Adversarial Machine Learning in Smart Energy Systems

Martin C. Bor
Angelos K. Marnerides
SCC, Lancaster University
Lancaster, United Kingdom
m.bor@lancaster.ac.uk
angelos.marnerides@lancaster.ac.uk

Andy Molineux, Steve Wattam
Upside Energy Ltd.
Manchester, United Kingdom
andym@upsideenergy.co.uk
stevew@upsideenergy.co.uk

Utz Roedig
School of Computer Science and Information Technology
University College Cork
Cork, Ireland
u.roedig@cs.ucc.ie

ABSTRACT

Smart Energy Systems represent a radical shift in the approach to energy generation and demand, driven by decentralisation of the energy system to large numbers of low-capacity devices. Managing this flexibility is often driven by machine learning, and requires real-time control and aggregation of these devices, involving a diverse set of companies and devices and creating a longer chain of trust. This poses a security risk, as it is sensitive to adversarial machine learning, whereby models are fooled through malicious input, either for financial gain or to cause system disruption. We show the feasibility of such an attack by analysing empirical data of a real system, and propose directions for future research related to detection and defence mechanisms for these kind of attacks.

1 INTRODUCTION

Smart Energy Systems represent a radical shift in the approach to energy generation and demand, driven by the decentralisation of a large number of smaller units of power, like Electric Vehicles (EVs), Uninterruptible Power Supplys (UPSs), Photo Voltaics (PVs), and heat pumps. Managing this flexibility requires real-time control and aggregation of these devices, involving a diverse set of companies and devices and creating a longer chain of trust. This is often driven by Machine Learning (ML), which poses a security risk, as it is sensitive to Adversarial Machine Learning (AML), whereby models are fooled through malicious input, either for financial gain or to cause system disruption.

In this paper we investigate methods an adversary may use to game a distributed smart energy system targeting the ML and decision making elements of the system. The specific contributions of the paper are: (1) *Vulnerability Identification*: We provide a description of data inputs and decision logic within an aggregator employing ML, (2) *Attack Examples*: We showcase attack examples evaluated in a practical distributed energy system, and (3) *Countermeasures*: We discuss potential future directions for development of a protection framework for defending against these attacks.

In the next section we describe related work. Section 3 gives a description of the architecture of distributed energy systems, and the system used for our study. Section 4 describes two examples of attacks on a real system, and sec. 5 gives possible countermeasures. Section 6 discusses future work and concludes the paper.

2 RELATED WORK

AML in energy systems has been introduced under different flavours in past literature. Most common was the category of false data injection attacks targeting the grid's power state optimisation process for affecting its overall resilience [6], and the potential impact on real-time market operations [9]. Through a number of studies, bad data injection attacks were demonstrated that operated in a stealthy fashion [3, 5]. Several solutions have been proposed that exploit the advancements of ML, like distributed Support Vector Machine (SVM) [4], or a synergy of k-Nearest Neighbour, SVMs, and Sparse Logistic Regression [7]. Nonetheless, [2, 8] highlight the loopholes present within a large set of ML algorithms used for a range of energy-explicit applications. In our work we consider data from a real operational system, whereas most studies were restricted on numerical results and simulations.

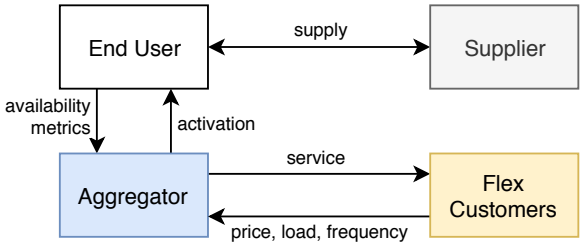

Figure 1: System architecture of an aggregator.

3 SMART ENERGY SYSTEMS

The role of the aggregator is to combine small loads into a 'virtual energy store', which can then be used for a wide range of flexibility services. Many of these services must operate near real time in order to respond to deviations, and contracts are available requiring different response rates.

To provide the services, an aggregator must: (1) predict availability for services given the capabilities of the devices under their control, (2) respond to signals that require dispatch of assets at given power levels, (3) provide an audit trail of power output for billing and regulatory requirements. Each of these is a potential target for performance figure manipulation, or service denial.

A general system architecture of an aggregator is shown in fig. 1. End users have one or more assets on a site that can be used for flexibility.In the case of decentralised system, these customers are connected to the electricity network via a supplier, who charges them for electricity. Assets communicate using low-level protocols (RS485, CANBUS, MODBUS), at a regular interval with the aggregator, typically over the internet, often using control hardware installed by the aggregator.

4 EXAMPLE ATTACK SCENARIO

We look at two scenarios to demonstrate the feasibility and impact of an attack: A 'lying' device, that systematically misreports behaviour, and a direct attack on an IDS designed to filter out anomalous input.

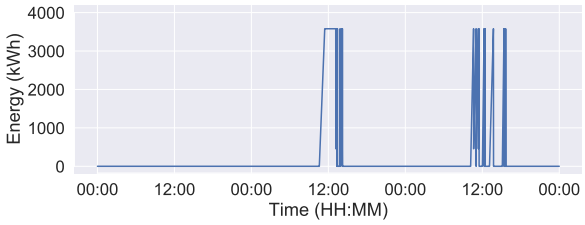

Figure 2: Three days of available energy for DSR, showing anomalous high figures in the last two days.

Modbus-TCP Injection. For Frequency Response (FR), an asset owner provides its assets to respond to a *frequency event*, that is when the AC grid frequency goes outside the dead band. To increase the likelihood for being picked to provide FR, an asset can overstate its reliability, by declaring its availability, even when it is under maintenance. Even when the asset communicates via an on-premise gateway, provided by the aggregator and which is tightly locked down, communications between the asset and gateway are typically performed over simple, unencrypted protocols. It is trivial to inject false Modbus TCP messages [1], or use an off-the-shelf emulator to simulate a malicious asset entirely, complete with a full audit trail of plausible data. This error is unlikely to be caught unless significant discrepancies exist between the site's metered electricity usage and the logs submitted as evidence of FR participation.

IDS Evasion. Figure 2 shows the energy available for FR as reported by a single UPS for three consecutive days. On the first day, the UPS operates normally, reporting a median available energy of 0.82 kWh. However, on the second and third day, just around noon (12 o'clock), the asset reports a capacity of over 3500 kWh, which is in excess of the nameplate capacity of the UPS' battery storage. If the aggregator would have nominated this UPS to perform FR, and a frequency event would have occurred, the asset owner would have been paid out around 4300% more than it would have with the actual amount of service provided.

An adversarial ML would, in this context, be capable of learning the thresholds at which the rule-based warnings are fired by inspecting the responses from the server: if these withdraw the device from service then a rule has been tripped.

5 COUNTERMEASURES

The scenarios described above is a simplified example. A simple, rule based, system to alert on anomalous values would be enough to detect this (malicious) activity. Attacks, however, are often much more sophisticated, resembling more noise, and requiring a better detection mechanism than looking for outliers. Additionally, a hand-crafted rule-based approach would be intractable, given the sheer amount of different assets, each with their own profiles, limits, and requirements.

Therefore, a better approach is an anomaly-based Intrusion Detection System (IDS) using binary classifier that is trained on the existing logs of device behaviour. This approach would seek to label incoming data as benign or malicious based upon features extracted real-time from the data stream. Through gradual refinement of feature extraction and expansion of the training dataset, this approach would yield a confidence score that may be used to tune precision/recall, leading to fewer false alerts and lower maintenance overheads than a manual system.

The IDS can also use a much wider range of features: besides the features derived from the metrics reported by the asset (energy, capacity, temperature, etc.), other features like the communication security (e.g. negotiated cipher suite, certificate fingerprints), and network properties (e.g. packet size, packet arrival time) can be used to detect malicious behaviour. Besides incoming data, *outgoing* data from the control algorithm can also be used as features for the IDS. This approach would enable the detection not only of the incoming, malicious data, but also the anomalous response sought by the attacker.

6 CONCLUSIONS

In this paper we illustrated the case of a ML-based aggregator in a smart energy system. We showed from empirical data that an attack on such a system is (potentially) trivial, using a variety of methods. Though many theoretical approaches to attacking and securing these systems have been proposed, we are in a position to test these using an established, practical, system. Further work will involve the inspection, classification, and formalism of attacks within the smart energy space, and production of a machine-learning based intrusion detection system, tuned to advance the current capabilities of the rule-based system.

ACKNOWLEDGMENTS

This work was supported by the Innovate UK Knowledge Transfer Partnership (KTP) Project partnership number 10838.

REFERENCES

[1] Bo Chen, Nishant Pattanaik, Ana Goulart, Karen L Butler-Purry, and Deepa Kundur. 2015. Implementing attacks for modbus/TCP protocol in a real-time cyber physical system test bed. In *Communications Quality and Reliability (CQR), 2015 IEEE International Workshop Technical Committee on*. IEEE, 1–6.

[2] Yize Chen, Yushi Tan, and Deepjyoti Deka. 2018. Is Machine Learning in Power Systems Vulnerable? *CoRR* abs/1808.08197 (2018). arXiv:1808.08197 http://arxiv.org/abs/1808.08197

[3] M. Esmalifalak, Z. Han, and L. Song. 2012. Effect of stealthy bad data injection on network congestion in market based power system. In *2012 IEEE Wireless Communications and Networking Conference (WCNC)*. 2468–2472. https://doi.org/10.1109/WCNC.2012.6214211

[4] M. Esmalifalak, L. Liu, N. Nguyen, R. Zheng, and Z. Han. 2017. Detecting Stealthy False Data Injection Using Machine Learning in Smart Grid. *IEEE Systems Journal* 11, 3 (Sep. 2017), 1644–1652. https://doi.org/10.1109/JSYST.2014.2341597

[5] Y. Huang, M. Esmalifalak, H. Nguyen, R. Zheng, Z. Han, H. Li, and L. Song. 2013. Bad data injection in smart grid: attack and defense mechanisms. *IEEE Communications Magazine* 51, 1 (1 2013), 27–33. https://doi.org/10.1109/MCOM.2013.6400435

[6] Yao Liu, Peng Ning, and Michael K. Reiter. 2009. False Data Injection Attacks Against State Estimation in Electric Power Grids. In *Proceedings of the 16th ACM Conference on Computer and Communications Security (CCS '09)*. ACM, New York, NY, USA, 21–32. https://doi.org/10.1145/1653662.1653666

[7] M. Ozay, I. Esnaola, F. T. Yarman Vural, S. R. Kulkarni, and H. Vincent Poor. 2012. Smarter security in the smart grid. In *2012 IEEE Third International Conference on Smart Grid Communications (SmartGridComm)*. 312–317. https://doi.org/10.1109/SmartGridComm.2012.6486002

[8] M. Ozay, I. Esnaola, F. T. Yarman Vural, S. R. Kulkarni, and H. V. Poor. 2016. Machine Learning Methods for Attack Detection in the Smart Grid. *IEEE Transactions on Neural Networks and Learning Systems* 27, 8 (Aug 2016), 1773–1786. https://doi.org/10.1109/TNNLS.2015.2404803

[9] L. Xie, Y. Mo, and B. Sinopoli. 2011. Integrity Data Attacks in Power Market Operations. *IEEE Transactions on Smart Grid* 2, 4 (Dec 2011), 659–666. https://doi.org/10.1109/TSG.2011.2161892