have included a real case study in Appendix A, showing that attack strengths as low as 100W per meter create a significant attack impact on the AMI utility. In our model, we consider an unbounded δ_{avg} value to show that our method reduces undetectable strategy space of attack strength.

Attack Types: We consider three different attack types. The adversary seeks to falsify original data points $P^i_{act}(t)$ representing actual energy consumption at time t by some factor δ_t , where $\delta_t \in [\delta_{min}, \delta_{max}]$ and the long term average value of δ_t is δ_{avg} (avg. margin of attack strength).

(i) Additive Attacks: Here the smart meters seek to increase the data from its original values, such that $P^i(t) = P^i_{act}(t) + \delta_t$. Motivation of such attacks are discussed in [20].

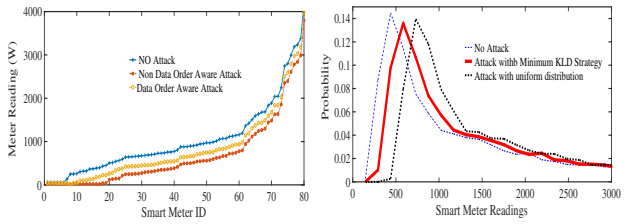
(ii) Deductive Attacks: Smart meters seek to decrease the data from its original values, such that $P^i(t) = P^i_{act}(t) - \delta_t$; this is equivalent to electric theft and the most commonly seen attack type [34].
(iii) Alternating Switching: In such an attack, every compromised meter alternates between launching additive and deductive attacks with the same margin of false data at different times of the day to take advantage of dynamic pricing/demand response of electricity. When the prices are high (due to higher demand), it launches a deductive attack, while compensating with an equal margin additive attack when the pricing is low (due to lower demand), causing the mean consumption trends from individual compromised meters practically unchanged. This is device level equivalent to a camouflage attack reported in [2] from two sets of meters in the *same time*, thus blinding a micro-grid level anomaly detector. However, our variation of camouflage attack is launched from the same end point meter to camouflage the end device (meter) level detectors.

Stealthy Attack Distribution Strategies: Now we focus on 'how' false data is introduced in the smart meters data streams. Apart from a non-stealthy random bias, we analyze our solution against four stealthy strategies, viz. (i) the data order aware, (ii) incremental ramp, (iii) KLD minimization (iv) persistent strategies. AMI applications are not real time systems; they can tolerate some delay. Therefore, if there is some timing delay due to coordination for the stealthy strategies, it is still practical. We assume that a reasonably organized attacker will have an idea of the data distributions and mechanisms used by usual anomaly detectors, and craft the following strategies accordingly:
(i) Data Order Aware Strategy: It is a stealthy falsification strategy that minimizes the chance of detection against mechanisms utilizing proximity (e.g., Euclidean L_2 distance) between the reported and original data distribution, while keeping the same δ_{avg}. Additionally, this strategy makes sure that the maximum and minimum values in the original and falsified distribution are not different, to prevent obvious statistical outliers.

The following strategy is implemented in the following manner: At any time slot t, the adversary sorts the actual recorded data vector from its compromised set of devices such that $P^{(1)}_t(act) \leq, \cdots, P^{(m)}_t(act), \leq P^{(M)}_t(act)$; as well as its corresponding bias vector $\delta^1_t(min) \leq, \cdots, \leq \delta^M_t(max)$. Under an additive attack, the minimum actual data is changed with the highest $\delta_t(max)$, while the maximum observed data is modified with lowest $\delta_t(min)$, and so on like an inverse matching, such that $P^{(1)}_t(act) + \delta_t(max), \cdots, P^{(M)}_t(act) + \delta_t(min)$, subject to the fact that it does not violate bounds on the historical distribution. For a deductive attack, the maximum recorded data is modified by matching with the maximum bias $\delta_t(max)$, while the lowest actual recorded data is altered with the lowest $\delta_t(min)$. For alternating switching attack, the additive and deductive attacks alternate with the strategy mentioned above.

In Fig. 2(a), the blue line corresponds to the non-attacked value of compromised meters. The yellow and red lines correspond to a realization of falsified data under a data order aware and non-data order aware strategy *with same* δ_{avg}*=200W and* ρ_{mal} = 40% for 'deductive' attacks from Texas dataset. The same revenue impact is achieved with both strategies, but chances of detection (using proximity/distance/similarity) are smaller in data order aware strategy.

The width of the interval of $\delta_t \in [\delta_{min}, \delta_{max}]$ is known as the *aperture of attack*. The aperture is varied as necessary to minimize the euclidean and KLD.
(ii) Incremental/ Ramp / Boil-frog Strategy: This strategy involves a very gradual increase in of δ_t bias over time, until intended δ_{avg} is reached. This attack strategy is termed as boil-frog in AI security and ramp attack in cyber physical system (CPS) security. The strategy causes all temporal metrics to record minimal changes that evolve over time to bypass detection.
(iii) KLD Minimizing Strategy: The falsified data is injected in a manner which minimizes the KLD, while preserving the target δ_{avg}. Fig. 2(b) shows an illustration for a single meter where the adversary crafts a distribution (bold red line) that minimizes the KLD; thus being closer to the actual data distribution (blue line) than to a uniform random bias attack (gray line), even when the $\delta_{avg} = 200$ for both attack strategies.
(iv) Persistent Strategies: In Section 7.5, we provide a list of strategies launched by an adversary that knows our defense model and launches evasion attacks. We show performance under such evasion attacks, my showing the extent to which undetectable strategy space is reduced, and break even time of adversary.



(a) (b)

Figure 2: (a) Data Order Aware (b) KLD Minimizing

Attacker Knowledge and Assumptions: We are aware that adversaries will have full knowledge of our defense mechanism when published. In Section 7.5, we showed the performance of our method in terms of reduction in undetectable attack strategy, assuming the adversary has knowledge of historical data and our method.

Our base assumption is that the defender has a significant portion of training set that is not attacked. As for this paper, we have only focused on evasion attacks that occur in the testing phase and ignored possibility of poisoning attacks. This is because reducing the undetectable attack strategy space itself is quite a challenging problem. Active poisoning attacks will be part of our future work.

Finally, since CPS application security is still a new field, datasets containing real attacks are unavailable. Therefore, we have parameterized the entire attack strategy space in terms of attack strength, scale, strategies, and types, and reported the failure bounds. This should alleviate concerns over non-availability of real attacks.

5 PROPOSED FRAMEWORK

The technical contribution of the paper is divided into three parts: (i) *Theoretical Intuition ((Sec 5.1 to 5.4))*: which describes a high level intuition of the proposed theory; (ii) *Proposed Methodology (Sec 5.5 to 5.9):* which describes our proposed information-theoretic framework for the trust scoring; (iii) *Parameter Optimization and*

Trade-offs (Sec 6): describes the trade-offs associated with choosing various optimal parameters in the proposed methodology.

We first discretize the power consumption random variable into bins/categories/species, find the relative abundances, and introduce the notion of Renyi Entropy and Hill's diversity index. We discuss why KLD fails to detect attacks in our threat model. Then, we show that Hill's index is not enough, but a promising starting point for solving our problem. Subsequently, we show how very low to high strength margins of attacks of various attack types, introduce different kinds of subtle shifts in the distribution of species and identify the 'factors' that need to be tracked/added into the existing Diversity and Renyi entropy. Then, using a bottom-up approach we propose the trust scoring classification framework that embeds those factors such that produce linearly separable scores, that can indicate smart meters whose data is falsified. Finally, we discuss the learning of parameters from data which maximizes the separation between scores of compromised and honest labeled meters.

5.1 Theoretical Intuition

In this subsection, we provide a theoretical intuition of why our framework is required and at a high level why it works.

Discretizing Data Distributions: The range of dataset from the smallest (C_{min}) to the highest recorded value (C_{max}) is gathered over all devices. The continuous range $[C_{min}, C_{max}]$ of reported data is divided into R number of discrete partitions of equal width termed as species width (denoted by sw watts). Each of these discrete partitions is individually visualized as distinct outcome categories known as 'species' or bins. The terms outcome category and species ID are used interchangeably.

Each outcome category is uniquely identified by a species ID $s \in \{1, R\}$. The formation of species categories and corresponding species width sw is common for all devices, whose value needs to be optimized. For AMI, if $sw = 100W$, then 0W-100W forms the first species ($s = 1$), 101W-200W forms the second species and so on. The total number of species R, depends on the species width sw and the data range. The optimal species width is a parameter that affects classification will be discussed later.

Probability of Relative Abundance of Species: We visualize the species as a random variable associated with a probability. This probability simply indicates the probability of reported data belonging to one of these species categories, and is termed as the 'probability of abundance' p_s^i of each species. Mathematically, we use the Bayesian interpretation of posterior relative abundance to calculate each of the p_s^i:

$$p_s^i = \frac{\eta(s)^i + 1}{R + \sum_{s=1}^{R} \eta(s)^i} \tag{1}$$

is the probability of abundance of the s-th species in i-th meter, where $\eta(s)$ is the total frequency of datapoints observed in that S, and $\sum_{s=1}^{R} p_s^i = 1$. Mathematically, for a given smart meter i, the probability of relative abundance of species is a column vector is denoted as $\boldsymbol{p}$, such that $\boldsymbol{p}^T$ is the corresponding row vector $\boldsymbol{p}^T = \{p_1^i, \cdots, p_s^i, \cdots, p_R^i\}$ collectively denote the relative abundances of all species for a given meter. From now on, since the following method applies to any meter i, we will drop the suffix i from our notations for simplicity.

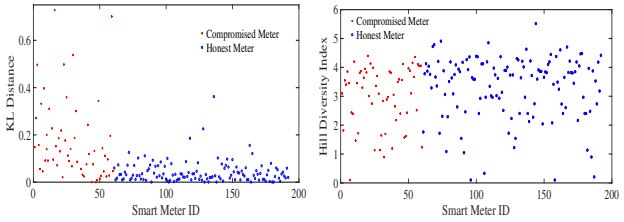
Weaknesses of KLD Approach: Information theory plays a central role in most learning and classification methods for security. However, in this section, we show that the typical information-theoretic approaches, do not work effectively in classifying devices that launch data falsification due to the high variance, shifting trends, low margins of attack, data order aware strategy for all attack types.

The H^i and the KLD^i denote the Shannon entropy and Kullback-Leibler Divergence of a smart meter i:

$$H = -\sum_{s=1}^{R} p_s log(p_s); \qquad KLD = -\sum_{s=1}^{R} p_s log\left(\frac{q_s}{p_s}\right) \tag{2}$$

where, R is the number of bins in the discretized data distribution and s denotes any such bin, p_s is the probability of observing a datapoint in the s-th bin, q_s denotes the same observed in a different time or space. Both [1, 14] use various bin partitions and then use KLD between the two distributions before and after an attack to find compromised meters.

In this paper, for the context of small margins of false data in smart metering data, we verify how KLD based methods, that has its roots in Shannon entropy perform. Shannon Entropy views each species as distinct without relationship among them and also treats each species as contributing equally to the overall score. However, in reality, such measures cannot track subtle changes in them. Fig. 3(a), shows the result where we partition the data into several small partitions and take the Kullback-Leibler Divergence of the distributions over all the partitions before and after the attack for a margin of false data of 200W. This resulted in scores that are not linearly separable as shown in Fig. 3(a).



(a) (b)

Figure 3: Failure to Classify (δ_{avg} = 200W, Texas Data, Additive): (a) KLD (b) Hill's Diversity Index

5.1.1 Renyi Entropy and Hill's Diversity Entropy. The Renyi entropy introduces the notion of 'order' q in the entropy measure. The 'order' mathematically allows one to embed the importance of certain rare or abundant outcome categories (or species) in the $\boldsymbol{p}$. As a special case, the Renyi entropy of $q \to 1$ converges to Shannon entropy. The Renyi entropy is mathematically defined as:

$$H_q(\boldsymbol{p}) = \frac{1}{1-q} ln\Big(\sum_{s=1}^{R} p_{(s)}^q\Big) \tag{3}$$

The exponential of the Renyi entropy of the q-th order is also known as the Hill's Diversity index of the order $q \in \mathbb{I}$ for such a community and is given by:

$$e^{H_q} = D^{(q)} = \Big(\sum_{s=1}^{R} p_{(s)}^q\Big)^{\frac{1}{1-q}} \tag{4}$$

where q is known as 'order of diversity', which controls the sensitivity of the diversity index measure to the most common or the

rarest species. A diversity index of $q = 0$ is completely insensitive to species frequencies and treats all species equally. All species $q < 1$ favors disproportionately the rarer species. All species $q > 1$ disproportionately favors the most common species. If $q = 1$, the diversity index proportionally favors species relative abundances. While this offers more math provisions for capturing subtle changes, the Fig. 3(b), shows no visible difference still in the attacked and non-attacked meters for $\delta_{avg} = 200$. Now let us discuss the effect of each attack on the species distribution.

5.1.2 Effect of Data Falsification: A Diversity Parallelism. Now we provide intuition on the effect over the dynamics of species abundance, under data falsification, with a low δ_{avg} and data order aware strategies. At the same time, the effect of intermediate and high δ_{avg} cannot be ignored, given that the defender has no idea on the value of δ_{avg} in the real world. We show how additive, deductive, switching attacks, introduce subtle *shifts* in various directions. We assign some nomenclature for these shifts, which needs to be accounted for in the trust classification method for successful identification.

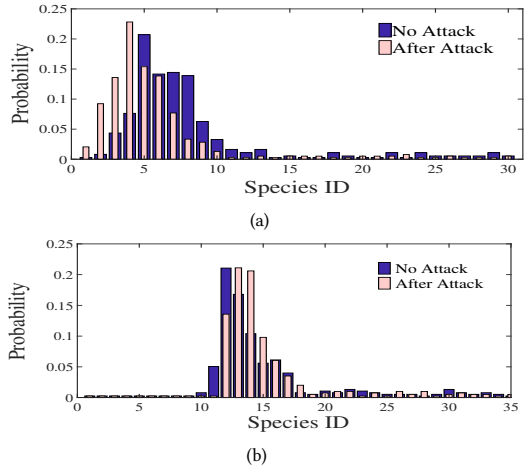


Figure 4: Effects of Attacks on Species Distributions: (a) Deductive (b) Additive

Deductive Attacks: Fig. 4(a), shows an illustrative example of a *deductive* attack with a $\delta_{avg} = 140W$. The 'blue bars' denote the probabilities of relative abundance 'before the attack'. The higher the blue bars, the least rare (or most abundant) the species, while the species with smaller blue bars, are more rare (or less abundant). The 'pink bars' in Fig. 4(a), denote the new probabilities of relative abundance of each species 'after the attack'. A comparison of blue and pink bars shows that there is a minimal horizontal shift in the species distribution. Certain species such as species IDs 1,2,3,4 that are intermediately rare before the attack become more abundant after the attack. However, this increase in abundance is contributed by a mixture of both abundant and rare species categories between 5 and 20, which experience a slight decrease in their probabilities of abundance (vertical changes). Hence, we conclude that change in rarer species is significant and can be visualized as an incline shift (*change in intermediately rare species*) which includes a mixture of horizontal (increased quantity diff aka number) and vertical shifts (increased quality difference aka probability) in the rarer species. The concentration of the resultant distribution is less, due to a decrease in the rarity of species. Any further increase in δ_{avg} is going to make the distribution, less concentrated.

Additive Attacks: Fig. 4(b), shows an illustration for an *additive* attack with $\delta_{avg} = 120W$. The blue and pink bars correspond to $\boldsymbol{p}$ before and after attacks. Here, the rarer species 1-10 almost disappear, while the most of the abundant species (11-16) become even more abundant. Between species ids $20 - 30$ (which are rarer), there is a no vertical increase in probabilities, barring a few with a slight increase. This gives the opposite effect compared to the deductive attack. The datapoints after attack become more concentrated into a smaller number of species categories because of an increase in the most abundant species, while decreases are seen in rarer species, which needs to be captured. However, when δ_{avg} is just large enough, this trend is reversed, and the distribution becomes less concentrated again (though in the opposite direction), and there is a small vertical change in each species, but spanned over a large number of rarer species (horizontal shift). This effect is shown in Fig. 5(b), for an additive attack of $\delta_{avg} = 240W$ that causes small increases in the probabilities across *many* rare species, which needs to be captured.

Alternating Switching Attacks: Fig. 5(a), is the trickiest attack type and shows two important but minor shifts in the distributions before and after an alternating switching attack of $\delta_{avg} = 180W$. The shift is in the *intermediately rare species* on either side of abundant species (resultant incline shift). In contrast, if there is an additive attack of high magnitude then the relative abundance of higher species will increase and lower species will decrease. However, the net effect in the change of relative abundance may cancel out. Therefore, it's true that relative abundance that measures the vertical heights of the species change is not enough. We also need a measure that captures the horizontal shifts and incline shifts.

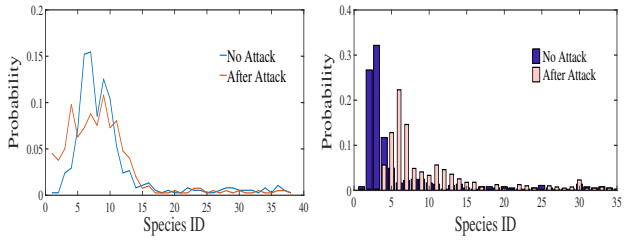


Figure 5: Effects of Attacks on Species Distributions: (a) Alternate Switching Attack (b) Large δ_{avg} Additive

5.1.3 Summary of Conclusions:. From the above study, we make the following conclusions: (i) Small δ_{avg} causes changes in a small number of intermediately rarer species. (ii) Higher δ_{avg} causes small changes in a large number of rarer species. (iii) The vertical shift in terms of the change in probabilities of abundance is the third thing that needs to be kept track of (introduced later as expected self-similarity).

From the above, it is clear that the number of rarer species experiencing small changes is a critical factor that needs to be embedded in the scoring framework. This is loosely referred to as 'quantity' of rare species since the adversary can craft an attack introducing very small vertical changes across all/large numbers of species. Additionally, the rareness of a particular species itself is important, because,

for very low δ_{avg}, the intermediately rare species are likely to get more effected than the rarest of species. This is loosely referred to as 'quality' of the change in species.

For lower δ_{avg}, the horizontal shifts are less pronounced but vertical and incline shifts (intermediate rare species) are most important and vice-versa. The quantity of rarer species and quality depicting the extent of change both needs to be embedded, which introduces a notion of dual weights that needs to be embedded in this existing diversity index measures.

5.1.4 Functional Form of Modified Hill's Index:. Following the above conclusions, we introduce modifications into existing Hill's Diversity Index measure in a way that embeds new variables that can track changes in the number of rarer species (horizontal changes), the change in the individual species abundance with itself (vertical changes), in a way that gives magnified importance to those changes observed in intermediately rare species (incline shifts) as compared to other species. All three factors ensure that the resultant modified diversity score (say rD) can separate the compromised meters from honest ones. Mathematically, the original equation of Hills Diversity is given by the following:

$$D^{(q)} = \left(p_1^q + \cdots + p_s^q + \cdots + p_R^q\right)^{\frac{1}{1-q}} \tag{5}$$

Every term in the summation in Eqn. 5 dictates the contribution of each species category i, to the diversity of a meter under consideration. Considering the conclusions drawn from the summarized effects of attacks, we need to add *atleast* two more factors in Eqn. 5 to each term in the summation of the following function:

$$rD^{(q)} = \left(x_1.y_1.p_1^q + \cdots + x_s.y_s.p_s^q + \cdots + x_R.y_R.p_R^q\right)^{\frac{1}{1-q}} \tag{6}$$

where x_s and y_s are analogous to weights, that embed the effect of the number of rare species, and vertical and incline changes as a combination respectively. Hence, at the end of the day, we would seek to get a formula for a modified diversity index score which confirms to the functional form in Eqn 6. We will later verify, how the final diversity index score confirms to this basic functional form. While building the functional form, we have to be aware of how different attacks, might exploit vulnerabilities, and also take steps to reduce false alarms.

Intuition on Diversity order q: We know that all p_s is between 0 and 1. Any p_s when powered by $q > 1$, the species with higher probabilities (abundant) show a lower resultant decrease than for lower probability species, that show a higher decrease for the same q. This is however, reversed when $q < 1$; where the rarer species see more increase compared to abundant species, when powered by $q < 1$. Since the data falsification attack has relationships with number as well as change of probabilities in the rarer species, this gives the intuition that optimal value of q might be between 0 and 1, which we will verify later.

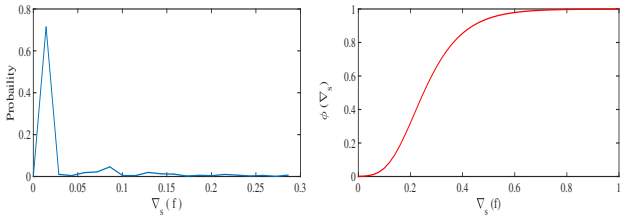
5.2 Modified Diversity Index based Trust Score

5.2.1 Forming Species Self Similarity Matrix:. We build a square matrix $\boldsymbol{D}$ of RxR dimensions known as the *species self similarity matrix*, where only diagonal entries are non-zero, and quantifies the effective level of similarity (or difference) of the relative abundance of a species with itself between the current time window (where the diversity index is being calculated) compared to past windows. The past may be previous years's history or a shorter term history of a set of previous consecutive time windows. For smart city application context, we use consecutive time windows, given the observation that shifting trends in data, can diametrically change the self similarity of species without presence of attacks over yearly time horizons.

To build $\mathbf{D}$, the simplest approach would be an absolute difference between the relative abundance of each species category between the current and the previous time window. Mathematically, let matrix $\mathbf{p}(f-1)$ denote the species abundance in previous time window $f-1$ for the i-th meter and the same at the current time window f is denoted by $\mathbf{p}^*(f)$. Then, the most simple self similarity matrix could be $\mathbf{S(f)} = |p(f-1) - p^*(f)|$, where:

$$\boldsymbol{p}(f-1) = \begin{bmatrix} p_1 & 0 & \dots & 0 \\ 0 & p_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & p_R \end{bmatrix}_{R\times R} \quad \boldsymbol{p}^*(f) = \begin{bmatrix} p_1^* & 0 & \dots & 0 \\ 0 & p_2^* & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & p_R^* \end{bmatrix}_{R\times R}$$

However, we found two problems with this approach. (1) this will fail to detect incremental ramp or boil-frog attack strategies, that cause very small vertical changes over time. Hence, we need to look over a longer time horizon for 'sustained' vertical changes. (2) there could be false alarms, since some of the meters may show a higher change in the legitimate difference of relative abundance in species without attacks in pairs of windows. Without any transformation, it creates a higher change in the eventual trust score even under benign changes, which needs to be avoided.



(a) (b)

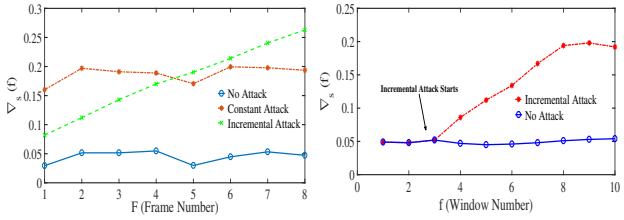
Figure 6: (a) Distribution of Benign Sample $\nabla_s(f)$ (b) The ϕ transformation function (Texas Dataset)

This gives an intuition that once an idea on the bounds of legitimate vertical changes is learned, changes beyond that can be overweighed, while changes below those bounds can be discounted. These two aspects are embedded in the following way: Let the difference between relative abundance vector between any two consecutive windows be denoted by $\epsilon_s(k) = p_s(k-1) - p_s(k)$, a shorter term similarity. Then we keep a long term memory of ϵ_s for each species represented by:

$$\nabla_s(f) = \sum_{k=f-F}^{f} \epsilon_s(k) \tag{7}$$

such that $\nabla_s(f)$ keeps the cumulative sum of the differences observed between pairs of time windows for a sliding frame containing F previous windows. When there are no attacks, $\nabla_s(f)$ has no increasing trend (see Fig. 7(a)) and the values are typically very small (See Fig. 6(a)). Infact, across an appropriate frame length (F), the ∇_s flattens out (blue lines in Figs. 7(a) and 7(b)). In contrast, for incremental attacks, there is a small monotonic increasing trend

in ∇_s (green and red lines in Fig. 7(a) and 7(b) respectively). For all other strategies, the average ∇_s is larger, under attacks.



(a) (b)

Figure 7: (a) Effect of Varying Frame Length (b) Reference Frame Tracking under Incremental Ramp Strategy

The species self-similarity matrix is given by $\boldsymbol{D}(f)$ such that each diagonal element is computed through a function of the form $(\phi(\nabla_s(f))$, such that the diagonal elements in $\boldsymbol{D}(f)$ is a mapping that takes the ∇_s across the frame within each species as the input and mathematically written as:

$$\boldsymbol{D}(f) = \begin{bmatrix} |\phi(\nabla_1) & 0 & \dots & 0 \\ 0 & \phi(\nabla_2) & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \phi(\nabla_R) \end{bmatrix}_{R \times R}$$

where each entry

$$\phi(\nabla_s) = \frac{1}{(1 + A_b e^{-B_b(\nabla_s)})^{1/\nu}}; \quad \phi(\nabla_s) \in [0, 1] \tag{8}$$

is a generalized sigmoidal function which inputs the vertical change over a frame of length F at a time window f. The ϕ transformation produces the necessary weighing that reduces the false alarm rate while not sacrificing missed detection. Here the B_b is a growth rate parameter controlling the value of ∇_s for which the $\phi(.)$ function reaches its max value, while ν is a displacement parameter that controls the value of ∇_s, where the $\phi(.)$ function enters the exponential growth phase. The A_b is a parameter that decides the initial y-intercept, when ∇_s is zero. Fig. 6(b), shows the ϕ function.

5.2.2 Expectation of Temporal Self-Similarity:. Now we quantify the overall average change in the similarity of the i-th meter. Let $\boldsymbol{p}$ denote the probability abundance vector of species calculated over a time window in near history (ideally just before attack starts) and the $\boldsymbol{D}$ is a probabilistic measure related to that p (given the design of $\boldsymbol{D}$), so that we get something similar to a second-order expectation where the random variable is itself a probability vector $\boldsymbol{p}$. Mathematically, we do the following operation:

$$\Big[E(D)\Big]_{Rx1} = \Big[\boldsymbol{D}\Big]_{R \times R} \Big[\boldsymbol{p}\Big]_{Rx1} \tag{9}$$

where $E(D)$ is an Rx1 matrix where each element represents the expectation (average) change in the self similarity (in terms of probability of species abundances) of the corresponding species between over this time frame. Each element of the $E(D)$ is of the form $(\phi(\nabla_s) * p_s)$, which gives an idea on the index of vertical change within each species s between two time frames. Let us call $\boldsymbol{p}$ as the reference probability vector.

Now it could be tricky to get the correct reference vector belonging to a frame just before attacks, especially if incremental attack strategy inflicted. However, $\nabla_s(f)$ also allows us to pinpoint this by backtracking the $\nabla_s(f)$ variation (See Fig. 7(b)) and the p is built from before the window just before the change-point of $\nabla_s(f)$.

5.2.3 Diversity Order Embedding:. From theoretical intuition, one requirement was to magnify changes in intermediately rarer species, which we accomplish here. We add the order into the expectation of similarity in a similar way that appears as a power in the Hill's Diversity Index, in order to achieve the embedding of non-uniform vertical change such that we get the following:

$$[M]_{Rx1} = [E(D)]^q \tag{10}$$

$$\boldsymbol{M} = \begin{bmatrix} \Big(\phi(\nabla_1).p_1\Big)^q \\ \vdots \\ \Big(\phi(\nabla_s).p_s\Big)^q \\ \vdots \\ \Big(\phi(\nabla_R).p_R\Big)^q \end{bmatrix}_{R \times 1} \tag{11}$$

5.2.4 Magnifying quantity of species with changes:. From theoretical intuition, we need to finally ensure that very small incremental changes but happening in many unique IDs of rare species, has importance in the resultant functional form of modified diversity index that we are striving to achieve. We put more emphasis on the shifts in rarer species as a weight to each of the species in the $R \times 1$ matrix, $[E(D)]^q$. Note that we need a scalar value for the diversity index trust score and the quantity weight matrix needs to be a $1 \times R$ matrix for the scalar to exist. Hence, we seek to design a weight vector that is $1 \times R$ dimension.

$$[W]_{1xR} = \Big[J\Big]_{1xR} - \Big[\boldsymbol{p}^T\Big]_{1xR} \tag{12}$$

where $[J] = [1 \dots 1]_{1xR}$ is a matrix containing all 1's for R columns, and the intuition is that one minus a rare value will be a high value and too many of these occurrences will push the resulting scalar to a higher portion in the number line. Hence, the resulting weight factor is given by:

$$\boldsymbol{W} = \Big[(1 - p_1) \quad \dots (1 - p_s) \quad \dots \quad (1 - p_R)\Big]_{1 \times R} \tag{13}$$

5.2.5 Final Modified Diversity Index Trust Score:. The diversity index based trust score of the q-th order for a smart meter i is given by the multiplication of $\boldsymbol{W}$ and $\boldsymbol{M}$. The reason they are multiplied is to achieve the functional form we had earlier in Eqn. 6, as we shall see next.

$$TR^i(q) = W \times M = \Big(([J] - [\boldsymbol{p}^T]) \times \Big[\boldsymbol{D}\boldsymbol{p}\Big]^q\Big) \tag{14}$$

One can verify how Eqn. 14 confirms to the original functional form of mathematical abstraction of a diversity index score for identifying data falsification. By plugging Eqn. 11 and Eqn. 13 into Eqn. 14, we get a scalar value due to matrix multiplications of dimensions 1XR and Rx1, which gives:

$$\text{TR}^i(q) = (1 - p_1).(\phi(\nabla_1))^q.p_1^q \quad + \dots + (1 - p_R).(\phi(\nabla_R))^q.p_R^q$$

If we assume $(1 - p_s) = x_s$ and $(\phi(\nabla_s))^q = y_s$, then the above reduces to the desired abstraction of the mathematical functional form, we expressed as required earlier in Eqn. 6.

Hence, to conclude the modified diversity index score of a meter i that can detect compromised meters is:

$$rD^i = \Big(([J] - [\boldsymbol{p}^T]) \times \Big[\boldsymbol{D}\boldsymbol{p}\Big]^q\Big) \tag{15}$$

where $rD^i > 0$, if $q > 0$. The whole exponent factor of $\frac{1}{1-q}$ in the original functional form is ignored since it does not provide any added classification advantage as far tracking changes. Another important point is the nature of change in diversity score after the attack is launched, and its effect on the final distribution of rD^i values of compromised versus honest meters. Due to the nature of Eqn. 15, where changes in each species are added up, the meters launching data falsification will experience an increase in the diversity scores after the attack. In contrast, the non-compromised meters will exhibit a lower diversity score than the compromised meters. We will verify this in the experimental results section.

6 PARAMETER LEARNING AND THRESHOLD

Now that we have the architecture of our base model, we need to provide a generalizable way of learning various parameter values given any dataset. Our approach towards this is a supervised one, where we divide the training set into two parts: first, without any attacks; the second containing attacks from a subset of meters we choose and program them to simulate a limited set of attacks. Our method learns parameters according to a target objective function that maximizes the difference between the diversity index scores of the honest and malicious classes in the training set. Later on, we use cross-validation set to find a threshold and then apply it on a testing set for performance evaluation.

6.1 Training Set Details

We use the full year of 2014 as the training set for Texas dataset. The attack starts after the end of 6-th month. The malicious class labels contain the following attack features: An additive attack with $\delta_{avg} = 100$W, $\rho_{mal} = 30\%$, with an incremental ramp strategy that increases by 20W every 15 days. The idea is that if it detects for the smallest and slowest moving attack, it will be able to detect anything stronger. Other parts of the threat model are not used for training, since we need to verify that our method is *generalizable* to detect 'mutated' and 'unknown' attack realizations that it was not trained on.

6.2 Decision Variables

The controllable decision variables are namely A_b, B_b, ν, sw, q and F which are candidates for optimization. Among these, parameters strongly related to the dataset are B_b and ν, others are weakly related to the dataset. Note that the δ_{avg} and ρ_{mal} are uncontrollable decision variables which are beyond defenders knowledge. However, it is known that if we observe a linear separability between diversity index scores of a compromised and honest set of devices, for a lower δ_{avg}, it will automatically hold for higher δ_{avg} values by virtue of our scoring design. Therefore, during learning, we train with only select candidates of δ_{avg} that are below the desired lower bound of sensitivity δ^{dlb}_{avg}. For tractability of search space, we partition the candidate species widths and candidate δ_{avg} into discrete partitions with upper and lower bounds δ_1 and δ_P.

6.3 Objective (Error) Function

The objective function (or the error/loss function) should maximize the separation between compromised and honest devices, in terms of the distribution of their diversity index scores. Hence, we used the squared difference of average of diversity index scores between the compromised and honest sets in the training set. Intuitively, that combination of parameters/decision variables that maximizes this objective function is the optimal parameter set.

$$e = max\Big(\frac{\sum(rD^h)}{N-M} - \frac{\sum(rD^m)}{M}\Big)^2 \tag{16}$$

$$\text{s.t.} \quad A_b > 0; \qquad 0 < B_b < 1; \qquad 0 < \nu < 1$$

$$\text{s.t.} \quad 0 < q < \infty; \qquad w_1 < sw \leq \delta^{dlb}_{avg}; \qquad 1 \leq F < F_{max}$$

It might seem that there too many variables to optimize. However, in reality, the search space of sw, q, A_b turns out to be bounded and small, once we apply the following pruning logic and design considerations: The candidate species width sw is upper bounded by the desired lower bound sensitivity of attacks δ^{dlb}_{avg}, which is small, making the sw range limited. Furthermore, given the role of the Renyi diversity order, we can prune the search space of diversity order to $q \in (0, 1]$; an explanation provided earlier in theoretical intuition. Appendix C.2 shows that this intuition matches the outcome of the optimization.

The optimization can be solved using a grid search; or an efficient method like gradient descent which scales well when there are many parameters with a wide search space. For gradient descent to work, the error function needs to be transformed into a convex function. Our objective function is a concave function with a global maxima. Such functions can be converted into a convex function using the negative logarithm of the original objective function, and then apply gradient descent. However, accuracy depends on the smoothness of the convex function. In our implementation, the number of parameters is limited, and has a smaller search space either by design or through pruning. Hence, we solved our optimization, using a grid coordinate search method. The explainability of the relationship of each parameter with the objective (error) function is shown in detail in Appendix C.

For Texas data, we found the following (near) optimal parameter values: $\nu = 0.05, B_b = 0.1, q = 0.55, sw = 100, A_b = 0.3$. To cross-check for parameter values for a different dataset, we repeated this process for over the Irish dataset. The first 7 months of the dataset were used as training set, and attack labels were introduced after the end of the 3rd month, using the same attack features as the Texas dataset. We solved the parameters separately and found $\nu = 0.03, B_b = 0.12, q = 0.5, sw = 100, A_b = 0.3, F = 8$ and window length is 15 days. We can observe that ν and B_b are slightly different (due to dataset specifics), while other parameters are closer due to their relationship with attack model and underlying theory.

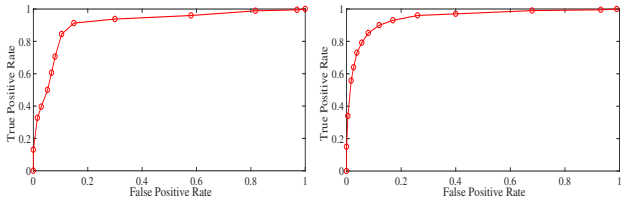
6.4 Threshold Selection

Cross-validation ascertains whether the optimal values generalize well or not to maximize the linear separation of scores, and also learn a classification threshold that generalizes during the testing set. We use a Receiver Operating Characteristics (ROC) curve to get the full spectrum of possibilities of false alarm (FA) to true positive (TP) rates. From this, based on the defender's desirable maximum tolerable false alarm rate, the corresponding threshold giving that FA rate is chosen, and then applied to the testing set for security performance evaluation.

Cross-validation Dataset: For Texas Dataset, we used 2015, partitioned into 12 partitions for cross-validation. For Irish dataset, we used 6 partitions, starting from the 8-th month of 2009. We

average the parameter outputs to provide more accurate estimate of model prediction performance. For Texas dataset, we got: $v = 0.04, B_b = 0.08, q = 0.55, sw = 100, A_b = 0.29$, while For Irish dataset, we got $v = 0.03, B_b = 0.1, q = 0.5, sw = 100, A_b = 0.31$. We used these values to retrofit in the model and generated the diversity index scores of both classes. Then thresholds are varied according to desired false alarm rate.

ROC Curve: Figs. 8(a) and 8(b), shows the ROC curve under a $\delta_{avg} = 100$ from cross-validation, with an AUC of 0.89 and 0.93 respectively. In general, the ROC curves for various δ_{avg} can be plotted. A utility can use his desired maximum allowable false alarm rate and find the corresponding threshold using this ROC.



(a) (b)

Figure 8: ROC in Cross-validation: (a) Texas (b) Irish

7 EXPERIMENTAL VALIDATION

This section includes cross-validation and testing set results of both Texas and Irish datasets for the smart metering application. The experimental result section is divided into the following subsections: (i) Attack Implementation on Test set description; (ii) Performance results (iii) Cost Benefit Analysis (iv) Comparison with other works (v) Verifying robustness to ramp strategy and attack scales (vi) Performance under complete knowledge of defense mechanism

7.1 Attack Implementation on Testing Set

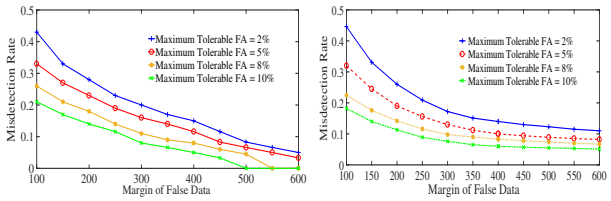
For each attack type, and strategy (discussed in the threat model) we did the following: For the Texas dataset, the 2016 year's data (having a duration of a year), we had five attack start points interspersed approximately by two months to cover the entire testset duration. Similarly, for the Irish dataset, the final five months of the 2010 data were used as a test set, with two attack start points interspersed in a two-month duration. This is done to show that regardless of the start point of attack, the reported missed detection is unbiased. Hence, five (or two) versions of the attacked testing set are obtained for each attack type for Texas (and Irish) datasets respectively.

In each version, we had six different sets of compromised meters per attack scale value (to remove compromised meter selection bias), making a total of 30 (or 12) versions. Each such version is attacked with the indicated several different δ_{avg} (from the compromised ones of course), and then fed to the diversity index model. Then, the final result on missed detection and false alarms is reported by combining the results from all these versions. For reporting baseline false alarm rate (where there are no attacks throughout the year or test duration), we counted the false alarms accordingly. Additionally, note that we have parameterized the space of attack strengths and scales covering all possible values. There is no availability of real attack dataset in this area, but our implementation included the gold standards for performance evaluation covering any gaps that might otherwise exist. Note that deductive, alternating switching attacks attack types, KLD minimizing strategies were not used for training. We put these in test set only to understand whether the method generalizes to previously unseen attacks.

7.2 Performance Results

Instead of ROC curve, we show (i) missed detection (MD) rates across a wide δ_{avg} range, for different thresholds based on user's tolerable FA rate; (ii) the *base rate FA*, which is false alarm rate, when there are no attacks throughout the test set; because most companies have a concern on lowering FA rates (because the prior probability of an actual attack is low). The ROC curve from cross-validation, is used to pick four corresponding thresholds that gave 2%,5%,8% and 10% FA rate; which are then applied to the test set.

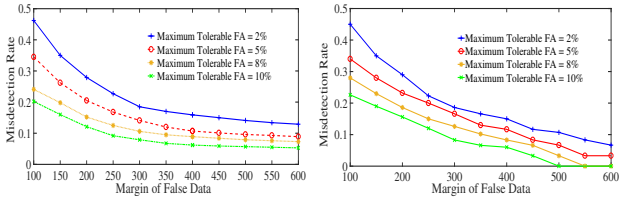
7.2.1 ***Generalizing against untrained Attacks:.*** We first show performance under previously unseen attack types (deductive and alternating switching) across varying δ_{avg} values and the new $\rho_{mal} = 40\%$; threats which did not feature in the training phase, using a data order aware strategy.



(a) (b)

Figure 9: Deductive Attacks: MD rates over Varying Max. Allowable FA ($\rho_{mal} = 40\%$): (a) Texas Data (b) Irish Data

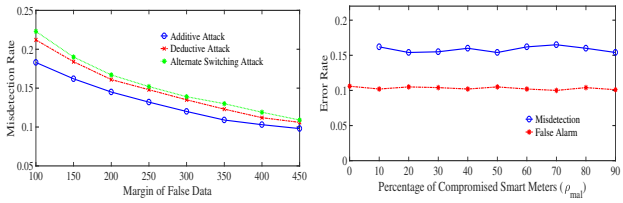
Figs. 9(a) and 9(b) show the MD rates across various δ_{avg} against 'deductive attacks' under the Texas and Irish datasets respectively. Each line corresponds to a performance given by different thresholds corresponding to that particular tolerable base FA rate. Similarly, Figs. 10(a) and Fig. 10(b) show the MD rate 'alternating switching attacks' for Irish and Texas datasets, respectively. To verify, that performance is also valid over the additive attacks and ramp strategy (on which we trained our model) please refer to Appendix D and Appendix E. We will show in Sec. 7.4, at these δ_{avg} values our missed detection rates are much smaller than to previous research.



(a) (b)

Figure 10: Alternating Switching: MD rates over Varying Max. Allowable FA($\rho_{mal} = 40\%$): (a) Irish Data (b) Texas Data

Performance against untrained KLD minimization strategy Fig.11(a) at tolerable FA rate of 10%, is shown for the 3 attack types. The performance is slightly worse compared to the data order aware strategy. The increase in mis-detection rate on average for the KLD minimizing strategy across all attack types and δ_{avg} values, is 7.3% keeping the same FA rate. The Fig. 11(b) shows that our method scales well and is invariant to changing ρ_{mal}.



(a) (b)

Figure 11: Performance: (a) KLD Minimizing Strategy (b) Invariance to Attack Scales

7.2.2 ***False Alarm Performance:*** A concern on anomaly based scoring frameworks are false alarms and their costs. A summary of *base rate false alarm performance in the testset* is included in Table 1. The Texas dataset has more shifting trends (due to renewable penetration), thus it has more base rate FA than Irish dataset.

Table 1: Base Rate False Alarm Percentages in test set

Tolerable FA Threshold	Irish Test Set FA	Texas Test Set FA
2%	2.11%	2.60%
5%	5.33%	6.25%
8%	8.86%	9.37%
10%	10.58%	10.93%

7.3 Cost Benefit Usability of our Performance

Here we analyze the *costs of MD and FA rates* from the perspective of real life usability. Once inferred as attacked, an audit trail is done by utilities on each device for confirmation. According to [37], audit inspections are billed for a median cost of CA = \$141 per device, while [36] reported the average time to inspect each meter device is 55-65 minutes. Audits are an annual affair in many companies and our test set is also for one year. There are two options for audit for a utility: (1) a utility wide audit (expensive), (2) an audit on those devices detected as positive (less costly). Let different utilities have different tolerable false alarm that vary between 2% to 10%. There is a loss due to audit on false alarms but a gain for detecting compromised meters successfully. We consider here only the monetary value per Kilo Watts hour (KWH) of electricity that is falsified. The effective profit/loss per year can be calculated as:

$$NProfit = \frac{\delta_{avg} \times \eta \times E \times 365}{1000} \times (M - md) \tag{17}$$

where M is the number of meters compromised, md is number of missed detections, η is the number of reports/day, E = \$0.12 per KWH is average cost of electricity in USA (could be as high as \$0.38 in some states). On the other hand, the cost of false alarms per year is: $L = CA * fa$ and $NetBenefit = NProfit - L$ where CA is the cost of audit/meter, and fa is the number of false alarms.

In Table 2, we provide the practical implication of our performance under user tolerable false alarm rates of 2% and 10%, with ρ_{mal} = 40%, in terms of monetary benefit. Given the numbers, a 2% tolerable FA is more profitable for Irish data, while 10% tolerable FA is more profitable for Texas data, for the same δ_{avg}. Since the difference in losses is not drastic, our recommendation for utilities is to choose 10% tolerable rate, since it will give much lower MD when attack actually occurs. Since the Irish data has a large micro-grid, the benefit is large, underscoring that the benefit is scalable.

Table 2: Profit/Loss Per Year with our Framework

Tolerable FA Threshold	δ_{avg}	$NetBenefit$: Irish	$NetBenefit$: Texas
2%	100	+ 21,219.12	+ 4,868.88
10%	100	+ 16,922.34	+ 5,597.16
2%	400	+ 141,686.64	+ 30,413.04
10%	400	+ 138,441.06	+ 32,087.47

7.4 Comparison with Previous Research

We compare our performance with 3 categories of existing methods: (i) classical ML, (ii) information-theoretic, (iii) statistical learning. Classical ML uses SVMs [10], decision/regression trees (DRT) [9]. The [10] outperforms [9], hence we compare our work with [10]. For information theoretic approaches [1, 14], we chose to compare with [1] (though mainly it showed the Texas data results) since it reports for various δ_{avg} unlike [14]. Statistical learning based method [3] outperform [18, 25, 26] and hence is chosen for comparison. The Fig. 12, shows a comparison of our method with existing works under our threat model (assuming deductive attacks over Irish dataset since its common to all previous works). We can observe that the MD rate of our method (blue- solid line) is much lower compared to other works especially for lower δ_{avg}, with a threshold corresponding to 10% acceptable FA. This is fair comparison since Refs [1, 3, 10] have FA rates $\geq$ 10%, even for attacks with $\delta_{avg} > 350$W (See Appendix B).

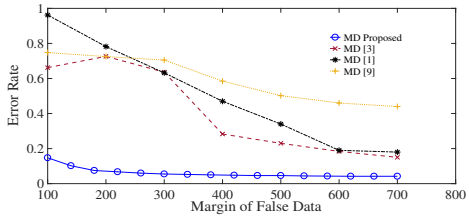
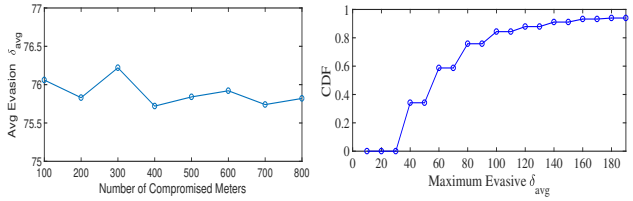


Figure 12: Performance Comparison with Existing Research

7.5 Quantifying Impact of Evasion Attacks with Knowledge of our Model

Now we analyze how adversary can bypass our detector according to knowledge of our the model. **Knowledge of Species Width and Threshold:** Here adversary's optimal strategy is to use a tailor made δ^i_{avg} for each compromised meter, that just refrains from crossing the threshold (by back calculating deviation from the original diversity index just before attacks start). The breakdown δ_{avg} has to be at-least equal to or less than the optimal species width, else it will change species memberships, to help the classification. The Fig. 13(a) shows the average upper bound evasion δ_{avg} is around 76 watts across randomized compromised meter sets. Fig. 13(b) is a CDF which proves that 84% of those 383 compromised meters have a maximum evasion δ_{avg} of less than 100W (93% of them are below 150W). The median absolute deviation of this upper bound is about 31 Watts. This is enough to show that we reduce the impact of the undetected attack, even when attacker has this knowledge.



(a) (b)

Figure 13: Performance Under Evasion Attacks: (a) Upper Bound breakdown δ_{avg} (b) CDF of evasion (M = 383)

Knowledge of Frame Size and Sigmoid parameters: The adversary can learn the exact frame length and sigmoidal parameters ν and B_b, to make sure that the total ∇_s for each species across the frame F is below the point that fails to trigger the exponential growth of the ϕ function, with an incremental attack. The increment per day can be lesser than intended δ_{avg} divided by window length multiplied by F or an increment/day such that ∇_s is safely below 0.5. With this, we found that it takes *120 days to reach δ_{avg} of 100W.* To reach 400W (where existing works have high MD), it will need 480 days. Thus, we reduce attacker's break-even time. The knowledge of A_b, does not help the adversary in evasion.

Knowledge of q: If adversary knows the exact value of q, understands the role of q, it can introduce an attack, avoiding the intermediate rare species magnification, by distributing the false data points onto the other rarer species, by introducing a proportional change (assuming knowledge of historical distribution) in the all species equally (See Fig. 14(a)). However, note that q is just one of parameters, and our method also looks at number of species that contain changes however small, with the W matrix and also uses an expectation of the similarity $D(f)$ across all the species, and these gets added up. Hence, an adversary with this attack will not be able to evade detection. The proof is shown in Fig. 14(b).

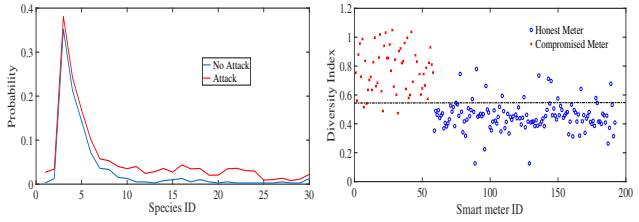


Figure 14: (a) Attack Crafted Using q knowledge (b) Corresponding Performance under q knowledge attack

8 CONCLUSIONS

In this paper, we offered a novel information-theoretic anomaly scoring technique that showed successful detection of smart meters launching data falsification with very low to high attack strengths and attack scales are possible, using AMI as proof of concept. The proposed method's accuracy generalizes well across two different datasets, with completely different years of data collection, countries, sizes of micro-grids. The conclusion is that the method is a way of inferring security status in terms of data integrity where inherent variances are higher than impactful attack strengths. Additionally, we conclude that for a cognizant attacker, the undetectable strategy space in smart energy AMI is reduced from what was achieved by previous works, without a drastic increase in false alarms. As part of future work, we will study how to strengthen the model under training data poisoning attacks and give theoretical estimations of expectation of change in diversity index score as a function of various attack parameters, and check on whether retraining over the untrained attacks improves missed detection performance.

ACKNOWLEDGMENTS

This work is funded by NSF grants SATC-2030611, SATC-2030624, OAC-2017289

REFERENCES

[1] S. Bhattacharjee, A. Thakur, S. Silvestri, S.K. Das, "Statistical Security Incident Forensics against Data Falsification in Smart Grid Advanced Metering Infrastructure", *ACM CODASPY*, pp. 35-45, 2017.

[2] S. Bhattacharjee, S.K. Das, "Detection and Forensics under Stealthy Data Falsification in Smart Metering Infrastructure", *IEEE Transactions on Dependable and Secure Computing*, Early access, 2018 (DOI: 10.1109/TDSC.2018.2889729).

[3] S. Bhattacharjee, A. Thakur, S. K. Das "Towards Fast and Semi-supervised Identification of Smart Meters Launching Data Falsification Attacks" *ACM ASIA CCS*, pp. 173-185, 2018.

[4] B. C. Costa, B. L. Alberto, A. M. Portela, W. Maduro, and E. O. Eler, "Fraud detection in electric power distribution networks using an ANN based knowledge-discovery process," *Int. J. Artif. Intell. Appl.*, vol. 4, no. 6, pp. 17-23, 2013.

[5] K. Foo, "Ghost Talk: Mitigating EMI signal injection attacks against analog sensors" *IEEE Symposium on Security and Privacy (Oakland, CA)*, May 2013.

[6] K. Fu, W. Xu, "Risks of Trusting the Physics of Sensors", *Communications of the ACM*, Vol. 61 No. 2, pp., 20-23, Feb 2018.

[7] 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*, Vol. 28(3), pp. 2959-2966, Aug. 2013.

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

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

[10] P. Jokar, N. Arianpoo, V. Leung, "Electricity Theft Detection in AMI Using Customers' Consumption Patterns", *IEEE Trans. on Smart Grid*, Vol. 7(1), 2016.

[11] L. Jost, B. Tungurahua "Entropy and Diversity", *Wiley Synthesizing Ecology.* Vol. 113(2), pp. 363-375, 2006.

[12] K. Khanna, B. K. Panigrahi and A. Joshi, "AI-based approach to identify compromised meters in data integrity attacks on smart grid," *IET Generation, Transmission & Distribution*, vol. 12, no. 5, pp. 1052-1066, 2018.

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

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

[15] 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.

[16] T. Leinster, M. Meckes, "Maximizing Diversity in Biology and beyond", *Entropy*, Vol 18, 2016.

[17] S. McLaughlin, D. Podkuiko, P. McDaniel, "Energy theft in the advanced metering infrastructure", *Proc. of Critical information infrastructures security (CRITIS'09)*, Springer-Verlag, pp. 176-187, 2009.

[18] D. Mashima, A. Alvaro, "Evaluating Electricity Theft Detectors in Smart Grid Networks", *Springer Heidelberg*, pp. 210-229, 2012.

[19] R. Mohassel, A. Fung, F. Mohammadi, K. Raahemifar, "A survey on Advanced Metering Infrastructure", *Elsevier Journal of Electrical Power & Energy Systems*, 63:473-484, Dec. 2014.

[20] A. Rad, 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] A. S. Rawat, P. Anand, H. Chen and P. K. Varshney, "Collaborative Spectrum Sensing in the Presence of Byzantine Attacks in Cognitive Radio Networks," in IEEE Transactions on Signal Processing, vol. 59, no. 2, pp. 774-786, Feb. 2011.

[22] S. Salinas, M. Li, and P. Li, "Privacy-preserving energy theft detection in smart grids: A P2P computing approach," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 9, pp. 257-267, Sep. 2013.

[23] T. Trippel, O. Weisse, W. Xu, P. Honeyman and K. Fu, "WALNUT: Waging Doubt on the Integrity of MEMS Accelerometers with Acoustic Injection Attacks," *IEEE European Symposium on Security and Privacy (EuroS & P)*, pp. 3-18, 2017.

[24] R. Sevlian and R. Rajagopal, "Value of aggregation in smart grids", *IEEE SmartGridComm*, pp. 714-719, Oct. 2013.

[25] E. Werley, S. Angelos, O. Saavedra, O. Cortes, A. Souza, "Detection and Identification of Abnormalities in Customer Consumptions in Power Distribution Systems", *IEEE Trans. on Power Delivery*, Vol. 26(4), 2011.

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

[27] W. Zhang, S. K. Das and Y. Liu, "A Trust Based Framework for Secure Data Aggregation in Wireless Sensor Networks," 2006 3rd Annual IEEE Communications Society on Sensor and Ad Hoc Communications and Networks, Reston, VA, 2006, pp. 60-69.

[28] M. Sinai, N. Partush, S. Yadid, E. Yahav, "Exploiting Social Navigation", Black Hat, 2015.

[29] https://www.smartgrid.gov/project/pecan_street_project_inc_energy_internet _demonstration.html

[30] Irish Social Science Data Archives, Available at: http://www.ucd.ie/issda/data/

[31] https://www.epri.com/#/pages/product/000000000001026553/

[32] http://energy.gov/sites/prod/files/oeprod/DocumentsandMedia/14-AMI_System_Security_Requirements_updated.pdf

[33] https://krebsonsecurity.com/2012/04/fbi-smart-meter-hacks-likely-to-spread/

[34] https://www.maximintegrated.com/content/dam/files/design/technical-documents/white-papers/smart-grid-security-recent-history-demonstrates.pdf

[35] https://web.archive.org/web/20170131081134/https://www.smartgrid.gov/files/The_Smart_Grid_Promise_DemandSide_Management_201003.pdf

[36] https://www.ovoenergy.com/help/smart-meter-installation#how-long-will-i-have-to-wait-until-i-get-my-smart-meter

[37] https://www.fairwarning.com/blog/power-audit-trail-data-security/

[38] https://www.miningreview.com/top-stories/research-study-quantifies-energy-theft-losses/

[39] https://www.npr.org/sections/money/2011/10/27/141766341/the-price-of-electricity-in-your-state

[40] https://electricitywizard.com.au/electricity/electricity-cost/how-much-does-electricity-cost

[41] https://www.cnet.com/news/money-trumps-security-in-smart-meter-rollouts-experts-say/

[42] https://resource.optimalnetworks.com/blog/2015/01/07/cost-security-audit

APPENDIX

A IMPACT CASE STUDY OF LOW MARGINS OF FALSE DATA

The impact of an attack on the utility not only depends on the average margin of false data (δ_{avg}), but also to the (i) real tampering incidence relative to the network sizes, (ii) time until the detection of the attacked meter, and (iii) widely varying prices across countries.

The size of real world utility's metering network and the realistic number of tampered meters are large. For example [38], real meter audits from three countries (USA, Canada, Australia) for utility deployments reported the rates/numbers of tampered meters, and the corresponding sample and population losses. The USA study (APS Arizona) is the most statistically scientific study that involved one utility with 868,000 meters. This study concluded 0.72% was the tamper rate with high statistical certainty, given the observations from the representative samples used, which comes to about 6000 meters of that utility. An additional 1% was believed to be tampered indirectly through diversion. On the other hand, the Canadian study did not reveal the total number of meters, but involved several utilities across the country to find 1.36% of tampered meters. In a real utility network, the raw number is quite large which makes total losses very high. Such studies are done once in decades due to the effort and cost, hence the detection times are also very large [38].

Consider the average margin as $\delta_{avg} = 100$W which is seemingly very low. The average electricity price in USA is $E = \$0.12$/kwH [39]. If this attack goes undetected for just a year (i.e., $D = 365$ days), the dollar loss/year is calculated as:

$$TI = (\delta_{avg} * M * \eta * D * E)/1000$$

where $\eta = 24$ is the reports/day (typically hourly). (Note that we chose δ_{avg} with data sampling rates equal with existing works for fair performance comparison). Under the USA (Arizona) study, the loss will be $\$660,048$/year at the corresponding lower bound tampering rate (0.72%), for $E = \$0.12$/kWH, $\delta_{avg} = 100$W. Including the diversion and theft together (at 1.72%), the loss/year is $\$1,569,399$ at $\delta_{avg} = 100$. At $\delta_{avg} = 350W$, where previous methods fail, the loss/year is $5,492,898$. It is easy to conclude that the lower bound impact of stealthy margins is very expensive. Furthermore, in many countries, the cost of electricity is not so cheap. For example, in Australia, where $E = \$0.34$/kwH [40], the same attack will cost three times more. However, these rates are only the tip of an iceberg. For example, in Puerto-Rico AMI utility attack (PREPA), about 10% of the meters were found compromised [33] among about 1.3 million customers, facing a $400 million loss.

B PERFORMANCE OF PREVIOUS WORKS

The table 3 includes the self reported numbers of what attack parameters they used and their corresponding reported performance. However, these numbers are not directly comparable in a fair manner with the our method. The [14] is not included in the table since the δ_{avg} assumed are too high and is easier to detect compared to other methods. The [22], is not included in the table, since the method used a synthetic dataset of just 25 users, which does not make the missed detection rate fairly comparable with ours or other methods. We found that for most existing methods, the FA performance is better for Irish dataset, due to high shifting trends. Most of the papers used as a random bias scaling attack. However, these numbers degrade especially when subjected a stealthier threat model. The [1], also shows seemingly good numbers but the δ_{avg} is much higher in this case. This work fails at very high δ_{avg} as well as low $\delta_{avg} \leq 500W$. In [10], the ρ_{mal} considered was extremely small, so the detection rate reported was inflated. When the ρ_{mal} was increased, the detection rate success dropped (i.e., the missed detection rate increased greatly). Specifically, under our threat model with The false alarm rates reported by previous works do not report the base rate false alarms. The [3] shows that classifier at 300W, has a high missed detection rate of 30%, and drops to 52% at 250W, but with high false alarm rates of 29% and 40%. If the FA is bounded like in our method, these missed detection rates will be much higher than what were reported. This is true for all the other works as well. In contrast, our method produces a missed detection rate in the range of 15%-22% regardless of attack type for a false alarm rate of 10% when the attack strength is lower than a factor of 4 (at $\delta_{avg} = 100$) in competing approaches.

Table 3: Self Reported Performance of Previous works

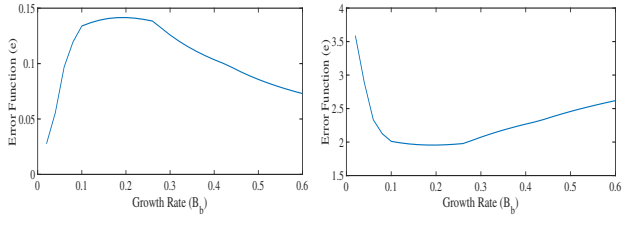
Parameter	FGKT [3]	CPBETD [10]	ARMA [18]	Entropy [1]
FA	29%-5%	29%	33%	11%
MD	30%-4%	24%	28%	8%
δ_{avg}	300W-700W	400W	NA	700-800W
ρ_{mal}	10% − 60%	0.72%	NA	$\leq$ 40%
TTD	$\leq$ 10 days	1 mo	1 mo	1 mo

C EXPLAINABILITY OF DIFFERENT PARAMETERS FOR OPTIMIZATION

Since it is difficult to visualize more than 3 dimensions, we show the effect of each parameter by varying only that while keeping constant other decision variables (both controlled and uncontrolled). This will prove that the error function has a global maxima and an equivalent convex function with global minima is achievable. We prove that for each parameter, an optimal answer is possible. All figures shown in this section of appendix is corresponding to the Irish training dataset.

C.1 Effect of B_b on error function

Figs. 15(a) shows the relationship of our objective (error) function and growth parameter B_b, keeping other factors constant. Fig.15(b) is proof of corresponding convex equivalent that allows gradient descent to be applied to quickly converge to the optimal values.

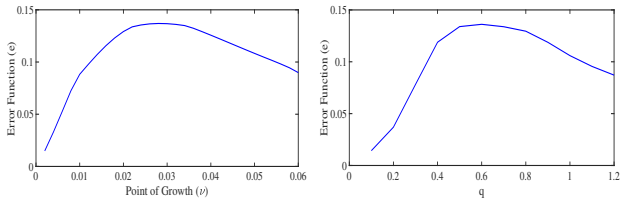


(a) (b)

Figure 15: Effect of B_b (a) Objective Function ; (b) Convex Equivalent of objective e

C.2 Effect of ν and q on error function

Figs. 16(a) and 16(b) show the objective function with changing displacement parameter ν and diversity order q, proving that a unique global optimal exists for each of these parameters.



(a) (b)

Figure 16: Objective Function: (a) Effect of ν, (b) Effect of q

C.3 Effect of Species Width sw

The optimal species width depends on δ_{avg} which is unknown to the defender. Hence, there is no practical way of scientifically calculating the species width *in the optimal sense*. Additionally, the false alarm is subjective to utility, and we need to be able to bypass this difficulty. The upper bound of the species width has to be equal to or lower than the initial desired lower bound (dlb) of δ_{avg} that the defender seeks to detect, else the method will not work. As long as the problem is in this constrained space, the following is true:

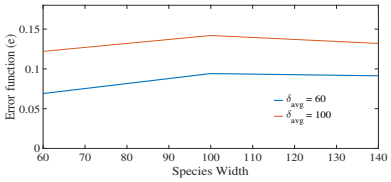
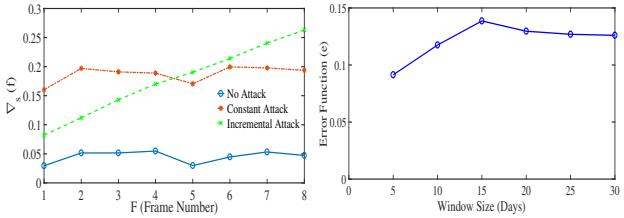


Figure 17: Species width versus objective function

From the perspective of our objective function, if the species width is too small, the legitimate changes will cause the diversity index score of honest meters to increase, thus making the difference between the compromised and the honest set minimum. However, if the species width is too high, then our model will miss the changes in the compromised set, and both sets will have the same range of diversity index score, but in a different region of the diversity index axis. Therefore, some intermediate optimal value would exist for sw which would produce a global maxima, showing the optima exists. This illustration is shown in Fig. 17, where $50 < sw \leq \delta_{avg}^{dlb}$ and given this constraint, we can see that for all δ_{avg} equal to or below this, the objective function is maximized at $sw = 100$.

C.4 Optimal Frame Length and Window Size

We felt that the explainability of this is not apparent, if explained with the objective function. Hence, we use a different y axis, which indirectly effects our objective function. The Fig. 18(a), shows how to find get an optimal frame length F for our model. The frame length should be atleast 5 and atmost 8, keeping in mind the incremental change. Above $F = 8$, the answers tend to become sub-optimal (since legit changes creep into the $\nabla_s(f)$), although the performance degradation is very slow. However, keeping in mind the threat of incremental attacks, we used the upper bound of optimal frame length $F = 8$. Information theoretic approaches and entropies are steady state measures and involve probabilities of categories to converge to their true value and therefore an adequate minimum time window is necessary. However, if the time window is too large the detection of meters will be delayed. Hence, we studied the effect of the time window sizes on the modified diversity index score, keeping other factors fixed. Fig. 18(b), shows that the optimal window size that maximizes the error function is a window length of 15 days (from cross-validation). The errors tend to increase slightly with increasing window length, because in a large time window, legitimate shifting trends in the consumption mimic attacks; thus increasing the false alarms slightly. Hence, for all results, we keep a sliding frame size as 15 days.

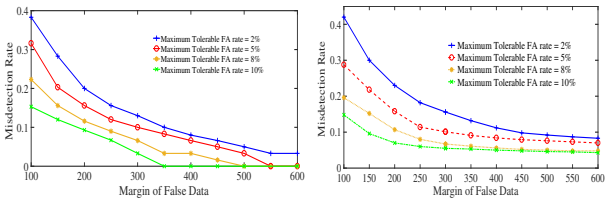


(a) (b)

Figure 18: (a) Optimal Frame Length (b) Optimal Window Size

D ADDITIVE ATTACKS IN TEST SET

The following figures Figs. 19(a) and 19(b) show performance in the testing set for additive attacks (which were trained on) for Texas and Irish datasets respectively. The conclusion is that they have a slightly less mis-detection rate, because these attacks were trained on during parameter learning.



(a) (b)

Figure 19: Additive Attacks: MD rates over Varying Max. Allowable FA (ρ_{mal} = 40%): (a) Texas Dataset (b) Irish Dataset

E RAMP ATTACK SENSITIVITY ANALYSIS

Fig. 20, shows the performance for under incremental ramp attack strategy for various δ_{avg} for Texas dataset for all three attack types (shown for tolerable FA of 10%).

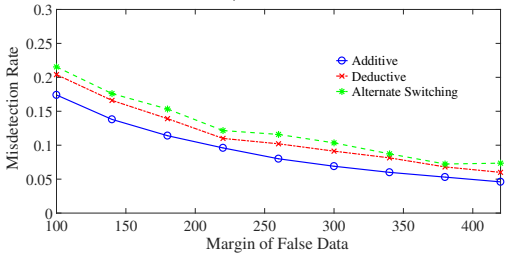


Figure 20: Performance against Ramp Attack Strategy