Deep Analytics for Management and Cybersecurity of the National Energy Grid

Ying Zhao[1][0000−0001−8350−4033]

Naval Postgraduate School, Monterey, CA 93943, USA
yzhao@nps.edu

Abstract. The United States's energy grid could fall into victim to numerous cyber attacks resulting in unprecedented damage to national security. The smart concept devices including electric automobiles, smart homes and cities, and the Internet of Things (IoT) promise further integration but as the hardware, software, and network infrastructure becomes more integrated they also become more susceptible to cyber attacks or exploitation. The Defense Information Systems Agency (DISA)'s Big Data Platform (BDP), deep analytics, and unsupervised machine learning (ML) have the potential to address resource management, cybersecurity, and energy network situation awareness. In this paper, we demonstrate their potential using the Pecan Street data. We also show an unsupervised ML such as lexical link analysis (LLA) as a causal learning tool to discover the causes for anomalous behavior related to energy use and cybersecurity.

Keywords: Big data platform · deep analytics · cybersercurity · usage patterns · anomaly detection · lexical link analysis · causal learning

1 Introduction

The United States' energy grid is evolving towards smart grid of future, which incorporates the digital technology to improve reliability, security and efficiency of the electric system through bi-directional information exchange, distributed generation, and storage resources for a fully automated power delivery network. The smart and integrated grids as shown in Fig. 1 seek efficiency through common communication standards and integrated networks, meeting the demand for the rapid growth in a cost-effective manner [16]. To further the concept smart devices including electric automobiles, smart homes and cities, and the Internet of Things (IoT) promise further integration. Better ways to manage the energy grid through better tools can also lead to better manage our energy resources and reduce greenhouse gasses.

This energy and smart grid not only need tools for better management and automation, but also face risks and vulnerability that bring unprecedented challenges and damage to national security:

1. Threats typical to smart grids and IoT devices are that they are not only vulnerable to physical faults and attacks but to cyber attacks as in the Internet, since the rise of the Internet and the integration via the Open Systems

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50426-7_23

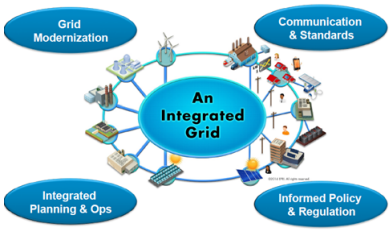

Fig. 1. The concept of smart or integrated grids seeks efficiency through common communication standards and an integrated grid (Electrical Power Research Institute) [20]

Interconnection model (OSI) allows an integration of standards throughout the different types of networks and devices Threats and vulnerabilities therefore are similar. The energy assets would naturally be susceptible to cyber attacks such as a Distributed Denial of Service (DDOS), worms, viruses and similar cyber exploits which might be the reasons for the Ukrainian power grid problems [16] and the Massachusetts gas explosions [18]. An attacker could take control of the company's "Supervisory Control and Data Acquisition" (SCADA) distribution management system [16]. The U.S. Energy Information Administration (EIA) notes that the rise of electric vehicles and smart devices has as one obstacle which is the cybersecurity [17]. Correlations between different categories of sensor big data could potentially act as "red flags" or early-warning signs of the cyber breach and vulnerability.

2. Distribution automation (DA) is a concept of smart grid which focuses on the operation and system reliability at the distribution level. The conventional centralized control management strategy is less effective for the smart grid due to the unidirectional power flow and requirement of control of a distributed grid. It is imperative to predict hotspots and areas of greatest concern which will lead to a reduction in waste and inefficiency or expose security vulnerabilities. This also calls for big data and deep analytics to be operated in a distributed fashion.
3. Risks include unanticipated operational conditions, for example, for a grid-connected microgrid, severe weather conditions or grid blackouts may trigger an unintentional islanding accident [2], which threats the safety operation and causes technical challenges.

New and emerging technologies offer opportunities to better analyze energy sensor data to improve our understanding of where energy is wasted and how to identify and best respond to risks. The IoT significantly increases the volume and velocity of big data through the concept such as "smart cities." Big data analytics can reduce risks and show scalable solutions for detecting patterns and anomalies from collective intelligence and from the distributed data sources [1].

In the past, data mining techniques [11], energy and entropy theories, wavelet transform [6], machine learning algorithms such as predictive maintenance, electric device health monitoring, and power quality monitoring have been used for

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50426-7_23

energy and smart grids [1]. The support vector machine (SVM) is used for an islanding detection [2]. A deep learning is used for IoT device as deep sparse coding [3]. Localized fault characterization uses a hybridization of evolutionary learning and clustering techniques [4, 5]. SVM, AdaBoost, and extreme learning machine (ELM) are used for online detection of risky events in power system [7, 8] Innovation of big data also comes to energy grid management and cybersecurity for example, real-time social sensor data using Twitter, Facebook could provide new insight using the location data [9, 10]

Unique methods illustrated in this paper can be used for both energy management and cybersecurity because we show data collection and analysis from sensors as the key components for constantly monitoring and detecting the threats from collective and distributed data sources. Specifically, this paper addresses how a "Big Data Platform" (BDP), lexical link analysis (LLA) to discover anomalies, potential threats, and vulnerabilities using the Pecan Street data set as a use case. The key contribution of this paper is causal learning since human is necessary in the process for the validation of decision making.

2 Pecan Street Big Data

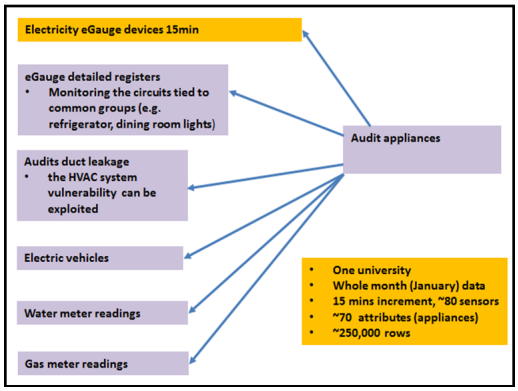

Fig. 2. Pecan Street Data

The data source used in this research was obtained by the Pecan Street organization [12]. Pecan Street collects energy usage for a smart city which means there is a conscious and curated effort to record the right data for energy consumption in a methodical manner. The organization host and maintain one of the largest databases of consumer electricity and water use in the world. 750 million records are collected daily as circuit-level use data from multiple sources available through their Dataport website. The data track appliance level consumer behavior.

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50426-7_23

We extracted sample research data from Pecan Street Dataport. Fig. 2 shows the depiction of a sample Pecan Street data. The sample data consists of participants' electricity usage data (per kWh) with 69 data fields including one key field (user id or data id and timestamp) and the remaining 67 fields listing various equipment used on site (e.g., furnace, kitchen, lights, dish washer, dryer, etc.). We selected one month of data consisting of 250,000 records in 15 min data blocks for 100 participants (users or data ids) as follows:

- air1: air conditioner 1
- air2: air conditioner 2
- air3: air conditioner 3
- aquarium1: aquarium 1
- bathroom1: bathroom 1
- bathroom2: bathroom 2
- bedroom1: bedroom 1
- ...

The "air1" field records electricity usage for air conditioner 1 for 31 days in for a January. Baselines could be set and monitored to find anomalies, for example, an insider threat could be the unauthorized running of energy grid servers in January not August which would increase the air conditioner usage. Fig. 3 shows an example of Pecan Street data.

3 Big Data Platform (BDP)

The Big Data Platform (BDP), which has been developed by the Defense Information Systems Agency (DISA) [19], runs on on Amazon Web Services (AWS) including a mix of big data standard and customized tools for data ingestion, management, security, exploration, and analysis. These functions are supported by open source tools including Apache Spark [25], Apache Storm [26], Hadoop Map/Reduce, Kibana [27], NodeJS [28], and R-Shiny [29].

BDP is designed for real-time processing of Big Data beginning at ingestion and ultimately presenting useful data visualizations that may alert decision makers of energy leaks and security vulnerabilities. The BDP has the strict compliance with the DISA security standards which provide a secure system security that can be an advantage to store big data such as Pecan Street and National Energy Grid. There are also analytics in BDP which can perform more complicated calculations on a larger data set. For the Pecan Street sample data set, we first applied BDP to provide initial useful information. We later applied unsupervised machine learning algorithms k-means and lexical link analysis (LLA) to discover patterns and anomalies in the data set.

4 Application of BDP

We first ingested and parsed the Pecan Street data into the BDP system. After ingesting, the data were available to the analytics tools inside BDP such as

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50426-7_23

Unity. Kibana is used to create a display of metrics, heatmaps, graphs, and charts. The BDP system is designed to display results quickly for real-time big data so that trends and outliers can be discovered quickly. BDP uses a catalogue or taxonomy shared by multiple users in the same domain. For example, in a typical cybersecurity environment, people often use similar network monitoring tools to collect data and monitor activities, therefore, the data fields are similar and can be shared across multiple locations. The characteristic applies to the energy grid as well. This is an unique advantage of using BDP. Fig. 4 shows the average electricity usage each hour over one month for the Pecan Street data set. It is interesting to note that the spikes in use are not regular. Fig. 5 shows a dashboard of graphs and metrics representing electricity usage over 24 hours for different areas of the data set. These could be updated in near real-time for monitoring activities should energy grid data hosted in such a secure data center. For example, why do the outside lighting plugs have higher average electricity usage around the noon time?

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
dataid	local_15min	use	air1	air2	air3	airwindowunit1	aquarium1	bathroom1	bathroom2	bedroom1	bedroom2	bedroom3	bedroom4	bedroom5	car1	clotheswasher1
93	1/1/2013 0:00	0.856266667	0.0414													0
93	1/1/2013 0:15	0.850666667	0.041													0
93	1/1/2013 0:30	0.857533333	0.040933333													0
93	1/1/2013 0:45	0.846266667	0.036933333													0
93	1/1/2013 1:00	0.5238	0.0366													0
93	1/1/2013 1:15	0.445533333	0.038													0
93	1/1/2013 1:30	0.447533333	0.038													0
93	1/1/2013 1:45	0.5276	0.038133333													0
93	1/1/2013 2:00	0.5628	0.038333333													0
93	1/1/2013 2:15	0.544333333	0.038866667													0
93	1/1/2013 2:30	0.547066667	0.038666667													0
93	1/1/2013 2:45	0.550266667	0.038133333													0
93	1/1/2013 3:00	0.5446	0.0382													0
93	1/1/2013 3:15	0.513866667	0.038066667													0
93	1/1/2013 3:30	0.442533333	0.038													0
93	1/1/2013 3:45	0.448666667	0.038													0
93	1/1/2013 4:00	0.4476	0.038													0
93	1/1/2013 4:15	0.436	0.038													0
93	1/1/2013 4:30	0.4458	0.038													0
93	1/1/2013 4:45	0.447466667	0.038													0
93	1/1/2013 5:00	0.441933333	0.038													0
93	1/1/2013 5:15	0.501933333	0.038													0
93	1/1/2013 5:30	1.143733333	0.043733333													0.003733333
93	1/1/2013 5:45	1.102666667	0.043266667													0.0034
93	1/1/2013 6:00	0.5606	0.038066667													0

Fig. 3. Pecan Street data example

4.1 Unsupervised Machine Learning

We first applied the K-means clustering algorithm from MATLAB and clustered the 250K records into 10 clusters as shown in Fig. 6. K-means requires a chosen k and k=10 in our case for simplicity. Fig. 6 is a radar graph showing cluster center values, i.e., average usages within clusters for the 67 areas labeled in the circle. These clusters represent the discovered patterns. The characteristics of the clusters show behavior patterns of the users and time periods in which characteristics of usage patterns can be summarized in the following examples:

- Cluster 7 (series7): Average high usages within the cluster attribute to the areas of “use”, “grid”,”drye1”, “furnace1”, “poollight1”,and “waterheater1”.
- Cluster 6 (series6): Average high usages attribute to the areas of “use”, “car1”, “gen”, and “grid”.

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50426-7_23

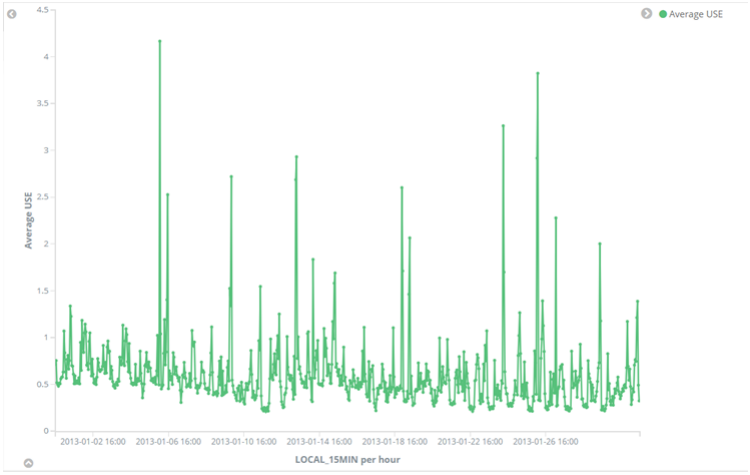

Fig. 4. Average electricity usage each hour over one month

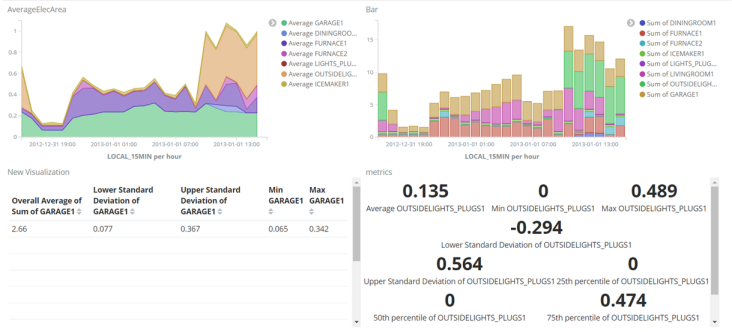

Fig. 5. A dashboard shows graphs and metrics representing electricity usage over 24 hours for different areas of the data set. BDP allows to update such graphs in real-time for monitoring activities.

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50426-7_23

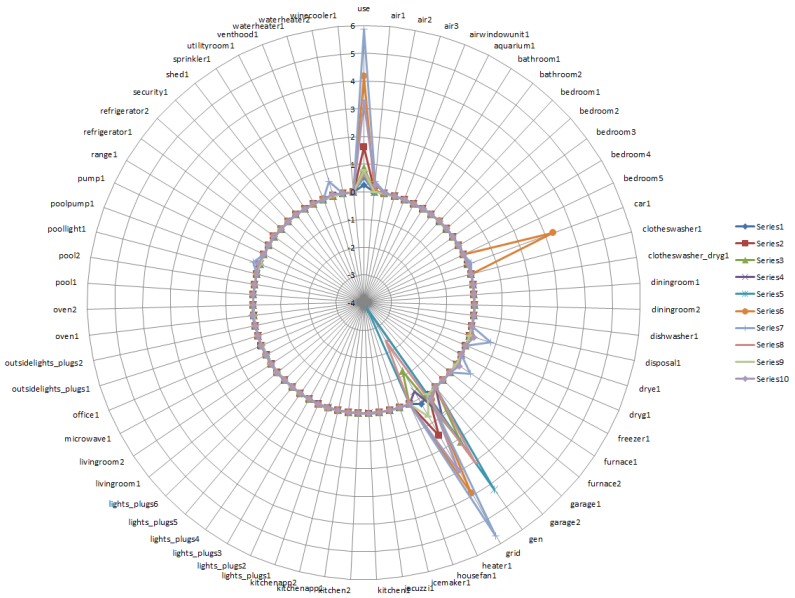

Fig. 6. Unsupervised learning and cluster characteristics using the k-means algorithm and displayed using a radar graph and k-means.

– Cluster 5 (series5): Average high usages attribute to the areas of "gen" and "grid" (negative – giving back to the grid).

We then computed an anomaly index value for each of the data points, which is the minimum distance of a data point to the 10 cluster centers. The higher an anomaly index, the far away is the corresponding data point from the 10 patterns ("normal behaviors"). Fig. 7 shows the value of the anomaly index for each cluster. Cluster 7, 6, and 5 have the highest 3 values of anomaly indexes, which are the potential candidates for further investigation for the areas of energy management and cybersecurity.

The anomaly detection system detected 3 anomaly profiles which gives three different reasons for these users to be different from the population as a whole. The information shows human analysts behavioral patterns for attributes and source of the anomaly, for example "gen" means there is a generator at home and negative "grid" means the generator gives energy back to the grid. The combination might indicate a different usage pattern for some users. If some anomaly patterns are trending collectively from various locations, they are the opportunities for human decision makers and consequences of the anomalies detected can be alerted to the human analysts via the BDP server.

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50426-7_23

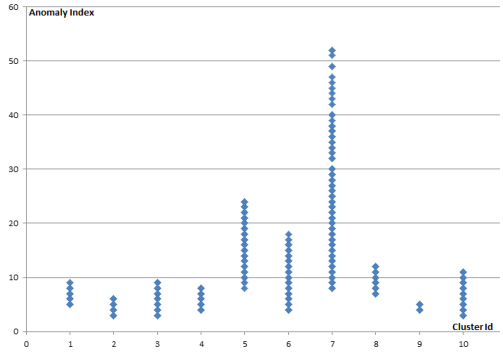

Fig. 7. The values of the anomaly index for the clusters

5 Lexical Link Analysis (LLA)

LLA is an unsupervised ML method [13, 14] which describes the characteristics of a complex system using a list of attributes or features, or specific vocabularies or lexical terms. Because the potentially vast number of lexical terms from big data, LLA can be viewed as a deep model for big data. LLA can describe a system using feature pairs as bi-gram lexical terms extracted from data. LLA automatically discovers word pairs, and displays them as networks.

Bi-grams allow LLA to be extended to numerical or categorical data. For example, using structured data, such as attributes from the Pecan Street data set, we discretize numeric attributes and categorize their values to paired features. The feature pair model can further be extended to a context-concept-cluster model [21]. A context can represent a location, a time point, or an object shared across data sources. For example, for the Pecan Street data, the data id and time point can be contexts.

5.1 LLA Outputs for the Pecan Street Data Set

In order to use LLA, we first generate word feature networks for the data set. The value for an attribute in Fig. 3, such as "grid" is discretized into three bins when applying LLA as a word feature: 1) less than (lt) the mean ($\bar{m}$) of the feature minus one standard deviation ($\bar{m} - \sigma$), 2) between (bt) the mean minus one standard deviation ($\bar{m} - \sigma$) and the mean plus one standard deviation ($\bar{m} + \sigma$), and 3) more than (mt) the mean plus one standard deviation ($\bar{m} + \sigma$). A node in LLA represents a discretized feature. For example, *grid_mt_*1.8 means if the "grid" (i.e., grid usage of electricity in a 15 minutes interval for a data id) is more than 1.8.

Probability and lift are the two measures in LLA defined in Equation (1) and Equation (3) to measure the strength of an association between two word features.

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50426-7_23

$$prob_{ij} = \frac{word\ features\ i,j\ together}{word\ feature\ j} \tag{1}$$

$$prob_i = \frac{word\ feature\ i}{all\ word\ features} \tag{2}$$

$$lift_{ij} = \frac{prob_{ij}}{prob_i} \tag{3}$$

Fig. 8 shows the output of LLA for a word feature *grid_mt_*1.8's associations with other features using the "lift" as the association strength measure listed as follows (filtered using "lift" > 4):

- *drye*1_*mt*_1.0: "dryer1" (dryer 1)'s usage of electricity is more than 1.0 in a 15 minutes interval
- *car*11_*mt*_2.0: "car1" (car 1)'s usage of electricity is more than 2.0 in a 15 minutes interval
- *air*1_*mt*_0.6: "air1" (air conditioner 1)'s usage of electricity is more than 0.6 in a 15 minutes interval
- *air*2_*mt*_0.6: "air2" (air conditioner 2)'s usage of electricity is more than 0.6 in a 15 minutes interval
- *waterheater*1_*mt*_2.0: "waterheater1" (water heater 1)'s usage of electricity is more than 2.0 in a 15 minutes interval
- *poolpump*1_*mt*_1.5: "poolpump1" (pool pump 1)'s usage of electricity is more than 1.5 in a 15 minutes interval
- *poolpump*1_*bt*_0.6_1.5: "poolpump1" (pool pump 1)'s usage of electricity is between 0.6 and 1.5 in a 15 minutes interval
- *oven*1_*mt*_0.3: "oven1" (oven 1)'s usage of electricity is more than 0.3 in a 15 minutes interval
- *dataid*_5357: data id (user) 5357

Fig. 11 shows *gen_mt_*1.7 (i.e., a generator, such as solar, alternative, and renewable energy with inverter interfaced distributed generators (IIDGs), generates electricity more than 1.7 in a 15 minutes interval) is associated with *grid_lt_*-0.9 (i.e., grid usage of electricity is less than -0.9, negative, giving back to the grid in a 15 minutes interval for a data id).These results are similar to the k-means result in Fig. 6.

LLA allows a drill down search as shown in Fig. 9. When clicking both nodes *grid_mt_*1.8 and *waterheater*1_*mt*_2.0: 373 data records in the Pecan Street data set have both characteristics *grid_mt_*1.8 and *waterheater*1_*mt*_2.0 and they are listed in the LLA search result. 373 data records have the characteristics *waterheater*1_*mt*_2.0, 100% of them also have the characteristics *grid_mt_*1.8. 16,434 data records have the characteristics *grid_mt_*1.8 out of the total 120,847 data records. So the lift is 7.4.

LLA also discovers interesting associations, for example, *grid_mt_*1.8 is associated with a specific user (data id) of "5357" in Fig. 8. As another example, Fig. 11 shows the time points of a day are associated with *gen_bt_*0.6_1.7 (i.e., generater) generates electricity between 0.6 and 1.7 in a 15 minutes interval).

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50426-7_23

5.2 Discussion: Discovering Causal Associations Using LLA

A unique requirement of anomaly detection for energy management and cybersecurity is causality analysis because human analysts need to understand causes behind any observable anomaly effects. This calls a systematic approach of deep analytics that is also causality analysis, i.e., linking an anomaly effect, e.g., grid

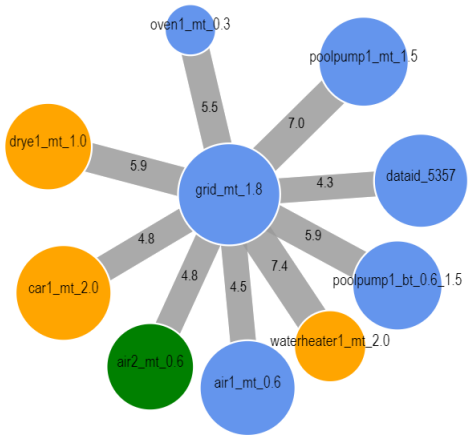

Fig. 8. Causal level 1

Fig. 9. Drill down

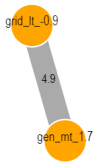

Fig. 10. Grid usage of electricity is less than -0.9: negative, giving back to the grid

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50426-7_23

usage of electricity in a 15 minutes interval is more than 1.8 (*grid_mt_*1.8), to the causes, e.g., specific users or time points. The key factors for causal learning includes the three layers of a causal hierarchy [23, 24] - association, intervention and counterfactuals.

The common consensus is that data-driven analysis or data mining can discover initial statistical correlations and associations from big data. Human analysts need to validate and understand if the associations make sense and what are the real causes and effects.

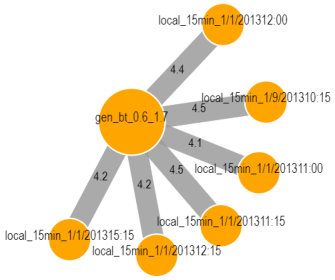

Fig. 11. Local time causal relations

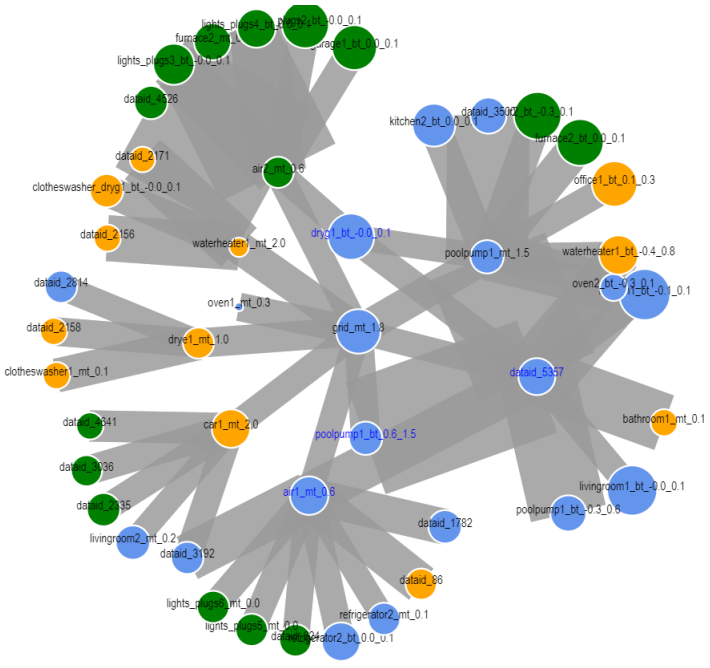

Fig. 12. Causal associations level 2

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50426-7_23

In a real-life application, one often wants to predict causes based on the data of effects, i.e., computing and validating the probability (P) of a potential cause (C) given an effect (E), i.e., $P(C|E)$. Effects are often observable data, e.g., *grid_mt*_1.8 in the Pecan Street data set. Causes, e.g., specific users and time points need to be discovered from the observable data. $P(C|E)$ is difficult to discover because causes are often hidden, anomalous, and capricious. In machine learning practices, the associations, correlations or probabilistic rules are typically cross-validated using separate or new data sets. Causal learning requires the intervention and counterfactual reasoning. An intervention reasoning tries to answer the question: What will happen if one takes an action? For example, instead of examining $P(C|E)$, if E is actionable or $P(C|do(E))$ [23] can be examined. The intervention more than just mining the existing data.

Counterfactual reasoning tries to answer the question: What if I had acted differently? If $P(C|E)$ is high-probability rule discovered from data, $P(C|Not\ E)$, $P(Not\ C|E)$, and $P(Not\ C|Not\ E)$ are the counterfactuals needed in the reasoning. Traditionally, the counterfactual is defined as the effect of an action for an entity and for the same entity without the action.

LLA calculates the lift measure that is one of the counterfactual reasoning in causal learning [22]

In the Pecan Street data set, although the linked features as shown in Fig. 8 make sense to human analysts, the specific user or data id or time points might be more detailed causes for energy management and cybersecurity. In Fig. 12, each cause feature nodes can be expanded to another level to reveal more causes such as more data ids (users) linked to the first level causes as shown in Fig. 8.

In LLA,when $lift_{E,C_i} > 1$, C_i is a potential cause for E. However, if another cause C_j is a confounder of C_i and E, then $lift_{C_j,C_i} > 1$. So if $lift_{C_j,C_i} > 1$ for some C_j and $lift_{C_j,E} > 1$, then C_j not C_i is the cause of E. In Fig. 12, only *dataid*_5357 directly links to *grid_mt*_1.8 and the first level features *poolpump*1_*mt*_1.5, *poolpump*1_*bt*_0.6_1.5, *air*1_*mt*_0.6, and *air*12_*mt*_0.6. *dataid*_5357 is a real cause and we can eliminate *poolpump*1_*mt*_1.5,
*poolpump*1_*bt*_0.6_1.5, *air*1_*mt*_0.6, and *air*12_*mt*_0.6. Other causes
*waterheater*1_*mt*_2.0, *drye*1_*mt*_1.0, *car*1_*mt*_2.0, and *oven*1_*mt*_0.3 are independent causes with no confounders.

6 Conclusion

We demonstrated that BDP and deep analytics using the Pecan Street data for anomaly detection and causality analysis of resource management, cybersecurity, and energy network situation awareness. We also demonstrated unsupervised learning algorithms to discover the usage patterns and anomalies. We also defined an anomaly index and showed its values for the clusters and time points. We showed LLA as an innovative approach to discover causal associations. The information can help business users to see the patterns and detect abnormal activities for the management and cybersecurity of an energy grid.

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50426-7_23

7 Acknowledgement

Thanks to the support of the Office of Naval Research for supporting the research, Quantum Intelligence, Inc. for support and collaboration, and the Pecan Street project for providing the data set. The views and conclusions are those of the authors and should not be interpreted, expressed or implied of the U.S. Government.

References

1. Zhang, Y., Huang, T., and Bompard, E.F.: Big data analytics in smart grids: a review. Energy Inform 1, 8 (2018). Retrieved from https://doi.org/10.1186/s42162-018-0007-5
2. Rezaul, A. M., Muttaqi,K. M., Bouzerdoum, A.: Evaluating the effectiveness of a machine learning approach based on response time and reliability for islanding detection of distributed generation. IET renewable power generation vol. 11-11 (2017)
3. Zico, K. J., Batra, S., Ng, A. Y.: Energy disaggregation via discriminative sparse Coding. In: Proceedings of the 23rd International Conference on Neural Information Processing Systems, vol 1, Vancouver, pp 1153–1161 (2010)
4. De Santis, E., Rizzi, A., Sadeghian, A.: A Learning Intelligent System for Classification and Characterization of Localized Faults in Smart Grids. 2017 IEEE Congress on Evolutionary Computation (CEC), San Sebastian, 5-8 June 2017
5. Wang, X., McArthur, S., Strachan, S., Kirkwood, J., Paisley, B.: A data analytic approach fault diagnosis and prognosis for distribution automation. IEEE Transactions on Smart Grid, 9(6),6265–6573 (2017)
6. Mishra, D.P., Samantaray, S.R., Joos, G.: A combined wavelet and data-mining based intelligent protection scheme for microgrid. IEEE Transactions on Smart Grid 7(5):2295–2304 (2016)
7. Wang, J., Xiaofu, X., Zhou, N., Li, Z., Wang, W.: Early warning method for transmission line galloping based on SVM and AdaBoost bi-level classifiers. IET Generation, Transmission Distribution 10 (14), 3499—3507 (2016)
8. Zhang, Y., Yan, X., Dong, Z. Y., Zhao, X., Wong, K. P.: Intelligent early warning of power system dynamic insecurity risk toward optimal accuracy-earliness tradeoff. IEEE Transactions on Industrial Informatics 13(5), 2544—2554 (2017)
9. Bauman, K., Tuzhilin, A, Zaczynski, R.: Using social sensors for detecting emergency events: a case of power outages in the electrical utility industry. ACM Transactions on Management Information Systems 8:2–3 (2017)
10. Sun, H., Wang, Z., Wang, J., Huang, Z., Carrington, N.L., Liao, J.: Data-driven power outage detection by social sensors. IEEE Transactions on Smart Grid 7(5):2516–2524 (2017)
11. Singh, S., Yassine, A.: Mining energy consumption behavior patterns for households in smart grid. IEEE Transactions on Emerging Topics in Computing, 7 (3) 404–419 (2017)
12. Pecan: https://www.pecanstreet.org/. Last accessed 13 Apr 2020
13. Zhao, Y., Gallup, S.P. and MacKinnon, D.J.: System self-awareness and related methods for improving the use and understanding of data within DoD. Software Quality Professional, 13(4): 19-31 (2011) Retrieve from http://asq.org/pub/sqp/

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50426-7_23

14. Zhao, Y., Mackinnon, D. J., Gallup, S. P.: Big data and deep learning for understanding DoD data, Journal of Defense Software Engineering. Special Issue: Data Mining and Metrics (2015). July/August 2015, page 4-10,Lumin Publishing ISSN 2160-1577. Retrieved from http://www.crosstalkonline.org/storage/flipbooks/2015/201507/index.html
15. Zhang Z.: Smart Grid in America and Europe: Similar Desires, Different Approaches. Public Utilities Fortnightly, 149, 1, 2011
16. Vianna, G.: Vulnerabilities in the North American Power Grid: Global Security Studies, Fall 2016, Volume 7, Issue 4 (2016)
17. Chase, N, (2018). Autonomous Vehicles: Uncertainties and Energy implications: 2018 U.S. Energy Information Administration Independent Statistics & Analysis. Retrieved from https://www.eia.gov/conference/2018/pdf/presentations/nicholas_chase.pdf
18. Associated Press (September 16, 2018), The Latest: Pressure Sensors Focus of Gas Explosions Probe. Retrieved from https://www.nytimes.com/aponline/2018/09/16/us/ap-us-gas-explosions-the-latest.html
19. BDP: https://www.disa.mil/newsandevents/2016/Big-Data-Platform. Last accessed 13 Apr 2020
20. IoT:http://smartgrid.epri.com. Last accessed 13 Apr 2020
21. US patent 8,903,756: System and method for knowledge pattern search from networked agents. 2014. Retrieved from https://www.google.com/patents/US8903756
22. Zhao Y., MacKinnon, D.; and Jones, J.: Causal Learning Using Pair-wise Associations to Discover Supply Chain Vulnerability. In the Proceedings of the 11th International Conference on Knowledge Discovery and Information Retrieval (KDIR 2019). September 17-19, 2019, Vienna, Austria. Retrieved from https://www.insticc.org/Primoris/Resources/PaperPdf.ashx?idPaper=80705
23. Mackenzie, D. and Pearl, J.: The Book of Why: The New Science of Cause and Effect. Penguin, New York (2018)
24. Pearl, J.: The Seven Pillars of Causal Reasoning with Reflections on Machine Learning. (2018). Retrieved from http://ftp.cs.ucla.edu/pub/stat_ser/r481.pdf
25. Apache Spark: https://spark.apache.org/. Last accessed 13 Apr 2020
26. Apache Storm: http://storm.apache.org/. Last accessed 13 Apr 2020
27. Kibana: https://www.elastic.co/products/kibana. Last accessed 13 Apr 2020
28. NodJS: https://nodejs.org/en/. Last accessed 13 Apr 2020
29. R-shiny:https://shiny.rstudio.com/. Last accessed 13 Apr 2020

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50426-7_23