A Clustering Approach for Consumer Baselining and Anomaly Detection in Transactive Control

Nasheen Nur

nnur@uncc.edu

University Of North Carolina at
Charlotte
Charlotte, North Carolina, USA

Siddharth Sridhar*

siddharth.sridhar@pnnl.gov

Pacific Northwest National
Laboratory
Richland, Washington, USA

Seemita Pal*

seemita.pal@pnnl.gov

Pacific Northwest National
Laboratory
Richland, Washington, USA

Aditya Ashok*

aditya.ashok@pnnl.gov

Pacific Northwest National
Laboratory
Richland, Washington, USA

Vinay C Amatya*

vinay.amatya@pnnl.gov

Pacific Northwest National
Laboratory
Richland, Washington, USA

ABSTRACT

The arrival of smart grid technologies has enabled the implementation of advanced grid control applications aimed at improving overall grid reliability and operational efficiency. The new controls are supported by advanced digital devices and communication networks that often possess inherent vulnerabilities, exposing the grid infrastructure to threat actors. Transactive control is one such smart grid application, wherein active participation of end-use customers in electricity markets is enabled by providing the flexibility to consume energy based on the dynamic price of power. Manipulation of signals exchanged in this control system will diminish the benefits of the application, thereby highlighting the need for control system cybersecurity. In this paper, we present a data-driven approach to building bad-data detection algorithms for transactive control security. Specifically, our solution characterizes end-user market participation using clustering techniques to create a baseline for normal system operation. We then show the application of these clusters to detect malicious behavior during real-time system operation. Our initial results show the promise this approach has towards enhancing the security of transactive control systems.

*PNNL is a multiprogram national laboratory operated by Battelle for the U.S. Department of Energy under contract No. DE-AC05-76RL01830. This work was sponsored by the U.S. Department of Energy as part of the Grid Modernization Laboratory Consortium (GMLC) program.

ACM acknowledges that this contribution was authored or co-authored by an employee, contractor, or affiliate of the United States government. As such, the United States government retains a nonexclusive, royalty-free right to publish or reproduce this article, or to allow others to do so, for government purposes only.

e-Energy '19, June 25–28, 2019, Phoenix, AZ, USA

© 2019 Association for Computing Machinery.

ACM ISBN 978-1-4503-6671-7/19/06...\$15.00

<https://doi.org/10.1145/3307772.3331028>

CCS CONCEPTS

• **Computing methodologies** → **Cluster analysis**; **Anomaly detection**; • **Computer systems organization** → **Embedded hardware**; • **Security and privacy** → *Artificial immune systems*; • **Applied computing** → *Computer-aided design*.

KEYWORDS

Anomaly Detection, Transactive Control, Clustering

ACM Reference Format:

Nasheen Nur, Siddharth Sridhar, Seemita Pal, Aditya Ashok, and Vinay C Amatya. 2019. A Clustering Approach for Consumer Baselining and Anomaly Detection in Transactive Control. In *Proceedings of the Tenth ACM International Conference on Future Energy Systems (e-Energy '19)*, June 25–28, 2019, Phoenix, AZ, USA. ACM, New York, NY, USA, 6 pages. <https://doi.org/10.1145/3307772.3331028>

1 INTRODUCTION

Transactive control is a smart grid control application that allows energy consumers to modulate their energy consumption in response to the real-time price of energy. This application, which combines demand-side management with traditional supply-side generation control, has been shown to benefit grid operational efficiency and reliability through use cases such as peak load reduction. Consumers also benefit from reduced energy costs by shifting energy consumption to periods of low-cost energy [12].

Adoption of such systems is gaining traction, with utilities implementing variations of demand-side management to leverage its benefits. The *gridSMART* demonstration project involving American Electric Power (AEP) Ohio featured the implementation of a dual-auction market, wherein energy consumption by residential air conditioning units was controlled using pricing signals [15]; analysis of data collected over a two-year period revealed that the system reduced

peak load and energy cost by 5%. The New York Independent System Operator (ISO) offers monetary benefits to customers to participate in reliability- and economy-based demand response programs, with an already enrolled capacity upwards of 1000 MW [10].

The arrival of price-responsive smart devices that can monitor and control energy consumption of the loads has spurred widespread implementation of transactive control. These devices, including smart energy meters and thermostats, are also capable of two-way communication, enabling the exchange of energy bids and price with the distribution system operator (DSO). These encrypted exchanges typically occur over wide-area networks that leverage public network infrastructures [15]. In several recent incidents, threat actors have been able to compromise systems by getting around encryption measures through credential theft and zero-day vulnerabilities [5]. Therefore, despite encryption, the exposure of these systems to the internet makes them vulnerable to compromise by adversaries. In addition, to the sophistication required to compromise cyber systems, the adversary also possesses in-depth understanding of the physical system being controlled. This means a successful compromise of a utility's transactive control network could provide the adversary with control over large amounts of load to strategically manipulate and severely impair grid operation.

Traditional approaches to securing control systems have involved the use of host and network-based cybersecurity solutions such as, firewalls and intrusion detection systems [11, 14, 16]. As much as these information technology (IT)-based security mechanisms are important, defense algorithms built into the operations technology (OT) will add defense-in-depth.

In this paper, we present a novel detection algorithm to be built into the operations technology to secure the transactive control system against data integrity attacks. The methodology employs machine learning to create a baseline of the usual temporal behavior of participating consumers and use it to monitor online performance. Deviations from the baseline could then be marked as anomalies that need further investigation for potential cyber compromise. This paper primarily focuses on the methodology for baselining and presents preliminary results from its application to anomaly detection. We explore the viability of our approach as a means of differentiating types of transactive control system disturbances to focus specifically on detecting cyber-attacks (i.e. ramping attack, scaling attack). We evaluate various scenarios as disturbance discriminators and conclude with the practical implications for deploying our system as an enhancement to existing smart grid architectures.

The remainder of the paper is organized as follows: Section 2 gives a general overview of transactive control and

describes the necessity of the presented research. We introduce our clustering methodology in Section 3, followed by experimental setup and analysis of results in Section 4. The paper concludes with future directions in Section 5.

2 OVERVIEW OF TRANSACTIVE CONTROL

Figure 1: Overview of transactive control

Transactive control is a market-based control application that uses energy pricing as a signal to influence energy consumption. In addition to offering lower energy costs to the consumer, it provides significant benefits to the utility, including peak load shaving and optimal use of renewables. Fig. 1 depicts the implementation of the AEP style of transactive control. Here, the consumers submit a bid to the DSO every 5 minutes, indicating the quantity of energy required and the price they are willing to pay. The DSO aggregates bids from all consumers participating in the market and presents an aggregated demand curve to the ISO. The ISO runs an optimization algorithm that takes into account the available generation, overall demand, transmission constraints, and other factors to arrive at a bus-specific locational marginal price (LMP). The LMP, which is the energy price for this cycle, is conveyed to the DSO, from where it reaches the consumers. At the consumer end, loads are cycled based on the LMP and the acceptable energy price configured by the customer.

Market-based signals directly influence grid operation in this closed-loop control system. A sophisticated adversary could exploit this influence to inject data manipulation attacks that could destabilize operation. For example, if the LMP communicated to the end consumer is scaled down to create the illusion of cheap energy costs, the flexible loads in the system could turn on, increasing load on the grid. This

scenario could, in turn, result in operational violations in the form of thermal overloads and voltage sags. The security of these systems therefore becomes critical to ensure normal operation.

The IT-based approach to protecting these systems employs firewalls and link encryption to secure networks and data. To enable defense-in-depth and bolster end-to-end security, it is important to design security into the transactive control application as well [2, 9]. The approach explored in this paper involves the use of machine learning to baseline consumer behavior in the transactive control system and use it identify deviations from normal operation.

3 METHODOLOGY FOR BASELINING CONSUMER BEHAVIOR IN TRANSACTIVE CONTROL

Baselining is an offline operation that processes historical bidding data from the transactive control system to characterize normal consumer behavior. The objective of the methodology is to identify a suitable approach to baselining consumer behavior to help detect cyber-threat-related anomalies. Toward this end, we adopt a three-step process, shown in Fig. 2, to characterize consumers' bidding patterns. Specifically, the methodology takes historical bidding data from consumers as input, and outputs clusters of consumers grouped together by similarity in bidding patterns. These clusters will be used during online operation to identify consumer bids that are out of character, potentially indicating manipulation from a threat actor. The details of the methodology are as follows:

Figure 2: Methodology for baselining consumer behavior

3.1 Step 1 - Data Preprocessing

The objective is to clean up the dataset before providing as input to Step 2. Specifically, the bid data is separated based on the following factors:

- *Season* - Weather variation between seasons will affect energy consumption, thereby influencing bidding patterns. Consequently, bidding data sorted by season and provided as input to the methodology will help develop season-specific clusters that are better suited to characterize consumer behavior for the period. The season split used for this effort is as follows: June–August is summer, September–November is fall, December–February is winter, and March–May is spring.
- *Day of Week* - Similarly, energy consumption patterns also differ by day of the week. For this effort, we use the broad classifications of weekday and weekend.
- *Time of Day* - Time of day directly affects energy consumption and bidding patterns. In this paper, we split bid data into blocks of time ranging from 30 minutes to 8 hours, from which the ideal split, as revealed by our performance measure, will be used to generate signatures.

3.2 Step 2 - Clustering

Preprocessed bid data from step one is provided as input to clustering. The following key decisions are made in this step:

- *Clustering Algorithm* - The suitability of the *k-means* [8], *Agglomerative* [13], and *DBSCAN* [6] algorithms is evaluated for this application. *K-means* is one of the fastest clustering algorithms, which scales well to a larger number of samples. In this case, the large number of samples is a large time-series dataset. However, the challenge with the *k-means* algorithm is the requirement to identify a good initial estimate of the number of clusters to prevent converging to local minima. Unlike *k-means* clustering, the *Agglomerative* method does not require a good initial estimate for the number of clusters. However, this algorithm, which adopts a variance-minimizing approach, is computationally expensive. *DBSCAN* is a density-based clustering algorithm that works well on non-linearly separable distributions. However, the algorithm could fail to identify clusters if the datasets are too sparse.
- *Number of Clusters* - The number is varied from an arbitrary lower limit of five to the number of participating consumers, as it is expected that the number of clusters will increase with the number of participating consumers.
- *Distance Measures* - The choice of distance measures is crucial as it determines how the similarity of two elements is calculated, which strongly influences the clustering results. The Euclidean [1] and Dynamic Time Warping (DTW) [3] approaches are two most-commonly-used distance measures for time-series clustering. Euclidean distance calculates the point-to-point

distance between two time series, while DTW is capable of doing the same thing for time series with varying speed.

A brute-force enumeration using all combinations of the above variables and blocks of data from time-based preprocessed dataset is made, and the performance of each combination is measured with the Caliński-Harabasz (CH) score [4]. The CH score is a good indicator of the optimality of the number of clusters, with a higher value indicating better performance.

3.3 Step 3 - Signature Generation

The results of the clustering step will help identify a specific combination of clustering algorithm, number of clusters, and distance measure that has the highest performance scores. The application of this combination to a dataset will output a season-, day-of-week-, and time-of-day-specific cluster set, identifying consumers that belong to each cluster. The idea is that consumers clustered together are identified to have similar bidding patterns and may thus be represented by a signature. In a two-dimensional space, a signature curve is the nearest representation of a bidding pattern characterizing all bidding patterns belonging to that cluster.

The clusters and their signatures could then be employed for comparison with the actual bids exchanged during online operation. This paper primarily focuses on the methodology for generating the clusters and their signatures. However, preliminary results from their application to anomaly detection are introduced in Section 4.

4 CASE STUDY AND EXPERIMENTAL RESULTS

In this section, we first introduce the experimental setup that was used for generating the required dataset. Next, we demonstrate the application of the baselining methodology to the dataset to generate the clusters and their signatures. Finally, we introduce preliminary results from the application of the clusters to anomaly detection.

4.1 Experimental Setup for Dataset Generation

Our case study was performed using a transactive control dataset generated from an environment consisting of MATPOWER, GridLAB-D, and the Framework for Network Co-Simulation (FNCS). MATPOWER was used to model the bulk energy market, while GridLAB-D was used to model the distribution system down to the house level, including transactive control. Specifically, transactive control was implemented at one load bus, which used GridLAB-D prototypical feeders to create a system with three feeders and 2236 houses. FNCS provided a message bus to allow MATPOWER and GridLAB-D to exchange information seamlessly. In MATPOWER, the LMPs were computed every 5 minutes, based

on the aggregated demand curve calculated from the individual demand bids sent from the houses. The LMPs were then dispatched to the consumers to suitably cycle their loads. A year's worth of simulation was run using this setup, from which the consumer-specific bid data was extracted and inputted to the methodology.

4.2 Application of the Baselining Methodology

The three-step baselining methodology was applied to the bid dataset obtained from the experimental setup.

4.2.1 Step 1 - Data Preprocessing. The dataset was first divided by season and day-of-week per the guideline defined in Section 3. For time-of-day split, the dataset was divided into blocks of time ranging from 30 minutes to 8 hours. Combinations were created using these blocks of data along with the clustering algorithm, distance measure, and number of clusters. The CH score was used to measure the performance of these combinations. The results revealed one common trait across all combinations of clustering algorithm, distance measure, and number of clusters. A time-of-day split of 30 minutes provided the best performance. This is evidenced by Table 1, wherein the scores for different time splits using the *k*-means algorithm and Euclidean distance are presented for summer and winter weekdays. It was observed that as the duration of time-split increased from 30 minutes to 8 hours, the score worsened.

Table 1: Variation of Performance with Time-of-Day Split

Duration of Time-Split	Caliński-Harabasz Score	
	Summer	Winter
30 minutes	8156.45	2990.9
1 hour	5656.15	2006.71
2 hours	4145.91	1499.25
3 hours	3312.05	1206.36
4 hours	2757.12	1019.07
6 hours	2422.04	877.43
8 hours	2159.94	787.36

4.2.2 Step 2 - Clustering. In addition to blocks of time for identifying the best time-of-day split, the other variables for the brute-force enumeration included the clustering algorithm, distance measures, and number of clusters. Table 2 presents the results from the brute-force analysis, with performance measured using the CH scores. The table shows results of the analyses from a specific period of operation: 8:30 to 9:00 AM on a summer weekday. Specifically, the table only shows a filtered subset of results—that is, only the top five combinations as measured using the score. From the table, it is observed that the *k*-means algorithm with the Euclidean distance measure outperforms all other combinations. In addition, the table also tells us that for this time period, a cluster count of 50 provides the best results.

Table 2: Results of Brute-Force Enumeration

Algorithm	Distance	Number of Clusters	Calinski-Harabasz Score
K-means	Euclidean	50	1089.02
Agglomerative	Euclidean	50	832.94
Agglomerative	DTW	170	549.18
K-means	DTW	50	288.741
DBSCAN	DTW	N/A	220.188

It was also observed that the *k*-means clustering algorithm with the Euclidean distance measure outperformed other combinations for all periods of analysis. However, the optimal number of clusters varied for these different time periods. Table 3 shows the variation in the optimal number of clusters required to characterize consumer behavior for the different time periods. It was observed that the number of clusters required represents consumer behavior not only depending on the time, but also on the number of consumers participating in the transactive market. Per our model, more consumers participate in the market during a summer weekday, which results in a larger number of clusters during the period. The degree of influence of the time of operation and count of participating consumers on the number of clusters will be studied in our future work.

Table 3: Optimal Number of Clusters by Time Period

Time of Day	Optimal Number of Clusters			
	Summer Weekdays	Summer Weekends	Winter Weekdays	Winter Weekends
1:00 PM–1:30 PM	270	50	70	50
1:30 PM–2:00 PM	270	50	110	70
2:00 PM–2:30 PM	290	50	140	70
2:30 PM–3:00 PM	290	70	170	90
3:00 PM–3:30 PM	270	50	210	90

The previous step provides season, day-of-week, and time-of-day-specific clusters that can be used to generate signatures. The signatures are obtained by averaging the bidding characteristics of all the consumers in a cluster at every point in the time period. Fig. 3 shows examples of good clustering with low variance (a) and suboptimal clustering with a large variance (b). In Fig. 3 (a), a case where 50 clusters were used to represent all the houses, the cluster signature, represented by the bold red line, is able to very closely match the bidding patterns of the 49 houses in that cluster. However, in Fig. 3b, a suboptimal case in which only five clusters were used to represent all the houses in the system, the signature fails to accurately represent the 508 houses present in the cluster.

4.3 Application of Clustering to Anomaly Detection

(a) Example of good clustering **(b) Example of suboptimal clustering****Figure 3: Good clustering vs. suboptimal clustering** **Figure 4: Detection of scaling attack using detection threshold**

The consumer baselining methodology is intended to be used at the DSO to identify manipulations of consumer bidding patterns. Application of the methodology in the offline stage would provide the DSO with a time-of-operation-specific cluster portfolio that maps consumers to specific clusters, and identifies a suitable signature that best represents the bidding pattern of the houses in that cluster. During online operation, the bidding pattern of consumers will constantly be compared to the signature of the cluster to which they belong, and significant deviations from the signature marked as anomalies.

For a proof-of-concept demo, a sample cluster was selected for developing a threshold of acceptable bid values. At every

point in time, the highest bid value from the historical dataset was selected as the upper bound and vice versa. These upper and lower bounds form an envelope of acceptable bid values (shaded area in Fig. Figure 4) around the cluster signature. The rule employed for detection is that if a bid value from a house within the cluster falls outside the envelope, it is an anomaly.

An attack dataset for a scaling attack, wherein bids from consumer to the DSO are scaled by a magnitude of 30% was generated and tested with the anomaly detection block introduced above [7]. When plotted against the envelope in Fig. 4, it is observed that the attack values fall well outside the threshold, and would hence be flagged as anomalies.

The approach to anomaly detection presented in this paper is very basic. A systematic design process should first consider a thorough enumeration of impactful attack models and ranges. This insight should then be used in the generation of the cluster signature and detection envelope through a systematic process that aims to guarantee minimal false-positive and negative rates. These steps are a part of our future work.

5 CONCLUSION AND FUTURE WORK

Transactive control is a price-based load regulation scheme that offers benefits to system operators and consumers alike. Its implementation depends on the exchange of signals through wide-area communication channels, whose inherent cyber vulnerabilities could be exploited by sophisticated threat actors to affect grid operation. This paper introduces a methodology to detect data integrity attacks involving manipulation of consumer bids. The approach leverages clustering techniques to classify and baseline normal consumer bidding behaviors during the offline phase. The baseline can then be used for anomaly detection during normal operation to identify consumer bids that are out of character. Our future work includes thorough investigation of the application of the baselining methodology to anomaly detection.

REFERENCES

- [1] Howard Anton and Chris Rorres. 1994. *Elementary Linear Algebra*, John Wiley & Sons. Inc., New York (1994).
- [2] A. Ashok, M. Govindarasu, and J. Wang. 2017. Cyber-Physical Attack-Resilient Wide-Area Monitoring, Protection, and Control for the Power Grid. *Proc. IEEE* 105, 7 (July 2017), 1389–1407. <https://doi.org/10.1109/JPROC.2017.2686394>
- [3] Donald J Berndt and James Clifford. 1994. Using dynamic time warping to find patterns in time series.. In *KDD workshop*, Vol. 10. Seattle, WA, 359–370.
- [4] Tadeusz Caliński and Jerzy Harabasz. 1974. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods* 3, 1 (1974), 1–27.
- [5] Defense Use Case. 2016. Analysis of the cyber attack on the Ukrainian power grid. *Electricity Information Sharing and Analysis Center (E-ISAC)* (2016).
- [6] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise.. In *KDD*, Vol. 96. 226–231.
- [7] Yu-Lun Huang, Alvaro A Cárdenas, Saurabh Amin, Zong-Syun Lin, Hsin-Yi Tsai, and Shankar Sastry. 2009. Understanding the physical and economic consequences of attacks on control systems. *International Journal of Critical Infrastructure Protection* 2, 3 (2009), 73–83.
- [8] James MacQueen et al. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, Vol. 1. Oakland, CA, USA, 281–297.
- [9] S. Sridhar, A. Hahn, and M. Govindarasu. 2012. Cyber-Physical System Security for the Electric Power Grid. *Proc. IEEE* 100, 1 (Jan 2012), 210–224. <https://doi.org/10.1109/JPROC.2011.2165269>
- [10] Mathangi Srinivasan. [n. d.]. New York ISO. https://www.nyiso.com/public/webdocs/markets_operations/services/market_training/workshops_courses/Training_Course_Materials/NYMOC_MT_ALL_201/Demand_Response.pdf accessed 2018-11-02.
- [11] Chee-Wooi Ten, Junho Hong, and Chen-Ching Liu. 2011. Anomaly detection for cybersecurity of the substations. *IEEE Transactions on Smart Grid* 2, 4 (2011), 865–873.
- [12] John S Vardakas, Nizar Zorba, and Christos V Verikoukis. 2015. A survey on demand response programs in smart grids: Pricing methods and optimization algorithms. *IEEE Communications Surveys & Tutorials* 17, 1 (2015), 152–178.
- [13] Joe H Ward Jr. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association* 58, 301 (1963), 236–244.
- [14] Dong Wei, Yan Lu, Mohsen Jafari, Paul Skare, and Kenneth Rohde. 2010. An integrated security system of protecting smart grid against cyber attacks. In *2010 Innovative Smart Grid Technologies (ISGT)*. IEEE, 1–7.
- [15] Steven Widergren, Kris Subbarao, Jason Fuller, David Chassin, Abhishek Somani, Cristina Marinovici, and Janelle Hammerstrom. 2014. *AEP Ohio gridSMART Demonstration Project Real-Time Pricing Demonstration Analysis*. Technical Report.
- [16] Saman Zonouz, Katherine M Rogers, Robin Berthier, Rakesh B Bobba, William H Sanders, and Thomas J Overbye. 2012. SCPSE: Security-oriented cyber-physical state estimation for power grid critical infrastructures. *IEEE Transactions on Smart Grid* 3, 4 (2012), 1790–1799.