Tackling Energy Theft in Smart Grids through Data-driven Analysis

Anish Jindal*, Alberto Schaeffer-Filho†, Angelos K. Marnerides*, Paul Smith‡,
Andreas Mauthe§, and Lisandro Granville†
* School of Computing & Communications, Lancaster University, UK.
† Institute of Informatics, Federal University of Rio Grande do Sul (UFRGS), Brazil.
‡ AIT Austrian Institute of Technology, Austria.
§ University of Koblenz, Koblenz, Germany.
(email: a.jindal3@lancaster.ac.uk, alberto@inf.ufrgs.br, angelos.marnerides@lancaster.ac.uk, paul.smith@ait.ac.at,
mauthe@uni-koblenz.de, granville@inf.ufrgs.br)

***Abstract*—The increasing use of information and communication technology (ICT) in electricity grid infrastructures facilitates improved energy generation, transmission, and distribution. However, smart grids are still in their infancy with a disparate regional role out. Due to the involved costs utility providers are only embedding ICT in selected parts of the grid, thereby creating only partial smart grid infrastructures. We argue that using the data provided by these partial smart grid deployments can still be beneficial in solving various issues such as energy theft detection. In this paper, we focus on various data-driven techniques to detect energy theft in power networks. These data-driven detection techniques (at the smart meter as well as the aggregated level) can indicate various forms of energy theft (e.g. through clandestine connections or meter tampering). This paper also presents two case studies to show the effectiveness of these approaches.**

***Index Terms*—Energy theft, smart grid, data analysis.**

I. Introduction

In recent years, electrical energy grids have undergone a major modernization process leading to improved energy generation, transmission, and distribution. Modern energy grid infrastructures are referred to as smart grids, since they allow a more resilient, secure, and reliable electricity supply for end-users while enabling a more fine grained control and better adjusted demand response, and targeted intervention in the face of challenges [1]. The Electric Power Research Institute (EPRI) predicts that a fully developed smart grid can save anywhere between $1.3 to $2 trillion, in comparison to the deployment costs, which are between $338 and $476 billion over 20 years [2]. According to the Global Smart Grid Federation Report, the deployment of smart meters alone would save up to $5 billion in Australia and £7.3 billion in the UK in the next two decades [3]. However, given the high investment required and the amount of existing legacy equipment, the adoption of smart grids has not been uniform across the globe, despite their well-recognized benefits.

Whilst the maturity of smart grid deployments in developed nations has made significant progress in the past few years [4], developing countries still rely on less sophisticated infrastructures. These typically make use of information and communication technology (ICT) at just the consumption level with some automation in the transmission and distribution level. However, it can still be effectively used for carrying out important tasks such as accounting, optimization and theft detection.

Energy theft, in particular, causes major losses for grid utility providers. Due to theft, utility providers lose $89.3 billion annually worldwide [5]. Proportionally these losses are comparatively high in developing countries. For example, $58.7 billion of such losses occur in developing markets [5].

In this paper we argue that, even in partial smart grid deployment scenarios, *data-driven techniques* (based on the collection of data at only a single level) can be useful for tackling the problem of energy theft detection. Hence, it is useful to consider such techniques on different granularity of data collection levels. The major contributions of the study presented in this paper are summarized as follows: the different viewpoints on data-driven techniques used for theft detection in smart grids are presented, highlighting the usefulness of such techniques. We present detailed case-studies at different levels of granularity using available real-world data. This study shows how data-driven techniques can be used to identify energy theft, even in less-advanced infrastructure deployments.

The rest of this paper is organized as follows. In Section II, we describe related data-driven schemes to automatically detect anomalies in energy consumption profiles, which can be used to detect energy theft. In Section III, the background on smart grid infrastructure and smart metering is presented. Section describes the datasets used and methodology followed in order to detect energy theft using data-driven analysis. In Section V, we present case studies evaluation. Finally, in Section VI we discuss final remarks and provide an outlook.

II. Related Work

Energy losses can occur due to *technical* problems or *non-technical* issues. While the former is usually caused by physical factors during energy distribution, the latter is mainly due to energy theft [6]. An important approach to mitigate theft is to use data-driven energy theft detection mechanisms that rely on energy consumption measurements. The data can be gathered by smart meters that are deployed in each household as well as from smart meter gateways on the aggregate level at either a regional (e.g. city-wide) or a neighborhood microgrid

level. In general, data-driven schemes can be further classified into smart meter profiling and aggregate profiling on the basis of considered level of granularity. In the following, the main research in this area is being introduced.

A. Smart Meter Energy Consumption Profiling

Smart meter energy consumption profiling is the basis for energy theft detection mechanisms that rely on statistical data-driven methods to characterize the normal behavior of consumers. These methods can be used to identify unusual energy utilisation, *i.e.*, to accurately localize energy theft. In regard to this, Singh *et al.* [7] utilised the concept of relative entropy to track the energy consumption variations in probability distributions obtained from different consumers. In another approach, Jokar *et al.* [8] utilised the consumers' consumption patterns to detect theft in the Advanced Metering Infrastructure (AMI) by modelling the predictability in customers' usual and abnormal consumption behaviours. In previous work [9], we have developed a scheme to detect the anomalous behaviour in smart homes using smart meter data. To this end, metrics were derived from each energy measurement, such as *Renyi entropy* and the *mean time and frequency marginals*, and used within a simple k-means clustering scheme [9]. Although, such data-driven techniques are aiding towards capturing the partial non-stationarity in the smart meter measurements which could be crucial in energy theft detection. However, its underlying statistical foundations needs to be solid to adequately map such non-stationary measurements into meaningful statistical metrics.

B. Aggregate Energy Consumption Profiling

The combined use of aggregate-level with smart meter-level profiling schemes can arguably enrich the overall characterization of normal power consumption in power distribution networks, thus strengthening the basis for detecting outliers that could relate to the energy theft phenomena. In this regard, Jindal *et al.* [6] analysed the aggregated data of multiple households to detect energy theft in the local communities with high accuracy. Their study used decision trees to compute a predicted energy consumption values for the household and then trained the support vector machine (SVM) with multiple feature sets to find consumers with anomalous behaviour. Pulz *et al.* [10] used the social indicators extracted from the census data to analyse the correlation between losses and socio-economic indices for energy theft detection. Apart from using machine learning approaches, Leite and Mantovani [11] devised a non-technical loss detection mechanism by monitoring the variance of various regional values using multivariate control chart. Xiao and Ai [12] proposed an energy theft detection method on the basis of random matrix theory model to identify the correlations between power consumption and system operation under various scenarios of electricity usage.

These schemes show that aggregated data-driven approaches are useful to identify anomalous behavior of consumers. However, aggregated energy consumption measurements are high in volume and non-stationary [9]. Therefore, a statistical energy consumption profiling scheme needs to address these challenges to compose useful clusters.

III. Smart Grid Infrastructure

A smart grid is an enhanced version of the traditional electrical energy grid. In addition to transmission lines, substations, transformers, and other physical equipment, the smart grid includes an ICT infrastructure that is used to enable enhanced monitoring, control and adaptation functions. However, in the present deployment of smart grids, ICT is not embedded at every level of the grid. This scenario is more common in developing countries, where ICT is only enabled at the facilities like substations, end-user, etc, due to the fact that embedding ICT at every level can be too costly. A simplified view of a high-level grid architecture, showing its major components for data collection, is presented in Fig. 1.

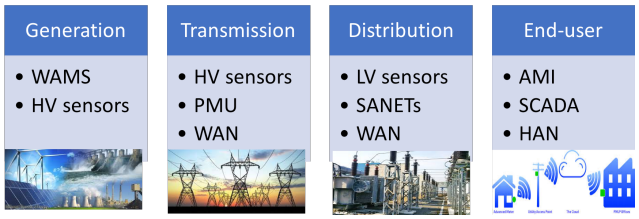


Figure 1: High-level view of the smart grid architecture.

The generation level include, wide area measurement systems (WAMSs) and other high voltage (HV) sensors which gather the data, while the transmission level includes HV sensors, phasor measurement unit (PMU) and wireless area network (WAN) communication infrastructure [13]. The distribution level in smart grid comprise of low voltage (LV) sensors, sensor and actuator networks (SANETs) and WANs to gather and communicate the data, whereas the end-users (or consumers including various sectors such as residential, commercial and industrial) have Advanced Metering Infrastructure (AMI), Supervisory Control and Data Acquisition (SCADA) and home area networks (HANs) [13]. A significant part of the ICT infrastructure deployed in smart grid architecture is the SCADA and AMI systems which are present on the end-user level of the grid network. These are distributed systems that are used to monitor, control, and manage automated processes and components in the energy grid. Apart from SCADA, the substations are also equipped with automation front-end, which provide a level of automation to the distribution substations. It is to be noted that the grid automation at sub-station and higher level does not contribute much to the purpose of detecting theft (which happens more at the end-user level). Moreover, the widespread roll out of SCADA and AMI systems at the end-user level makes them better candidates for data analysis as they gather variety of data at very short time intervals. More specifically, for the purposes of detecting and preventing energy theft across the grid, the systems deployed at end-user level such as SCADA and AMI in smart grid architecture alleviates the need of a fully embedded ICT into each and every level of the energy network. This can be done by performing a data-driven analysis at at different granular levels of data collection in order to detect energy theft in the electricity networks.

IV. Methodology and Data Description

To show the effectiveness of the data-driven techniques for identifying the energy theft, two case studies on different granularity levels of energy consumption profiling with partial ICT deployment in smart grids are discussed.

A. Case study 1: Smart meter energy consumption profiling

In this, the public residential dataset from Dutch Residential Energy Dataset (DRED) is considered [14]. This dataset consists of data collected from several sensors that measure energy, occupancy and ambient conditions in a household for a six months period from 5th July to 5th December 2015 (http://www.st.ewi.tudelft.nl/akshay/dred/). Details of this dataset are given in Table I with the frequency of collection.

Table I: DRED Dataset

Dataset	Description	Frequency
Aggregated	Aggregated consumption with date, time	1 Hz
Appliance	Appliance level consumption with date, time	1 Hz
BT	Beacon values sensed every 1 minute by the occupant's mobile phone. The details include Date, Time, Id, RSSI, Temperature, BatteryLevel, Proximity, Location	1 min
Wi-Fi	Access points RSSI values sensed every 1 minute by the occupant's mobile phone. The details include Date, Time, MAC id, AP name, RSSI	1 min
Temperature	Indoor and outdoor temperature with date, time, location	1 min
Occupancy	Occupancy data inferred using RSSI localization mechanisms with Date, Time, Location	1 min

This dataset monitored the consumption patterns of the appliances in a household along with the ambiance and environmental factors. The histogram of the mostly used appliances in the considered household is shown in Fig. 2.

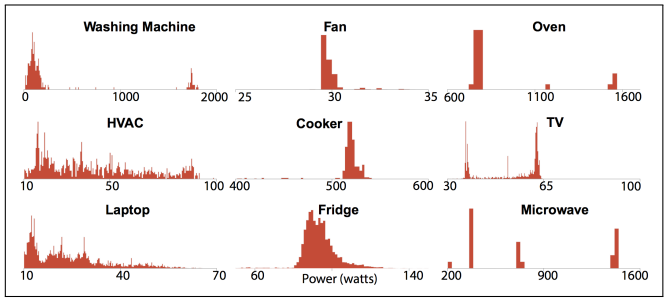


Figure 2: Histogram of different appliance loads (from [14]).

A critical way to detect anomaly/theft in the consumption at this level is to statistically analyse and profile the energy measurements in terms of their frequency components as extracted by the Fast Fourier Transform (FFT)-based analysis as given by,

$$y(z) = \sum_{j=0}^{n-1} x(j) e^{-i2\pi zj/n} \tag{1}$$

where, $x(j)$ is the time domain signal (j=0...n-1), $y(z)$ represents the transformed signal in frequency domain (z=0...n-1), and n is the length of the input signal. Other statistical methods can also be used in conjunction to the above for deeper analysis such as entropy of signals can be effectively examined for identifying any unusual variations [15]. In this case study, the spectral entropy of the signal is analysed which can be defined as:

$$H(t) = -\sum_{n=1}^{N} P(t,n) \log_2 P(t,n) \tag{2}$$

where, $H(t)$ is the spectral entropy at time t in time-frequency domain, $P(t,n)$ is probability distribution of the signal $S(t,n)$ (where $S(t,n) = |F(t,n)|^2$) and $F(t,n)$ is the discrete Fourier transform of the original signal.

B. Case study 2: Aggregated energy consumption profiling

In this case study, load demand from various homes is considered from the Open Energy Information dataset [16]. It contains the electric consumption data for various homes and appliances with respect to date and time for one year. A snippet of this dataset for a typical day's load in winter is illustrated in Fig. 3.

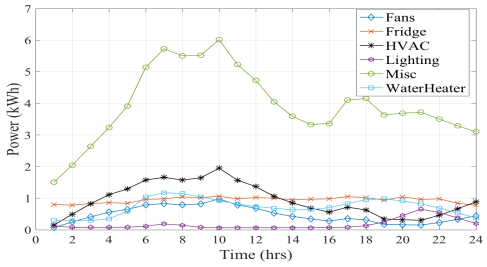


Figure 3: Energy consumption of various appliances.

200 homes from the given dataset were chosen at random in this case study for aggregated analysis of the consumption profiles. For the purpose of detecting theft using aggregated consumption profiling, a simplistic k-means clustering model was chosen to compute the baseline profiles. The premise behind using k-means clustering is that each consumer consumption profile would conform to a certain cluster consumption curve as similar households have similar energy usage pattern [17]. Initially, each consumption profile is assigned to a cluster profile at random after which the mean of the cluster centroid is computed. Then, each profile is re-assigned to the clusters on the basis of their closeness to the new cluster centroid. This closeness is given by the gap between the consumption value and the centroid value, which is calculated as below.

$$d(h,\mu) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(h_i - \mu_i)^2} \tag{3}$$

where, h represents the consumption value of a home and μ depicts cluster centroids respectively, and N is the size of a cluster. This process is repeated until the value of a centroid converges. After the successful generation of the baseline profiles, the energy consumption profiles of the households can be compared with these baseline models in order to find out any deviations from the normal.

V. Case Studies and Evaluation

The evaluation of the results pertaining to the case studies presented in the prior section, to show the effectiveness of data-driven techniques, are discussed in detail as follows.

A. Case study 1: Smart meter energy consumption profiling

As the dataset described in Section IV-A does not have any record of energy theft, therefore, we have simulated a theft scenario by injecting synthetic anomalies for the purpose of identifying energy theft in this case study. The hypothesis for this synthetic injection is that an attacker tampers with the home energy management system of his neighbourhood to reduce the value of energy consumption reported from his/her household [6]. The premise behind this hypothesis is that the attacker would spread a share of his appliance measurements through the similar appliances in other legitimate households in order to avoid detection by keeping the overall energy consumption to the actual value. For this purpose, we injected false data values (to decrease the load) for a customer for a single type of load (i.e., fridge) and compared with the Fourier-based transformation of the injected data with the consumption data of the legitimate customer. An illustration of the FFT-transformation of the normal load profiles of various appliances in a home is depicted in Fig. 4. It is possible through the detailed examination of the coefficients of the Fourier-based transformations from the appliance consumption profiles to identify whether extremely high frequencies are present within the dataset.

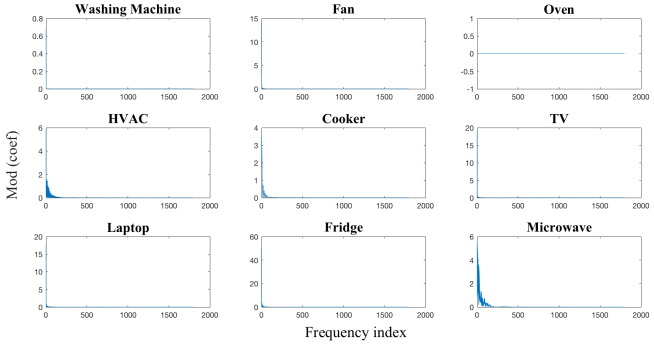
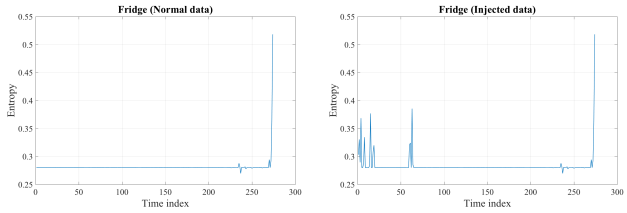


Figure 4: FFT coefficients of the appliances in DRED dataset.

By assessing the resulting transformations (from the normal and injected data) in terms of the appliance-specific measurements, we identified that there were slight variations in the injected data, which reflects on the possibility of theft. In order to further validate the hypothesis, the spectral entropy of the signals is analyzed as mentioned in Section IV-A. This analysis revealed some major changes in the spectral entropy of the normal and injected data in the appliance at particular time-slots (which portray theft) where the consumption values were modified as seen in Fig. 5. Therefore, we argue that it is possible to filter out the anomalous behavior of the customers using data-driven techniques and manually associate their power measurements with other aspects related to them.



(a) (b)

Figure 5: Spectral frequency analysis of the appliance.

B. Case study 2: Aggregated energy consumption profiling

Using the k-means clustering model presented in Section IV-B, the consumption profiles of various homes are modeled into different clusters on the basis of the historic data. The output depicts various average consumption profiles which serve as baselines profiles for theft detection. For example, a model for computing the baseline profiles in 200 homes using different number of cluster centroids is depicted in Fig. 6. This figure shows the baseline curves computed for one day using the historic data.

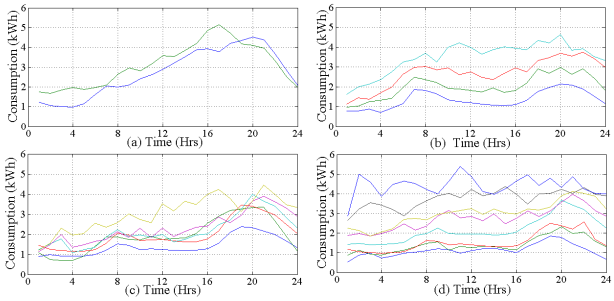
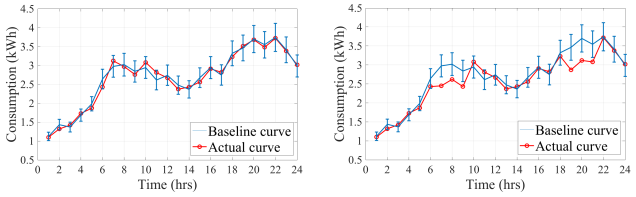


Figure 6: Cluster formation with different number of cluster centroids a) C=2 b) C=4 c) C=6 d) C=8 (adapted from [17]).

It can be seen in Fig. 6 that centroids with 4 cluster profiles are able to successfully give different load profiles that do not overlap with each other unlike the case with other cluster profiles. Thus, these could serve as the baseline load profiles for theft detection. To check if there is theft or not, the first step is to compare the load profile of the household with the baseline curves. Once the baseline curve is identified for a particular household, the anomaly can be identified if the consumption profile deviates for more than, say, 10%. An example of this comparison is shown in Fig. 7 where two scenarios are presented. The first scenario depicts the normal load consumption of the household, while in the second scenario, we injected false values (for simulating theft scenario) in order to reduce the reported energy consumption. One of the scenarios (normal load profile) shows that the actual consumption conforms to the baseline pattern whereas the other scenario depicts anomaly in the consumption pattern of the household as it deviates from the baseline curve. In this case, it can be inferred that the load pattern from 7am to 9am and 7pm to 9pm does not conform to the rest of the curve. Therefore, it can be said that this method successfully identified the time slots where false data values were injected.



(a) Normal load profile (b) Anomalous load profile

Figure 7: Analysis of the baseline model with actual values.

C. Discussion

The techniques we presented provide an insight on the importance of adequately characterizing the energy consumption at two major granularity levels (*i.e.,* smart meter and aggregate level). We argue that a building block towards energy theft detection is the data-driven composition of energy consumption profiling schemes that are able to flag anomalous patterns. Due to the dependency of raw energy consumption measurements (either on the smart meter or aggregate-level) on several factors (*e.g.,* temperature, humidity, and time of the year) that do not necessarily relate to the grid infrastructure, we suggest that such data-driven schemes should be broad and sophisticated enough to consider such factors. Moreover, the mathematical formulation of such schemes should also be able to conform with the highly dynamic and non-stationary nature of such measurements. Consequently, the resulting granular profiling of energy consumption would then be much more accurate and provide the ability to identify the root cause of an observed anomalous pattern. Overall, our case studies demonstrated the potential of identifying energy theft, a major challenge across the globe that could be intelligently confronted by employing data-driven schemes in conjunction with smart meters/AMIs that exist in a modern grid infrastructure. We showed here that how data at two granularity levels can be analysed and through correlation with additional (or situational) data, one might be able to develop a viable energy theft detection system.

However, other aspects also have to be addressed in the future. The data protection, the protection of privacy and the prevention of data misuse are all elements that have to be addressed when making use of data-driven schemes. While it is important that energy theft is identified and stopped, it is equally important that the data is not misused for any other purposes. Other, non-technical, aspects also have to be investigated, *e.g.*, related to urban design and socio-economic circumstances alongside wider infrastructure and community resilience aspects.

VI. Conclusion

Modern energy grid infrastructures (i.e. smart grids) offer improved energy generation, transmission, and distribution. However, the adoption of smart grids has not been uniform across the globe. While more developed nations are heavily investing in the modernization of their power grids, developing countries have a more patchy smart grid deployment. In this paper various data-driven techniques are investigated and applied to a set of relevant case studies to show how they can be used for the detection of electricity theft in smart grid infrastructures. In particular we discussed two data-driven energy theft detection schemes, at the smart meter level as well as at an aggregate level. We argue that the combination of per-household energy measurements as collected by individual smart meters and as aggregate measurements can be used to reliably identify anomalous measurements. Such anomalies can, for example, indicate the establishment of a clandestine connection or even meter tampering, which are common forms of energy theft.

Acknowledgement

This work was supported by ProSeG - Information Security, Protection and Resilience in Smart Grids, a research project funded by MCTI/CNPq/CT-ENERG (Grant # 404958/2013-3). This work has also received funding from the EU's Horizon 2020 research and innovation programme for "EASY-RES" project under grant agreement No 764090.

References

[1] K. Moslehi and R. Kumar, "A reliability perspective of the smart grid," *IEEE Transactions on Smart Grid*, vol. 1, no. 1, pp. 57–64, June 2010.

[2] E. P. R. Institute, "Estimating the costs and benefits of the smart grid," https://www.smartgrid.gov/files/Estimating_Costs_Benefits_Smart_Grid_Preliminary_Estimate_In_201103.pdf, 2011, last accessed 03 2019.

[3] "Global smart grid federation report," https://www.smartgrid.gov/files/Global_-Smart_Grid_Federation_Report.pdf, 2012, last accessed 03 2019.

[4] I. Colak, G. Fulli, S. Sagiroglu, M. Yesilbudak, and C.-F. Covrig, "Smart grid projects in europe: Current status, maturity and future scenarios," *Applied Energy*, vol. 152, pp. 58 – 70, 2015.

[5] C. P. Newswire, "World loses $89.3 billion to electricity theft annually, $58.7 billion in emerging markets," https://www.prnewswire.com/news-releases/world-loses-893-billion-to-electricity-theft-annually-587-billion-in-emerging-markets-300006515.html, 12 2014, last accessed 02 2018.

[6] A. Jindal, A. Dua, K. Kaur, M. Singh, N. Kumar, and S. Mishra, "Decision tree and svm-based data analytics for theft detection in smart grid," *IEEE Transactions on Industrial Informatics*, vol. 12, no. 3, pp. 1005–1016, June 2016.

[7] S. K. Singh, R. Bose, and A. Joshi, "Entropy-based electricity theft detection in ami network," *IET Cyber-Physical Systems: Theory Applications*, vol. 3, no. 2, pp. 99–105, 2018.

[8] P. Jokar, N. Arianpoo, and V. C. M. Leung, "Electricity theft detection in ami using customers' consumption patterns," *IEEE Transactions on Smart Grid*, vol. 7, no. 1, pp. 216–226, Jan 2016.

[9] A. K. Marnerides, P. Smith, A. Schaeffer-Filho, and A. Mauthe, "Power consumption profiling using energy time-frequency distributions in smart grids," *IEEE Communications Letters*, vol. 19, no. 1, pp. 46–49, Jan 2015.

[10] J. Pulz, R. B. Muller, F. Romero, A. Meffe, A. F. Garcez Neto, and A. S. Jesus, "Fraud detection in low-voltage electricity consumers using socio-economic indicators and billing profile in smart grids," *CIRED - Open Access Proceedings Journal*, vol. 2017, no. 1, pp. 2300–2303, 2017.

[11] J. B. Leite and J. R. S. Mantovani, "Detecting and locating non-technical losses in modern distribution networks," *IEEE Transactions on Smart Grid*, vol. 9, no. 2, pp. 1023–1032, March 2018.

[12] F. Xiao and Q. Ai, "Electricity theft detection in smart grid using random matrix theory," *IET Generation, Transmission Distribution*, vol. 12, no. 2, pp. 371–378, 2018.

[13] N. Kayastha, D. Niyato, E. Hossain, and Z. Han, "Smart grid sensor data collection, communication, and networking: a tutorial," *Wireless communications and mobile computing*, vol. 14, no. 11, pp. 1055–1087, 2014.

[14] A. S. Uttama Nambi, A. Reyes Lua, and V. R. Prasad, "Loced: Location-aware energy disaggregation framework," in *Proceedings of the 2Nd ACM International Conference on Embedded Systems for Energy-Efficient Built Environments*, ser. BuildSys '15. New York, NY, USA: ACM, 2015, pp. 45–54. [Online]. Available: http://doi.acm.org/10.1145/2821650.2821659

[15] C. Callegari, S. Giordano, and M. Pagano, "Entropy-based network anomaly detection," in *2017 International Conference on Computing, Networking and Communications (ICNC)*. IEEE, 2017, pp. 334–340.

[16] Open Energy Information, Available: http://en.openei.org/datasets/dataset/commercial-and-residential-hourly-load-profiles-for-all-tmy3-locations-in-the-united-states, Last accessed: Jan 2019.

[17] A. Jindal, M. Singh, and N. Kumar, "Consumption-aware data analytical demand response scheme for peak load reduction in smart grid," *IEEE Transactions on Industrial Electronics*, vol. 65, no. 11, pp. 8993–9004, Nov 2018.