F-DETA: A Framework for Detecting Electricity Theft Attacks in Smart Grids

Varun Badrinath Krishna, Kiryung Lee, Gabriel A. Weaver, Ravishankar K. Iyer and William H. Sanders
Information Trust Institute, Department of Electrical and Computer Engineering,
University of Illinois at Urbana-Champaign, 1308 West Main St, Urbana, IL 61801-2307, USA
E-mail: {varunbk, klee81, gweaver, rkiyer, whs}@illinois.edu

***Abstract*—Electricity theft is a major concern for utilities all over the world, and leads to billions of dollars in losses every year. Although improving the communication capabilities between consumer smart meters and utilities can enable many smart grid features, these communications can be compromised in ways that allow an attacker to steal electricity. Such attacks have recently begun to occur, so there is a real and urgent need for a framework to defend against them. In this paper, we make three major contributions. First, we develop what is, to our knowledge, the most comprehensive classification of electricity theft attacks in the literature. These attacks are classified based on whether they can circumvent security measures currently used in industry, and whether they are possible under different electricity pricing schemes. Second, we propose a theft detector based on Kullback-Leibler (KL) divergence to detect cleverly-crafted electricity theft attacks that circumvent detectors proposed in related work. Finally, we evaluate our detector using false data injections based on real smart meter data. In mitigating electricity theft, we show that our detector provides up to a 94.8% improvement over the most advanced detector that we know of in related work.**

I. Introduction

The intent of the *smart grid* is to improve the reliability, resiliency, sustainability, and energy efficiency of traditional power grids. The smart grid integrates a variety of digital computing and communication technologies with the power-delivery infrastructure. One important grid communication technology is the Advanced Metering Infrastructure (AMI), which provides a means for electric utilities to monitor the grid. Vulnerabilities in this infrastructure could allow cyber adversaries to compromise the grid in ways that can have adverse effects on utilities and consumers. These effects include electricity theft, disruption of electricity service, and damage to the electricity delivery infrastructure.

Although one should defend smart grids against a diversity of possible attacks (as described in [15]), we focus our attention on electricity theft, which is one of the most important problems faced by electricity suppliers and utilities around the world. According to the World Bank, electricity theft contributes to a loss in electricity delivery of over 25% of generated supply in India, 5% in Australia, 6% in China and the U.S., and 16% in Brazil [17]. Theft in these countries is almost always achieved by tapping into electric distribution lines. To detect these thefts, utility companies such as BC Hydro have been convincing consumers to install smart meters [4]. However, there has been some push-back as consumers have begun to realize that smart meters are vulnerable to cyber intrusions [5]. In 2010, the Cyber Intelligence Section of the FBI reported that smart meter consumptions were being under-reported in Puerto Rico, leading to annual losses for the utility estimated at $400 million [7]. In 2014, BBC News reported that smart meters in Spain were hacked to cut power bills [27]. Given that smart meters can be compromised, the smart meter roll-out efforts of utilities such as BC Hydro may only increase the attack surface for cyber intrusion-based theft methods.

To the best of our knowledge, our research is the first to systematically identify and classify electricity theft attacks targeted at utilities as well as consumers. Specifically, we classify an attack based on (1) its ability to evade detection methods currently used in industry and (2) its applicability in various electricity pricing schemes. Furthermore, the fundamental nature of our approach has allowed us to identify seven classes of attacks, only two of which have been presented in related work. Some of these classes may distribute the monetary loss across consumers, at no loss to the utility. All the attack classes identified in this paper can be launched by compromising the integrity of smart grid communication signals (e.g., price or consumption measurements). Only some of these attacks may also be achieved by tapping electric distribution lines. The identification and analysis of these attack classes guides the creation of detection approaches that drastically mitigate, if not completely eliminate, electricity theft.

It must be noted that smart meters, such as those manufactured by GE [6], are equipped with encrypted communication capabilities and tamper-detection features. However, reliance on these mechanisms alone is not sufficient to ensure total defense against cyber intrusions that exploit communication vulnerabilities. Methods to circumvent encrypted communications in smart grid protocols were recently presented in [16]. The logistics of how the attacker can get into a position where he or she is capable of modifying communication signals is not a focus of this paper and is discussed in [15], [16], [20], and [22]. *Our aim is to detect attacks under the conservative assumption that the attacker has successfully managed to compromise the integrity of smart meter consumption readings*. We validate the data reported to the utility by modeling the normal consumption patterns of consumers, and detecting deviations from this model. Our models are data-driven, and use readings from a real smart meter deployment in Ireland.

The rest of this paper is organized as follows. We discuss related work in Section II. The preliminaries in Section III provide background to understanding the attack model presented in Section IV. A topological representation of the electric distribution system is described in Section V. That tree-based representation guides the formal description of attack strategies in electricity theft. We make three major contributions in

this paper and the first is the classification of those attack strategies in Section VI. Second, we analyze the nature of anomalies introduced by those classes of attacks, to develop a non-parametric detection method using Kullback-Leibler divergence in Section VII. Finally, we evaluate this method in Section VIII, and show that it outperforms detectors in related work in mitigating electricity theft. We conclude in Section IX.

II. Related Work

In this section, we present prior and ongoing efforts by the research community and industry to detect and defend against electricity theft attacks. The detection of electricity theft is an active area of research in the control and communications communities, and is the subject of very recent papers ([2] and [3]). Surveys of electricity theft detection methods include those based on well-defined attack strategies [2], [3], [19] and general consumption behavior anomalies [15].

In [19], the authors evaluate a few different attack detection algorithms for a very specific electricity theft strategy in which the attacker does not change her consumption behavior, but reports lower consumption readings by compromising her own smart meter. In [2], the authors evaluate a different attack strategy wherein the attacker steals electricity from a neighbor at no loss to the utility. Those two papers failed to capture other possible attack classes, because the authors did not adopt a comprehensive and fundamental approach to classification. Therefore, we fill in the gap with a framework that provides a comprehensive classification for better defense. In [9], [10], and [24], the authors assume that smart meters have not been compromised, and use their readings to detect electricity theft. They do so by calculating the total power lost and estimating how much of the loss was due to electricity theft. Their methods fail under the realistic scenario that smart meters are hacked, and we address this gap. Motivating factors for attackers who steal electricity are discussed in detail in [10].

In [20], the authors describe how they simulated consumption patterns of loads in households and detected changes in these patterns to report electricity theft achieved by tapping power lines. They reduce false positive rates by fusing alerts reported by multiple sensors. Their approach is similar to ours in that they try to identify ways in which attacks can take place, and employ learning algorithms to detect attacks. The authors of [15], [20], and [21] all independently claim to have built comprehensive attack trees that span all possible electricity theft attacks. However, their attack trees all depend on existing technologies. In contrast, our work analyzes the fundamental necessary conditions for the execution of a successful electricity theft attack, which are equally applicable to future, unknown technological approaches for such attacks.

Industry has also invested in mitigating electricity theft. Utilities such as BC Hydro and CenterPoint have implemented tamper detection features on smart meters [18]. Unfortunately, penetration testing on a variety of different smart meters has shown that such features are ineffective [22], and that despite decades of work in tamper detection schemes (the first of which was the patent [23]), better protections against electricity theft are needed. BC Hydro has worked with start-up Awesense to go one step further than tamper detection by placing distribution grid meters—different from consumer smart meters—at key nodes on BC Hydro's distribution grid [18]. Although all these efforts have been tailored for line-tapping electricity theft, we show that this investment in distribution grid meters can also be effective against cyber intrusion-based theft attacks.

In this paper, we contribute to the state of the art and the state of the practice by evaluating cyber-based electricity theft attacks on the smart grid in the presence of security measures that are currently employed by industry. We are the first to employ and study the Kullback-Leibler (KL) divergence in the context of detecting electricity theft, but this method has been used in other anomaly detection applications in [1] and [13].

III. Preliminaries

We analyze changes in electricity consumption, price, and their reported values in discrete time. In our power network model, all electricity consumers have meters installed to measure their consumption. These meters are all electronic with network interfaces, not traditional analog meters. We assume that these *smart* meters have a fixed polling time period (Δt) that we treat as our basic discrete time unit. We label each time period using an integer $t \in \mathbb{Z}_{>0}$. The smart meter readings in our model represent the average demand during each time period t. The average demand can be multiplied by Δt to obtain the consumption for billing purposes.

We use $D_C(t)$ to denote the *average demand* of a consumer C during the time period t, where $D_C(t) \in \mathbb{R}_{\geq 0}$. $D'_C(t)$ is the average demand *reported* by the smart meter to the utility during time t. Under that notation, $D_C(t) \neq D'_C(t)$ would imply that the smart meters (or network communications) have been compromised or are malfunctioning. Unless otherwise stated, we assume that smart meters are correctly functioning, so the inequality would imply that they have been compromised.

To the best of our knowledge, we are the first to take into account electricity pricing schemes when studying electricity theft attack strategies. It is important to consider pricing schemes, because false data injections in a certain service area (or against a certain set of consumers) could be tailored to the specific pricing scheme in that area (or pricing scheme adopted by those consumers). We consider flat-rate, time-of-use, and real-time pricing schemes. In *flat-rate pricing*, the price stays constant throughout a billing cycle spanning T time periods. This is the traditional pricing scheme and is prevalent all over the world. In *time-of-use* (TOU), the price of electricity varies according to a plan. Certain hours of the day are designated as *peak*, *partial-peak*, or *off-peak*, and the prices for these hours are published by the utility before consumers agree to sign up for this scheme. Finally, in *real-time pricing* (RTP), the prices change in a non-deterministic manner that captures the dynamic market trends in electricity demand and supply.

Let $\lambda(t)$ denote the electricity price during the time period t, where $\lambda(t) \in \mathbb{R}_{\geq 0}$. Note that the price does not necessarily change between smart meter polling periods, and that price updates are usually less frequent than polling reports. We assume that the price update period is $k\Delta t$, where $k \in \mathbb{Z}_{>0}$.

IV. Attack Model

The attacker in this paper is an electricity consumer in the electric distribution grid who consumes more electric energy than she pays for. We follow network security naming

convention and refer to her as Mallory, the attacker whose intentions are malicious. Mallory's intention in stealing electricity is to make a monetary profit at the expense of the utility or her neighbors. Broadly speaking, an attacker can steal electricity by manually tapping electric power lines or by electronically injecting false readings. Our focus is on the latter method, which assumes that *either the smart meter or the communication link has been compromised, and the attacker is now an insider in the system*. This is reasonable to assume given evidence of real hacking incidents in [5], [7] and [27].

In the attack model, we assume that the smart meters are in either of two states: compromised or correctly functioning. If meters are faulty without the intervention of an attacker, it is possible that electricity can be unaccounted for. However, this can be easily detected and investigated, as we later show in Subsection V-B. Therefore, our attack model is concerned only with faults that are intentionally introduced by an attacker.

If the billing cycle contains T time periods, then a single attacker A can successfully steal electricity (in other words, execute an electricity theft attack) if and only if the following condition holds:

$$\sum_{t=1}^{T} \lambda(t)[D_A(t) - D'_A(t)] > 0 \tag{1}$$

This is a simplified form of the expression that describes Mallory's profit or monetary advantage α, which is given by the difference between what the utility should bill her based on actual consumption, B_{Utility}, and what the utility actually bills her based on reported consumption, B'_{Utility}:

$$\begin{aligned} \alpha &\triangleq B_{\text{Utility}} - B'_{\text{Utility}} \\ &= \sum_{t=1}^{T} \lambda(t) D_A(t) \Delta t - \sum_{t=1}^{T} \lambda(t) D'_A(t) \Delta t \\ &> 0 \end{aligned} \tag{2}$$

Here the units may be given as follows: λ is in \$/kWh, D is in kW, Δt is in hours, and α is in \$ (dollars). Since Δt is a positive constant, it does not factor into (1). Mallory's objective is to maximize α subject to the constraint that her attack must go undetected.

Our principled approach to searching for attack strategies begins with the observation that it is necessary for Mallory to under-report her consumption at some time t. This is formally stated in the following proposition:

Proposition 1: If constraint (1) holds, then $\exists t$ such that $D'_A(t) < D_A(t)$.

Proof (by contradiction): Assume $\forall t, D'_A(t) \geq D_A(t)$, then:

$$\sum_{t=1}^{T} \lambda(t)[D_A(t) - D'_A(t)] \leq 0 \tag{3}$$

which contradicts (1). ■

Therefore, any strategy that does not under-report a value at any time cannot be an electricity theft attack. Under-reporting can be achieved either by compromising the integrity of the smart meter readings or by tapping the electric power line immediately upstream of the smart meter. In the latter case, the smart meter measures only what is downstream of it and misses what was tapped out upstream of it, as shown in Fig. 1. As a result, the meter, though not compromised, reports values that are lower than what is actually consumed. *Although, Proposition 1 is evident, formalizing it provides structure to the exercise of identifying attack classes.* Furthermore, *Proposition 1* helps identify whether or not an integrity attack on smart meter communications leads to electricity theft.

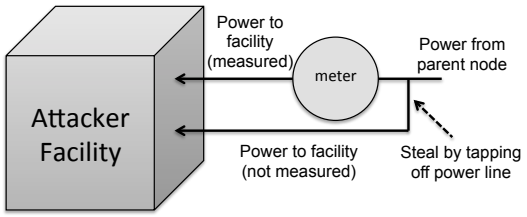


Fig. 1: Illustration of how an attacker can under-report her consumption without needing to compromise her meter.

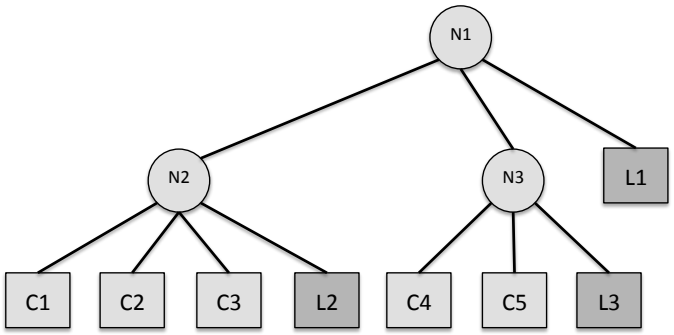


Fig. 2: Illustration of a radial power network topology as a n-ary tree. Circles represent internal nodes $N1 - N3$. Squares represent leaf nodes that include end-consumers $C1 - C5$ and network losses $L1 - L3$. In this example, $D_{N1}(t) = D_{N2}(t) + D_{N3}(t) + D_{L1}(t)$ and $D_{N3}(t) = D_{C4}(t) + D_{C5}(t) + D_{L3}(t)$.

V. Electric Distribution Grid Topology Representation and the Balance Check

Certain attacks can be launched and detected using information about the electric distribution grid topology. *Most of these topologies in practice are radial, so we only assume radial topologies in this paper*. A radial topology can be represented as an unbalanced n-ary tree, where n represents the maximum number of consumers, or leaf nodes, connected to a single node. Another common topology is the loop system, which was designed to improve the reliability of power delivery. This system is essentially radial, as the loop is only closed during a fault (see [12]). Therefore, under the assumption of a radial topology, power to a consumer at any one time is supplied through a single path from the distribution substation, which we refer to as the *root node* of the n-ary tree. Through a series of transformers and protective equipment, power is supplied from the root node to the leaf nodes. This root node would typically lie in a substation that connects the transmission (high-voltage) electric grid with the distribution (low-voltage) electric grid.

Since active power is additive, the total average power that is supplied at a node in the tree at time t is equal to the sum of the average demands at all its child nodes at time t. This is

illustrated in Fig 2, where network losses are also modeled as leaf nodes. This illustration helps describe the balance check, which is an important detection mechanism used in industry.

A. The Balance Check

Let $D_N(t)$ represent the demand at internal node N at time t. In addition, let C (L) be the set of all consumer (loss) nodes that are descendants of N. Those loss nodes model line impedances and transformer losses. Then:

$$D_N(t) = \sum_{c \in C} D_c(t) + \sum_{l \in L} D_l(t) \tag{4}$$

This equation is the first step to arriving at the balance check, and is the foundation for patent [8]. The balance check is used by utilities, such as BC Hydro, to check if readings add-up correctly. The check is performed at internal nodes in the distribution grid topology, which we refer as *balance meters*.

In [8], D'_N, which is the balance meter reading at N, is compared with the readings reported by smart meters located at each $c \in C$. Utilities and the authors of [8] assume that there is no electricity theft if the following condition is satisfied:

$$D'_N(t) = \sum_{c \in C} D'_c(t) + \sum_{l \in L} D_l(t) \tag{5}$$

In this paper, we show that the above assumption is false, and identify attacks that are effective despite having satisfied the condition. Note that the losses are not reported, but calculated by utilities based on known values of distribution system component specifications, such as line impedances. That calculation is discussed in [24]. Thus, we do not require a symbol like D'_l for reported loss readings.

Assuming that these balance meters are trusted, the reported meter measurement at node N, which is D'_N, is equal to D_N. Thus, we can combine (4) and (5) to obtain the simplified balance check for electricity theft detection:

$$\sum_{c \in C} D'_c(t) = \sum_{c \in C} D_c(t) \tag{6}$$

B. Detecting Faulty or Compromised Meters

Let W denote the event that *a meter at a specified node reports a balance check failure*. The following points would help identify a faulty or compromised meter.

If W is true for an internal node, it must be true for all its ancestors, all the way up to the root node. An alarm should be raised for investigation if W is true for an internal node and false for its immediate parent node. Such a situation would imply that at least one of the two meters is faulty or that at least one of them has been compromised. Note that the inverse of this implication is not true. If W is false for a node, then it is still possible that W is true for the node's parent. This would imply that W is true for at least one of the parent's other child nodes.

If a parent of internal nodes has W true, and all of its child nodes have W false, then an alarm should be raised for investigation. This situation implies that at least one of the children, or the parent itself, are faulty or compromised.

TABLE I: Attack Classification

Attack Class	1A	2A	3A	1B	2B	3B	4B
Possible despite Balance Check	N	N	N	Y	Y	Y	Y
Possible with Flat Rate Pricing	Y	Y	N	Y	Y	N	N
Possible with TOU Pricing	Y	Y	Y	Y	Y	Y	N
Possible with RTP	Y	Y	Y	Y	Y	Y	Y
Requires ADR	N	N	N	N	N	N	Y

C. Investigating Failure of the Balance Check

If the balance check fails at a balance meter that is correctly functioning and has not been compromised, then it implies that electricity has been stolen. If the balance meter were malfunctioning or compromised, the failure of a balance check may not indicate electricity theft. Even so, the meter should be investigated and fixed by the utility. Such investigations, currently made periodically [11], can be made less frequently, and in a systematic manner. In finding the faulty meter, the worst case search order is $O(N)$. Using the following approaches that exploit the tree structure of the topology, the effort and investment incurred by the utility can be minimized.

Case 1 (every internal node has been instrumented with a meter): Finding the deepest meter in the tree that reports a failure of the balance check would identify the geographic neighborhood that needs to be investigated. If all the internal nodes are trusted and functioning correctly, then the deepest internal nodes to report this failure would have a limited number of consumer leaf nodes connected to them. These leaf nodes would then need to be manually inspected. One or more of these consumers would be an attacker.

Case 2 (at least one internal node has not been instrumented with a meter): In this case, the utility could send a serviceman with a portable meter and then perform a search similar to the Breadth First Search tree traversal algorithm. Starting with the root node, the serviceman would check each child of the node and see if the readings match the readings of the smart meters of consumer nodes that are descendants of the child, accounting for the nodes that represent losses due to impedance in the tree representation. After each check, he only would investigate the subtree of the node whose check failed. The other subtrees would not need to be investigated.

While it is true that the failure of the balance check implies that investigation is needed, it would be false to say that if the balance check is satisfied, then there has been no theft. Unfortunately, that was not explicitly recognized in [8] and other related work. We recognize that, and show that there are theft attacks that can circumvent the balance checks.

VI. Classification of Attacks

Our primary classification of attack strategies is based on those that fail the balance check (*Attack Classes 1A–3A*), and those that successfully circumvent it (*Attack Classes 1B–4B*). We identified seven classes of attacks in this paper, and summarized them Table I, based on whether they are possible under different pricing schemes, and whether they require Automated-Demand Response (ADR) to be in place. We now define the attack classes, and will later describe ADR in the context of Attack Class 4B. We hypothesize that electricity theft attacks in practice may be a combination of one or more of these seven attack classes.

A. Attacks that can be Detected using Balance Checks

In this subsection, we describe theft Attack Classes 1A, 2A, and 3A, for which the balance check in (6) can be used to locate where the theft is occurring within the power network. As acknowledged in [19], this method does not identify the individual attacker, but it drastically reduces the search space for the investigation, which can then be done manually. Although the simplified balance check in (6) works against these attacks, it is possible for the balance meters to be compromised. A single attacker at a leaf node of the tree would only need to compromise the balance check meters in the direct route to the root node. This direct route is the depth of the branch of the tree that supplies Mallory. The tree depths, which we have seen in the distribution grid models that we have used in previous work, range from 5 to 135. For an unbalanced tree, the number of meters that Mallory needs to compromise would be $O(N)$ in the worst case, where the tree is a linear structure. If the tree were balanced, the number of meters she needs to compromise would be $O(\log(N))$.

1) Attacks under Flat-rate Pricing Schemes: In our attack model, under the flat-rate pricing system, smart meter consumption readings reported to the utility are compromised, while pricing signals are not compromised. The assumption that pricing signals are not compromised is reasonable under a flat-rate pricing scheme, since price signals are fixed and predecided. Thus, any deviation from the fixed price is suspicious can be monitored.

The difference between what Mallory pays and what she should pay, as given in (1), quantifies the loss for the utility during the billing cycle. Depending on how the utility business model is structured, the price of the stolen electricity is either paid for by the utility itself or jointly paid as service fees by all the consumers in the system.

Under the flat-rate pricing scheme assumed in [19], the attack condition (1) is reduced to the following:

$$\sum_{t=1}^{T} D'(t) < \sum_{t=1}^{T} D(t) \tag{7}$$

Within this scheme, the attack can take two approaches:

Attack Class 1A: The LHS of (7) is not varied, in which case the reported measurements $D'(t)$ do not deviate from Mallory's typical consumption. Instead, Mallory consumes more electricity than is typical, so the RHS of (7) is increased. This attack is potentially limited only by the capacity of the power network to meet Mallory's demand. Therefore, a large amount of electricity can potentially be stolen.

Attack Class 2A: The RHS of (7) is not varied, in which case Mallory does not change her typical behavior. Instead, the LHS of (7) is artificially decreased in order to satisfy the condition. This strategy is the only attack strategy that is presented in [19] and its severity is more limited than that of Attack Class 1A.

To quantify the limited amount of electricity that can be stolen under Attack Class 2A, we define a threshold $\tau \geq 0$ below which reported consumptions $D'(t)$ can be correctly classified as a theft attack. As an example in [19], τ can be defined as the minimum of daily consumption averages over a fixed number of days. The smallest value that τ can possibly take is 0. Therefore the maximum electricity that Mallory can steal is her typical consumption. In practice, she would steal less as τ would be greater than 0.

All the detection algorithms proposed in this paper, and in [19], are based on analyzing the change in the pattern of reported smart meter readings under the attack. Under Attack Class 1A, there is no change in the reported readings pattern, and therefore the attack would go completely undetected. Most electricity thefts seen in the world today can be classified as Attack Class 1A, implemented by tapping the electrical power lines. The attack class can be detected by the balance check.

2) Attacks under Time-of-Use and Real-Time Pricing Schemes: Under variable pricing, we default to (1), since $\lambda(t)$ is not constant. Attack Classes 1A & 2A are still possible in this context (see Table I). In addition, the following attack is possible.

Attack Class 3A: In this attack, Mallory does not steal any electricity, but shifts her load in a way that she still makes a monetary profit. As marked in Table I, this attack is not possible with a flat rate structure as it requires Mallory to report that her consumption happened at a time when the price was low, when it really happened when the price was high.

We now formalize Attack Class 3A. Consider two time periods t_1 and t_2 such that $\lambda(t_1) < \lambda(t_2)$. This can happen when t_1 is an off-peak period and t_2 is a peak-period in a time-of-use (TOU) pricing system. At time t_1, Mallory may over-report her off-peak electricity consumption such that the difference $D'_A(t_1) - D_A(t_1) > 0$. Similarly, she may under-report her consumption during the peak period so that the difference $D_A(t_2) - D'_A(t_2) > 0$. If the differences match, then $D'_A(t_1) + D'_A(t_2) = D_A(t_1) + D_A(t_2)$, so the total demand is equal to the reported demand and no electricity is stolen. Even so, by making it appear as though Mallory has shifted her load from the peak to the off-peak period, she can profit without having stolen any electricity. Although we have used TOU pricing to illustrate this attack, equivalent arguments can be made for real-time pricing.

B. Attacks that Circumvent Balance Checks

In this subsection, we identify classes of attacks that go undetected by balance meter checks. Consider an electric distribution network node to which $M + 1$ consumers are connected: Mallory A and a set of M innocent neighbors $N = \{N_1, N_2, ..., N_M\}$. Then, the balance check constraint at this node at time period t is an instance of (6) given by:

$$D_A(t) + \sum_{n \in N} D_n(t) = D'_A(t) + \sum_{n \in N} D'_n(t) \tag{8}$$

The following proposition helps us identify theft strategies that meet the above constraint:

Proposition 2: If both constraints (1) and (8) hold, then $\exists (n, t)$ such that $D'_n(t) > D_n(t)$ where $n \in N$.

Proof: Given that (1) holds, we know from *Proposition 1* that $\exists t$ such that $D'_A(t) < D_A(t)$. For every such t, if we can prove that $\exists n$ such that $D'_n(t) > D_n(t)$, then we are

done. Again, we prove by contradiction: Assume at time t that $\forall n \in N, D'_n(t) \leq D_n(t)$, then:

$$[D_A(t) - D'_A(t)] + \sum_{n \in N} [D_n(t) - D'_n(t)] > 0 \qquad (9)$$

which contradicts (8). ■

Proposition 2 tells us that in order to be successful with a theft attack that evades the balance check, it is necessary for Mallory to ensure that the consumption of at least one of her neighbors is over-reported. This can be equivalently achieved by compromising a neighbor's smart meter or by physically tapping into the neighbor's electrical system. Both methods could ensure that the neighbor pays for Mallory's electricity.

Note that by "neighbors" we refer to the specific set of consumers who share the same parent node in the distribution grid topology. The balance meter is located at that parent node. Physically, the parent node may be a bus or a transformer.

Let L_n denote the monetary loss incurred by a neighbor n, who has been targeted by this attack. Then,

$$L_n = \Delta t \sum_{t=1}^{T} \lambda(t)[D'_n(t) - D_n(t)] \qquad (10)$$

The structure of (8) shows that the amount of electricity Mallory can steal at time t is maximized when she steals as much electricity as she can from all her neighbors. If she compromises more than one neighbor, then the total amount of electricity she steals in T time periods is given by $\Delta t \sum_{n \in N} \sum_{t=1}^{T} [D'_n(t) - D_n(t)]$, and the total monetary worth of this electricity is given by α in (2) as $\alpha = \sum_{n \in N} L_n$. Note that if $D_n(t) = D'_n(t)$, then Mallory has not stolen electricity from neighbor n at time t.

If the balance check were in place, Attack Classes 1A, 2A, and 3A could only be implemented if Mallory performed the additional step of over-reporting a neighbor's consumption. With that additional step, Attack Classes 1A, 2A, and 3A are renamed to *Attack Classes 1B, 2B, and 3B*, where the letter 'B' is used to delineate the fact that they circumvent balance checks. Those attacks are summarized in Table I, and illustrated later while discussing false data injections in Fig. 3. We now describe a fourth class that involves compromising the electricity price in addition to consumption readings.

Attack Class 4B: This class of attacks illustrates how a neighbor can receive a lower electricity bill than expected despite having electricity stolen from him. This strategy can only work in a real-time pricing setting in which consumers are equipped with Automated Demand Response (ADR) interfaces. ADR encourages electricity consumers to adapt to changes in signals from the utility, most importantly the electricity price. This ensures that demand is adjusted to meet supply constraints. The most well-known implementation of ADR, known as OpenADR, is based on the *Energy Market Information Exchange* (EMIX) specification [25].

In Attack Class 4B, Mallory actively decreases her neighbors' demand by compromising their ADR interfaces. Here, she increases her consumption in proportion to the amount by which she decreases her neighbors' consumption. She effects the decrease of her neighbors' consumption by increasing electricity price seen by the neighbors' ADR systems. Since consumption is typically modeled as a monotonically decreasing function of the price, an ADR system of a consumer would be programmed to automatically consume less if the electricity price has increased. The Consumer Own Elasticity model [26] is an example of such a monotonically decreasing function that captures how much a consumer would decrease his consumption in response to a given increase in the price.

Therefore, the attack is designed in such a way that for some time t and neighbor n, $D_n(t) < D'_n(t)$, $D_A(t) > D'_A(t)$ and $\lambda(t) < \lambda'_n(t)$, where $\lambda'_n(t)$ is the compromised electricity price seen by neighbor n. Based on his meter's readings, the unsuspecting n thinks he has consumed $D'_n(t)$ and expects to pay a bill of B_{Expected} during a billing cycle of T time periods. Instead, he is sent a lower electricity bill by the utility, B_{Utility}, which leads him to believe that he benefited by a positive quantity ΔB:

$$\begin{aligned} \Delta B &= B_{\text{Expected}} - B_{\text{Utility}} \\ &= \Delta t \sum_{t=1}^{T} \lambda'_n(t) D'_n(t) - \Delta t \sum_{t=1}^{T} \lambda(t) D'_n(t) \qquad (11) \\ &> 0 \end{aligned}$$

This is interesting since, in reality, he lost a positive amount to Mallory, L_n, given by (10). The notation holds good in the case where this attack has been launched against multiple neighbors. Those neighbors who have not been compromised would have $D_n(t) = D'_n(t)$, and $\lambda(t) = \lambda'_n(t)$.

VII. Data-Driven Detection Methods

We have discussed three attack classes that fail balance checks and four that circumvent these checks. Mallory could use any combination of these attack classes in her actual attack, so detection methods cannot be designed to target individual classes. With F-DETA, we propose a holistic approach to detection that works on *all* the classes.

The five general steps for detecting all seven attack classes are to (1) use a model to estimate the expected consumption of all consumers for the upcoming time period; (2) evaluate whether the actual readings obtained are anomalous based on what was expected; (3) identify whether the anomalies indicate that the consumer is an attacker (abnormally low readings) or a victimized neighbor of an attacker (abnormally high readings) as per *Proposition 2*; (4) use external evidence (severe weather conditions, holiday periods, special events, etc.) to determine whether the anomalous consumption may be a false positive; and (5) systematically investigate the anomaly, if there is no reason to suspect a false positive, by checking the integrity of smart meters as described in Section V-B.

Although F-DETA is a general framework that does not prescribe specific detection methods, we evaluate detection methods proposed in the literature and propose a new method in this paper that leverages Kullback-Leibler (KL) divergence to drastically mitigate the amount of electricity that can be stolen. The methods in the literature that we evaluate are specifically those presented by the authors of [2] and [19]. The KL divergence method *complements* those detection methods proposed in the literature in a way that can mitigate all the attack classes that we have identified.

A. Assumptions and Scope of Evaluating Detection Methods

In the context of this paper, a detection method is a centralized online algorithm that would run at an electric utility's control center. The detection methods discussed in this section look for anomalies in the smart meter readings that are reported to the utility and stored at the control center. We assume that the meters perform accurate measurements. This assumption is justified by a study [11] that concluded that 99.96% of electronic smart meter readings were within $\pm 2\%$ of the actual value, and that 99.91% were within $\pm 0.5\%$. Therefore, an attacker cannot leverage measurement errors inherent to smart meters to steal a significant amount of electricity. *While the measurements are accurate, the reported readings may have been compromised by an attacker*.

If the smart meters can be compromised, it is reasonable to assume that the meters performing the balance check at the internal nodes of the network topology can also be compromised. We assume that the balance meter at the root node of the network alone can be trusted. This assumption is easily justified if this meter is located in a substation on the same premises as the utility-owned control center. The control center may be primarily tasked with substation automation, but it can also be used to detect electricity theft. Since the balance meter at the root node and the control center are co-located, the meter may directly feed into the servers at the control center using dedicated communication infrastructure that is not exposed to external attacks.

Under the assumption that the root node balance meter is trusted, Attack Classes 1A–3A are automatically detectable using the balance check. Therefore, no further work needs to be done in order to detect these classes of attacks. However, Mallory can still use Attack Classes 1B–4B to steal electricity from her neighbors. *We focus on detecting Attack Classes 1B, 2A, 2B, 3A and 3B*. While Attack Classes 1A–3A can be detected using the balance check, Attack Classes 2A and 3A can also be detected using the same data-driven methods that we use to detect Attack Classes 2B and 3B. *Attack Class 1A cannot be detected by data-driven methods since the reported consumption in this class of attacks is in no way abnormal*. Also, we have no data to support an ADR-based system in the case of Attack Class 4B. In order to study Attack Class 4B, we would need to make assumptions of how each consumer in the dataset changes consumption in response to changes in real-time electricity prices. It would also require the simulation of a real-time electricity market, and we leave this as a separate study for future work. This paper is fully based on real data and empirical models derived from this data.

B. Nature of Anomalies due to Attack Injections

Abnormally high consumptions may indicate that the owner of the smart meter is a neighbor of the attacker in one of Attack Classes 1B–3B. Under Attack Class 1B, Mallory reports normal consumption readings but consumes more electricity than reported. This extra electricity is billed to her neighbors, whose consumptions are over-reported. In order to identify Mallory, we would need to identify her neighbors and then manually validate all meters connected to their parent node. As mentioned earlier, Attack Class 1B is the most severe of all classes, as Mallory can steal an arbitrary amount of electricity from her neighbors. The only limit on how much she can consume is determined by the physical limits of the electrical conductors in the distribution lines that connect her facility to the grid. The solution we provide in this paper for this class of attacks leverages Kullback-Leibler divergence, and drastically improves upon work by the authors of [2].

Abnormally low consumptions, as would be reported under Attack Classes 2A and 2B, are a characteristic of the attacker, and would help us identify Mallory herself. Since consumption readings are under-reported in both Attack Classes 2A and 2B, we group the two classes together as 2A/2B and apply the same techniques to detect abnormally low consumption.

Attack Classes 3A and 3B both involve load shifting from a period of high prices to a period of low prices. We show that abnormal consumption *for a given price* indicates that an attack from either of these two classes is happening, and we group the classes together as 3A/3B for detector evaluation.

C. Single-Reading Anomaly Detection

We know from *Proposition 1* that, in order to launch a successful theft attack, $\exists t$ such that $D'_A(t) < D_A(t)$ and we know from *Proposition 2* that $\exists (n, t)$ such that $D'_n(t) > D_n(t)$ for $n \in N$. So at each time period t, we need to check how far the smart meter reading $D'_c(t)$ is from our expected value of $D_c(t)\ \forall c \in C$. The larger the difference, the larger the amount of electricity that can be stolen.

The authors of [2] detect single-reading anomalies using the Auto-Regressive Integrated Moving Average (ARIMA) model. In that method, historic data (from the training set) is used to forecast the next consumption reading in the time series $D_c(t)$. The authors proposed the ARIMA-based detector as a first-level check on the range of smart meter readings using the confidence interval obtained from the ARIMA model.

D. Multiple-Reading Anomaly Detection using Kullback-Leibler Divergence

The authors of [2] identified a clever realization of Attack Class 1B, which they called the *Integrated ARIMA attack*. This attack injects false data in such a way that the readings are not anomalous when each reading is observed individually. The anomalous behavior is apparent only when one observes a set of multiple consumption readings that deviate from the historic trends. In this paper, we standardize the size of the set to a full week of 336 half-hour readings. That is a good choice of size, as consumers' weekly consumption patterns tend to repeat. Daily trends also exist; however, they are not as pronounced as weekly trends, because of differences in schedules between weekdays and weekends.

It may appear that the use of multiple readings has a limitation in that an entire week of data needs to be collected before an anomaly can be detected. There are two counter-arguments to this point. The first is technical; the new week vector can be completed with trusted data from a week in the training set (historic readings). As new consumption readings are recorded, they will replace the historic readings in the week vector. If the week vector contains sufficiently anomalous readings right at the beginning, it may appear appear anomalous before a full week of new data has been collected. This approach was used

by the authors of [3] to calculate the time-to-detection. The second counter-argument is based on litigation proceedings; if an electricity thief has been caught, the fines imposed may exceed the value of electricity stolen in a single week. Under that assumption, the week-long upper-bound on the time-to-detection may be acceptable.

The Kullback-Leibler divergence (KLD) is useful for comparing the distribution of a set of measurements against the distribution of a baseline model. *This method does not assume any underlying parametric distribution*. It is ideal for detecting the Integrated ARIMA attack, because it can detect the change in consumption pattern distributions caused by the attack. The implementation and evaluation of this method in the context of electricity theft is a main contribution of this paper.

KLD is calculated as follows. For each consumer, we construct a training matrix X with M rows (one for each week in the M-week training set period) and 336 columns (one for each half-hour of the week). We then calculate a histogram of all values of X using B bins. We refer to the corresponding distribution as the *X distribution*. Using the $B+1$ bin edges, we iterate through each row i of X (denoted by X_i, where $i \in \{0, 1, ..., M-1\}$, and calculate the probability that values in X_i will lie within these $B+1$ bin edges. Those probabilities give us what we call the X_i *distributions*. It is essential to use the exact same bin edges determined from the X *distribution* while calculating the X_i *distributions*.

We construct a vector K of KLD measures from a consumer's training set. Each element K_i in K is equal to the KLD between X_i and X:

$$K_i = \sum_{j=1}^{B} p(X_i^{(j)}) \log_2 \frac{p(X_i^{(j)})}{p(X^{(j)})} \tag{12}$$

where $p(X_i^{(j)})$ refers to the number of values in X_i that belong to bin j, normalized by the total number of values in X_i. Similarly, $p(X^{(j)})$ is the corresponding relative frequency for values in X that belong to bin j. We refer to the distribution of K as the *KLD distribution*.

A week of consumption readings is deemed to be anomalous if it deviates too much from the historic distribution. We set thresholds on the *KLD distribution* at the 90$^{\text{th}}$ percentile and 95$^{\text{th}}$ percentile for the sake of illustration. For a new week of consumptions, we construct a week vector X_A, and calculate K_A using (12). Let the null hypothesis be defined as the event that a new week of consumption readings is not anomalous. If $K_A > 90^{\text{th}}$ percentile, we say that the week was anomalous, or the null hypothesis was rejected at an upper-tail significance level of $\alpha = 10\%$. The $\alpha = 10\%$ is a more aggressive detection boundary than the $\alpha = 5\%$ corresponding to the 95$^{\text{th}}$ percentile. This is because more values are likely to be flagged as anomalous. However, it is also subject to a higher false-positive rate, as normal consumptions may also be flagged as anomalous. A successful detector has both a high attack detection rate as well as a low false-positive rate.

VIII. Evaluation of Detection Methods

In this section, we use real smart meter data to quantify the benefits of the KLD detector proposed in this paper with respect to approaches proposed by the authors of [2]. We evaluate the detectors using false data injection and define the *attack vector* as the set of readings produced by this injection.

A. Description of the Dataset

The dataset we used was collected by Ireland's *Commission for Energy Regulation* (CER) as part of a trial that aimed at studying smart meter communication technologies. [1] This is the largest publicly available dataset that we know of and was used by the authors of [2] and [3]. Since the dataset is public, we were able to compare our results against theirs, using the same data. The dataset is an anonymized collection of smart meter readings from consumers, collected at a half-hour time resolution, for a period of up to 74 weeks. Our set of 500 consumers includes 404 residential consumers, 36 small and medium enterprises (SMEs), and 60 unclassified by CER.

We assume that the dataset obtained from CER is free from integrity attacks. However, there are anomalous consumption behaviors in the dataset. These behaviors might reflect periods when consumers were traveling (leading to abnormally low consumption) or hosting parties (leading to abnormally high consumption). Such events lead to false positives if the detection strategy classifies them as suspected attacks.

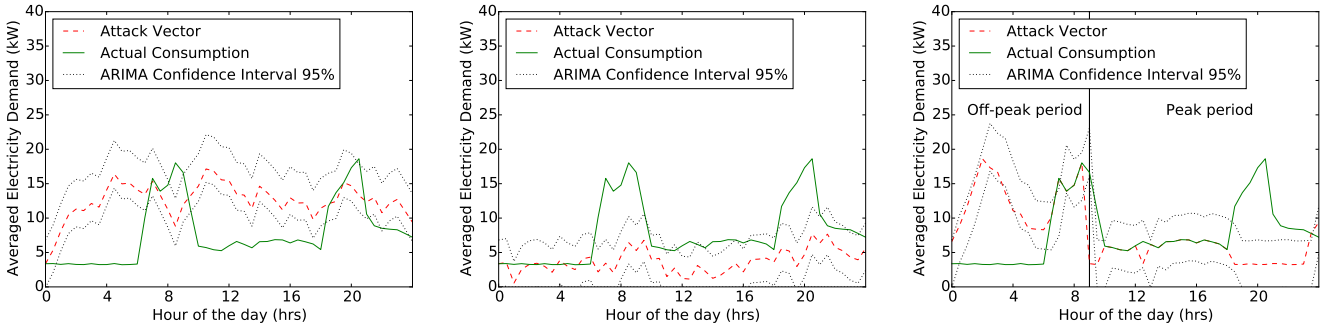
We divided the 74 weeks of consumption data obtained from the CER dataset into two sets: a *training set* of the first 60 weeks, and a *test set* of the remaining 14 weeks. Note that anomalies in the training set are not labeled, so we do not have ground truth on which readings are anomalous. Thus, our algorithm is essentially unsupervised, and our training set serves to build a model of the consumption patterns while accounting for the possibility of anomalies in it. The test set is used to evaluate false positives.

As the dataset does not contain the electric distribution network topology, we do not know which consumers share a parent node in the topology. As a result, the placement of balance meters is unknown, so we make the conservative assumption that *the balance meter at the root node is the only meter that has been deployed in the electric distribution network*. Irrespective of the network topology, all the consumers must be ultimately connected to the root node, so the sum of the readings from all consumers can be used to enforce the balance check at the root node. Equivalently, a more complex assumption can be made to say that *Mallory has compromised all balance meters in the topology except the one at the root node*. We justified this assumption in Section VII-A.

B. Electricity Theft Attacks by False Data Injection

The various attack classes discussed in this paper can be realized by Mallory in many ways as long as they obey the class definitions. For example, Mallory can under-report her consumption readings in Attack Classes 2A/2B by setting all reported readings to zero. Thus, Mallory maximizes the amount of electricity that she can steal. However, it is easy to detect such an attack, and we want to inject attacks that are not as easy to detect. We inject attacks using random numbers, as this ensures that deterministic patterns do not emerge, leading to easy detection. Note that the reported attack consumption

[1] The providers of this data, the Commission for Energy Regulation, bear no responsibility for the further analysis or interpretation of it.



(a) Attack Class 1B (subject is a neighbor) (b) Attack Classes 2A/2B (subject is the attacker) (c) Attack Classes 3A/3B (subject is the attacker)

Fig. 3: Illustration of Attack Classes 1B and 2A/2B using the *Integrated ARIMA attack*, and Attack Classes 3A/3B using the *Optimal swap attack*. In (a) the consumption of one of Mallory's neighbors is over-reported, in (b) Mallory's own consumption is under-reported, and in (c) Mallory's own highest consumptions are swapped into the off-peak period.

poisons the utility's ARIMA model, so the confidence intervals follow the attack vector, and not the actual consumption. Fig. 3 illustrates the injections for Consumer 1330 in the dataset.

The injections were performed on the 500 consumers in the dataset. For each consumer, we injected attack vectors according to methods that will be described in this section. For the Integrated ARIMA attack, we generated 50 different attack vectors from the Truncated Normal Distribution, to reduce bias in the samples obtained from the distribution (at least 30 vectors are recommended for statistical significance). From these 50 attack vectors, we evaluated all the detectors using the worst-case vector that provided Mallory with maximum profit. The computation that went into producing this paper required that 74 CPU cores be run for a total period of 4 weeks.

1) Attack Class 1B Injection: If we assume that Mallory can compromise a smart meter, it is also reasonable to assume that she can passively monitor it and build the same models of the data that we have built from the training set. If we were to install the detector in [2] that uses the ARIMA model to create confidence intervals for detection, we could assume that Mallory can do the same. We refer to that detector as the *ARIMA detector*. By replicating the confidence intervals at her end, Mallory can ensure that her attack vector lies within the confidence interval and thereby avoids detection. In particular, she can maximize how much she steals by setting the vector at the ARIMA confidence threshold, so it does not exceed it. That attack is referred to in [2] as the *ARIMA attack*, and it can be mitigated by placing checks on the mean and variance of a set of readings. With these additional checks in place, we refer to the detector described in [2] as the *Integrated ARIMA detector*. As shown in [2], Mallory can circumvent the *Integrated ARIMA detector* by launching an instance of *Attack Class 1B* called the *Integrated ARIMA attack*.

The *Integrated ARIMA attack* was identified and described in detail in [2]. The false readings are injected from a Truncated Normal Distribution in a way that the neighbor's readings are over-reported, while remaining within the ARIMA confidence interval. At the same time, the mean and variance of the false readings do not exceed thresholds based on historic data. This attack is illustrated in Fig. 3(a).

2) Attack Classes 2A/2B Injection: We show that both the ARIMA attack and the Integrated ARIMA attack can be implemented to realize Attack Classes 2A/2B. That was not shown in [2]. Mallory under-reports her own consumption readings to reduce her bill. For Attack Class 2B, she also over-reports consumption readings for her neighbors to circumvent the balance check. In the case of the ARIMA attack, Mallory's attack vector would be set to the lower confidence threshold (or zero, whichever is greater). The ARIMA attack can be detected by the Integrated ARIMA detector, which in turn can be circumvented by the Integrated ARIMA attack as described by the authors of [2]. There is only one difference in injecting the Integrated ARIMA attack for Attack Classes 2A/2B, and that is that the mean of the attack vector generated by the Truncated Normal Distribution is equal to the minimum of the means (as opposed to the maximum as for Attack Class 1B) in the training set. That attack is illustrated in Fig. 3(b).

3) Attack Classes 3A/3B Injection: In Attack Classes 3A/3B, Mallory reportedly swaps her load to exploit variable pricing. Injecting an attack vector requires us to assume either time-of-use (TOU) or real-time pricing. We assume TOU as there is currently no real-time pricing plan for end-consumers in Ireland (which is where our data comes from). Based on Nightsaver plans offered by various Irish utilities, we assume two TOU periods: a peak period from 9:00am to midnight and an off-peak period from midnight to 9:00am. We evaluated this choice of peak period for our dataset and found it to be suitable because 94.4% of consumers had higher consumption during the peak period on over 90% of the days in the training set.

Assuming the 9:00am to midnight peak period, we inject an instance of Attack Classes 3A/3B that we call the *Optimal Swap attack*. We take a week of readings from the test set and swap the highest readings from the peak period with the lowest readings in the off-peak period. Note that this attack does not affect the mean or variance of the data for the day (or for the week). It does not even affect the distribution of readings, so the KLD detector would not work if it were designed to compare the new week's vector with previous week vectors. *The only change in the dataset is the temporal ordering of readings*. That ordering exploits the changing electricity price, as Mallory would pay less for her highest consumption readings in the day. The Optimal Swap attack would, however, require Mallory to be able to perfectly predict future high consumption readings so she can swap them with current low

consumption readings in the early, off-peak period of the day. An imperfect prediction could also produce the same result, but the prediction would need to be good enough to ensure that the distribution of readings is not significantly changed. For our study, we used prior knowledge of the values in the test set to make a perfect prediction, and inject the Optimal Swap, or worst-case scenario. Note that we injected the swapping of small values for large ones in a way that minimized errors due to exceeding the confidence intervals of the *ARIMA detector*. That is illustrated in Fig. 3(c).

C. Evaluation Metrics

We quantify how well the ARIMA detector, Integrated ARIMA detector, and KLD detector can detect the realizations of Attack Classes 1B, 2A/2B, and 3A/3B (this selection was explained in Section VII-A) using the two following metrics.

Metric 1: the percentage of consumers for whom the detector successfully detected the attack. For Attack Class 1B, this directly translates to the percentage of neighbors who were protected from the attack by the detector.

Metric 2: the maximum amount of electricity stolen over a period of one week, as a result of the attacks going undetected as per Metric 1. For Attack Class 1B, Metric 2 is the sum of the electricity stolen from all consumers for the period of one week. For Attack Classes 2A/2B, Metric 2 is the maximum amount of electricity that was stolen by a single attacker by under-reporting her own consumption. We include in Metric 2 the monetary profit associated with the electricity stolen. Recall that this profit can be attained by a realization of Attack Classes 3A/3B by leveraging variable electricity prices, without stealing any electricity on the whole.

In order to obtain the monetary gain for the attacker, we assume the following pricing system, which is based on the rates set by Electricity Ireland [14] (as our dataset comes from Ireland). The peak price is 0.21\$/kWh (0.195€/kWh), and is valid from 9:00am to midnight. The off-peak price is 0.18\$/kWh (0.172€/kWh), and is valid from midnight to 9:00am. Adopting this TOU pricing scheme allows us to make a fair comparison between Attack Classes 1B-3B, as all of these classes work under TOU pricing (refer back to Table I). Note that the attack classes also work under RTP, but TOU is a far more widespread scheme than RTP, and data is available to make realistic assumptions about prices for TOU.

D. Illustration of the KLD Detector

In Fig. 4, we illustrate the *X distribution* and X_i *distribution* for Consumer 1330 in our dataset. There is one X_i *distribution* for each week in the training set; we illustrate the first of them (X_1). The strong similarity of the shapes of the distributions is an indication that the first training week resembles the expected distribution across all weeks. Most training weeks would have similar shapes, while anomalous weeks would have different shapes. For example, the distribution for the Integrated ARIMA attack vector is significantly different from the *X distribution*, as shown in Fig. 4(a). The difference between each X_i *distribution* and the *X distribution* is given by the KL divergence K_i as defined in (12). The distribution of K_i is illustrated in Fig. 4(b) with 90$^{\text{th}}$ and 95$^{\text{th}}$ percentile points marked. For this illustration, and for our evaluations, we used 10 bins. Fewer bins produce more false negatives and fewer false positives. The impact of the number of bins on the results is a study to be included in extensions of this paper.

TABLE II: Results for Metric 1: Percentage of consumers for whom the detector successfully detected the attack

Electricity Theft Detector	1B	2A/2B	3A/3B
ARIMA detector	0%	0%	0%
Integrated ARIMA detector	0.6%	10.8%	0%
KLD detector (5% significance)	90.3%	72.6%	72.8%
KLD detector (10% significance)	88.9%	83.6%	79.8%

TABLE III: Results for Metric 2: Maximum gains for attacker in one week as a result of circumventing theft detectors

Electricity Theft Detector	Attack Class	1B	2A/2B	3A/3B
ARIMA detector	Stolen (kWh)	362,261	2,687	0
	Profit (\$)	71,707	542	14.3
Integrated ARIMA detector	Stolen (kWh)	79,325	1,541	0
	Profit (\$)	15,413	297	14.3
KLD detector (5% significance)	Stolen (kWh)	4,129	1,541	0
	Profit (\$)	808	297	14.3
KLD detector (10% significance)	Stolen (kWh)	5,374	237	0
	Profit (\$)	1,049	49	14.3

E. Treatment of False Positives

False negatives occurred when the KLD detector failed to detect an attack at a given significance level of $\alpha = 10\%$ and 5%. False negatives imply that the detector was not aggressive enough. False positives occurred when the KLD detector claims to have detected an attack for a week of normal consumption readings. False positives imply the detector was too aggressive. If the KLD detector produced either a false negative or a false positive for a consumer, we declared that it failed for that consumer. Further, if the detector failed for a consumer, we declare that Mallory's gain for that consumer was maximized. By making that assumption, *we penalize all false-positives to the greatest extent possible, assuming that they are totally unacceptable to the utility*. In future work, we will study cost models for false positives, and benefits associated with allowing a limited number of false positives.

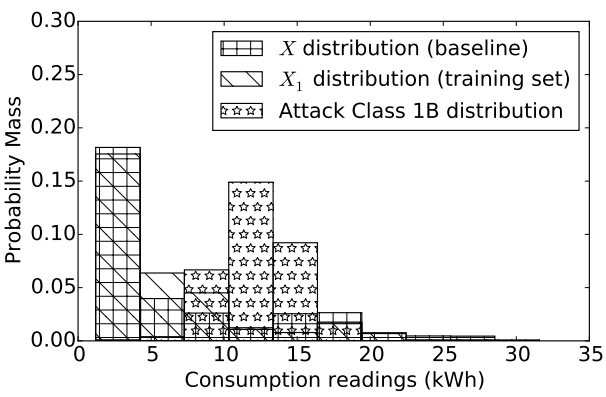
F. Results

The detectors proposed in this paper, and in the previous literature, are aimed at mitigating attacks. Some attacks can be detected, while others are constrained to operate in a mitigated manner so as to circumvent a detector.

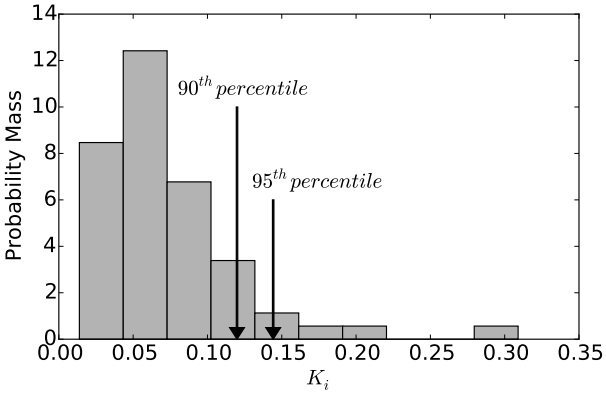
1) Results for Attack Class 1B: We evaluated the KLD detector in comparison with the ARIMA detector and the Integrated ARIMA detector proposed in [2], for the Integrated ARIMA attack as a realization of Attack Class 1B.

By design, the Integrated ARIMA attack evaded the ARIMA detector and the Integrated ARIMA detector. However, as shown in Table II, the KLD detector detected the attack (without false positives) for 90.3% and 88.9% of consumers at the 5% and 10% significance levels, respectively. Therefore, with the KLD detector, we provide a solution to detect the Integrated ARIMA attack, addressing the obvious gap in [2].

The electricity stolen because of the failure of the detectors is given in Table III. Our results for the attacker's gain are



(a) Comparison of X, X_i and Attack distributions



(b) Distribution of K_i (see Equation (12))

Fig. 4: Illustration of the KLD detector. The X_i distributions from the training set overlap with the X distribution in most cases. The "Attack Class 1B distribution" refers to the distribution of consumption readings due to the Integrated ARIMA attack, which is clearly different from the baseline (with a KL Divergence of 0.765, which is greater than the 95^{th} percentile point of 0.144).

larger than those obtained by the authors of [2], because we ran our study for 500 consumers while they ran theirs for 450. The electricity stolen as a result of evading the Integrated ARIMA detector was less than that stolen as a result of evading the ARIMA detector by a factor of 77.5% in [2], and 78.1% in our study (scaling the simulation did not change this percentage reduction significantly). This quantifies the improvement of the Integrated ARIMA detector on the ARIMA detector in mitigating theft. By adding the KLD detector as an additional layer of detection, an improvement of 94.8% on the Integrated ARIMA detector was seen in the worst case attack. As a result, *the KLD detector almost completely mitigated theft through the Integrated ARIMA attack (Attack Class 1B).*

Notice, from Table II, that the KLD detector at the 5% significance level outperforms the detector at the 10% significance level for Attack Class 1B. The contrary is true for the other attack classes, because at the 10% significance level, the detector is more aggressive. The reason for that unexpected result for Attack Class 1B is that the detector at the 10% significance level was too aggressive, which resulted in false positives. Since we heavily penalized false positives (discussed in Section VIII-E), the detector's performance was degraded.

2) Results for Attack Classes 2A/2B: We evaluated the four detectors against the Integrated ARIMA attack as a realization of Attack Classes 2A/2B. The attack was supposed to circumvent (or be completely mitigated by) the Integrated ARIMA detector. However, it was detected for 10.8% of consumers because the consumption readings were so low to begin with, that the random numbers generated by the Truncated Normal Distribution failed to maintain a high-enough average to go undetected by Integrated ARIMA detector. The KLD detector performed much better, as given in Table II.

Since the gains from Attack Classes 2A/2B are obtained by under-reporting the attacker's consumption, one may suspect that Mallory would stand to gain the most by being the largest consumer. However, the maximum amount of electricity stolen in a week due to the ARIMA attack was achieved by the second largest consumer in our dataset (Consumer 1330), and the corresponding value was $2,687$kWh. We found that this was because the lower bound on the confidence interval for Consumer 1330 was lower than it was for the largest consumer in our dataset (Consumer 1411). Thus, the readings for Consumer 1330 could be further under-reported.

The largest amount of electricity stolen in a week due to the Integrated ARIMA attack was achieved by the eleventh largest consumer in our dataset (Consumer 1333), and the corresponding value was $1,541$kWh. In comparison, $1,382$kWh was stolen from Consumer 1330 due to the Integrated ARIMA attack. Upon investigating this unexpected result, we found that it was because the minimum of average values for the weeks in the training set for Consumer 1333 was lower than it was for Consumer 1330. This implies that the mean of the Truncated Normal Distribution for Consumer 1333 was correspondingly lower than it was for Consumer 1330, providing more room for Mallory to steal electricity.

At the 5% significance level, the Integrated ARIMA attack for Consumer 1333 was not detected by the KLD detector in at least one of the 50 simulation trajectories. Hence we say that the detector failed for that attack, and the worst case electricity stolen remains $1,541$kWh. At the 10% significance level, however, the attack for Consumer 1333 was detected. The worst-case amount of electricity stolen despite this detector setting was only 237kWh, and this is 84.6% less than what it was under the Integrated ARIMA detector (see Table III).

The lesson from these results is that *Mallory does not need to be the largest consumer in order to gain the most from theft*. It is also evident that the amount of electricity that can be stolen by circumventing the KLD detector via Attack Classes 2A/2B is an order of magnitude less than the amount for Attack Class 1B. This validates our earlier claim that Attack Class 1B is indeed the most advantageous attack class for Mallory.

3) Results for Attack Classes 3B: All four detectors would fail to detect the Optimal Swap attack as it does not alter the distribution of consumption readings for the week. In order to detect the attack, the KLD detector needs to be modified by conditioning on the electricity price. Since we have two prices, we can split the X distribution into two distributions, one for peak period consumption readings and one for off-

peak consumption readings. This idea can be extended from two distributions to multiple distributions, each conditioned on an electricity price in the case of RTP systems. By using this method of conditioning, we believe the KLD detector can also be used to detect Attack Class 4B.

In the case of Attack Classes 3A/3B, the Optimal Swap attack was not detected by the ARIMA detector and Integrated ARIMA detector, as shown in Table II. The KLD detector (conditioning on prices) was again successful in detecting this attack. The maximum monetary benefit was for Consumer 1254 (amounting to 14.3$), whose Optimal Swap attack circumvented all four detectors by not significantly altering the distribution of consumption readings in the week. Even so, the benefit was so small, that we believe Mallory would not invest the effort required to inject an instance of Attack Classes 3A/3B alone. She may, however, inject an attack that combines Attack Class 3B with Attack Classes 1B and/or 2B.

IX. Conclusion

In this paper, we developed a comprehensive, theoretical framework to provide a defender with insights into strategies for electricity theft attacks on smart grids. We applied this framework to a large-scale smart-meter dataset to devise data-driven detection mechanisms that mitigate these attacks. We evaluated a detector based on KL divergence in comparison with detectors proposed in related work, and presented insights on why some consumers are better candidates for attackers than others. We showed that the KL divergence method was successful in detecting attacks for most consumers. Where the detector was unsuccessful, it placed large constraints on the attacker, so that evading the detector resulted in a drastically mitigated attack. We plan to extend our work to study other datasets, and to account for the presence of multiple attackers.

Acknowledgment

This material is based upon work supported by the Department of Energy under Award Number DE-OE0000780[2] and the Siebel Energy Institute. The smart meter data used is provided by the Commission for Energy Regulation, and accessed via the Irish Social Science Data Archive at www.ucd.ie/issda.

References

[1] M. Afgani, S. Sinanovic, and H. Haas, "Anomaly detection using the Kullback-Leibler divergence metric," in *proceedings of ISABEL'08*, Oct 2008, pp. 1–5.

[2] V. Badrinath Krishna, R. K. Iyer, and W. H. Sanders, "ARIMA-Based Modeling and Validation of Consumption Readings in Power Grids," in *Proceedings of Critical Information Infrastructure Security (CRITIS)*. Springer Verlag, 2015.

[3] V. Badrinath Krishna, G. A. Weaver, and W. H. Sanders, "PCA-Based Method for Detecting Integrity Attacks on Advanced Metering Infrastructure," in *Proceedings of Quantitative Evaluation of Systems (QEST)*. Springer Verlag, 2015, pp. 70–85.

[4] BC Hydro. (2014) Smart metering program. BC Hydro. [Online]. Available: https://www.bchydro.com/energy-in-bc/projects/smart_metering_infrastructure_program.html

[5] Coalition to Stop Mart Meters in BC. (2014, October) Theft of power through hacking of smart meters. [Online]. Available: http://www.stopsmartmetersbc.com/2014-10-23-theft-of-power-through-hacking-of-smart-meters/

[2]The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

[6] ComEd. (2014) Safeguarding data through smarter technology. Commonwealth Edison Company. [Online]. Available: https://www.comed.com/documents/technology/grid_mod_fact_sheet_security_2014_r2.pdf

[7] Cyber Intelligence Section. (2010, May) Smart grid electric meters altered to steal electricity. Federal Bureau of Investigation. [Online]. Available: http://krebsonsecurity.com/wp-content/uploads/2012/04/FBI-SmartMeterHack-285x305.png

[8] E. de Buda, "System for accurately detecting electricity theft," U.S. Patent 12/351,978, 2010.

[9] S. Depuru, L. Wang, and V. Devabhaktuni, "Support vector machine based data classification for detection of electricity theft," in *Proceedings of IEEE PSCE'11*, 2011.

[10] S. Depuru, L. Wang, V. Devabhaktuni, and N. Gudi, "Measures and setbacks for controlling electricity theft," in *Proceedings of North American Power Symposium (NAPS)*, 2010.

[11] Edison Electric Institute, "Smart meters and smart meter systems: A metering industry perspective," p. 35, March 2011. [Online]. Available: http://www.eei.org/issuesandpolicy/grid-enhancements/documents/smartmeters.pdf

[12] J. D. Glover, M. Sarma, and T. Overbye, *Power System Analysis & Design, SI Version*. Cengage Learning, 2011.

[13] J. Harmouche, C. Delpha, and D. Diallo, "Faults diagnosis and detection using principal component analysis and kullback-leibler divergence," in *In proceedings of IEEE IECON'12*, Oct 2012, pp. 3907–3912.

[14] E. Ireland. (2015, November) Valuesaver nightsaver electricity price plan. Electric Ireland. [Online]. Available: https://www.electricireland.ie/switchchange/detailsValueSaverNightSaver.htm

[15] R. Jiang, R. Lu, L. Wang, J. Luo, S. Changxiang, and S. Xuemin, "Energy-theft detection issues for advanced metering infrastructure in smart grid," *Tsinghua Science And Technology*, vol. 19, no. 2, pp. 105–120, April 2014.

[16] P. Jovanovic and S. Neves, "Practical cryptanalysis of the open smart grid protocol," in *Proceedings of Fast Software Encryption*. Springer Berlin Heidelberg, 2015, vol. 9054, pp. 297–316.

[17] R. Katakey and R. K. Singh. (2014, June) India fights to keep the lights on. Bloomberg Businessweek. [Online]. Available: http://www.businessweek.com/printer/articles/205322-india-fights-to-keep-the-lights-on

[18] P. Kelly-Detwiler. (2014, April) Electricity theft: A bigger issue than you think. Forbes. [Online]. Available: http://www.forbes.com/sites/peterdetwiler/2013/04/23/electricity-theft-a-bigger-issue-than-you-think/

[19] D. Mashima and A. A. Cardenas, "Evaluating electricity theft detectors in smart grid networks," in *Proceedings of RAID'12*. Springer, 2012.

[20] S. McLaughlin, B. Holbert, S. Zonouz, and R. Berthier, "AMIDS: A multi-sensor energy theft detection framework for advanced metering infrastructures," in *Proceedings of SmartGridComm*, 2012, pp. 354–359.

[21] S. McLaughlin, D. Podkuiko, and P. McDaniel, "Energy theft in the advanced metering infrastructure," in *Critical Information Infrastructures Security*, E. Rome and R. Bloomfield, Eds. Springer Berlin Heidelberg, 2010, vol. 6027, pp. 176–187.

[22] S. McLaughlin, D. Podkuiko, S. Miadzvezhanka, A. Delozier, and P. McDaniel, "Multi-vendor penetration testing in the advanced metering infrastructure," in *Proceedings of ACSAC'10*. New York, NY, USA: ACM, 2010, pp. 107–116.

[23] S. McWilliams, "Tamper protection for an automatic remote meter reading unit," U.S. Patent 4,357,601, 1982, Bell Telephone Laboratories.

[24] D. N. Nikovski, Z. Wang, A. Esenther, H. Sun, K. Sugiura, T. Muso, and K. Tsuru, "Smart meter data analysis for power theft detection," in *Proceedings MLDM'13*. Springer-Verlag, 2013, pp. 379–389.

[25] OASIS Consortium. (2012, January) Energy Market Information Exchange (EMIX) Version 1.0. [Online]. Available: https://www.oasis-open.org/committees/tc_home.php?wg_abbrev=emix

[26] R. Tan, V. Badrinath Krishna, D. K. Yau, and Z. Kalbarczyk, "Impact of integrity attacks on real-time pricing in smart grids," in *Proceedings of ACM CCS'13*. New York, NY, USA: ACM, 2013, pp. 439–450.

[27] Ward,Mark. (2014, October) Smart meters can be hacked to cut power bills. [Online]. Available: http://www.bbc.com/news/technology-29643276