

Outlier Detection in Smart Grid Communication

Nelson Makau Mutua
Faculty of Information Technology
Brno University of Technology
Brno, Czech Republic
Email: imutua@fit.vut.cz

Petr Matoušek
Faculty of Information Technology
Brno University of Technology
Brno, Czech Republic
Email: matousp@fit.vutbr.cz

Abstract—Industrial Control System (ICS) networks transmit control and monitoring data in critical environments such as smart grid. Cyber attacks on smart grid communication may cause fatal consequences on energy production, distribution, and eventually the lives of people. Since the attacks can be initiated from both the inside and outside of the network, traditional smart grid security tools like firewalls or Intrusion Detection Systems (IDS), which are typically deployed on the edge of the network, are not able to detect internal threats. For this reason, we also need to analyze behavior of internal ICS communication.

Due to its nature, ICS traffic exhibits stable and predictable communication patterns. These patterns can be described using statistical models. By observing selected features of ICS network communication like packet inter arrival times, we can create a statistical profile of the communication based on the patterns observed in the normal communication traffic. This technique is effective, fast and easy to implement. As our experiments show, statistical-based anomaly detection is able to detect common security incidents in ICS communication. This paper employs selected network packet attributes to create a statistical model for anomaly detection using the Local Outlier Factor (LOF) algorithm. The proof-of-concept is demonstrated on IEC 60870-5-104 (a.k.a. IEC 104) protocol.

Index Terms- anomaly detection, communication pattern, smart grid, IEC104, statistical model, ICS.

I. INTRODUCTION

ICS communication provides proper functioning of critical infrastructure systems. These systems are naturally exposed to external threats including cyber attacks [1]. Traditionally, industrial systems are well protected against external threats through the use of firewalls and IDS devices that filter communication between ICS systems and Internet traffic, making direct attacks on smart grid communication a rare case. However, attackers can gain access to the system by sending malware to a user via an infected e-mail attachment [2].

The importance of the research is motivated by recent cyber attacks against critical infrastructure systems. One of the attacks against Ukrainian power company happened in December 23, 2015 when BlackEnergy malware caused power disruption to 225,000 customers, lasting up to 6 hours [2]. This attack included multiple stages starting with “spear phishing” e-mails targeting a staff to gain access to the corporate network of the power company. Once inside the power company network, attackers gathered credentials and used VPNs to get access to the internal network. More recent ransomware attack against the Colonial Gas Pipeline in the U.S. happened in May

2021 [3]. Similarly to BlackEnergy attack, it was also initiated from an infected internal station.

Compared to standard information and communication systems, smart grid communication exhibits stable, periodical, and regular communication patterns since the communication occurs between devices with no or little human interference. Typically, a controlling station periodically requests status data from a field device like the Programmable Logical Controller (PLC) or Remote Terminal Unit (RTU) in order to provide a real-time view on industrial processes.

To detect internal threats, we need (i) to regularly monitor smart grid communication and (ii) observe suspicious patterns that occur in the network traffic. One solution is to employ ICS monitoring using extended IPFIX protocol[4] that retrieves monitoring data about active ICS communication. To detect unknown adverse events or unusual behavior, we can observe statistical patterns using ICS flow data.

In this research, we closely examine statistical distribution of inter-arrival times of IEC 104 packets. This is a part of my PhD study that is focused on anomaly detection of ICS communication using statistical methods. The main idea behind the research comes out from stable communication patterns that are typical for ICS communication and can be observed on packet and flow levels. Within the PhD. research, we plan to apply various statistical methods on smart grid communication and evaluate their efficiency for covering common ICS attack vectors defined by MITRE ATT&CK for ICS matrix¹.

A. Contribution

This research paper presents a proof-of-concept to outlier detection in smart grid communication. We observe packet inter-arrival times of IEC 104 communication and create a statistical profile of the normal traffic. Then we apply the Local Outlier Factor (LOF) algorithm to detect outliers which represent anomalies. Our preliminary results proves viability of this approach for statistical-based anomaly detection in ICS communication.

B. Structure of the Paper

This paper is structured as follows. In Section II we give an overview of published works related to statistical anomaly detection in ICS networks. Next, we briefly introduce IEC 104

¹See https://collaborate.mitre.org/attackics/index.php/Main_Page [06/2021]

protocol, statistical features and the LOF algorithm. Section IV shows our preliminary results. The last section concludes the work and discusses future steps.

II. RELATED WORK

Exploitation of timing attributes such as average packet inter-arrival times and the number of packets or bytes transmitted in a certain interval for anomaly detection has been studied in the past. Barbosa et al. [5], [6] investigated the use of spectral analysis to uncover traffic periodicity. Udd et al. [7] examined a TCP sequence prediction attack for the IEC 104 protocol. Other works mostly focused on flooding and DoS attacks [8]. We explore what attack vectors described by MITRE ATT&CK for ICS can be covered by LOF method.

To enhance detection ability of cyber attack, deep packet inspection (DPI) implemented in IDS tools and timing models have been proposed. Sayegh et al. [9] modelled the inter-arrival times between signatures (i.e., packet sequences) and validated this technique with large amount of injected signatures. Barbosa et al. [10] proposed to model the period of repeated requests in an orderless group. They evaluated the approach on Modbus and MMS datasets without attacks and set relaxed thresholds to avoid high false positive rates. This prevents the detection of subtle changes within a single period which is covered by our approach.

More recently, sequence-aware approaches have been explored. Yang et al. [11], Goldenberg and Wool [12], and Kleinmann and Wool [13] used finite automata to model message sequences of IEC 104, Modbus TCP and S7 respectively. Casselli et al. [14] modelled a sequence of messages using discrete-time Markov chains in order to detect sequence attacks. These approaches observe the order of messages. Since they require deep packet analysis, they have high demands on processing. Modeling of statistical behavior as proposed in this paper is fast and simple for implementation while giving comparable results.

III. ANOMALY DETECTION IN SMART GRID NETWORKS

A. Smart Grid Communication

Industrial protocols used in smart grid communication include protocol IEC 61850 (GOOSE, MMS), Modbus, IEC 104, DNP3, DLMS, and others [15]. These protocols transmit control and status data from industrial processes running on RTUs or IEDs. Protocols like IEC 104, DNP3, MMS or DLMS communicate using a *client-server model*. A master (controlling) station sends commands to a RTU slave (controlled station) in control direction while the slave delivers monitoring data in monitor direction. Protocols like GOOSE or Modbus use a *publish-subscribe mechanism* in which an application writes data into a local buffer, which is then periodically transmitted to a subscribed agent via L2 multicast.

1) *IEC 104*: IEC 104 protocol is an application protocol that consists of Application Protocol Control Information (APCI) and Application Service Data Unit (ASDU) sub-layers. It is implemented over TCP but for monitoring purposes we observe so-called virtual flows that represent records with a

single ASDU packet transmitted on wire. We focus on inter-arrival times but there are additional attributes that can be used for statistical model: ASDU size which is also stable for specific IEC 104 commands, frequency of selected commands, e.g., spontaneous events, activations, packet size, etc.

2) *Packets inter-arrival time*: Packets inter-arrival time is the time taken between the arrival of two subsequent packets. It is computed as a difference between timestamps of two these packets. Inter-arrival time is characteristic for ICS transmissions and exhibits stable and predictable patterns. Changes in inter-arrival times indicate an anomaly on the communication channel. As demonstrated in our previous research [16], additional features like packet size or modeling of exact IEC 104 commands do not improve the accuracy of the statistical approach. We plan to confirm this observation by future experiments. For modelling inter-arrival time of industrial protocols with master-slave communication profile, it is natural to observe statistical distribution of bi-directional transmissions as depicted in Fig. 1.

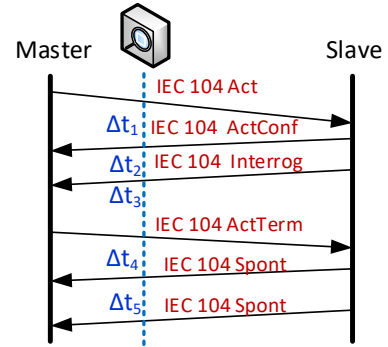


Fig. 1. Observing inter-arrival times in IEC 104 communication.

3) *Dataset*: In our experiments, we work with the IEC 104 protocol that is commonly used in smart grids for substation control. We utilise four datasets obtained from the Brno University of Technology testbed, see Tab. I. In order to test the ability of attack detection we created a set of ICS flow records with emulated attack and a AD testing tool ².

Dataset	Packets	Duration	Devices
10122018-104Mega	104,534	4h 53min	4
13122018-mega104	1,460,829	71h 17min	14
mega104-14-12-18	14,597	15h 38min	2
mega104-17-12-18	58,931	67h 55min	2

TABLE I

IEC 104 DATASETS FOR EXPERIMENTS.

B. Statistical Outlier Detection

Outlier detection is a statistical technique that discovers data points that are inconsistent with the rest of the data. In general, outliers and inliers are determined using a data distribution model. This paper focuses on unsupervised outlier detection.

Unsupervised methods do not require data labels. As a result, they are more adaptable. The fundamental idea behind

²The tool and dataset are available at <https://github.com/nelsonmakau/anomaly-detection-in-smart-grid-communication>.

unsupervised outlier detection is to score data points solely on the essential characteristics of the dataset. In general, density or distance are used to determine whether a data point is an inlier (normal) or outlier (anomaly). In this proof-of-concept study, we apply Local Outlier Factor (LOF) method [17].

The scientific research on statistical outlier detection provides two approaches to handle outliers in a dataset. First, outliers must be identified for further investigative process. Second, the data model should be designed to handle outlier data points accurately.

C. Detecting Outliers in Smart Grid

Outlier detection is a statistical procedure that finds suspicious events or items that differ from a dataset's normal form. Outliers are detected as data points that have a significantly lower density than their neighbors. The purpose of outlier detection is to detect rare events or unusual activities that differ from the majority of data points in a dataset [18]. Outlier detection can detect global or local outliers. For a global outlier, outlier detection considers all data points, and the data point is considered an outlier if it is far away from all other data points. The local outlier detection covers a small subset of data points at a time. A local outlier is based on the probability of data point being an outlier as compared to its local neighborhood which is measured by the k -Nearest Neighbors (kNN) algorithm. LOF is a density-based unsupervised anomaly detection method that computes a given data point's local density deviation with respect to its neighbors. LOF scores are computed for all data points according to parameter k (the number of nearest neighbors) as follows [19]:

Definition 1: $d(p, o)$ is the Euclidean distance between two data points p and o . The distance between two data points p and o is calculated using an Euclidean n -dimensional space:

$$d(p, o) = \sqrt{\sum_{i=1}^n (p_i - o_i)^2} \quad (1)$$

Let D be a dataset and k a positive integer. For a data point p , the k -distance (p) is the distance $d(p, o)$ between p and the farthest neighbor data point o by the following conditions:

- 1) At the least, k data points (records) $o' \in D \setminus \{p\}$ maintains that $d(p, o') \leq d(p, o)$.
- 2) At the most, $k - 1$ data points (records) $o' \in D \setminus \{p\}$ maintains that $d(p, o') < d(p, o)$.

Definition 2: k -Nearest Neighbors of p . The meaning of k -Nearest Neighbors of p is any data point q whose distance to the p data point is not greater than the k -distance(p). Those k -nearest neighbors of q form the so called k -distance neighborhood of p , as described below:

$$N_{k\text{-distance}(p)}(p) = \{q \in D \setminus \{p\} | d(p, q) \leq k\text{-distance}(p)\} \quad (2)$$

Definition 3: Reachability distance of p with respect to o . Let k be a positive integer. The reachability distance of data point p with regard to o is as follows:

$$\text{reach-dist}_k(p, o) = \max\{k\text{-distance}(o), d(p, o)\} \quad (3)$$

The principle of LOF reachability is depicted in Fig. 2, further details are available at [17].

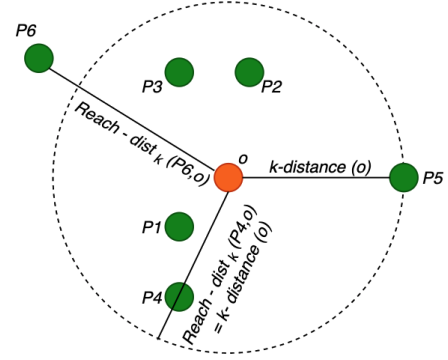


Fig. 2. Reachability distance for different data points

IV. PRELIMINARY RESULTS

We created a prototype² tool implementing LOF algorithm. We applied LOF on IEC 104 inter-arrival times in order to check if LOF modeling is suitable for ICS communication. We used LOF algorithm to learn inter-arrival time distribution of the transmitted IEC 104 packets by computing the k -distance, reachability distance and density of data points. The algorithm uses the learned model to raise an alarm for packets that significantly differ from the model. Figures 3 and 4 present a graphical representation of inter-arrival times of two datasets. The red points denote inter-arrival time distribution detected as anomaly (outliers) while the blue points represent normal communication.

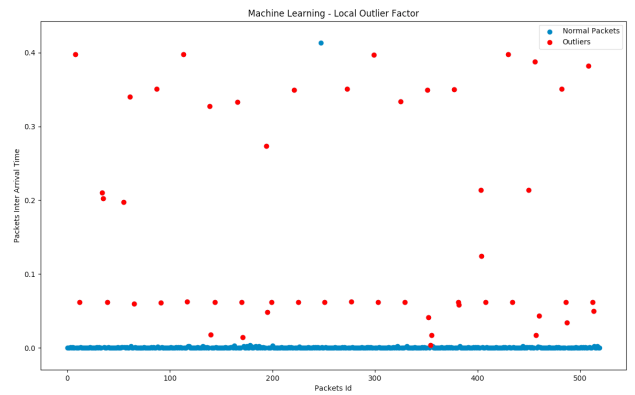


Fig. 3. Anomaly detection in 10122018-104Mega PCAP file

We discovered that applying LOF to a large data stream is extremely computationally inefficient and may lead to incorrect prediction results. We then applied the LOF on time windows where each data block contains 5000 inter-arrival time records.

The main advantage of the LOF algorithm is that it works well with stream data and detects outliers with respect to

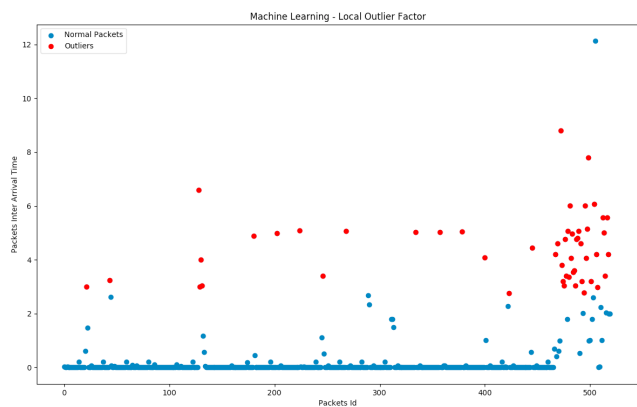


Fig. 4. Anomaly detection in mega104-14-12-18 PCAP file

density of their neighbouring data records. Also, this algorithm is able to detect outliers regardless of the data distribution of normal behavior, since it does not make any assumptions about the distribution of data records. Our preliminary results show that the proposed method works well with ICS data and is easy to operate.

Validation tests were applied to determine the viability of the proposed technique. We tested the stability of the approach using our datasets. For our validation tests, we divided normal communication into two parts. We used the first two thirds of the communication to train the model. Then, we tested these ranges on the last third of data. After testing, we classified all testing packets as normal communication. The results demonstrate high accuracy of the proposed approach. The method did not produce any false positive during the test.

V. CONCLUSION

In this paper, we presented a proof-of-concept method for anomaly detection of smart grid control protocols. We observed inter-arrival times of IEC 104 communication and applied LOF algorithm for outlier detection. The preliminary results show that LOF creates a stable statistical model for ICS traffic and is able to detect outliers caused by cyber attacks. The proposed security mechanism was tested and validated using our IEC 104 datasets. Experiments show the suitability, usability and high accuracy of the proposed method on smart grid communication. In the future work we plan to apply this approach to other ICS protocols like GOOSE, MMS and DLMS. We also plan to observe other statistical features like packet or flow size. Although our key goal was to test the applicability of the proposed solution, we also tested the performance of the developed tool. We gathered basic time statistics. The processing time of extracting key attributes from the PCAP file into a CSV file took an average of one minute and 10 seconds. This time had small deviations because it takes time to initialize Tshark. The time to learn a CSV file depends on the number of packets. At average, to learn 5,000 packets, 20 seconds were required.

The results are part of my PhD research that focused on statistical-based anomaly detection in ICS protocols. In the future work, we focus is extending detection model to common cyber attacks defined by MITRE ATT&CK ICS matrix.

ACKNOWLEDGMENT

The work is supported by the Brno University of Technology project "Application of AI methods to cyber security and control systems", no. FIT-S-20-6293.

REFERENCES

- [1] H. Leith and J. W. Piper, "Identification and application of security measures for petrochemical industrial control systems," *Journal of Loss Prevention in the Process Industries*, vol. 26, no. 6, pp. 982–993, 2013.
- [2] R. M. Lee, M. J. Assante, and T. Conway, "Analysis of the Cyber Attack on the Ukrainian Power Grid. Defense Use Case," Electricity Information Sharing and Analysis Center, Tech. Rep., March 2016.
- [3] J. Tidy, "Us fuel pipeline 'paid hackers \$5m in ransom'," 2021. [Online]. Available: <https://www.bbc.com/news/business-57112371>
- [4] P. Matoušek, O. Ryšavý, and M. Grégr, "Increasing Visibility of IEC 104 Communication in the Smart Grid," in *The 6th International Symposium for ICS & SCADA Cyber Security Research 2019*. BCS Learning and Development Ltd, 2019, pp. 21–30.
- [5] R. R. R. Barbosa, R. Sadre, and A. Pras, "A first look into SCADA network traffic," in *2012 IEEE Network Operations and Management Symposium*, April 2012, pp. 518–521.
- [6] R. R. R. Barbosa, "Anomaly detection in SCADA systems: a network based approach," Ph.D. dissertation, University of Twente, 4 2014.
- [7] R. Udd, M. Asplund, S. Nadjm-Tehrani, M. Kazemtabrizi, and M. Ekstedt, "Exploiting bro for intrusion detection in a scada system," in *Proceedings of the 2nd ACM International Workshop on Cyber-Physical System Security*, 2016, pp. 44–51.
- [8] S. Bhatia, N. S. Kush, C. Djameludin, A. J. Akande, and E. Foo, "Practical Modbus flooding attack and detection," in *Proceedings of the 12th Australasian Information Security Conference*. Australian Computer Society, Inc., 2014, pp. 57–65.
- [9] N. Sayegh, I. H. Elhajji, A. Kayssi, and A. Chehab, "Scada intrusion detection system based on temporal behavior of frequent patterns," in *MELECON 2014-2014 17th IEEE Mediterranean Electrotechnical Conference*. IEEE, 2014, pp. 432–438.
- [10] R. R. R. Barbosa, R. Sadre, and A. Pras, "Exploiting traffic periodicity in industrial control networks," *International journal of critical infrastructure protection*, vol. 13, pp. 52–62, 2016.
- [11] Y. Yang, K. McLaughlin, S. Sezer, Y. Yuan, and W. Huang, "Stateful intrusion detection for ics 60870-5-104 scada security," in *2014 IEEE PES General Meeting*. IEEE, 2014, pp. 1–5.
- [12] N. Goldenberg and A. Wool, "Accurate modeling of modbus/tcp for intrusion detection in scada systems," *international journal of critical infrastructure protection*, vol. 6, no. 2, pp. 63–75, 2013.
- [13] A. Kleinmann and A. Wool, "Automatic construction of statechart-based anomaly detection models for multi-threaded scada via spectral analysis," in *Proceedings of the 2nd ACM Workshop on Cyber-Physical System Security and Privacy*, 2016, pp. 1–12.
- [14] M. Caselli, E. Zambon, and F. Kargl, "Sequence-aware intrusion detection in industrial control systems," in *Proceedings of the 1st ACM Workshop on Cyber-Physical System Security*, 2015, pp. 13–24.
- [15] E. D. Knapp and J. T. Langill, *Industrial Network Security: Securing critical infrastructure networks for smart grid, SCADA, and other Industrial Control Systems*. Syngress, 2014.
- [16] P. Matoušek, V. Havlena, and L. Holík, "Efficient modelling of ics communication for anomaly detection using probabilistic automata," in *IFIP/IEEE Int. Symposium on Integrated Network Management*, 2021.
- [17] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "LOF: Identifying Density-Based Local Outliers," *SIGMOD Rec.*, vol. 29, no. 2, 2000.
- [18] A. Boukerche, L. Zheng, and O. Alfandi, "Outlier detection: Methods, models, and classification," *ACM Computing Surveys (CSUR)*, vol. 53, no. 3, pp. 1–37, 2020.
- [19] O. Alghushairy, R. Alsini, T. Soule, and X. Ma, "A Review of Local Outlier Factor Algorithms for Outlier Detection in Big Data Streams," *Big Data and Cognitive Computing*, vol. 5, no. 1, 2021.