Power Consumption Profiling using Energy Time-Frequency Distributions in Smart Grids

Angelos K. Marnerides, *Member, IEEE,* Paul Smith, *Member, IEEE,* Alberto Schaeffer-Filho, *Member, IEEE,* and Andreas Mauthe

***Abstract*—Smart grids are power distribution networks that include a significant communication infrastructure, which is used to collect usage data and monitor the operational status of the grid. As a consequence of this additional infrastructure, power networks are at an increased risk of cyber-attacks. In this letter, we address the problem of detecting and attributing anomalies that occur in the sub-meter power consumption measurements of a smart grid, which could be indicative of malicious behavior. We achieve this by clustering a set of statistical features of power measurements that are determined using the Smoothed Pseudo Wigner Ville (SPWV) energy Time-Frequency (TF) distribution. We show how this approach is able to more accurately distinguish clusters of energy consumption than simply using raw power measurements. Our ultimate goal is to apply the principles of profiling power consumption measurements as part of an enhanced anomaly detection system for smart grids.**

***Index Terms*—SCADA systems, Smart Grid, Clustering methods, Power measurement, Energy Time-Frequency Distributions**

I. Introduction

SMART grids are power distribution networks that depend on an increased level of automated monitoring and control [1]. To achieve this automation, new sensors and actuators are connected to Supervisory Control and Data Acquisition (SCADA) systems via wide-area communication networks. Alongside these SCADA systems is an Advanced Metering Infrastructure (AMI), which permits two-way communication between the smart meter at the customer premises and the utility company. Together, these systems can be used to collect usage data and monitor the operational status of the infrastructure. However, there are drawbacks of increased levels of automation in the smart grid, including higher system complexity, and a greater interdependency between several communication protocols and middleware components. Furthermore, a growing reliance on automation means that security threats can cause serious disruptions.

Hence, it is crucial that any challenge to the smart grid and supporting communications infrastructure is promptly detected and acted upon. To do this it is necessary to detect a range of challenges, including those that manifest as anomalies at the *network-level* (e.g., unexpected peaks in traffic volume) and the *appliance-level* (e.g., non-standard energy consumption and generation measurements). For example, data injection attacks may be used to change measurement values of some devices, in order to hinder the operation of the grid [2]. Further, Mohsenian-Rad and Leon-Garcia [3] describe *load altering attacks*, which attempt to cause circuit overflow or disturb the balance between power supply and demand in the grid. Approaches for detecting network-level anomalies in communication networks are relatively well-established. Furthermore, some research has focused on detecting attacks in SCADA communication protocols [4].

In this letter, we address the problem of detecting and attributing anomalies that occur in sub-meter power measurements, i.e., those for individual consumers of a smart grid. In particular, we aim to detect the effects of load altering attacks [3] that may be caused by an attacker sending direct load control (DLC) commands to appliances, resulting in unusual power consumption. For the clustering of power consumption we propose a method using a statistical description of power measurements. This is necessary because directly clustering raw power measurements, as is for instance employed by [5], leads to insufficient and inaccurate power consumption profiling, since the measurements are non-stationary. In order to describe the power measurements we apply the Smoothed Pseudo Wigner Ville (SPWV) energy Time-Frequency (TF) distribution, which results in a coherent statistical probability distribution. Next we extract a three-dimensional feature set that is based on the 3^{rd} order Rényi entropy and the mean frequency and time marginal values. These features are used as input to a clustering algorithm that can identify groups of residences with similar power consumption and anomalous outliers. The latter could indicate malicious behavior.

The three TF-based features that we have used in our analysis have, to the best of our knowledge, not previously been applied to power measurement profiling. Using these features we are able to clearly identify five power consumption clusters in a dataset that was used to validate our approach. These findings were cross-validated with the original raw power measurements. The overall aim of this work is to apply the principles of profiling power consumption as part of an attack detection system that draws on multiple data sources from the *network-* and *appliance-level* of a smart grid.

II. Profiling of Power Consumption Measurements

A. Dataset and Clustering Raw Power Measurements

The dataset[1] that we used for our analysis was gathered by the Smart* project [6], and contains anonymized average

This work was supported by the Brazilian MCTI/CNPq/CT-ENERG via the ProSeG project (grant 404958/2013-3), Capes/Brazil PVE (grant 13983130) and the EU-funded SPARKS project (grant 608224).

Angelos K. Marnerides is with the School of Computing & Mathematical Sciences, Liverpool John Moores University, Liverpool, UK, e-mail: a.marnerides@ljmu.ac.uk

Paul Smith is with the Austrian Institute of Technology (AIT), Vienna, Austria, e-mail: paul.smith@ait.ac.at

Alberto Schaeffer-Filho is with the Institute of Informatics, Federal University of Rio Grande do Sul (UFRGS), Porto Alegre, Brazil, e-mail: alberto@inf.ufrgs.br

Andreas Mauthe is with the School of Computing & Communications, Lancaster University, UK, e-mail: a.mauthe@lancaster.ac.uk

[1]Smart* Dataset: http://traces.cs.umass.edu/index.php/Smart/Smart

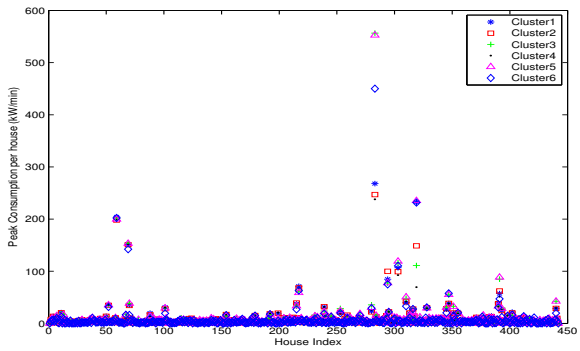

Fig. 1. Initial clustering with k-means by using raw power measurements.

power utilization measurements from 440 residential buildings in western Massachusetts, USA. The measurements were captured at one-minute intervals. In a similar fashion to related work [5], we initially attempted to cluster the raw power measurements with a well-known clustering scheme. We chose to use the k-means algorithm because of its minimal computational cost when compared to other schemes.

Initial analysis using the raw measurements indicated that k-means clustering could not provide a clear interpretation of the power measurements or appropriately group similar usage patterns that are present in the dataset. This is illustrated in Fig. 1. The graph shows that there is a number of examples where the power utilization of a given residence is similar to others, but they are not placed in the same cluster. We concluded that, because of the non-stationary nature of the measurements, this method of creating power consumption profiles is inappropriate. In order to overcome this problem, we looked to derive meaningful statistical features from each raw power measurement signal before applying k-means.

B. Determining Statistical Features for Effective Clustering

To determine a set of statistical features that can be used as input to a clustering algorithm, we compute for all the measurement signals, i.e., the power measurements from a single residence, their corresponding Smoothed Pseudo Wigner Ville (SPWV) Time-Frequency (TF) distribution. Determining a suitable probability distribution is a necessary first step in order to derive metrics from a stochastic and non-stationary process, i.e., the power measurements. The SPWV is a special case of the Wigner-Ville (WV) distribution, and is derived by the general class of the Cohen energy TF distributions [7]. The SPWV-TF was chosen because it is capable of handling non-stationary signals and can map their energy on the TF plane. In addition, the SPWV-TF also addresses the constraints related to auto, cross and interference terms that are not fully dealt with through the original WV distribution.

In more detail, let each power measurement signal for a given residence be denoted as $s(t)$, and its Hilbert-based analytical form be $s(u)$; the WV distribution which re-formulates the general Cohen distribution is then expressed as [7]:

$$WV(t,\omega) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} s^*(t-\frac{1}{2}\tau)e^{-j\tau\omega}s(t+\frac{1}{2}\tau)d\tau \quad (1)$$

Given this definition when $\tau \to 0$, the product of $s^*(t-\frac{1}{2}\tau)$ and $s(t+\frac{1}{2}\tau)$ contributes to the integral and is coherent. However, as indicated in [7], [8], the bilinear nature of the WV distribution has several shortcomings regarding windowing trade-offs, and there are cases when regional signal intensities are indicated, even though these are expected to be zero-valued. The interference terms that are not fully considered by the original WV distribution [7] may be dramatically reduced through a smoothing function $\zeta(\tau)$. The function $\zeta(\tau)$ can also be designed with explicit TF smoothing kernel functions in order to adapt sampling in real-time and enable the efficient online computation of the SPWV over a non-stationary measurement signal [8]. When $\zeta(\tau)$ is employed in the WV distribution in Eq. 1, we obtain the SPWV [7], [8]:

$$SPWV(t,\omega) = \int_{-\infty}^{+\infty} \zeta(\tau)s(t+\frac{\tau}{2})s^*(t-\frac{\tau}{2})e^{-j\pi\omega\tau}d\tau \quad (2)$$

Subsequently, we extract the descriptive statistics of the *time* and *frequency marginals* alongside the 3^{rd} *order Rényi entropy* in order to characterize the resulting probability energy TF distribution on the raw power measurements per residence. These features are selected as they empirically lead to better k-means clustering results when compared to other metrics derived by the SPWV distribution, such as the resulting total unit-less energy metric of the SPWV distribution and the 1^{st} order frequency and time moments.

The time (seconds) and frequency (Hz) marginals, which jointly describe the marginal probability distribution of the values contained within the overall SPWV probability distribution, are extracted as follows:

$$m_\omega(t) = \int_{-\infty}^{+\infty} SPWV(t,\omega)d\omega \quad (3)$$

where $m_\omega(t)$ denotes the time marginal and $m_t(\omega)$ is the frequency marginal expressed as:

$$m_t(\omega) = \int_{-\infty}^{+\infty} SPWV(t,\omega)dt \quad (4)$$

We compute the 3^{rd} order Rényi entropy since it has been shown to be a good discriminative feature in other disciplines [9]. In particular, Baraniuk *et al.* [8] were able to estimate the 3^{rd} order Rényi entropy R^{α}_{SPWV} using the following definition:

$$R^{\alpha}_{SPWV} = -\frac{1}{2}log_2 \iint_{-\infty}^{+\infty} SPWV^{\alpha}(t,\omega)dtd\omega \quad (5)$$

The Rényi entropy is a generalization of Shannon entropy, which provides a means of describing the level of complexity of a signal, and is measured in bits. The entropy order is denoted by the parameter α which in our case is $\alpha = 3$. When $\alpha = 1$ we recover the Shannon entropy as well as the Kullback-Leibler divergence [8]. Despite the fact that the general form of the Rényi entropy depends on the entropy order, we strictly compute it for $\alpha = 3$, since lower α values would negatively affect the complexity representation for any Cohen-based distribution [8].

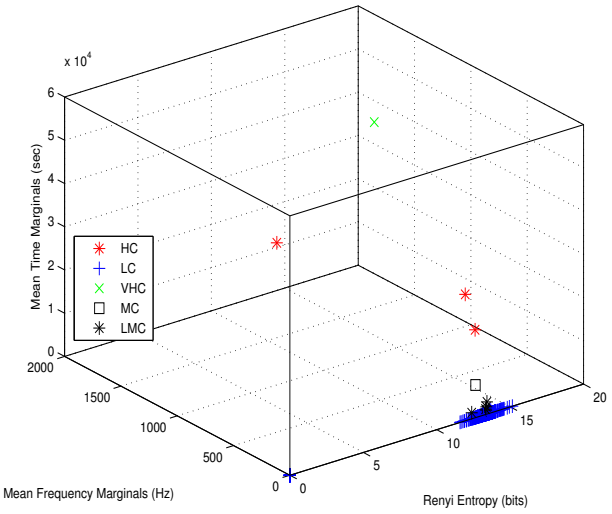

Fig. 2. Clustering of power consumption based on TF features under five clusters, namely: LC = Low Consumption, LMC = Low-to-Medium Consumption, MC = Medium Consumption, HC = High Consumption, VHC = Very High Consumption.

C. Applying k-means Clustering

Using the statistical features calculated earlier, we apply the k-means clustering algorithm in order to group similar power measurements, and possibly identify *outliers* that represent unusual consumption patterns that could indicate malicious behavior. We initially construct a composite i by j matrix X, where each row i represents a residence and each column j represents the computed feature. In this study, $i = 440$ and $j = 3$, since there are 440 residences and three features. Then, we apply k-means to identify the most representative k clusters. Initially, k random rows in X are selected and marked as the centroids for k clusters C, respectively $[C_1, C_2, \cdots, C_k]$. In our experiments, we determine the most representative number of centroids by minimising the sum of squares within a cluster using Eq. 6, where X_i denotes the actual row data vector that represents a single residence from a total of M residences, c_n is a given cluster centroid of the k resulting clusters, and $||X_i - c_n||^2$ is defined as the Euclidean distance between the data vector for a residence and its centroid.

$$\sum_{n=1}^{k} \sum_{i=1}^{M} ||X_i - c_n||^2 \tag{6}$$

In our experiments, we apply k-means iteratively, minimizing the sum of squares within a cluster on every iteration, in order to identify the optimal number of clusters.

III. Clustering Results Analysis

Fig. 2 depicts the final number of clusters and provides a visualization of the relationship between these clusters and their corresponding ranges of TF energy statistical features. There is a clear separation of per-residence power consumption into five distinct clusters:

1) **Low Consumption (LC)**: Defined by the majority, consisting of 429 houses where power consumption ranges between 0-10 kW/min;
2) **Low-Medium Consumption (LMC)**: Defined by a set of four residences where power consumption is between 10-70 kW/min with mild peak fluctuations;
3) **Medium Consumption (MC)**: A single residence that exhibits high peaks within the range of 45-120 kW/min;
4) **High Consumption (HC)**: Mapped to a set of three residences, in which each demonstrates mild to high fluctuations of the average peak power and most of their measured values range between 10-200 kW/min; and
5) **Very High Consumption (VHC)**: A single house with an initial consumption of 130 kW/min and extreme fluctuation on its power peaks that reaches up to 530 kW/min.

This clustering approach has provided a distinct separation of low power utilization users, which were placed in the LC cluster. Given the visual representation of this cluster, it can be seen that its most discriminative feature is the similar range of values in the mean frequency marginals of the SPWV distribution (i.e., the z-axis in Fig. 2). In particular, all houses that belong to the LC cluster have a range of 0-210 Hz with a mean time marginal that varied in the range of 0-1 seconds, and the Rényi entropy for the majority of these houses ranges between 12 and 16 bits. Moreover, none of the houses assigned to this cluster consumed large amounts of power per minute, as can be seen in Fig. 3.

In a similar fashion, the houses assigned to the LMC and MC clusters had as their most distinguishable feature the range of values for the mean frequency marginals of the SPWV distribution. Despite their similar values for the Rényi entropy and estimated mean time marginals, the members of the LMC and the single member of the MC cluster showed higher mean frequency marginals than the LC. This means that their raw power measurements had more frequent peaks than houses in the LC cluster. However, apart from higher frequency marginals in its corresponding SPWV distribution, the single house of the MC cluster had also a higher mean time marginal with a larger power consumption compared to the four residences in the LMC cluster, as shown in Fig. 4[2].

Members of the HC cluster exhibited higher mean frequency marginals, with medium-ranged mean time marginals that correspond to frequent fluctuations of peak power utilization. Nevertheless, none of the houses assigned to the HC cluster reached more than 200 kW/min on a single peak, and their peak fluctuations did not significantly exceed their average power peak (which was $\approx$ 142 kW/min). However, as indicated in Fig. 5, the single house that exhibited excessive power consumption (VHC cluster) had a sudden increase of power utilization that remained constant for a period of 12.5 hours (i.e., 750 minutes). The behavior demonstrated by this particular house can also be seen in Fig. 2, in which a single point indicates a higher mean time marginal than any other data point, as well as higher Rényi entropy.

Overall, the outcomes of our clustering approach demonstrate that power consumption profiles can be established based on energy TF features that are derived from the SPWV

[2]In Fig. 3, Fig. 4 and Fig. 5 we show only representative houses of the LC, LMC and HC clusters in order to more clearly demonstrate the differences in power consumption between houses in these clusters.

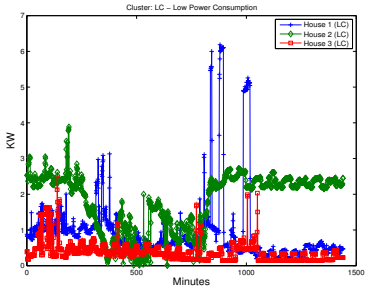

Fig. 3. Exemplar power consumption of three residences from the Low Consumption (LC) cluster.

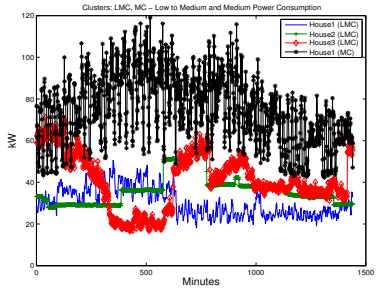

Fig. 4. Exemplar consumption of three residences from the Low-Medium Consumption (LMC) and one of the Medium Consumption (MC) cluster.

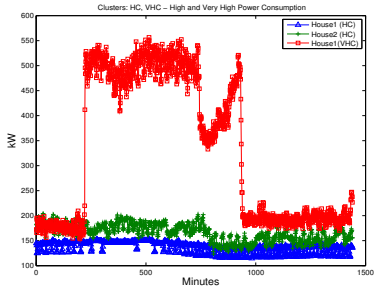

Fig. 5. Exemplar consumption from two residences of the High Consumption (HC) cluster and from the only member of the Very High Consumption (VHC) cluster.

distribution. Consumption profiles characterized by low, low-medium and medium utilization can be easily distinguished via the mean time and frequency marginals. Also, the separation of houses with higher power consumption (HC and VHC clusters) can be distinguished based on the mean time marginal and the Rényi entropy. Arguably, a shortcoming of our approach is its inability to place residences with all-zero measurements (shown at the origin of Fig. 2) in a unique cluster – these measurements could be indicative of malicious behavior. Experiments aimed at addressing this problem with different numbers of clusters led to over-fitting. This issue motivates our overall goal to integrate this profiling technique into a system that collectively examines both appliance- and network-level anomalies. For example, a lack of network packets from a smart meter could indicate that physical tampering has occurred, or network traffic is being blocked or manipulated at an intermediate point – data from sensors that measure these characteristics could be used to build a hypothesis regarding the root cause of an anomaly. Furthermore, based on the introduced scheme it will be possible to correlate and cluster network communication and appliance-related features from a number of sub-meters in order to adequately profile distinct situations that are either caused by local system failures or distributed power load attacks.

IV. Concluding Remarks

Resilient operation of smart grid communication networks must ensure the detection of anomalies that manifest at the *network-level* and the *appliance-level*. Although attack detection in SCADA communication protocols has attracted some attention recently, we advocate in this letter that it is also necessary to detect and attribute anomalies that occur in the appliance-level measurements. To achieve this we apply a Smoothed Pseudo Wigner Ville (SPWV) energy Time-Frequency (TF) distribution as a basis for creating statistical features that describe power measurements. A significant benefit of using SPWV is that the input timeseries can be non-stationary. This is in contrast to timeseries models that are used in commercial tools (e.g., IBM's SPSS[3]) that require the (weak) stationarity of measurements. This is a shortcoming that results in the need for transforming a signal, potentially leading to errors in later analysis stages. Having applied SPWV, we use the k-means clustering algorithm to identify outliers that could indicate malicious behavior. Our results are promising and we anticipate that profiling power consumption using the techniques introduced in this letter can be part of an enhanced anomaly detection system. Additionally, we foresee detailed power profiles being useful for grid capacity planning and generating accurate consumption forecasts that can be applied for energy trading [10]. Future work will investigate how to realize our approach in an online manner. Specifically, we will adapt the MATLAB-based routines that we have used for our analysis to a parallel programming model, such as MapReduce. Furthermore, we will investigate the performance benefits of data sampling, so that very large data volumes can be processed in near real-time.

[3]IBM SPSS Statistics: http://www-01.ibm.com/software/uk/analytics/spss/products/statistics/

References

[1] Y. Yan, Y. Qian, H. Sharif, and D. Tipper, "A survey on smart grid communication infrastructures: Motivations, requirements and challenges," *Communications Surveys Tutorials, IEEE*, vol. 15, no. 1, pp. 5–20, January-March 2013.

[2] P.-Y. Chen, S.-M. Cheng, and K.-C. Chen, "Smart attacks in smart grid communication networks," *Communications Magazine, IEEE*, vol. 50, no. 8, pp. 24–29, August 2012.

[3] A.-H. Mohsenian-Rad and A. Leon-Garcia, "Distributed Internet-based load altering attacks against smart power grids," *Smart Grid, IEEE Transactions on*, vol. 2, no. 4, pp. 667–674, Dec 2011.

[4] Y. Yang, K. McLaughlin, S. Sezer, T. Littler, E. Im, B. Pranggono, and H. Wang, "Multiattribute SCADA-specific intrusion detection system for power networks," *Power Delivery, IEEE Transactions on*, vol. 29, no. 3, pp. 1092–1102, June 2014.

[5] D. Gerbec, S. Gasperic, I. Smon, and F. Gubina, "Consumers' load profile determination based on different classification methods," in *Power Engineering Society General Meeting, IEEE*, vol. 2, July 2003.

[6] S. Barker, A. Mishra, D. Irwin, E. Cecchet, P. Shenoy, and J. Albrecht, "Smart*: An open data set and tools for enabling research in sustainable homes," in *Workshop on Data Mining Applications in Sustainability (SustKDD)*, August 2012.

[7] L. Cohen, "Time-frequency distributions – a review," *Proceedings of the IEEE*, vol. 77, no. 7, pp. 941–981, July 1989.

[8] R. Baraniuk, P. Flandrin, A. J. E. M. Janssen, and O. Michel, "Measuring time-frequency information content using the Renyi entropies," *Information Theory, IEEE Transactions on*, vol. 47, no. 4, pp. 1391–1409, May 2001.

[9] A. K. Marnerides, D. Pezaros, H. Kim, and D. Hutchison, "Internet traffic classification using energy time-frequency distributions," in *Communications (ICC), 2013 IEEE International Conference on*, June 2013, pp. 2513–2518.

[10] S. Al Kaabi, H. Zeineldin, and V. Khadkikar, "Planning Active Distribution Networks Considering Multi-DG Configurations," *Power Systems, IEEE Transactions on*, vol. 29, no. 2, pp. 785–793, March 2014.