UNIVERSITY OF CALIFORNIA
SANTA CRUZ

ANOMALY DETECTION OF SCADA NETWORKS THROUGH NETWORK MEASUREMENT STUDY

A dissertation submitted in partial satisfaction of the
requirements for the degree of

DOCTOR OF PHILOSOPHY

in

COMPUTER SCIENCE

by

Xi Qin

June 2022

The Dissertation of Xi Qin
is approved:

Alvaro Cardenas, Chair

Chen Qian

Yu Zhang

Peter Biehl
Vice Provost and Dean of Graduate Studies

Table of Contents

List of Figures

List of Tables

Abstract

Anomaly Detection of SCADA Networks through Network Measurement Study

by

Xi Qin

Despite all the increasing research efforts in industrial control systems (ICS), these systems still fail to defend themselves at the time of some high-profile cyber attacks. The most high-profile attack events include but not limited to the Stuxnet attack on an Iranian nuclear power plant in 2010, the Industroyer malware attack on the Ukrainian power grid in 2016, and the recent ransomware attack on the U.S. Colonial pipeline in May 2021, which severely limit the fuel supply to half of the east coast. The Supervisory Control and Data Acquisition (SCADA) networks expose themselves to a broader attack surface after the migration from serial communication network to TCP/IP compatible networks. Therefore, they are susceptible to cyber-attacks. Another main reason why the security and resilience of SCADA networks have limited improvements over the decade is that, the majority of the previous work does not have access to real-world systems or datasets. Because it is not possible to interrupt the production process with penetration tests, and not easy to earn the trust of the operators.

In the prelusive chapter of this dissertation, we first introduce the concepts of SCADA and industrial control protocols. Then we review the previous work divided by energy sectors in the critical infrastructure, which enables us to recognize contributions, identify limitations, raise research questions, and discover answers. With network captures from the SCADA networks in operational industrial control systems, specifically the power grid and the natural gas distribution network, we launch our project with the reversal of the SCADA network topology with

different levels of system knowledge, and show that even in the least bliss, one can still conduct network discovery to the majority of network nodes. The later characteristics we extract from the communication conversations between substations and control servers challenge the long-term understanding of the SCADA network in the security community. The primary industrial protocol under investigation is IEC 60870-5-104, an application-layer protocol designed to control and monitor the physical processes in federated SCADA networks.

With the knowledge base obtained from network characterization, then we propose network flow based anomaly detection method by applying unsupervised clustering of the network flows, and process-based anomaly detection. The anomaly detection is based on profiling process variables, by applying gradient boosting tree algorithm and deep neural networks on time series datasets. Both work flows are experimented with datasets divided by our system knowledge levels range from the system operators help verifying the majority of network topology and hardware devices, to no support at all. Approaching from the perspective of network measurements, our goal is to establish the normal behavior baseline of the anomaly detector by applying deep-packet inspection, and have captured several intriguing outliers and process anomalies that are not available in a simulation/emulation environment. After successfully training of the gradient boosting based detector, we use feature importance analysis to mitigate the existing limitation of black-boxed machine learning applications, and quantify the contribution of features leading to the detection result.

The contributions of this dissertation are as follows:

- Provide solid testimonies that shred the security community's consensus of SCADA networks being stable and predictable, from the overall network topology to the subtleties in the process variables

- Construct the first network characterization for an operational bulk power grid, that offers the first view of the unique difficulties in defending a federated SCADA network

- Implement the first process-aware anomaly detector for two operational SCADA networks, one bulk power grid and one gas pipeline network, that successfully identifies the process anomalies and potentially dangerous misconfiguration errors

- Present the discussion of the ambiguous understanding of false positives in the anomaly detection for ICS, with the valuable insight from the study of real-world datasets

To my father, Fei Qin,

my mother, Hua Kuang,

my husband, Xiaochu Yao,

all the people rooting for me,

because of whom I can accomplish this Ph.D.

Acknowledgments

I would like to first give the most sincere appreciation to my Ph.D. advisor, Dr. Alvaro Cardenas, for his continuous support and guidance all the way through. His inspirational motives, infinite passion and instant help, have built many valuable moments I will never forget.

Thanks to my other Ph.D. committee members for their valuable time and dedication, Dr. Qian, Dr. Zhang and Dr. Zambon who attended my advancement exam from Europe.

I am grateful to all the lab colleagues from Dallas to Santa Cruz for their support and help. Specifically, I'd like to thank Mustafa Faisal who tutored me in the first project. Thanks to Junia Valente who introduced the lab and various academic activities to me. Kelvin Mai, whom I closely worked with, together we got our first top-tier paper, who had rich industrial experience and taught me so much. Neil Ortiz influenced me to keep an artistic heart and always draw well-designed diagrams. I also appreciate the many fun discussions of network security with Luis Salazar.

I want to thank the fund sponsors, NSF, LAS (Laboratory for Analytic Sciences) and DHS for providing the research opportunity.

For the families, I am blessed to have such parents, who have been so caring and considerate, who have being paying attention, encouraged me and kept me from all kinds of anxieties. Thanks to my parents-in-law who fully understand and support the meaning of a Ph.D. life. Thanks to my husband, for his genuine support and love. With the significant impact from his amazingly positive and healthy work ethic and life altitude, I've grown into someone more productive and confident.

The department academic advisor for Ph.D. students, Alicia Haley, is the most

amazing staff I've met in all of my educational programs, without whom I may not smoothly get all logistics done. With her personal touches in the advising, Alicia has made my Ph.D. journey warmer and more accessible.

Thanks to all the warm and inspiration moments offered by my dear friends Xinyi, Jingchen, Man, Zehui, Yafei, Wenting, Jing (not ordered here), and also super grateful for the push power from students also doing oversea study all over the world in the Discord group study room.

The last thing to be thankful for is the series game of The Sims. I felt much better when playing the poor simulated college life (each time I have to use cheat codes to get my sim to graduate). Knowing that college life is not easy even in a game looses my nerves. Although Sims 4 is not the best product of Maxis Studio, still the aesthetic in their rendering has healed so many me in my own multi-universe.

Last of last, I would appreciate myself for keeping up the spirit, and solute with the song of "New Horizon" from Nintendo's Animal Crossing by Kapp'n.

Chapter 1

Introduction

The security and resilience of the Supervisory Control and Data Acquisition (SCADA) in industrial control systems (ICS) is an interdisciplinary research area, which includes but not limited to the subjects of network security, control theory and machine learning. Thus to perform research in this direction, one must understand the scope of ICS, the architecture of SCADA, the industrial communication protocols, network traffic analysis and statistical models. In this chapter, we present a concise demonstration of the key concepts in this area.

1.1 Industrial Control Systems

Industrial control systems, is a generic term for all kinds of control systems and the related infrastructures in both software and hardware to monitor and control industrial processes, including but not limited to energy resources (electricity, water, oil and gas), transportation regulation, commodity production and building automation. Usually, there is a control center room in the ICS which is remote in geographical distance from the local substations. The control server orchestrates local field devices in the substations from miles away through electronic communication messages. In the last century, this kind of communication channels were not connected to the Internet, which were mostly serial communications. In the recent decades, the operators have started to adopt the modern TCP/IP compatible protocols. So the messages of control commands are encapsulated into communication packets and sent to the field devices such as Programmable Logic Controllers. In the opposite communication direction, the operators collect passive measurement packets such as voltages and pressures from the sensors.

Figure 1.1 shows that the four core components in an ICS control loop. Let us assume the components transmit messages to each other at a certain time point k. Here is the big picture how the monitoring and controlling work.

1. The specific physical process generates a value of z_k for a certain signal. E.g. with all the connected loads, there is a consumed power signal;

2. The sensor connected to the generator collects this z_k and sends out the y_k ($y_k = z_k$) to the controller. E.g. a power meter measures the power consumed in real-time;

3. Receiving the sensor measurement, the controller sends the related control command u_k to the actuator. E.g. the controller sends a command of slowing down power generation after receiving a consumption power lower than a threshold value;

4. The actuators react to the control command u_k, and adjust the related mechanical device to reach the desired status by the controller. E.g. the valve is open by a mechanical spring and the pressure is lowered to the normal range.

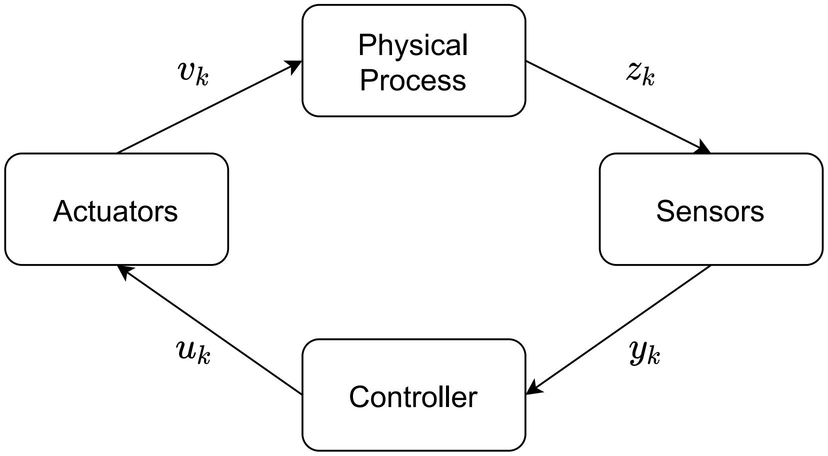


Figure 1.1: The four most important components of an industrial control system

1.2 SCADA

Operators monitor and control the gas network with the help of a SCADA system, which is one the most common type of ICS. A typical SCADA system has at least one control center and multiple geographically distributed remote stations. In the control center, the operator can see the status of the gas network through a Human Machine Interface (HMI), which watches the status of all the stations and is interactive for engineers. The control center may have other typical IT network services, such as a database for historical data storage, the file system for logging and administration documents, and a time server for network clock synchronization. The SCADA server communicates with Remote Terminal Units (RTUs), which is the local controller in each remote substation. In the ICS community, there is a common hierarchical design of the SCADA network connections, shown in Figure 1.2.

From top to bottom, Layer 5 and 4 refer to the enterprise IT networks, similar to all other IT networks of other industries. Layer 3 has this DMZ regulating the

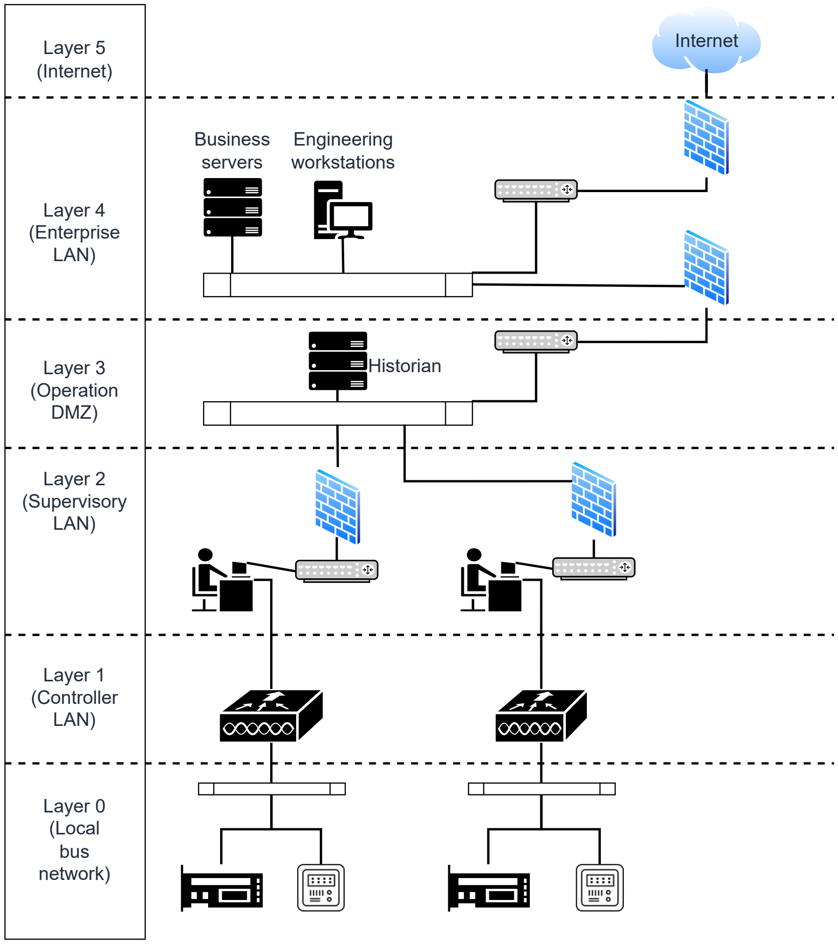


Figure 1.2: The five layers of SCADA systems in terms of network connections

system historians and SCADA related applications. Layer 2 is the actual supervisory network, which is usually a local network at the substations with HMIs. Engineers can monitor the status of the devices in this station here. In the Layer 1 of controller network, the operators utilize the Programmable Logic Controllers (PLCs), RTUs, Intelligent Electronic Devices (IEDs) to perform any active control or passive monitoring. These controller devices are also connected to the remote control center. RTUs/PLCs interact directly with sensors (e.g., pressure sensors) and actuators (e.g., valves) in Layer 0 through a local bus network. In the case of more complex stations such as gas turbine stations, RTUs communicate within the station with other PLCs, which are in charge of local control. But in the most common scenarios, RTUs connect to sensors and actuators with either analog or digital inputs and outputs. “Inputs” refer to incoming sensor values from the physical process, and “outputs” refer to setpoints.

1.3 Power Grid

Approaching the end of the last century, the U.S. National Academy of Engineering selected the top 20 engineering accomplishments of the twentieth century that have most improved people’s quality of life. Power grid is at the top of the list [19].

The power grid has three major components: generation, transmission, and distribution. It generates electricity, transmits the energy across broad geographic areas, and finally delivers to the consumers’ locations. In the transmission process, the electricity transmission powers initially carry the energy at a relatively high voltages at the degree of hundreds of kVs and then convert it to a lower voltages at tens of kVs. The transmission system can span very large regions, such as the state or even country level. The distribution network takes care of more locally regions,

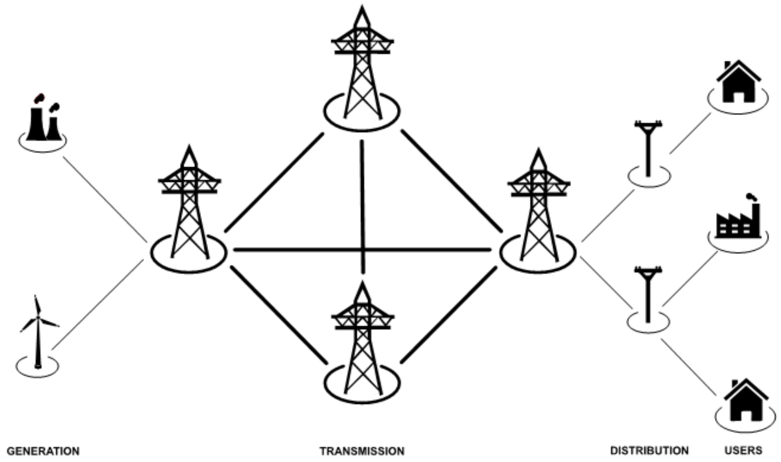


Figure 1.3: Abstract scheme of the power grid components

Table 1.1: The Impact Differences between the Transmission and Distribution

	Transmission	**Distribution**
Power [W]	10^9	10^6
Area [km^2]	> 4.67 million	> 10600
Voltage level [kV]	> 110	< 34.5

such as cities or districts. These differences are in Table 1.1. The generation and transmission parts together compose **the Bulk**. Therefore, the failure of the Bulk and the distribution network are not comparable in terms of the impacts. If the adversary takes down the Bulk, it leads to a country-level black-out and affects the whole population. While the distribution network's failure usually endangers merely a local area.

The measurements which the bulk system operator needs to collect from multiple electric companies (as well as the control commands they need to send) are carried by standardized industrial protocols. One of the most popular industrial protocols for bulk power system operators is the international standard IEC 60870-5-101 (IEC 101), meant for serial communications, and more recently IEC 60870-5-104 (IEC 104), which is an adaptation of IEC 101 for TCP/IP networks.

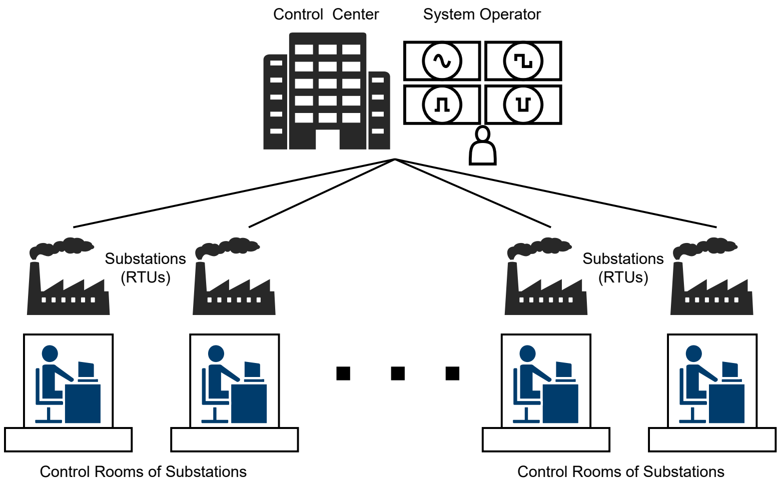


Figure 1.4: The structure of central and local administrators in a federated SCADA network

In this dissertation we study the IEC 104 SCADA network of a system operator as illustrated in Figure 1.4, orchestrating the operation of the **Bulk** power grid. Federated networks such as the one in our study, lead to interesting observations, because while all of the previous work in SCADA systems assumes that devices are configured and maintained by the same system administrator and therefore have predictable dynamics, we show here that in a federated system, SCADA networks have diverse behaviors. Even when devices operate erratically or do not follow the standards, the administrators of those devices do not respond to requests of the system operator to update their systems.

1.4 Natural Gas Networks

Natural gas is a fossil fuel used worldwide primarily for heating, power generation, and cooking. Gas delivery is organized hierarchically, with a country-wide transportation network, regional transport networks, and pipelines connecting consumers to the network as outlined by the [6], [20], and [62]. The United States has roughly three million miles of mainline in the gas network.

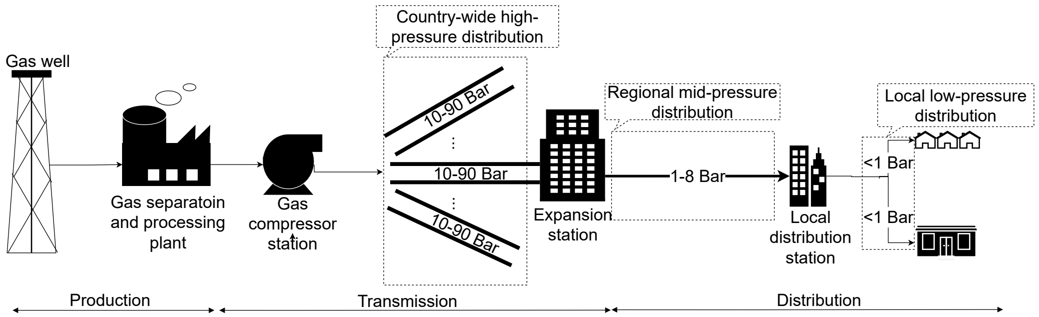

Country-wide high-pressure distribution
Gas well
10-90 Bar
10-90 Bar
10-90 Bar
Regional mid-pressure distribution
Local low-pressure distribution
Gas separatoin and processing plant
Gas compressor station
Expansion station
1-8 Bar
<1 Bar
<1 Bar
Local distribution station
Production
Transmission
Distribution

Figure 1.5: Gas network and the differences between the country-wide, regional and local distribution networks.

Generally, a natural gas network has three components, production, transmission, and distribution. With drilling and delicately breaking the rocks, gas wells release and let natural gas arise from oil/water and collect the gas. Then the separation and separation station further divides the liquid and gas and cleanses any impurity to improve the gas quality. The compressor station's function is to compensate for the friction loss of natural gas during forwarding through the metallic pipelines[6]. As shown in 1.5, gas naturally flows from high to low pressured sections of the network, taking the point of least resistance. The gas transportation network uses long-distance (country-wide) high-pressure pipelines, with pressures ranging between 10 and 90 bar. Regional transport networks utilize medium pressure pipelines and connect the high-pressure network to the local (low-pressure) distribution networks. The low-pressure gas distribution network operates at pressures ranging between 1 and 8 bar within a city or regional area. For every 250 to 500 households, local gas network operators deploy a distribution station [62], which in Europe takes the form of closets visible at many street corners. The final pipeline that connects distribution stations to individual consumers operates at pressures below 1 bar.

An automated control system ensures that gas pressure and flow remain within operational limits and according to the applicable standards and regulations by adjusting the state of valves at different points in the network. Closing a valve reduces the amount of gas that flows into a specific network branch, creating a higher pressure in the rest of the network. Likewise, opening a valve will allow more gas to flow into a specific branch of the network and thus lower the pressure in the rest of the network. The local network operator may deploy gas turbines to increase gas pressure.

1.5 The Protocol of IEC 60870-5-104

The International Electrotechnical Commission (IEC) is an international standards organization. They prepare and publish standards for a variety of industrial processes. One of the most popular SCADA communication standards for serial lines is IEC 60870-5-101 (IEC 101). As TCP/IP networks slowly replace serial communication links, standard bodies have developed new TCP/IP protocols compatible with legacy technologies. IEC 60870-5-104 is an extension and adaptation of the same IEC 101 message structure but carried over TCP/IP standard. We will refer to IEC 60870-5-101 [41] and IEC 60870-5-104 [42] as IEC 101 and IEC 104 respectively for the rest of this dissertation. IEC 104 is a protocol that is being gravely targeted and very well studied and exploited by the adversaries[8]. As security researchers, we should investigate and comprehend how the industrial control systems work under this protocol, and then we can evaluate how secure and resilient the system is. International Electrotechnical Commission (IEC) developed IEC 101 first in 1995 and made amendments in 2000 and 2001 for gradually upgrading to tele-control communications between control stations (e.g., SCADA centers) and Remote Terminal Units RTUs (e.g., field devices in RTUs). The connection was intended to be compatible with TCP/IP network. As a result, in 2000, IEC 60870-5-104 (IEC 104) came to the public in 2000 as a transportation method to transmit IEC 101 telecontrol messages over TCP/IP using port 2404. IEC 104 encapsulates modified IEC 101 telecontrol messages into a TCP packet. Later in 2013, IEC developed IEC 62341-5 [40] as a security extension, providing security features for IEC 104. IEC 62341-5 provide end-to-end encryption through the transport-layer-security TLS protocol, aiming to mitigate the replay and Man-in-the-middle attacks that IEC 104 channels are prone to. However, most vendors haven't adopted this security extension yet, which is probably because of the extra configuration effort and complexity based on the feedback from one SCADA engineer of our data provider.

The anatomy of the frame structure of one IEC 104 packet is in Figure 1.6. The TCP payload of an IEC 104 packet contains one or more **Application Protocol Data Units (APDUs)**. The number of APDUs is configurable, depending on the choice of the operators in the setting of IEC 104 protocol stack size. The Application Protocol Control Information (APCI) is the header of the message. Application Service Data Unit (ASDU) follows APCI, which is comprised of the sensor values and control messages. Depending on the APCI format type as defined by bit 0 and 1 of Control Field Octet 1, each APDU could consist of APCI only (without any ASDU), or APCI with ASDU. For example, an APDU that carries S-Format APCI has no ASDU while APDU that carries I-Format APCI will have ASDU.

There are three types of APDUs, i.e. IEC 104 messages, in Figure 1.7:

I-Format APDUs have the meaningful information from the field devices. ASDUs are indexed by a Data Unit Identifier (DUI) and by Information Objects (IO). Each IO represents a specific process variable measuring the value in a field device, assigned with a unique address called Information Object Address (IOA). IEC 104 only inherited 54 types from IEC 101's 127 types. **The**

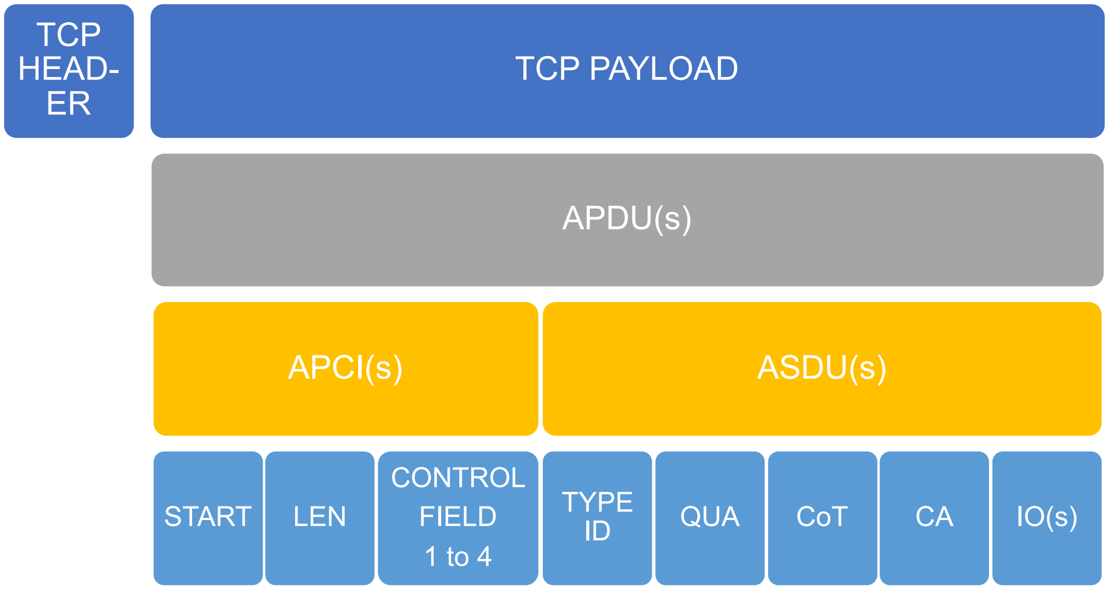


Figure 1.6: IEC 104 Packet Frame Structure

passive measurements of sensors and the active control commands contents are in this I-Format messages.

S-Format APDUs are acknowledgments after a certain number of I-Format APDUs have been received, marked by the receive sequence number (N(R)) from the receiving station back to the sending station for data loss protection purposes. While TCP ACK number indicates the sequence number of the next expected *byte* from the other end. S-format N(R) indicates the sequence number of the next expected APDU.

U-Format APDUs supply two options: 1. the start and termination of I-Format transmission via a STARTDT/STOPDT act message, which is acknowledged with a STARTDT/STOPDT con message; 2. keep-alive exchange the connection status requests with the TESTFR act/con messages.

Sending I-format is initiated when the control server sends a STARTDT act to the RTU, basically telling the RTU to start transmitting its sensor

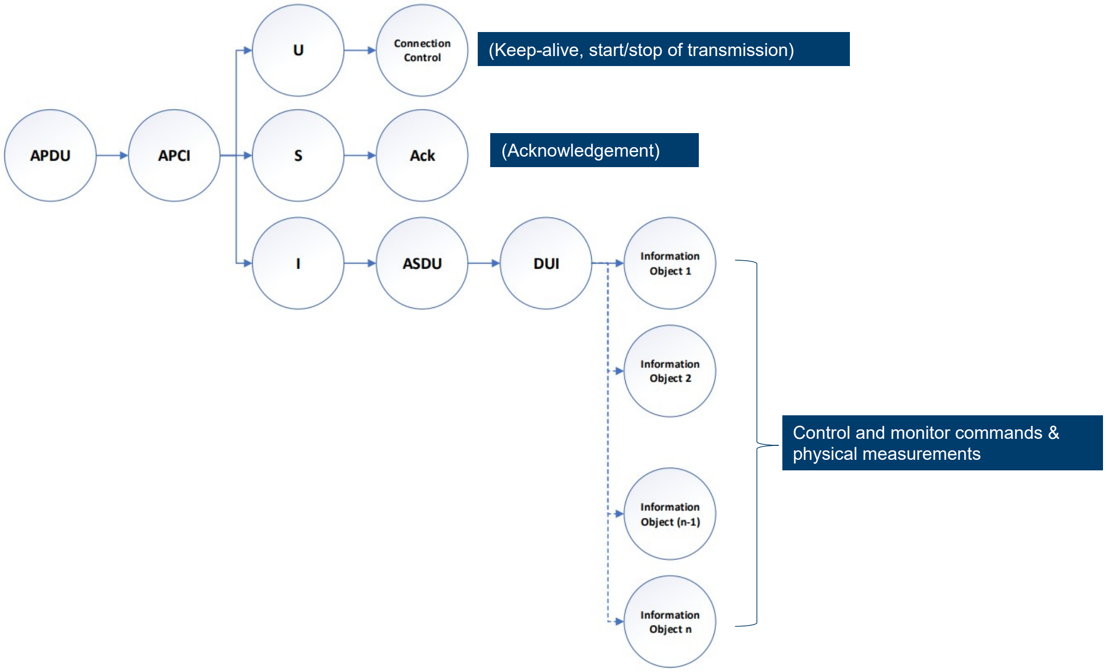


Figure 1.7: IEC 104 Packet Message Structure

readings. Then the RTU responds with STARTDT con, confirming that it will start transmitting I-format APDUs. Otherwise, no I-format APDUs can be transmitted since newly established (or switchover) connections are by default in a STOPDT state. TESTFR act/con APDUs are primarily used by both control server and outstation to keep the redundant connection from being disconnected, i.e. keep-alive messages.

To understand the deep-packet inspection in the later sections, we need to emphasize on some key terminologies: the ASDU type identifier (Type ID), the cause of transmission (CoT or CauseTx), the common address (CA), and the information objects (IO).

- **Type identifiers** define the format of the numeric values in the data units, i.e. digital or analog, normalized or not, with/without time tags, etc. And also the type of messages is a command or a measurement.

- **CoT** configures how the ASDU is transmitted, and it gives how fresh the APDU is transmitted. e.g., spontaneous, interrogation, and activation related.
    - Spontaneous is an aperiodic mode, but not totally irregular. There is a configurable threshold (also referred to as "deadbands" by SCADA engineers). If the difference between two consecutive measurements is larger than the threshold, the later one is polled. Otherwise the newer measurement is not polled.
    - Interrogation is an important mode for the remote control server to query the targeted RTU, and collect all present values of all the connected devices in this RTU. The ASDU type ID also defines the interrogation command. If a new connection is set up, the interrogation is must be done at the start of data transmission. The detailed frame structure is shown in Figure 1.8. This command is in our analysis interest in the later section, since it helps to separate different groups in all the RTUs.
    - Activation modes indicate the start of a new data transmission.
- The **common address** is a prefix to the IO address shared among all IOs in the ASDU. In the protocol principles, it is the fingerprint of a (virtual) device (e.g., an RTU) within the IEC 104 network. Although based on our experience, some operators do not follow this recommendation.
- An **IO** is the IEC 104 representation of a process variable, e.g., a sensor input reading or an actuator output value. The IO comprises, among other fields, an Information Object Address (IOA) and the actual variable data. Together, CA and IOA uniquely identify a process variable of a certain type

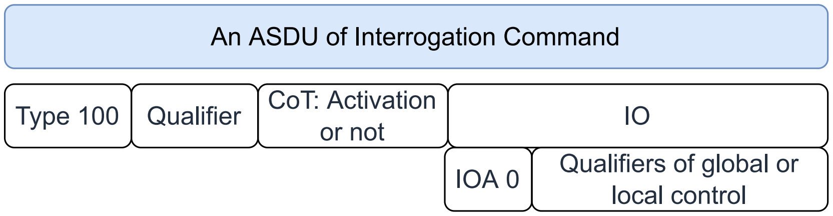

An ASDU of Interrogation Command
Type 100
Qualifier
CoT: Activation or not
IO
IOA 0
Qualifiers of global or local control

Figure 1.8: Frame structure of an interrogation command ASDU.

ID across the IEC 104 network.

In the high-profile power grid attack in Ukraine [36], the attackers injected the IEC 104 payload into the SCADA network and accomplished a black-out. In the most recent Industroyer2[66], the malware continued to focus on the exploit of IEC 104. Therefore, IEC 104 is a protocol that is being gravely targeted and very well studied and exploited by the adversaries. As security researchers, we must investigate and comprehend why the payloads in this protocol are so attractive to the adversary, and how the industrial control systems work under this protocol. With more contexts, we can help build a more robust defense mechanism for ICS.

1.6 Other Industrial Protocols

In the routine activity of ICS operational network (OT network), availability places the first in the security design principles, instead of the confidentiality in the typical IT network. This consideration affects the design of ICS protocol standards to be focused on the freshness of data exchange. The SCADA networks were usually private networks composed of dedicated leased lines provided by a telecommunications company; as such, SCADA networks are rarely connected to the public Internet. In the last decade, the communication technologies used for supervision and control of gas systems have migrated from serial links to IP-network protocols, adapting to the deployment of industrial IoT devices. The modern ICS protocols widely adopted are Modbus/TCP, DNP3, and IEC 104. Since the topic in this dissertation is not about the comparison of all these protocols, we will only briefly introduce the application scenarios instead of deep diving in the frame structures.

Modbus/TCP[64] are used between the PLC and field devices in a substation. Clients and servers listen to the port 502 for data exchange. Packets of this protocol have a straight-forward one-layer structure. There is an important field of function code, which contains the read or write operation to a data register. For example, one PLC as the server talks to a client device connected to a circuit breaker in Modbus/TCP packets. The server requests to read a register storing the breaker's status. Then the client sends the data at the desired address to the server as the response.

DNP3 [2] are widely implemented for communication between substations, RTUs, and even control stations. The default TCP port is 2000, and there are three layers (data link layer, transport function, and application layer) in a DNP3 packet. The similar field of function code describes the operation. For example, in one conversation between a controller and general bus linked to all the field devices, the controller requests to read the data for a certain group of event data, if the bus has data in the requested group, it responses with the present data.

We compare these two protocols with the protocol in this dissertation in Table 1.2.

Table 1.2: Comparison of three popular ICS protocols

Aspects	Modbus/TCP	DNP3	IEC 104
Messaging modes	Request and response	Other than request and response, additional support for authentication	64 choices of messaging modes (CoT)
Structure	Single-layer	Three-layer	Flexible, one or multiple layers
Data format	4	5	54
Parsing	Available in Pyshark APIs	Available in Pyshark APIs	Need to implement
Security	No specific security design	Authentication and limited access control	Authentication design in the protocol extension[40]

1.7 Outline of Dissertation

In this section, we present the outline of this dissertation and a short summary of each chapter. First, we provide a tutorial of the industrial control systems, SCADA, and the industrial protocols in chapter 1. When reviewing the previous work in 2, we collect the works from the defense and offense perspectives for the power grid and gas network. In chapter 3, we clarify the research problem that we are interested in, frame the exploration scope, and present the methodology and research questions. From chapter 4 to chapter 7, we try to answer our research questions by presenting the results from analyzing the network traffic from two different sectors of industrial control systems, i.e. the bulk power grid and the natural gas distribution network.

Chapter 1: Introduction - Industrial Control Systems. This is an interdisciplinary area where lots of areas resonate together. To properly protect its part, speaking in the language of industrial control systems (operations, systems, protocols, electronic devices) is very crucial. In this chapter, we tutor the reader about the generic concepts and terms in ICS, especially the structure of one important industrial protocol, IEC 104. As a complimentary reading, other popular relevant procotols are also introduced.

Chapter 2: Related Works - Security studies in ICS. In this chapter, we first provide an overview of the security challenges and trend in the ICS security. Then we start with presenting the well-known real-world attacks, and attack methods published previously against the power grid. In this research, we are try to distinguish the work with testbed/simulation/emulation setup and the work with operational ICS systems. We provide a taxonomy that organizes the previous work divided by testbed-based or real-world systems based for the first time. And we try to clarify for the research work in gas network that, a few papers on the study of consumption dataset is not about modeling the control and monitoring behaviors of SCADA network, which is different from what we do in this dissertation.

Chapter 3: Proposed Work - Defining our research scope, methodologies and threat model for research questions. In this chapter, we want to construct a systematic way to characterize the SCADA network and develop detection models following that. In all the experimental chapters, we follow our six-step methodology.

Chapter 4: Experiments, Results, and Discussion for the Analysis of the Bulk Power Grid - In this chapter, we first describe the experiments and used datasets along with data preprocessing technique as well as protocol fields selection through deep packet inspection (DPI). After that, the results of these experiments are presented. Following this, a detail discussion of these results are provided. Specifically, clustering techniques play an important role in recognizing the communication patterns of traffic flows, since most of the time, the problem in ICS systems is defined as an unsupervised learning problem when no label is available. Through clustering, we assign each sample instance to well-separated clusters. In this chapter, we analyze the impact of flow-level information in the clustering approach. After that, we look into the connections to discover the devices, and the physical dynamics within them. We identify interesting automatic generation control during a power generation event.

Chapter 5: Experiments, Results, and Discussion for the Analysis of the Gas Network - In this chapter, we also follow the six-step methodology, starting with the description of the network captures. The first discovery we make is the exciting exploration of unique RTU fingerprints. With two important IEC 104 features, we identify an ongoing maintenance in the network and successfully reverse the whole SCADA network topology.

Chapter 6: A Quick Comparison of the Two SCADA Network - We want to give the big picture of how these two networks are different or similar in terms of the capture location, the topology arrangements, and the utilization of control and monitoring messages. To be noted, the differences not only come from the different energy sectors, but also come from the different management styles and goals of operators.

Chapter 7: Anomaly Detection Design in the ICS SCADA Network - We profile the devices/process variables in the SCADA network for anomaly detection from two directions, profiling the device type and the device readings with time series modeling. we are able to get great results (99+% accuracy and performant precision-recall curves) from the type profiler with gradient boosting algorithms, and acceptable good results (20% false positive rate) from the reading profiler with deep neural networks. From the deep dive in the correlated process variables, we can know the root cause of anomaly (like the dangerous situation of pressure rising out-of-range). In this case, process-aware anomaly detection would be possible to implement. Process-aware approach can act as a core approach to augment detection ability of IDS. Here we apply such kind of specification approach for IEC 104. But, both protocol-level and configuration-level specifications are highly customizable to deploy such technique.

Chapter 2

Related Work

For a long period of time, the research community assumed SCADA and industrial control networks being more stable and consistent over time. However, more operators connect their network nodes and devices to the Internet in the most recent decade, which introduces the network to a larger attack surface and leads more potential threats to their network. Over the most recent decade, researchers have paid increasing attention to the security and resilience of SCADA and industrial control networks. On a global scale, the attention governments paying to the critical infrastructures attacks continue to grow, when administrations view the defense as part of their future military and political conflicts. Therefore, the importance of securing the energy systems will keep arising. Because of the confidentiality and availability issues of this type of infrastructure, it is almost impossible to perform any active penetration test onsite. Also because of the criticality of SCADA systems, combined with the conservative approach of industries operating our critical physical infrastructures, researchers usually cannot get access to SCADA networks to perform measurements or security experiments. Consequently, most researchers conduct attack and defense experiments on simulated/emulated test environments, or analyze network traces obtained by passively monitoring SCADA/industrial control networks, which is our approach in this dissertation. In the way of analyzing network traces, out of confidentiality constraints related to the criticality of production environments, availability of production datasets remains to be quite limited. Even in the few work that is fortunate to access the production datasets, the network scope that the datasets cover is usually localized to an individual station. In this chapter, we review the research work on the attack incidents and designs, and defend methods in different sectors of the critical infrastructures, i.e. the power grid and the natural gas pipeline network. For the power grid, we first start with a summary of recent attacks to power systems and other new potential attacks developed by security researchers, then demonstrate the evolution of technologies used for monitoring and controlling the power grid, and eventually discuss the novel directions of the power grid defense utilizing new unique properties from the smart grid. For the gas network, we briefly discuss about the research work in defensing the gas utility companies. But for the more critical operational gas pipeline network, there is rarely any computer science paper on the monitoring and control interactions. After each review of the power grid or gas network, we present the uniqueness of the work in this dissertation in comparison with the previous work.

2.1 Security and Resilience of the SCADA of Power Grid

In the United States, the administration recognizes that the aging infrastructure in the power grid has been increasingly vulnerable. Enacted by Congress and signed by the President, the Infrastructure Investment and Jobs Act (IIJA) came into play in 2021[63]. The IIJA assigns a $2.5-billion fund to prioritize the improvements for enhancing the resilience and reliability of the grid against threats like cyber and physical attacks, and natural disasters. As early as 2007, the Department of Energy and Idaho National Laboratory had a project named "Aurora generator test"[79]. In a testbed with one power generator and one distribution substation, the program drove the generator to a status of "out-of-phase" by switching breakers status at a high frequency, and the generator got destructive damage from the failure of synchronizing rotation speed. One of the most high-profile power grid attacks in real world should be a series of attacks on Ukraine power grid from 2015 to 2016 [36]. The adversary opened circuit breakers through remotely compromised the control room in the distribution network in 2015, leading to a blackout. In 2016, the attackers injected a malware Industroyer[8] into a transmission substation, let the malware carry over control commands, and eventually accomplished an automatic black-out. Under the circumstance of Russia-Ukraine wartime at the time of this dissertation, Industroyer had an updated version Industroyer2[66], which focused on the IEC 104 exploit with higher flexibility in configuration. This serial event comprises the core part of our motivation in this work. As illustrated in the introduction, the network structure is rather complex on the route from electricity generation to the consumer market. One of the most effective attack methods is the false data injection (FDI) attacks targeting the readings of physical sensors in the field. In FDI attacks, the adversary injects stealthy false data to the sensor measurements, and successfully evades from the intrusion detection system in the control room. The researchers have proved this attack is possible in either transmission substations[55], or in a nuclear generation plant[76]. From the field to electricity customers, there are mainly two directions to maneuver the electricity consumption. Modernized smart grid employs advanced metering infrastructures (AMI), where the electricity price adjusts itself based on the balance between demand and generation. The AMI systems is vulnerable from the electricity theft when the attackers can easily alter the consumption time series data in the meters, conceive the local utility company with the fabricated lower consumption, and get a smaller utility bill[14]. Other than direct electricity thefts, the most recent load-altering attacks[72][7] plays with the automatic program applied for consumers to have an adaptive load planning based on the peak hours. This attack with the consumer devices (instead of the field devices or meters) has two profit formulas for the adversary. In the first one, the adversary can obtain economic benefits or bring down partial components in the power grid, by causing a load surge with the high-wattage consumer devices[71]. Secondly, the manipulation of consumer devices can also impact the electricity market in trading prices. The most recent attack on the consumer market by Shekari et al. drives a high wattage IoT botnet to increase the electricity demand suddenly, exploits the relationship between electricity prices and demand, and offers advantages of predicting the price surge to the adversary[70].

Although there is more work in recent years on the characterization and security analysis of industrial control systems or power grids specifically, previous work has limited access to the real-world systems or has a shortage of the critical system specifications. We classify of the most relevant work on the measurement and security studies of power grids (with a focus on the ones employing the protocol IEC 104) into the taxonomy as such:

1. Conducting their work on emulated/simulated networks in either laboratory environments, or purely software testbeds

2. Presenting analysis results with insufficient descriptions of the control system and the communication network

3. Studying a comparably smaller component (distribution network) in a power

grid

As the result of our literature review, we present a taxonomy as shown in Table 2.1. While the simulation system certainly gives the researchers freedom to manipulate the devices as needed, still researchers should be cautious when using this type of dataset to approach the first network characterization step in anomaly detection design. Because the limited scope of the simulation will directly narrow the scope of system events in the normal behavior baseline [52, 39, 59]. Most simulated environments assume a stable SCADA server interacting with only a few substations with mostly expected behaviors. While the work in this dissertation studies the real-world SCADA networks with dual SCADA servers and up to hundreds of RTUs, having complex system configuration events.

A few papers do claim they have access to a real-world operational system, however, there's no such system information in their presentation. Wressnegger et al. [81] states vaguely that their dataset is captured from a power plant, but we don't find any specification of the protocols or network topology of such a system in their paper. Yang et al. [82] mentions their work is based on a real-world system using IEC 104 also with no traces of the system and network information in their published paper.

The work that is most closely related to ours is Formby et al. [31, 29] and Irvene et al. [45]. This link of work studies a few real-world power grid distribution substations under the standard of DNP3 industrial control protocol. Still, as explained in the introduction, the distribution network is a relatively smaller component comparing to the bulk in the power grid. Furthermore, Formby et al. do not perform deep-packet inspection of the protocol, limiting their analysis only to TCP dynamics. Irvene et al. also mainly parses the DNP3 packet headers to report histograms of the message types observed in the datasets, but do not

Table 2.1: Related Work on the security study of the power grid

Authors	Facility	Protocols	Data Type
Formby et al. [29, 31]	Power distribution substations	DNP3	Real-world
Irvene et al. [45]	Power distribution substations	DNP3	Real-world
Barbosa et al.[13]	Water treatment and distribution	SNMP, Modbus/TCP	Real-world
Villez et al. [76]	Nuclear power plant	unknown	Simulated LabVIEW environment
Liu et al. [55]	Power substations	unknown	Simulated power flow with IEEE Test Systems
Lin et al.[52]	Power grid	IEC 104	Emulated power grid testbed with real devices
Lin et al. [53]	Power grid	DNP3	Emulated power grid testbed with real devices
Yang et al. [82, 83]	SCADA testbed	IEC 104	Testbed

study the payload.

2.2 Security and Resilience in the SCADA of Natural Gas Pipeline Network

While the natural gas network may have been less targeted as the power grid, the Colonial Pipeline ransomware attack happened to the IT network on May 8, 2021 disrupted the gasoline supplies throughout the East Coast. The attackers (identified as DarkSide) broke into the company's network as a result of an employee's password leak for the VPN account, and compromised multiple computer systems. The company had to shut down the pipeline until May 12 to prevent the spread of the ransomware, and paid 4.4 million dollars of ransom. With this background, in September 2021, multiple federal agencies and industry stakeholders presented specific pipeline cybersecurity issues to Congress in a federal program[69], to emphasize the urgent need to strengthen the national effort of pipeline protection.

There is a lot of work for scholars in the energy and resources, environmental, or geographical communities to analyze the natural threats and hazards to the gas pipeline. The major direction of gas network security in computer science revolves around the gas consumption dynamics of data from gas utility companies.

These papers use statistical models to detect aberrant data samples in the time series of consumed gas flow, including but not limited to Bayesian classifiers[4] and deep learning neural networks[33]. While these detectors are effective in detecting anomalies, these publications are focused on utility users and do not provide expert information for readers to grasp the more crucial monitoring and controlling operations in a distributed gas network. Wang et al. designed a detector for FDI attacks, specifically topology attack, in a simulated gas distribution network[77] with the IEEE 118-bus system. However, the modeling simplified the network as two pipelines with 14 nodes, and also compressed gas stations as discrete nodes without complex interactions with the SCADA servers. This limitation may cause the model has an incomplete baseline for the network state estimation. Therefore, to our best knowledge, the work of gas pipeline network characterization and process anomaly analysis in this dissertation, is the first engineering work analyzing the SCADA network of a real-world operational gas pipeline network. Similar to the uniqueness and potential contribution demonstrated in 2.1, only after characterization of an operational gas network, can we construct a defense method that has the capability of distinguishing routine activities, systematic configurations, and actual attacks.

As far as we are aware, the only analysis of an IEC 104 network used in an operational environment is our work in[56] and [57] describe the first network measurement and security analysis of an IEC 104 network in the bulk power grid. Another of our paper [65]is the second analysis of an IEC 104 network used in a real-world system and the first measurement study of the SCADA in a gas pipeline network. Moreover, our dataset (500GB) of the gas network is larger than previous work, which captured more than three months of network activity for hundreds of stations. To our best knowledge, this dataset of industrial control systems is the largest ever studied. The most comparable size of previous work is the dataset (335 GB) from [46]. Their dataset comes from only four substations in a power grid distribution network with DNP3 as the industrial control protocol.

Chapter 3

Proposed Work

3.1 Problem Statement

We observe three problems in the domain of industrial control system security. Problem one: Attacks to industrial control systems are real. Stuxnet attack [26] on a nuclear power plant, the Industroyer malware attack on the Ukrainian power grid in 2016 [8], and the recent ransomware attack to the U.S. Colonial pipeline in May, all had severe consequences, despite all the research efforts in industrial control systems or cyber-physical systems in general by the time of the attack. Problem two: The Supervisory Control and Data Acquisition (SCADA) network still requires more in-depth investigation to understand the basic system behaviors. The critical infrastructures, such as the power grid and oil/gas systems, used to isolate their networks from the Internet. Within the recent decades, they have gradually migrated from the traditional serial communication network to the TCP/IP compatible networks. Consequently, their SCADA networks and field devices are connected to the Internet and exposed to a broader attack surfaces. Problem three: There is only limited or almost none data access to such real-world networks.

Considering the critical and essential functionalities of these infrastructures, it is hardly possible for security professionals to interrupt the normal operations or perform any security experiments onsite in an active way. Instead, the security research of such systems usually takes two paths: 1. researchers build a simulation/emulation testbed mimicking some certain system and conduct experiment there; 2. the operators provide the researchers with a dataset of network traffic and the researchers study the system behaviors passively from the network captures. Even so, most of the time it's difficult to earn the trust from the operators and obtain the data access.

3.2 Scope and Methodology

In this dissertation, we propose to develop the ICS (Power Grid and Natural Gas) Anomaly Detection design based on the network measurement study. The first step is to profile the network behavior under regular operation by deep-packet inspection of the network capture datasets from two real-world SCADA networks in a power grid and a natural gas system. Next, we hope to identify the outliers deviating from the expected behaviors if there's any. After that, we aim to extract the system operational events behind the outliers.

We apply three levels of characterization for the two SCADA networks in my projects. The dashed lines matching the X-axis in Figure 3.1 show how much knowledge we obtain before our analysis, mostly from the conversations with the network operators. The dashed lines matching the Y-axis indicate current progress of each project. For each project, we follow these steps:

1. Dissect the network traffic captures in PCAP files

2. Divide into flows/connections

3. Recognize endpoints and devices
4. Compute message statistics
5. Extract physical dynamics
6. Identify ordinary and abnormal events

Some security researchers may pose the question, why is the network characterization necessary to defend a SCADA network in industrial control systems? From the view of the adversary, he/she needs to understand what is the most effective strategy to disturb and alter the system state. As a result, the more system specification he/she obtains, such as the control and monitor routines or the mapping of the field devices, the higher attack power he/she is in charge of. From the perspective of the defense team, it's more ensuring to investigate the network connections and master what are the most vulnerable access points in the loop, what could be the potential attack path and where could be the attacker's most favorable target for effective attacks. This includes and not limited to the regular operational routines and the possible system upgrade and configuration events.

3.3 Threat Model

Inspired by the Industroyer malware, it is possible for the adversary to compromise the computer system in the SCADA network, and to inject either control commands or passive sensor measurements into the communication channels between RTUs and SCADA servers. Therefore, our threat model assumes the adversary has access to send malicious control command and to plant fabricated measurement into the point variables collecting sensor readings, as shown in Figure 3.2. He/She also may have in-depth knowledge of the SCADA network and the physical functions of control systems components, which grants his/her capability of effectively disrupt the physical process.

3.4 Research Questions

With the above proposed research plan, we intend to answer the following research questions.

For the sections of network characterization, we will answer:

1. Will all the TCP connections stay alive for a long time once established?
2. Are real-world SCADA networks following the consensus of being stable and predictable over time?
3. What types of operational connections we can learn about the network operation?
4. What new physical behaviors we can extract from deep-packet inspection?

For the sections of anomaly detection, we will answer:

- Can we learn more about the process anomalies from the measurement insights? How different are the process anomalies in different control systems?
- How can we utilize the system insights of the physical processes, to strengthen the robustness of the anomaly detection design, i.e. to lower the false detection rate?

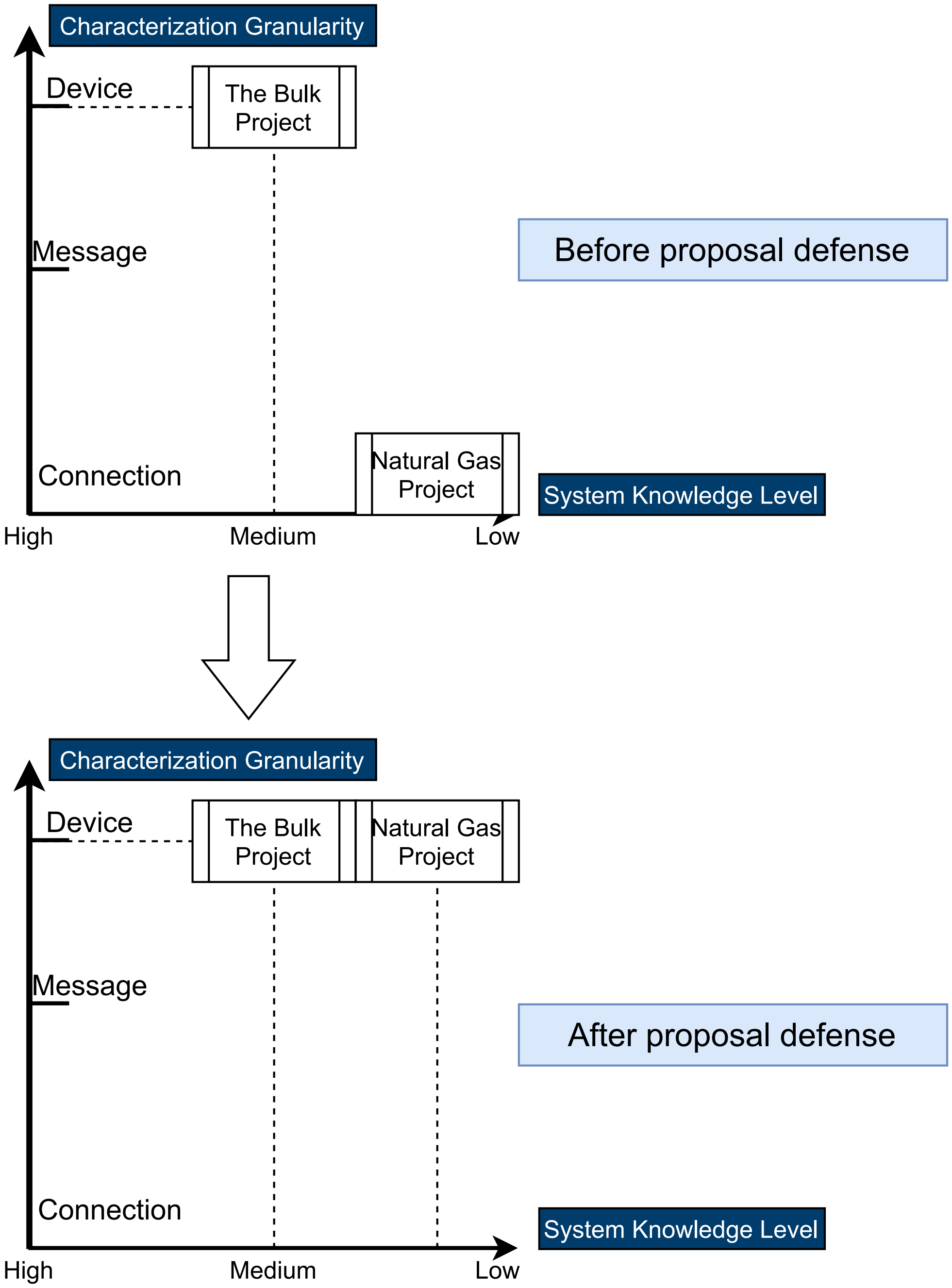


Figure 3.1: The overview of the work in this dissertation, with the research scope and the methodology as two dimensions

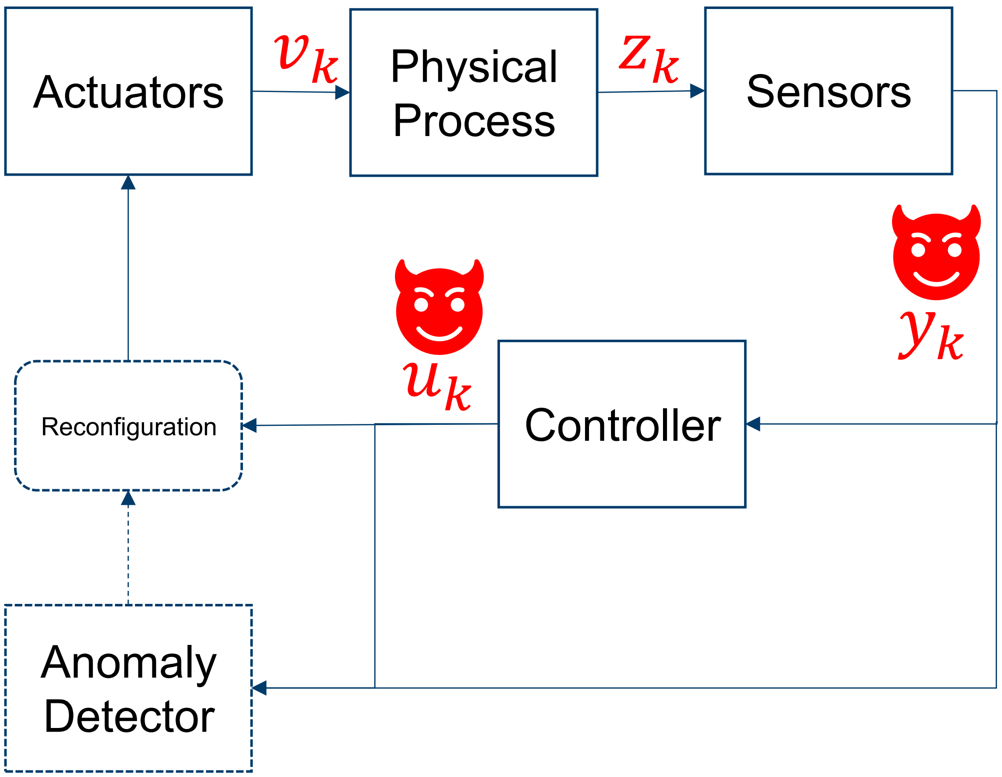


Figure 3.2: Attack locations in the control loop of a typical ICS

Chapter 4

Characterization of the Bulk Power Grid SCADA Network

4.1 Description of the Network Capture Dataset

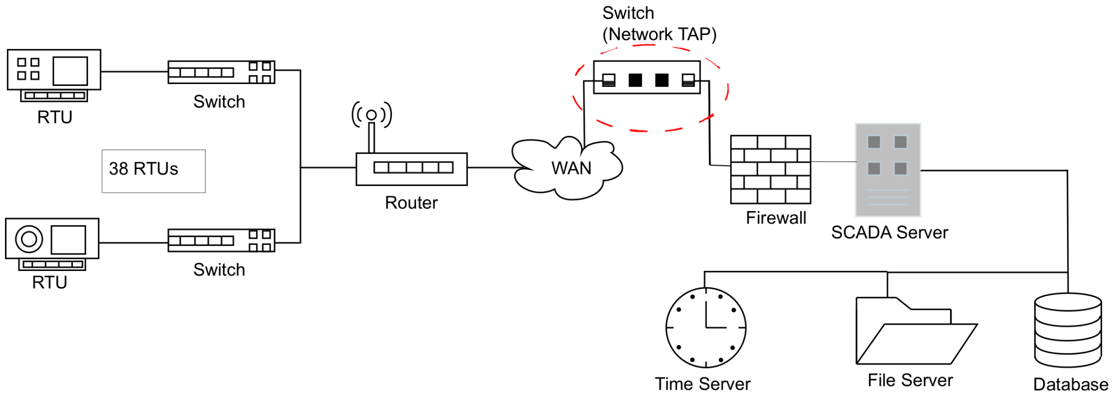


Figure 4.1: Abstract visualization of the SCADA network capture location, noted in the dashed oval

Figure 4.1 gives a rough location where the operators sniff the network traffic and obtain the capture dataset. The capture location is between the firewall and the switch routing out to different substations. In this bulk power grid, each substation can have one or multiple RTUs and each RTU can be identified by its static IP address. We omit the drawing of local field devices that each RTU is in charge of since it's irrelevant information here. But one should assume that within each substation, each RTU controls tens to hundreds of field devices. The substations communicate with SCADA servers within the private network under IEC 104 for control and monitor purposes. The operators performed the sniffing in discontinuous days of two consecutive years, as shown in Table 4.1. As a result, we have the benefits to observe the network's invariant and transitions over two years. In the rest of the dissertation, we refer to the capture obtained in the first year as "Y1" and the one in the second year as "Y2". Datasets 1 to 5 are from "Y1". Datasets 6 to 8 are from "Y2".

Table 4.1: IEC 104 Traffic Information of All the Datasets

Dataset	Number of Packets	Proportion	Packet/sec	Duration (H:M:S)
1	495,950	32.3%	76	01:48:18
2	592,055	31.7%	74	02:12:55
3	883,209	32.0%	75	03:14:06
4	517,908	32.4%	75	01:54:15
5	29,554	32.3%	74	00:06:36
6	346,712	4.55%	96	01:00:00
7	278,867	3.68%	77	00:59:59
8	837,871	3.98%	85	02:44:06

4.2 SCADA Network Topology Reversal

We can see that the control room of the system operator has 4 control servers: C1, C2, C3, and C4. We also observed a total of 27 substations (identified in the Figure as S1-S27).

Most substations are next to a power generator (identified as ovals) and some substations only deal with transmission equipment (identified as semi-circles). This makes sense as the role of IEC 104 for this particular operator, is to monitor and control generators (via AGC). The few substations that do not have generators provide auxiliary network measurements of the bulk power system. Each substation has one or more RTU, and because RTUs are called Outstations in the IEC 104 standard, we identify them in the figure as O1-O58. We can see that each pair of servers (C1/C2 and C3/C4) maintains a primary and a secondary connection to each outstation (as expected by Fig. **??**). Finally, each RTU collects measurements from a variety of field devices, from sensors in generators, to circuit breaker information, frequency sensors, etc. These devices are enumerated in the "cloud" attached to each Outstation (RTU).

In order to test our first hypothesis, we first look at the changes of the network over a year. Fig. **??** illustrates several changes from Y1 to Y2. We can see in red the substations and outstations removed from Y1, and in green, the new outstations that had been added to Y2. The arrows associated with each "cloud" indicate changes in field device measurements that we observed between Y1 and Y2. An upward arrow indicates that we observed more IOAs in Y2 than in Y1, and an downward arrow indicates we observed less IOAs in Y2 (the number of IOAs observed in Y1 are in red and the number of IOAs seen in Y2 are in green).

We asked the bulk system operator about these changes and their answers are summarized in Table 4.2. There are four different reasons for having new outstations in Y2. The first reason is that there are new substations that came online in Y2. In particular O50, which is associated with substation S24, and O53, associated with substation S27. The power grid operator told us that adding new substations over the years is not uncommon, and in fact this trend is accelerating with the addition of renewable energy. The second reason for additions is that substations with serial links (IEC 101) were updated to TCP/IP networking with IEC 104; these correspond to O52/S23 and O55/S26. The third addition occurred because O54/S25 was undergoing maintenance during the first year of the capture, and that is why we did not see it in Y1. The final reason for additions is a simple one, many substations have backup outstations that can talk to the control servers. In the first year we captured a different set of outstations communicating with the servers, but in the second year we captured their alternate outstation; these include O51, O56, O57, and O58, and similarly some of the removed outstations like O28 were replaced by these redundant RTUs, while others such as O15 have a backup outstation (O9 in this case) which still represents the substation to the control servers. Perhaps the most surprising finding was the removal of O2/S2; the operator told us that this substation had lost their connection and therefore was not monitored by the system operator, but this does not mean the substations was completely unsupervised, as it still presumably has the main connection to the SCADA server of the power company managing the substation. Another reason S2 was not essential for the operator is because it is not a generation substation (i.e., it does not have a generator that can be controlled by their system) and therefore it is one of the auxiliary substations that send data complementing their view of the grid, but the missing data from S2 did not represent a critical component for the operation of the AGC algorithm.

Overall, we see that 7 substations out of 27 (26%), and more precisely, 14 outstations out of 58 (25%) remained connected and reporting the same number of IOAs in a year. So the answer of whether Hypothesis 1 is validated in this network is not clear; on one side, most of the network changed between two years; however, we can see that the server configuration remains stable, and over 1 out of 4 of the devices in the field remains stable.

Table 4.2: Outstations added or removed between Y1 and Y2.

Outstation	**Added/Remove**	**Description**
O50, O53	Added	New substations
O52, O55	Added	Updated from 101 to 104
O51, O56, O57, O58	Added	Backup RTU
O54	Added	Under Maintenance in year 1
O15, O20, O22, O28, O33, O38	Removed	Redundant RTU in operation
O2	Removed	Substation without supervision

4.3 Connection Analysis

We confirm our theory that the short-lived flows exist for the purpose of secondary connections. The followings are the processing procedures, also depicted in Figure **??**:

1. Filter out retransmission packets; retransmission packets also match the SYN-RST/SYN-FIN pairs, but they are not the kind of short-lived flows we target.

2. Only keep TCP flows with either source port or destination port as 2404 (port for TCP flows transmitting IEC 104 data).

3. Divide flows into short-lived flows with matching SYN-FIN/SYN-RST pairs and long-lived flows.

Consequently, we've answered the research question one that the TCP connections don't retain the stable connections, and there is a significant amount of short-lived connections lasting less than one second.

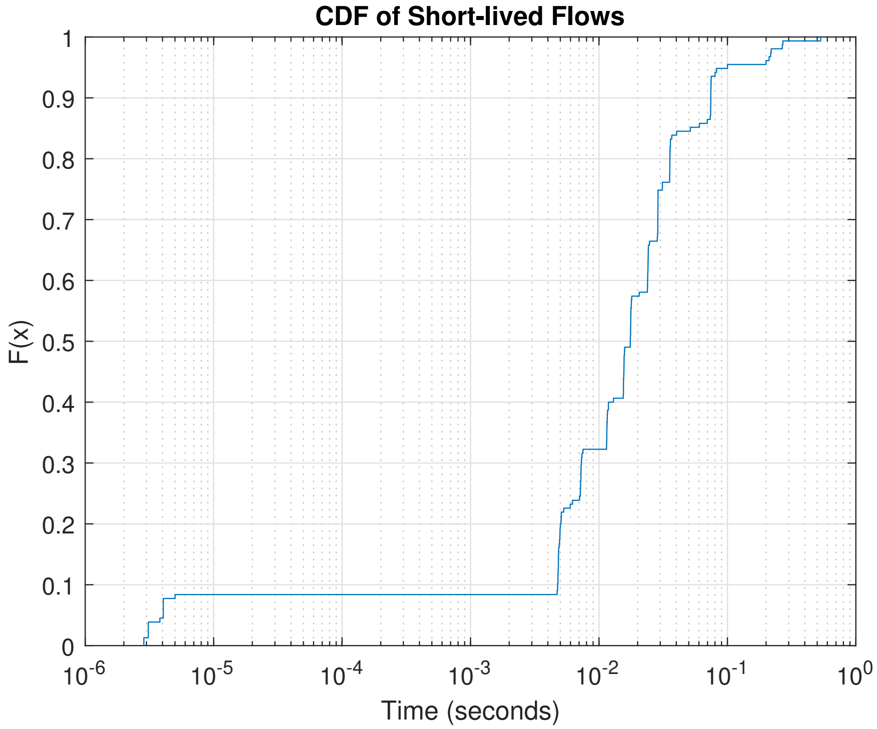


Figure 4.2: Observation of the time duration of TCP short-lived flows

4.4 Distribution of Control and Monitor Message Types

In an operational network where there's only one single administrator, the network might configure the majority measurements under the same data type with the same communication message type. However in this network there are multiple administrators for different RTUs. They may choose message types following their preference, such as the two message types for floating-point values with or without time tag, respectively.

From Table 4.4, we observe that IEC 104 traffic in this bulk power grid is heavily measurement-centric with 99.75% is for monitored measurement. If the reader wants to learn the rest I-Format message types, please refer to Table .1 in

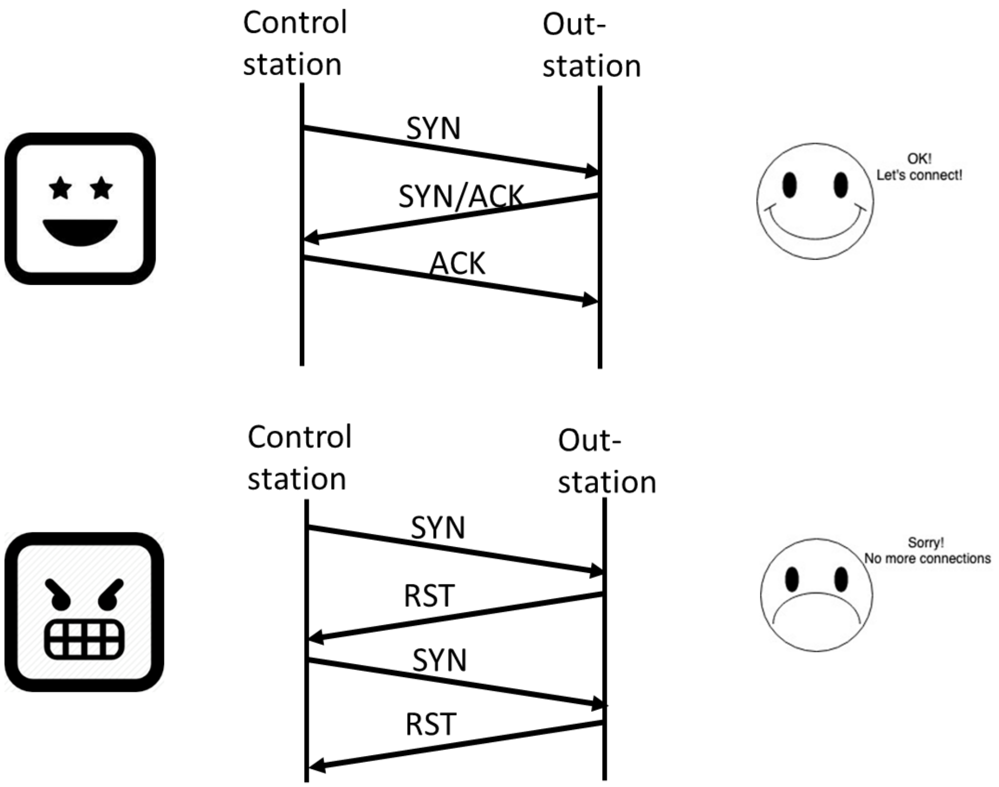


Figure 4.3: The cause of TCP short-lived flows: RESET in the TCP hand-shake

Table 4.3: Comparison of the number of TCP short-lived flows and long-lived flows in two years

Year	Y1	Y2
Count of Less-than-one-second Short-lived Flows(proportion)	31614(99.8%)	7937(93.5%)
Count of Longer-than-one-second Short-lived Flows(proportion)	63(0.2%)	549(6.5%)
Count of Short-lived Flows (proportion)	31677 (74.4%)	8486 (93.8%)
Count of Long-lived Flows (proportion)	10898 (25.6%)	560 (6.2%)

the appendix for details.

Table 4.4: I-format IEC 104 Message TypeIDs and Proportions

ASDU ID	Meaning	Percentage
I_{36}	Measured value, short floating point number with time tag	65.13%
I_{13}	Measured value, short floating point number	31.70%
I_9	normalized measurement	2.70%
I_{50}	Set-point command for short floating point number	0.23%
I_3	Double-point information	0.14%
I_5	Step position information	0.14%
I_{100}	Interrogation command	$< 0.01\%$
I_{103}	Clock synchronization	$< 0.01\%$
I_{30}	Single-point information with time tag	$< 0.001\%$
I_{70}	End of initialization	$< 0.001\%$
I_{31}	Double-point information with time tag	$< 0.001\%$
I_1	Single-point information	$< 0.001\%$
I_7	Bitstrings	$< 0.0001\%$

4.5 Community Discovery of IEC 104 Connections

After understanding the message types, we want to use clustering algorithms to identify the community groups inside of all the IEC 104 connections, for a bigger picture of the communication patterns and profiles that can give us insights into the operation of the network.

We define as a *session*, all the packets that are sent in one direction between the same end points. Originally we considered in total of 10 statistical features to investigate, including the transmission direction (is the message coming from the control center or from the outstations?), average inter-arrival times, total bytes, total number of packets, and even some features that looked into the APDU information such as the count of IOAs or the distribution of APDUs by type (U/S/I). Using the Silhouette score for each individual feature[67], we pick the features that generate relatively high Silhouette scores and reduce the feature space dimensionality from ten to the following five features:

- Δt_i: average inter-arrival time between two consecutive packets
- num_i: total number of packets sent in the same direction by two end points.
- $percentage_I$: the percentage of I-format data units.
- $percentage_S$: the percentage of S-format data units.
- $percentage_U$: the percentage of U-format data units.

We use K-means++ clustering [47] on these features. To select the number of clusters K we use the Elbow method on the sum of squared error [74], the explained variance[35], and Silhouette scores[67]. These methods suggest that a good number of clusters is K=5. In addition, we use Principle Component Analysis (PCA) [32] to project and visualize our results to a the lower dimension (2D plane). Our clustering results can be seen in Fig. 4.4. while retain the original variance in the dataset as much as possible, we use principle component analysis (PCA)[32], which results in Fig. 4.4.

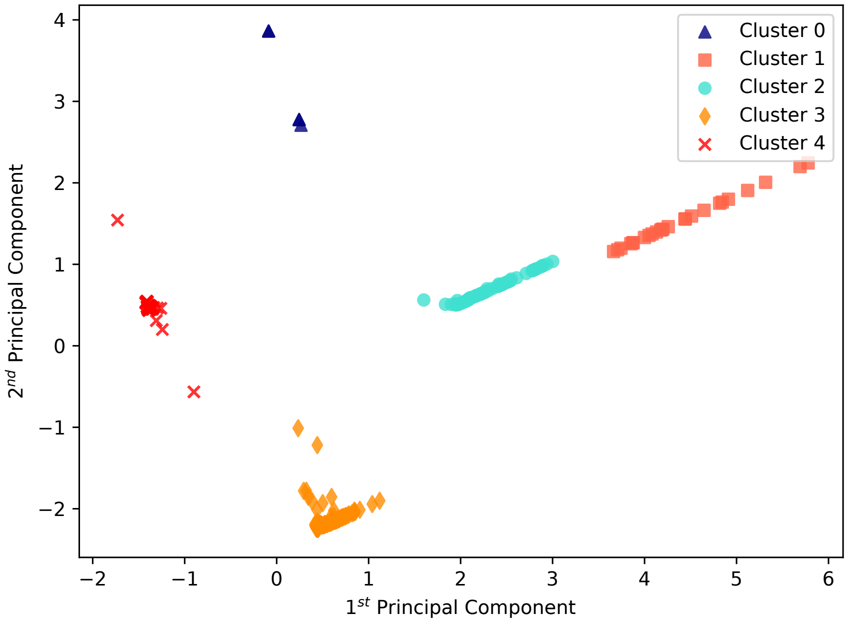


Figure 4.4: PCA of clustered IEC 104 sessions in Year 1

Inspecting the characteristics of each cluster, we find the following five representative behaviors: (1) Cluster 0 represents (extremely) long inter-arrival arrival times between packets; (2) Cluster 1 contains the largest amount of I-format packets, characterized also by being spontaneous transmissions (as opposed to periodic), (3) Cluster 2 represents the "average" case representing most outstations sending a regular amount of I-format packets, (4) Cluster 3 captures all the acknowledgements (S-format packets) sent from control servers to outstations, and (5) Cluster 4 represents the keep alive messages of the backup IEC 104 connection. Figure 4.5 summarizes these clusters and their percentages.

For the outlier group in cluster 0, we identify they are secondary connections and their communication pattern as in Figure 4.7 and compare it with all the other secondary connections. While the RTU might reject TCP connection attempts for backup IEC 104 channels, when the main connection is teared down, they readily accept the backup connection to the other control server to send I messages.

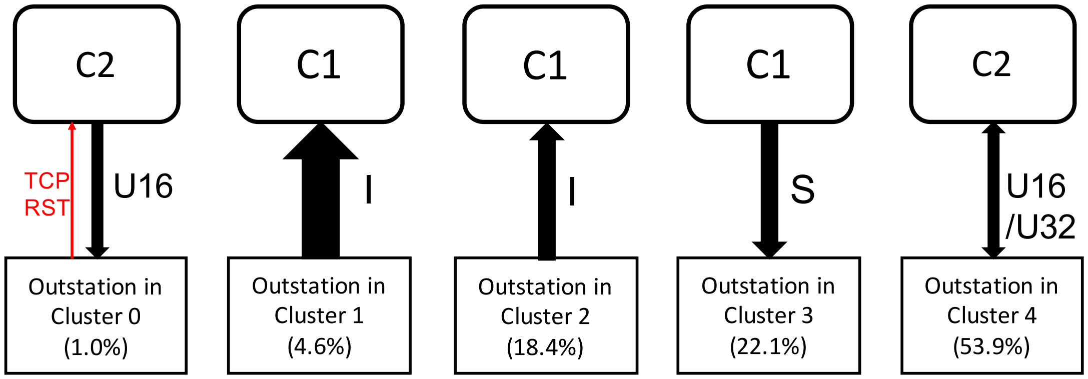
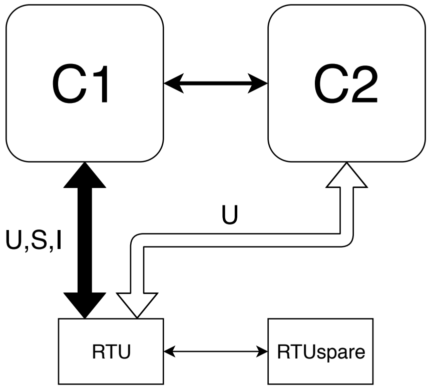

C2
C1
C1
C1
C2
TCP RST
U16
I
I
S
U16 /U32
Outstation in Cluster 0 (1.0%)
Outstation in Cluster 1 (4.6%)
Outstation in Cluster 2 (18.4%)
Outstation in Cluster 3 (22.1%)
Outstation in Cluster 4 (53.9%)

Figure 4.5: Communication patterns between outstations and control stations in each cluster



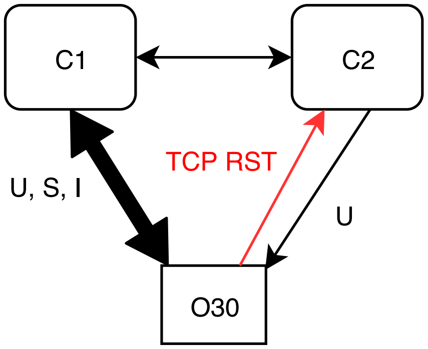


Figure 4.6: One standard communication pattern of the primary and secondary IEC 104 connections

Figure 4.7: Outlier communication pattern in clustering analysis.

In these outlier connections, control server sends the keep-alive message every 430 seconds and then the TCP connection is torn down; whereas the default universal timer for this type of packets is only 30 seconds. Figure 4.8 presents a standard testing procedure with keep-alive messages, where a station should reply a confirmation message to a request of testing activation.

The root cause is that, the central operator does not own or manage the configuration of the RTUs/outstations. It does not have regulatory power to demand these changes if the local operators don't follow a specific configuration recommendation. As long as the application-level behavior is satisfied, and all entities are in compliance with the general reliability standards, the central operator mostly won't proceed to more intervention to force the local operators to alter anything, even when the network behavior underneath has problems as such.

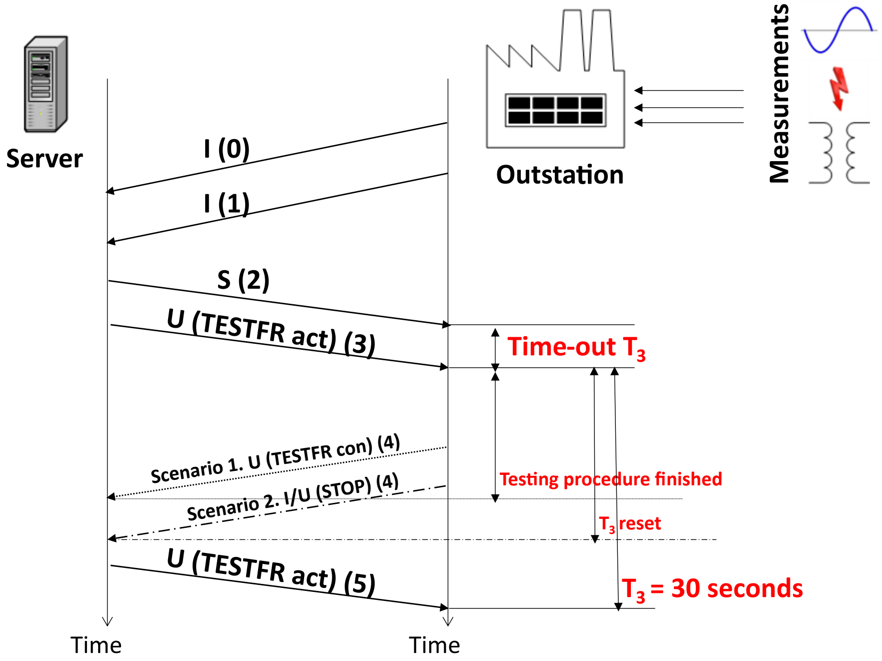


Figure 4.8: The standard setting for timer T_3 followed by the majority of RTUs

Cluster 1 contains connections with a large number of packets in I-format. Sessions in this cluster all have the spontaneous type of transmission. Spontaneous transmission type is one of the reasons behind the heavy traffic. The field device measurement values are volatile. For instance, a small change, 0.01 mW in power variable, stimulates the RTU to report an IEC 104 data unit. While in cluster light-I-message, around 35% of the sessions have other types of transmission, such as periodic type and data transmission activation type. The other reason is a critical control command in IEC 104, interrogation. Each time the control server interrogates an outstation, and the outstation must report the status of all the field devices. This command acts like a through activation and examination, which even extracts the status of in-sleep devices, thus causes a burst in traffic density.

Cluster 2 represents the "average” case where connections have a reasonable enough bandwidth when transmitting I-type messages.

Cluster 3 accumulates all sessions in the control direction from control servers to outstations. In these sessions, the control servers response to the I-format field devices measurement values from the outstations in S-format messages. Servers use one S message to acknowledge the receiving of every eight I-format measurements. For such a federated power grid with multiple servers, the formation of this cluster can contribute to distinguishing between the main and backup server for each outstation before any in-depth analysis. Because theoretically, the outstations only report I-format messages to the central server.

Following cluster 3, cluster 4 aggregates all sessions exchanging U-format control information. There are mainly three subtypes in U-format messages, the start, and stop of data transmission and the testing function for keep-alive messages. It’s very tricky for the operators to precisely capture the beginning and end of a data transfer. So we only have less than 6% of U-format messages are the start and stop signals, and all the rest are testing messages. When investigating both cluster 3 and cluster 4 in parallel, we can clarify the primary and secondary connections for each outstation.

With this section, we've answered the research question 3, where we can use unsupervised learning algorithm to learn the communication patterns of all connections at once.

4.6 Physical Dynamics Study

Our first finding was in the analysis of power, as we can see power fluctuations in Figure 4.9. This was a use case caused by a load connection failure, it means that there was a lost electric load at a certain time point, the frequency of the power grid will decrease because the electric generation is more than electric load. The system operator need to make generators reduce their production of electricity in order to stabilize the system balance via AGC messages. Once the load is reconnected to the power grid, the operators ramp the generation up again. These sequence of AGC commands and their effect can be seen in Figure 4.9. The $y_{1,2,3}$ signals are the sensor readings for power measurements at the generators. The u signal is the control signal for the control center taking charge of the power generator. This plot shows that the actuation process is quite responsive within seconds, when the generators adjust their power generation at the moment of receiving the control command.

We also examine the measured voltage values in the same RTU. One voltage time series jumps from zero kV to about 120 kV together with a circuit breaker status value from zero to two before the above AGC event happens. It means the operator tries to connect a generator to the power line, and the breaker closes to make this connection. The discovery of physical dynamics in such process variables can help collect the signatures for monitoring the local status among

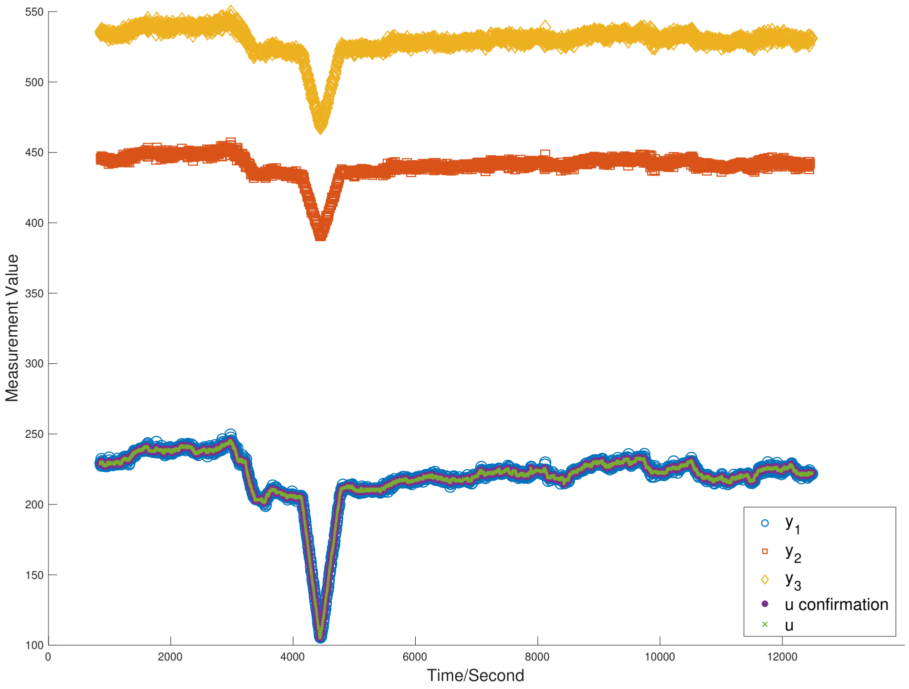


Figure 4.9: The bottom time-series is the AGC control commands, and the top two series show how generators react to the control actions.

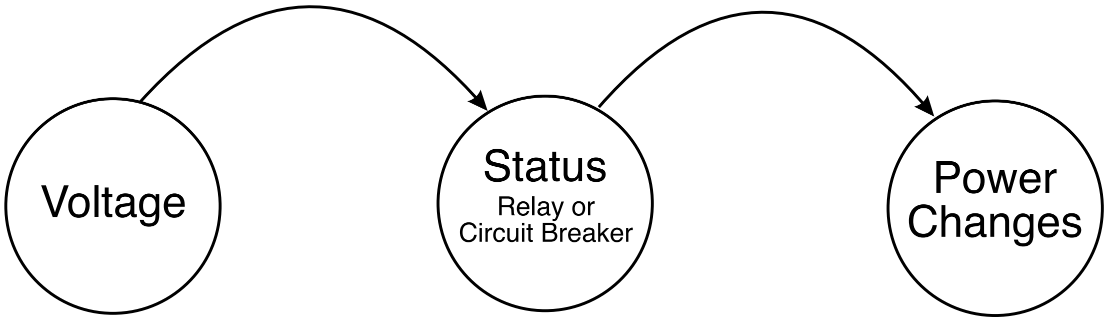


Figure 4.10: Signature creation of power system behavior:

these correlated variables within a remote station, as in Figure 4.10.

4.7 Measurement Transmission Causes

We now discuss the cause of transmission utilization rate per ASDU typeID and their timing characteristics.

According to Table 4.5, spontaneous and Periodic/Cyclic (Per/Cyl) COTs have the highest utilization rates. Based on the measurement's variation, the transmission can be handled by triggers as follows:

- **Periodically:** the transmission is carried based on a fixed periodicity in seconds.

- **Spontaneously:** it means every time the measurement changes, a data transmission for that value is performed.

As their names suggest, transmission of ASDU with spontaneous COT has no fixed time interval between any two consecutive ASDUs, whereas Per/Cyl ASDUs should have a fixed intervals, i.e., preset cycle. Therefore, one can expect that spontaneous can have wide range of time intervals while Per/Cyl have narrow range. If a measurement changes too much in time, spontaneous triggers can try to deliver too much data leading to big payloads and it is not possible in some

Table 4.5: Cause of Transmission utilization by typeID

	Cause of Transmission (COT)								
	Act	ActCon	ActTer	BackScan	Init	Inro	Cyc	RetC	Spont
I_{36}	0	0	0	4	0	0	0	0	1.58M
I_{13}	0	0	0	0	0	129	256K	0	513K
I_{9}	0	0	0	0	0	0	65K	0	0
I_{50}	2.8K	2.8K	0	0	0	0	0	0	0
I_{3}	0	0	0	0	0	134	3.3K	0	0
I_{5}	0	0	0	0	0	0	2	0	2.2K
I_{100}	66	64	63	0	0	0	0	0	0
I_{103}	0	0	0	0	0	0	0	0	26
I_{30}	0	0	0	0	0	0	0	0	12
I_{70}	0	0	0	0	12	0	0	0	0
I_{31}	0	0	0	0	0	0	0	2	9
I_{1}	0	0	0	0	0	8	0	0	1
I_{7}	0	0	0	0	0	0	0	0	1

Legends: M=Million; K=Thousand; Act=Activation; ActCon=Activation Confirmation; ActTer=Activation Termination; BackScan=Background Scan; Init=Initialization; Inro=Interrogated General; Cyc=Periodic/Cyclic; RetC=Returned by Local Command; Spont=Spontaneous scenarios when the communication is jeopardized. To avoid that, spontaneous data is triggered by data changes above a predefined threshold. The threshold can be fixed or dynamically calculated as percentage of its variations related to a reference value. In the other hand, big thresholds lead to poor quality of supervision since some important changes of variable's could be lost and there are only data transmissions when big amount of changes are produced. RTU configuration plays an important role here since balance between big bandwidth usage and reasonable reporting of data should be considered.

A plot of consecutive value changes on an IOA can disclose the possible value configured for that IOA, so there is shown in Figure 4.11 that a minimum value is held among time, indicating by a blank space between the bottom values of the data samples and the X-axis. We think this threshold value is potential in developing and defending stealthy false data injection attacks, where the volatile variations in the sensor values can be a noise signal.

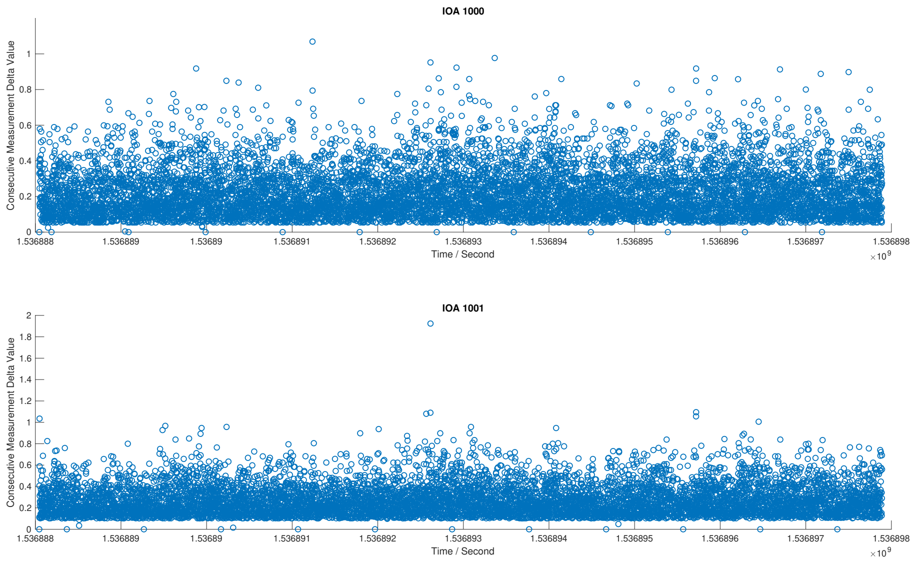


Figure 4.11: Different threshold values for measurement values inside two information objects in a RTU

The last two sections enable us to answer the research question 4. By performing deep-packet inspection of the values in the sensor readings, we can spot useful invariant and identify anomalies in the physical world. While these same anomalies can be spotted in the SCADA server, the rise of attacks to control systems like Industroyer [8], will motivate the identification of not only network-based anomalies, but also physical-based anomalies and their correlations. This type of anomalies essentially motivate this dissertation of the process/physics-based anomaly detection system.

4.8 Discovery of Protocol Noncompliance

At the beginning of our deep-packet inspection, we extract the process variables, i.e. IOs inside of data units of IEC 104 packets. When we compute the distribution of IOAs, we observe that the majority of IOAs are large numbers as shown in Figure 4.12. These large numbers are invalid addresses for IOs, because IOs have at most three octets for the address. Moreover, the payload content of these suspicious data unit are incomplete with none information inside of IOs, or the values simply appear to be random numbers not reflecting any real-world physical variables. The packets cut off right at the IO frames.

Checking with Wireshark's dissector, we find out these packets are malformed. Wireshark official appendix has the following description about malformed packets: Malformed packet means that the protocol dissector can't dissect the contents of the packet any further. There can be various reasons:

- Wrong dissector: Wireshark erroneously has chosen the wrong protocol dissector for this packet. This will happen e.g. if you are using a protocol not on its well known TCP or UDP port. You may try Analyze|Decode As to circumvent this problem.

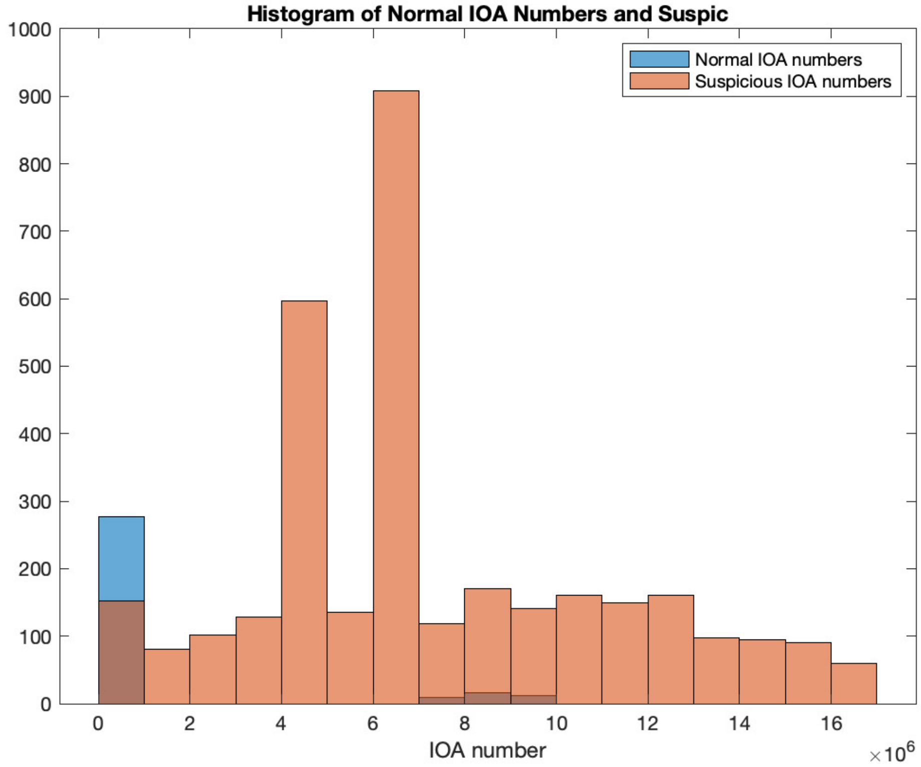


Figure 4.12: Irregular IOA ranges (extremely large) identifies malformed packets

- Packet not reassembled: The packet is longer than a single frame and it is not reassembled, see Section 7.8, “Packet Reassembly” for further details.

- Packet is malformed: The packet is actually wrong (malformed), meaning that a part of the packet is just not as expected (not following the protocol specifications).

- Dissector is buggy: The corresponding protocol dissector is simply buggy or still incomplete. In our dataset, the malformed issue reported by Wireshark matches the third reason: "Packet is malformed". The malformed field comes as the last field in an APDU:

In our dataset, the malformed information is like this:

```
"_ws.malformed": {
         "_ws.expert": {
           "_ws.malformed.expert": "",
           "_ws.expert.message": "Malformed Packet (Exception occurred)",
           "_ws.expert.severity": "8388608",
           "_ws.expert.group": "117440512"
         },
         "_ws.malformed": "Malformed Packet"
       },
```

To investigate the cause, we trace back to the network flows and the RTUs generating these packets. Precisely, outstations O37, O53, O58, and O28 had 100% invalid packets in all our traces. Eventually we identify two reasons why the Wireshark IEC 104 dissector cannot properly interpret these packets. First, outstation O37 used an IOA length of just two octets (instead of the standard three octets length for an IOA address). The second set of malformed packets came from outstations O53, O58, and O28, which used just one octet for the “cause of transmission” field, while the IEC 104 standard specifies that the cause of transmission field should be two octets. Validated by the system operator, these two scenarios happened because these stations went through a network upgrade from serial communication in IEC 101 to the TCP/IP network in IEC 104. For these specific stations, the upgrade was not configured correctly. Consequently the vendors of these stations are exchanging 104 packets in the 101’s legacy format.

Chapter 5

Characterization of the Gas SCADA Network

5.1 Description of the Network Capture

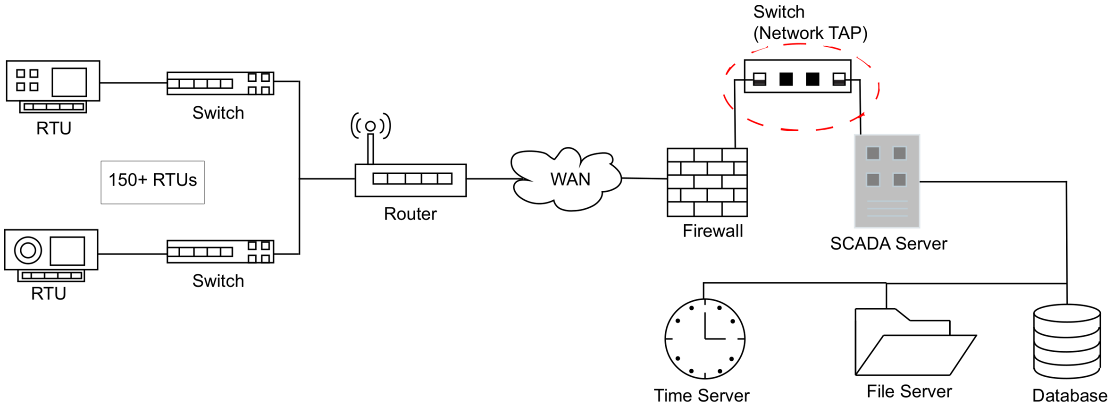


Figure 5.1: Abstract visualization of the SCADA network capture location, noted in the dashed oval

This capture of a natural gas distribution network occurred outside of the private network serving the operational network. It is between the firewall and the mirror port of the main network switch for the control center. As a result, the dataset includes traffic between devices and applications in the control center and the RTUs. The communication between the control room and the remote RTUs is through GPRS and certain leased lines.

The dataset covers 98 days in total, within which 79 days of capture being consecutive and seven days missing from the capture in the last month. The possible cause of the missing days potentially came from a prompt glitch in the switch's port mirroring process. As a result, the actual overall capture time duration is roughly 85 days. This giant dataset has 500 GB of raw PCAP files, consisting of 2,222,253,073 network packets. It is the largest industrial control system's dataset ever studied according to our investigation of the previous work. The projects with comparable sizes (around 200-300 GB) of the network captures have measured individual substations in a power grid distribution network [29][31][45].

There are in total of 397 IP addresses observed in this network. By inspecting the application-layer traffic, we identify the biggest group of 304 hosts are RTUs, 82 hosts are implicit workstations in the control room, four are HMI stations, two are SCADA servers, two are time servers, one is a printer, and the rest two are network switches. The only application-layer protocols deployed for communication with RTUs are IEC 104 (around 118 millions of packets) and Telnet (around 72 thousand of packets). If keeping the view within the control room, we spot the typical IT protocols like DHCP, DNS, HSRP, NetBIOS, NTP, SLP, SMB, SNMP, SSDP, SSH, X11), and an unrecognized proprietary industrial protocol for intercommunication between the two SCADA servers or between the workstations and SCADA servers. While we can recognize all the network services except this proprietary industrial protocol, the payloads encapsulated in these packets are simply the routine synchronization between SCADA servers and HMI workstations. The network characterization and anomaly detection in the following contents focus on

Table 5.1: Enumeration of the network services

OSI Layer	Protocol	# packets	% packets
Transport-layer	TCP	2155570963	96.82%
	Unknown proprietary protocol	40420388	1.82%
	UDP, ICMP	XXXX	1.36%
total	-	2226220420	100.00%

the operational technology (OT), so we only analyze the traffic behaviors related to the RTUs, i.e. the traffic in IEC 104 and Telnet.

To present the connections and their bandwidths in a complete view, we visualize the packet count of each flow in this chord diagram 5.2, where the width of each band between the two endpoints represents the total amount of IEC 104 packets in that flow. This visualization is able to demonstrate three folds of flow dynamics: 1. the connection quantity of all IEC 104 endpoints 2. active flow or not between two endpoints 3. the bandwidths of all flows. By a glance at this figure, we can observe the majority of RTUs have a single-connection with one of the servers C_2. Very few RTUs on roughly 3'clock direction have dual-connections with both servers C_1 and C_2. Both C_1 and C_2 are the primary SCADA control servers. In the middle of the capture time, the operators switch all the RTUs from the server C_1 to C_2 under some unknown considerations. There are a few noticeable groups based on the rainbow bands' thickness. We'll present the reason why some connections have more conversations than the others in the following section, where we use clustering algorithm to group these connections into community and analyze the community formation in details.

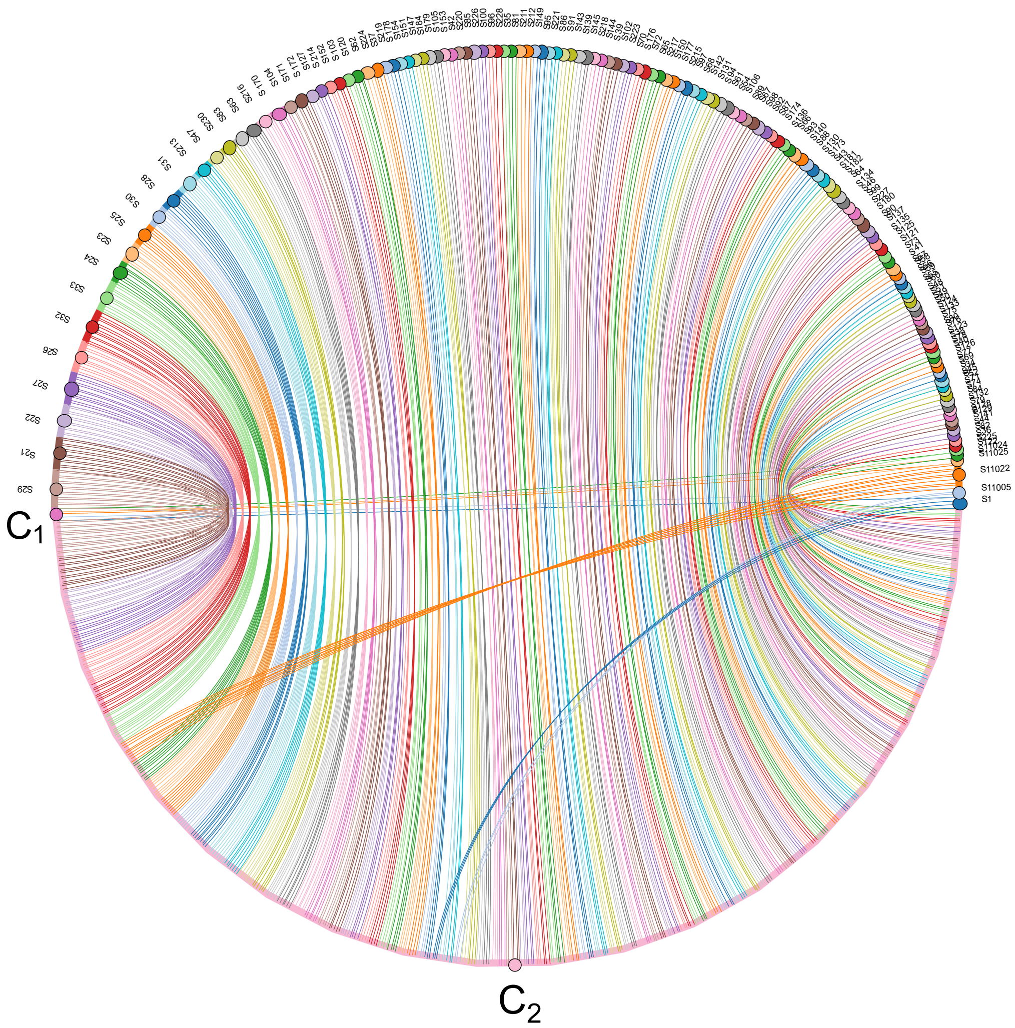


Figure 5.2: Complete chord diagram of all IEC 104 flows; channel bandwidths (the amount of packets) rank in descending order in the clockwise direction

5.2 Connection Analysis

We don't find a SCADA master backup server in the gas dataset based on the analysis so far. There's no short-lived TCP flows among IEC 104 endpoints triggered by secondary connections for the whole period of the gas dataset. But there are a few short-lived TCP flows caused by malformed packets. The reason for the malformed packets and the terminated TCP connections is under investigation. For the primary connections, the scenario is the same as the one in the bulk power grid. Once established, IEC 104 can maintain the stable and long-lived type of TCP connections.

5.3 Fingerprints of RTU by IEC 104 Features

From the start of endpoint recognition through IP addresses, we find 312 IPs potentially used for RTUs. However, the odd fact is that new IP addresses being discovered continuously for multiple days, as shown in Figure 5.3. This fact does not matches the consensus of SCADA network in the security community, otherwise IP addresses should all be discovered within minutes or even seconds.

Moreover, more than 90% of the initial set of IP addresses are different from the final set when the capture ends. Over the whole capture time, the number of active IPs transmitting IEC 104 messages changes, starting at 65 on the first day, ending on 153 on the last day, and having a peak of 217 on day 57. We find only 4 out of 65 RTUs that maintain the active status throughout the whole time span of the dataset (first day to last day). The other $153 - 4 = 149$ IP addresses are new; they don't exist when tracing back to day one and $65 - 4$ IPs were retired. We basically reach to an initial judgement that having 312 IP addresses does not directly imply the presence of 312 physical RTUs. Especially,

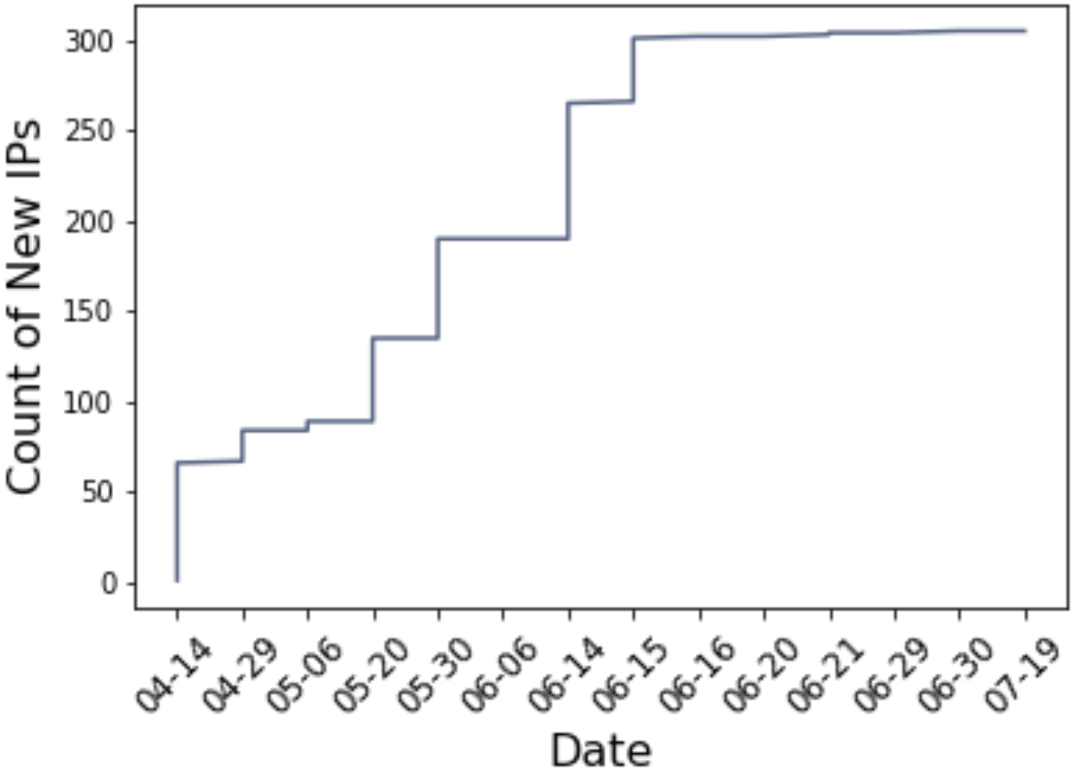


Figure 5.3: Count new IP addresses in the gas SCADA network

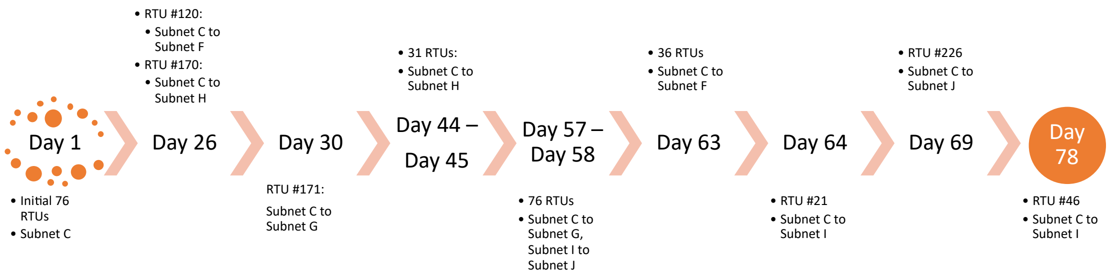


Figure 5.4: Timeline of IP reconfiguration activities in the IEC 104 network

we locate some RTUs reconfigured among different subnets through Telnet traffic. Therefore, we have a hypothesis that this network has been through a series of IP reconfiguration events. We extract the total line of reconfiguration in Figure 5.4.

To the best of our knowledge we are the first to define an algorithms capable of identifying the fingerprints of RTUs by their IEC 104 traffic, even when the IP address migration events change the static network configuration. efalg:rtu-identification-algorithm. With our expertise in the IEC 101/104 protocol specifications, we acknowledge that the protocol defines two of its primitives together have the representation of the physical stations and device variables in the real systems. If the operators follow the protocol's principle in this manner, then the combination of both common addresses and IOAs shall have a one-to-one mapping relationship with the physical system, projecting to a specific station and its process variables. Each I-format message consists of at least one ASDU message. Each ASDU message contains at least one information objects (IOs) holding the values for specific process variables. Each IO uses information object address (IOA) as the identity. In our dataset we observe that common addresses (CAs) are unique across all RTUs, allowing us to use them as RTU identifiers ($RTU = CA$). The algorithm outputs two dictionaries. $D_{ioa} : RTU \rightarrow \mathcal{P}(IOA)$ maps which IOAs belong to which RTU, and $D_{\Delta ip} : RTU \rightarrow \mathcal{P}$ records derived IP changes (from old to new ip, derived at given time). Given a method

Algorithm 1: RTU Identification and IP change detection Algorithm.

```
Input: Sequence of IEC 104 messages $msg_1 \ldots msg_M$
Init: $D_{ip} \leftarrow \{\}$ $D_{ioa} \leftarrow \{\}$ $D_{\Delta ip} \leftarrow \{\}$
For $i = 1$ to $M$
  $t_i, ip_i, \{(a_{i1}, o_{i1}), \ldots (a_{in}, o_{in})\} = parse(msg_i)$;
  $rtu = a_{i1}$ ;
  AddSet $\{o_{i1}, \ldots, o_{in}\}$ to $D_{ioa}[rtu]$;
  If $ip_i \neq D_{ip}[rtu]$ then
    AddElement $(D_{ip}[rtu], ip_i, t_i)$ to $D_{\Delta ip}[rtu]$;
    $D_{ip}[rtu] \leftarrow ip_i$
  EndIf
EndFor
```

$parse : IMsg \rightarrow TimeStamp, IP, \mathcal{P}(CA, IOA)$ that extracts the RTU's IP and associated (CA, IOA) pairs from an IEC 104 message. D_{ip} stores the last known IP of each RTU. Given an IEC 104 message, a_{ij} is the corresponding common address, and o_{ij} represents the information object address.

The algorithm uses $D_{ip} : RTU \rightarrow IP$ to store the last known IP of each RTU. With this algorithm, we are able to extract this series of IP reconfiguration events in Figure 5.4.

Finally we conclude there are 154 unique RTUs in this gas distribution network[1]. During the time period of this capture (98 days), 304 IP addresses were assigned to these RTUs. The subnets migrate from the initial two subnets C and I to four other subnets F, G, H, J. Based on the limited knowledge learned from the operator, this series of IP migration is part of the regular network maintenance. Therefore, we can learn a lesson from this observation that researchers cannot treat IP addresses as invariant in the SCADA networks. Researchers cannot depend on the set of IP addresses without further exploration of other network traffic to infer the static network layout. The examination of the endpoint IP dy-

[1] As defined in 1, we name the RTUs with their common address number, e.g. RTU #171 has common address as 171.

namics and the network services diversity can be beneficial to concluding explicit network topology. This section answers the research question two, that the operational SCADA network can contain a series of active configuration events last as long as 78 days. One can not simply assume the SCADA network doesn't evolve and retain the same over time any more.

As illustrated in Figure 5.5, at the beginning of our network capture, a subset of RTUs got connected to one control server. At the end of the capture, we can see more RTUs connected to a second control server. During our 100 days, we see a significant change in the SCADA network, and it is to be expected that such a significant change was scheduled during the summer, the period of low demand for gas (residential consumers use gas for heating) when potential outages have less severe consequences.

We now utilize the information of the network and look deeper into the IEC 104 payload.

5.4 Community Discovery of IEC 104 Connections

To understand how the supervisory control process of a natural gas distribution network works, we focus on classifications based on intuitive and straightforward features that look promising. A more extensive investigation of the many potential features and their usefulness in classification is left as future work. First, we look at the characteristics of individual RTUs on the transport layer; we use an agglomerating hierarchical clustering algorithm [78] with the pair-wise Euclidean distance as the similarity metric to cluster RTUs based on their total amount of TCP packets and the total size of TCP payloads. The result is in Figure 5.6 for

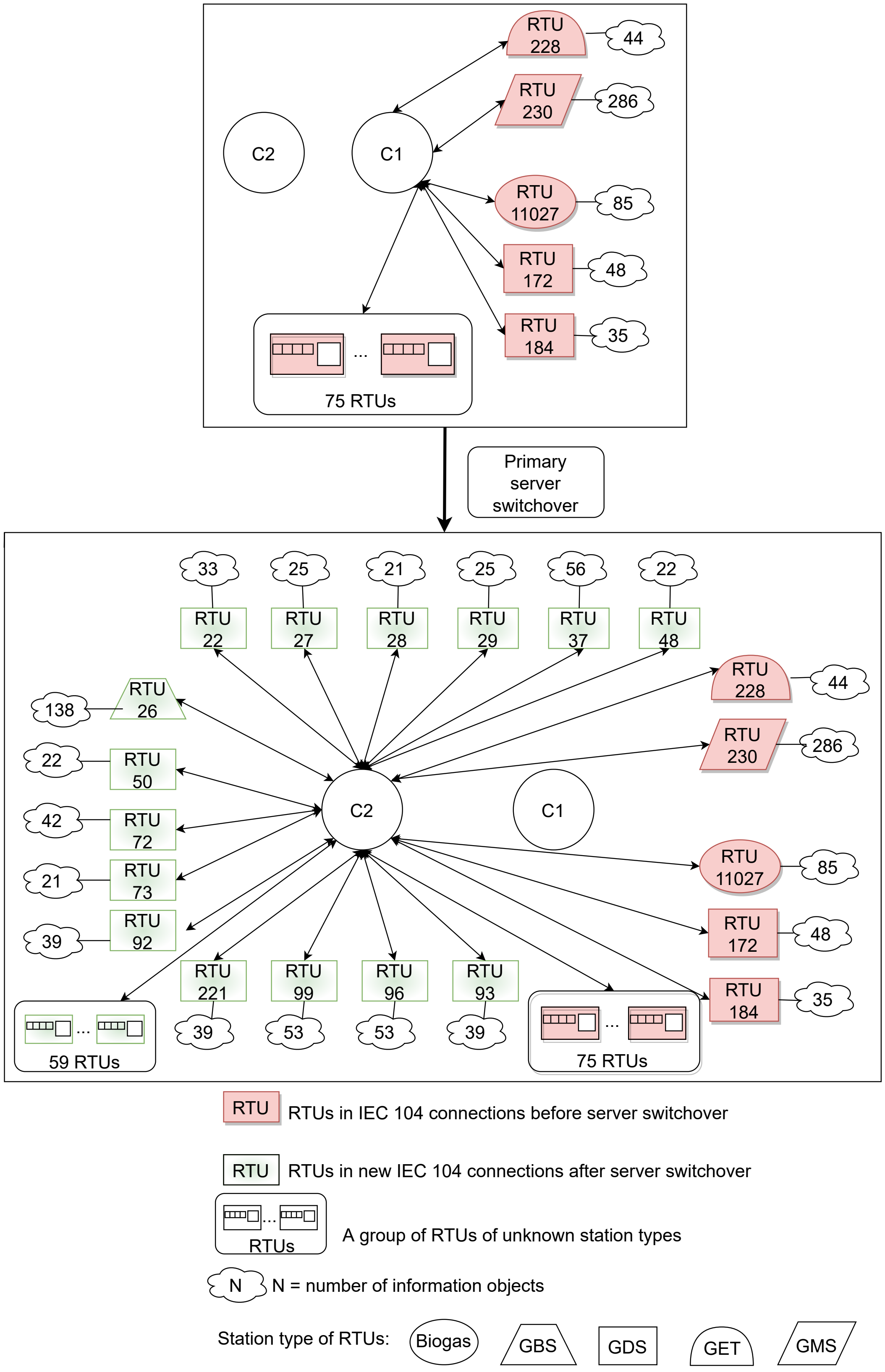


Figure 5.5: IEC 104 Network Topology before and after switching SCADA Servers

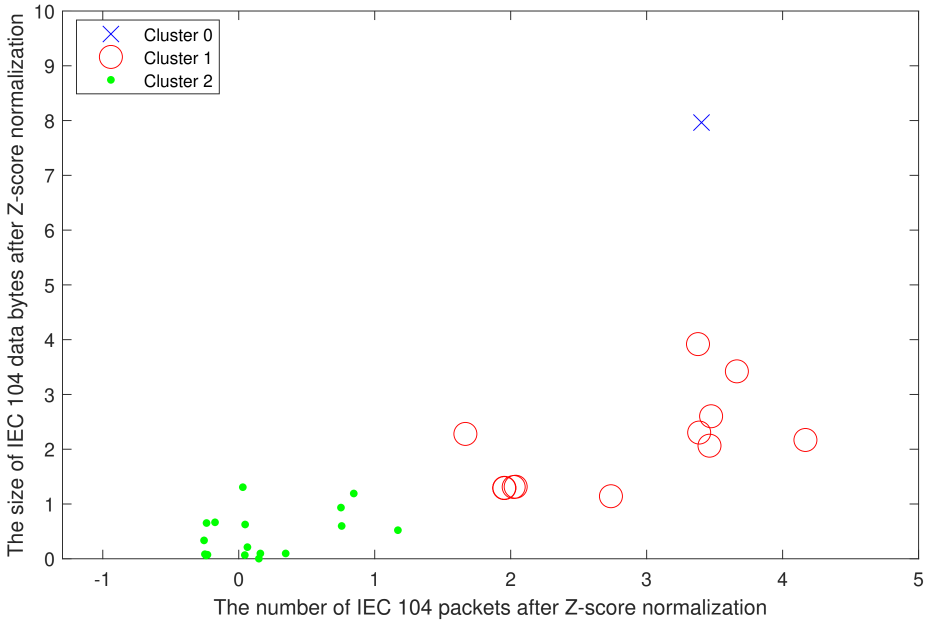


Figure 5.6: Clustering result of all IEC 104 flows

the total 158 flows.

First, we cluster RTUs based on flow statistics in the transport layer. Next, we explain clustering results by also looking at the application layer information. We further investigate the IEC 104 payload and classify RTUs by the type of commands they execute.

The flow traffic volume decreases from cluster 0 to cluster 2. Cluster 0 has only one IEC 104 flow between RTU 26 and control server C_2. This RTU is a clear outlier because it sends the largest TCP payload size (IEC 104 data) with over 8-million bytes in our dataset. In comparison, almost all other RTUs have flows with less than 2 million bytes (there are only 7 other RTUs with flows between 2 to 4 million bytes). The other two clusters have more regular behavior, with Cluster 1 aggregating 14 RTUs that are moderately more heavy-traffic and Cluster 2 with RTUs that send few and small packets.

We find that the main reason is because of a unique application of the IEC 104

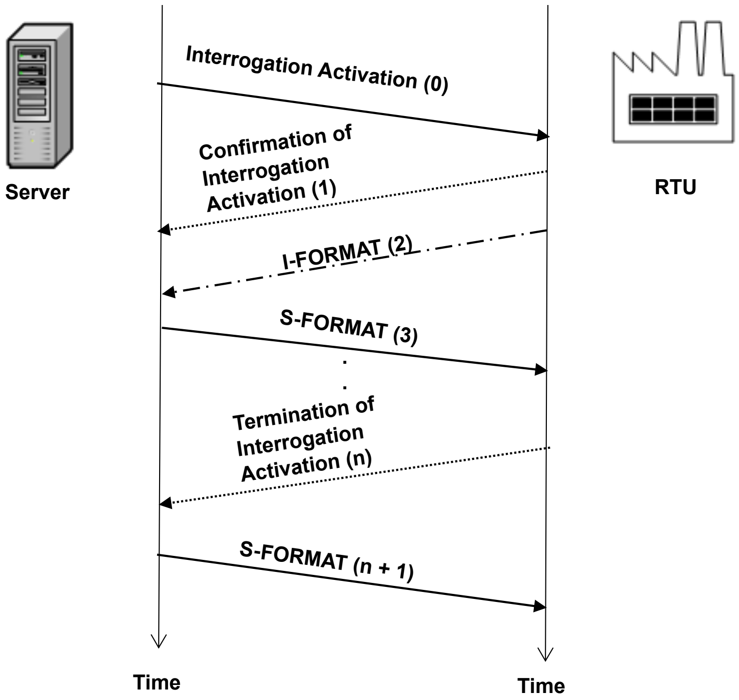

Server
RTU
Interrogation Activation (0)
Confirmation of Interrogation Activation (1)
I-FORMAT (2)
S-FORMAT (3)
Termination of Interrogation Activation (n)
S-FORMAT (n + 1)
Time
Time

Figure 5.7: The full process of the interrogation

interrogation command: when the control server sends an interrogation command to an RTU, the RTU has to report back immediately the status of all the process variables it has access to. In 5.7, we present a complete interrogation process between the control server and the RTU. After the RTU confirming launching the interrogation, it starts to collect all the current readings in the sensors and encapsulate these values into information objects in the IEC 104 packets (one or multiple packets, depending on the specific RTU's configuration). As one can imagine, this command triggers a burst in network traffic since all the process variables under one RTU send back their values at the same time. As shown in 5.8, RTUs in both Cluster 0 and Cluster 1 have interrogations every minute, while all the RTUs in Cluster 2 have interrogations every ten minutes.

Now to identify the flows differences between Clusters 0 and 1, we need to dig deeper into the payload of the IEC 104 flows. In particular, we find that

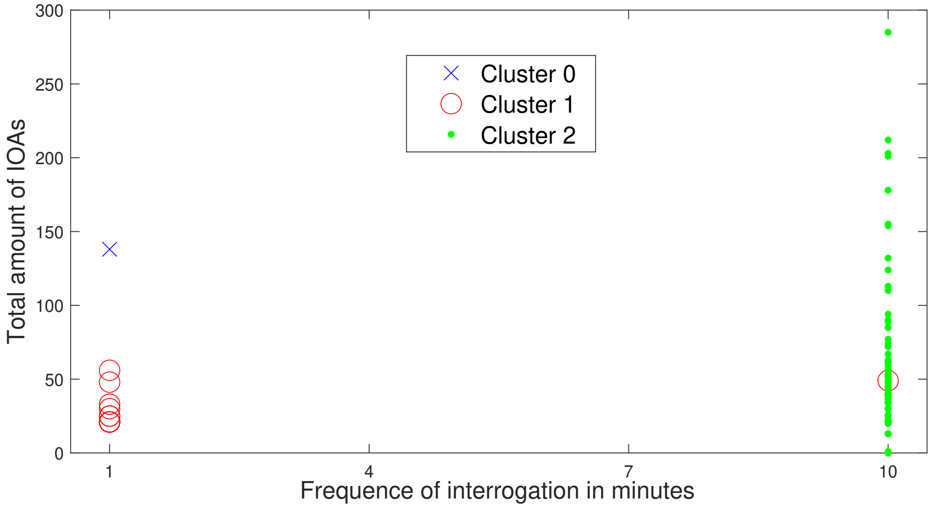


Figure 5.8: The frequency of interrogations differentiates the clusters.

RTU 26 (the outlier in Cluster 0) has 138 process variables (in contrast the highest number of variables in an RTU from Cluster 1 is 56; this can also be seen in the y-axis of 5.8). We find this particular use of interrogation commands in this gas dataset surprising, as our previous chapter reports that all process variables are automatically sent by RTUs to the control server, and the control server only sends interrogation commands when they establish a new connection.

From this observation of inner groups in the RTUs through analyzing statistical traffic performance and process variables, the lesson we learn is that the classification of IEC 104 connections are crucial to understand the roles of gas stations. While we can obtain partial ground truth through locating the automatic configuration sessions in other network service traffic, a majority of the RTUs still require our effort in the study of the operational network traffic. The intuitive features for clustering such as the network traffic volume are effective in categorizing the RTUs and isolating the outlier station. This section answers the research question three by locating the inter-communities in all the IEC 104 connections of this network.

After grouping, we will elevate the station-wise understanding into the device-wise recognition by a deep dive in the I-format types. We leave the extensive time series analysis of the process variables, i.e. sensor readings and actuator status, in the future work, which will also be the answer to research question four.

5.5 Distribution of Control and Monitor Message Types

Table 5.2: I-format IEC 104 Message TypeIDs and Proportions

ASDU ID	Meaning	Percentage
I_1	single-point information	*inrogen*, *req*, *retrem*
I_9	normalized measurement	*spont*, *inrogen*, *req*
I_{30}	single-point information with time tag	*spont*,*retloc*
I_{34}	normalized measurement with time tag	*spont*
I_{45}	single command to alternate single-point variables	*act*, *actcon*, *actterm*
I_{70}	end of initialization	*init*
I_{100}	interrogation command	*act*, *actcon*, *actterm*
I_{102}	read command	*req*
I_{103}	clock synchronization	*act*, *actcon*, *spont*
I_{200}	user-defined type	*spont*

Furthermore, thanks to our previous analysis in Telnet traffic in our paper [65], we find that the RTU in the outlier flow is the only “Testing Station”, which is used to detect gas leakage. This substation has several safety valves and sensors to keep the gas pressure within a safe range. Due to the critical requirement in the case of leak detection, it makes sense that the operator sets this station’s interrogation frequency to be as high as possible, so that the safety in the distribution lines can be guaranteed. Telnet traffic also provides the information of physical stations behind 20 RTUs as in Table 5.3.

Table 5.3: Identified RTU station types (from Telnet)

#	Name	Description
18	Distribution Station (GDS)	Where the low pressure local transport network becomes the last mile local distribution network
1	Measuring Station (GMS)	Measures pressure variables from surrounding locations/streets.
1	Expansion Station (GET)	Where the the local distribution network connects to the the high-pressure gas transport network. To our understanding, the high-pressure gas is expanded, i.e., to reduce the pressure to the operational values of the distribution network.
1	Biogas Generator Station (Biogas)	A third party company that produces and injects biogas into the network.
1	Testing Station (GBS)	Dedicated point for testing & measuring for gas leakage and over-pressure conditions in the distribution network, with safety outlet valves to reduce pressure in case of over-pressure.

Chapter 6

A Bulk Power Grid v.s. A Gas Distribution Network

After the separate investigation of the two SCADA networks, we would like to make a thorough comparison of the characteristics we learned from both measurement studies. Here we present some preliminary experiments and insights after the comparison at levels of network topology and messages. Other in-depth comparisons and the meaningfulness will follow. For clarification, we do not indicate the two systems we study represent all the power grids and natural gas distribution networks. Instead, we show that it is not rare that the network design fails to be compliant with the protocol standard, which introduces challenges in the specification-based intrusion/anomaly detection when using protocol standards for whitelisting. We also show that depending on the preferences of the operators and the needs of physical processes, even when two networks share the same communication protocol, the networks can have various behavior patterns, i.e. use the same protocol feature for different physical functions.

6.1 Visibility of the Capture Location

Both datasets of the bulk and the gas network are collected outside of the wide area network (WAN), which provides the communication channels for all remote stations. The main difference is that the bulk dataset capture location sits outside of the control room, while the gas one is inside of the control room. Consequently, the gas dataset contains all interactions among all the devices in the control center, other than the conversations between SCADA server and RTUs. This vantage point brings us the remote management and diagnostic traffic, i.e. the Telnet traffic used for network configuration and RTU maintenance. The details of benefits from remote management and diagnostic traffic study are in our paper of [65].

6.2 Network Topology

As shown in Table 6.1, the natural gas distribution network has much more endpoints than the bulk power grid. The scale difference makes sense if considering the natural gas distribution network mostly have stations propagating the pipeline expansion areas measuring gas pressures and setting up alarm signals for transmission safety purposes. The operator of the bulk power grid constructs two pairs of primary and secondary servers to control different geographical areas, while the gas network operator centralizes the orchestration in one primary server and switches the between two servers during the capture time.

Table 6.1: Endpoints in IEC 104 Communication

Dataset	Substations Count	SCADA Control Server Count	RTU Count
Natural Gas	154	2	154
Bulk Power Grid Year 1	42	4	38
Bulk Power Grid Year 2	44	4	40

When we design the anomaly detection system at the starting phase, we need a white list of the endpoints and their IP addresses allowed in the operational network. For the most proactive defense, we want to know how fast the monitoring system is able to detect the asset, i.e. the RTUs and the control servers. So that we can raise alerts as soon as new unauthorized endpoints connect to the network. As shown in Figure 6.1, for the power grid, we can identify 90% endpoints within 25 seconds in Y1. We can classify the rest 10% endpoints discovered later (within 2.5 hours) in the power grids into three types: 1. the RTUs that transmit only one IEC 104 packet during the whole capture; 2. the outlier RTU from the clustering analysis; 3. a RTU that has a primary server switch-over event. The reasons for the delayed discovery of these endpoints are as follows. First, type 1 RTUs with a single packet and type 2 the clustering outlier RTU are from the redundant backup stations. The operators don't assume these RTUs have stable communication channels all the time. The servers only ping them occasionally with keep-alive testing messages to confirm they are still capable of setting up connections. These keep-alive messages force these RTUs to come online in a glimpse. Second, our dataset happens to capture the start of IEC 104 data transmission for this type 3 RTU, after the capture has elapsed 2.43 hours. This RTU is only online for six minutes, during which it switches the primary server between C1 and C2. In Y2, we can locate 100% endpoints within 17 seconds. For the natural gas network, we discover 100% endpoints within 90 seconds. Therefore, other than the redundancy setting and server reconfiguration in the power grid Y1, the majority endpoints in both SCADA networks are discovered within 1.5 minutes, since most primary operational connections need to remain online stably for 24/7.

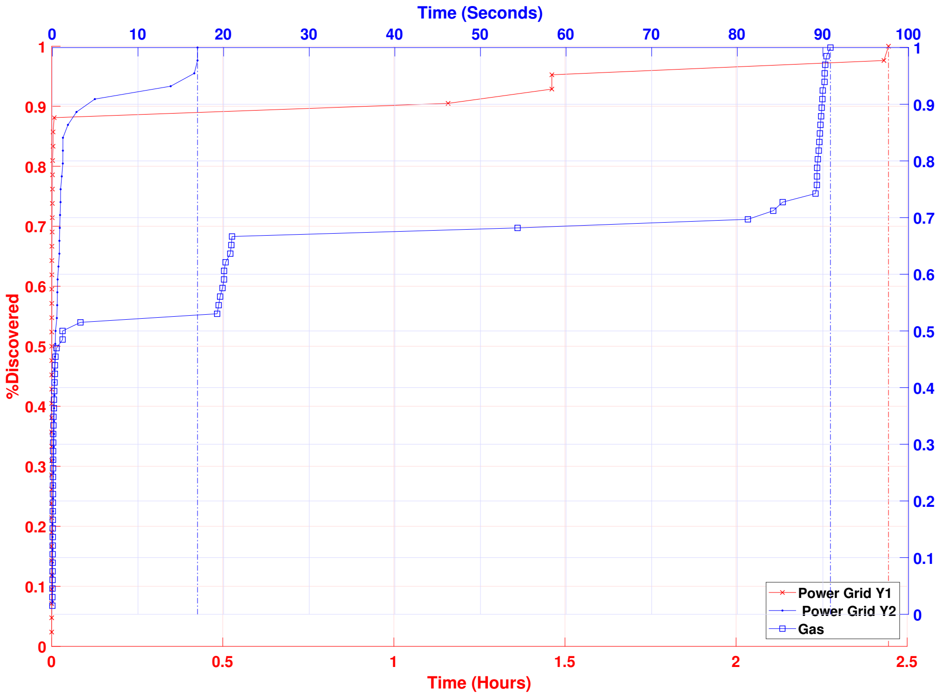


Figure 6.1: Endpoint IP addresses discovery speed comparison in two systems. Power grid can have a delayed discovery because backup RTUs that only communicate when testing.

6.3 Monitor and Control in IEC 104 Communication

6.3.1 Message Distribution

As introduced in Chapter 1, IEC 104 protocol has mainly three types of control and monitor messages. I format messages are for encapsulation of the control commands, sensor readings, and actuator states. S format messages acknowledge the receiving of the I format messages. U format messages are for the instantiation and termination of the data transfer in primary connections and the keep-alive information exchange in secondary links. Computation of the distribution of these three types can provide an overview of whether this SCADA network is measurement-centric or control-centric. It is helpful for the anomaly detection system to understand how often the data transfer and redundancy control messages should occur in the regular operation. So that the anomaly detection system can make a just verdict whether a new data transfer should open or not at this time.

From Table 6.2, the bulk power grid is more measurement-centric where the U messages take less than 5% over time. The natural gas network contains a much higher portion of U messages. Unlike the standard recommendation in IEC 104 to use keep-alives in secondary connections, the operator uses keep-alives in all primary links. Also, unlike the occasional testing in the secondary connections, they use it over the whole capture time. Other than the keep-alive messages, because of the IP reconfiguration events observed previously, there are much more re-initiation and termination events in U messages for all the RTUs' data transfer processes than those in the bulk power grid.

We will examine the deep reasons from the physical process that might trigger this setting in the following work.

Table 6.2: APCI distribution of all three datasets: for each dataset, the proportion of the count of I/S/U format IEC 104 data units (APDUs) in all APDUs

APCI format	Gas	Power Grid Year1	Power Grid Year2
I-format	28%	83.46%	84.71%
S-format	26%	12.02%	11.02%
U-Format	46%	4.51%	4.27%

6.3.2 Polling Patterns of I-format Messages

As introduced in section 4.7, we can learn the direct cause that triggers the specific I-format message (i.e. process values), or how often this conversation happens. We calculate the portions of different transmission causes for all I-format IEC 104 messages in the datasets, and present the results in figure 6.2 and 6.3. The proportions in the heat maps have been rounded to three decimal places. For instance, in the bulk dataset, the most messages are encapsulated in floating point values with a time tag, mostly (57.1%) in transmission only when the value change exceeds certain threshold (i.e. the spontaneous mode). Compared to non-cyclic mode, cyclic pattern only applies for 1.18% ($0.001 + 0.024 + 0.093 = 0.118$) of all the measurement values in the bulk dataset.

From the color shades of two heat maps, we can already observe the different distributions in these two systems. Specifically, the differences are:

- The bulk has 87.9% of the polled measurement values in spontaneous mode, while the gas network has only 11.6%;
- 9.3% of the measurement values in the power grid are collected periodically, while the gas network does not configure any polling periodically;
- The majority of process values in the bulk are analog, only around 0.1% digital values, while the gas has the digital ones as the largest category in 27%;
- The bulk has less than 0.01% measurement polled from the interrogation process, while interrogation in the gas network happens much more frequent in 47%;

We have our hypothesis with high confidence for each observation above:

- The majority process variable type in a gas distribution network is the gas valve. While the power grid measures the current, power, frequency, all in analog values. Therefore, digital values amount dominates in the gas network.

- Compared to the dynamics in electricity power or voltages, the changes in valve status (open/close) are much less frequent. Therefore, gas network attempts to use interrogation configured at a high frequency, instead of spontaneous or cyclic mode to guarantee the freshness of process values.

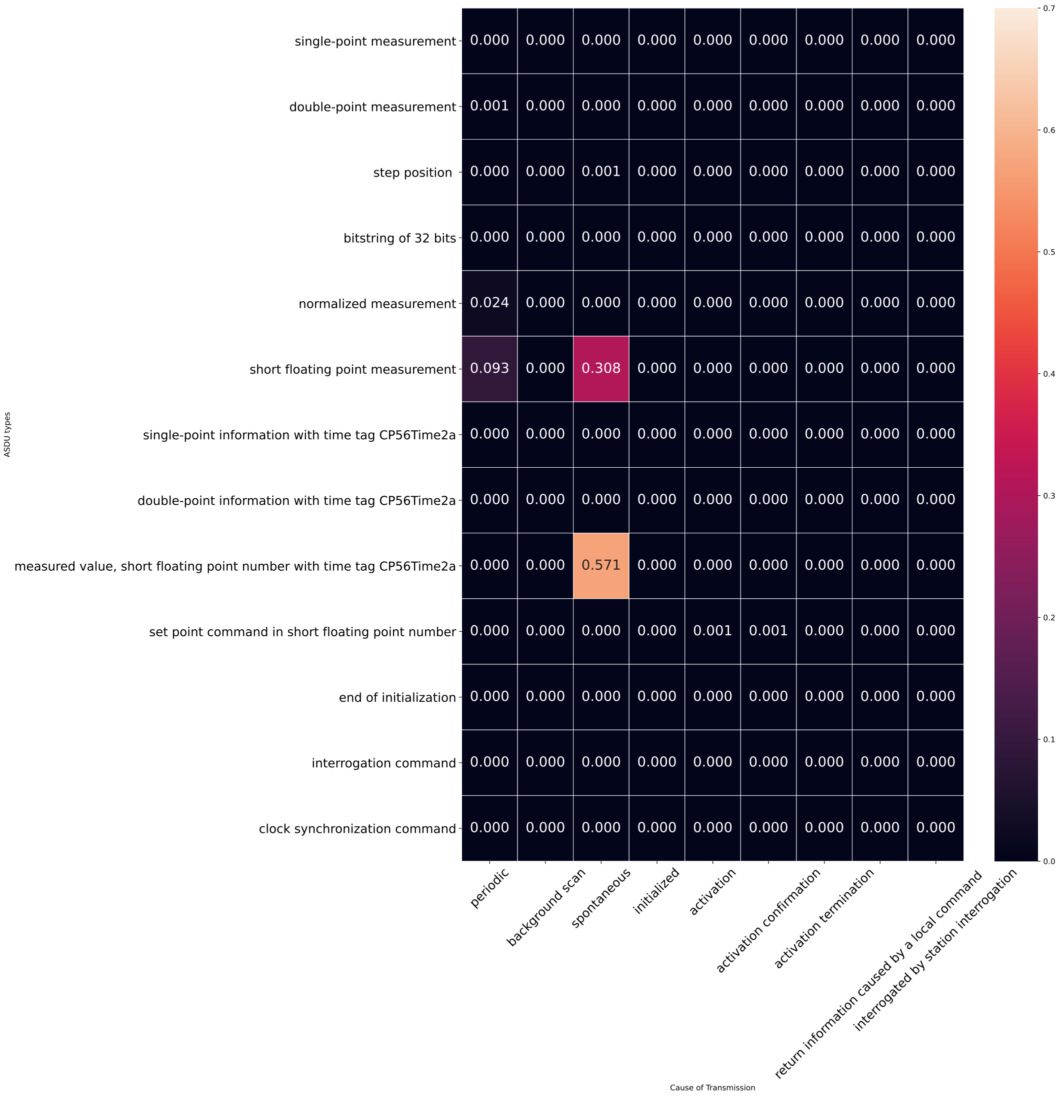


Figure 6.2: Distributions of transmission causes for I-format messages, the bulk power grid

6.4 Clarification

By making this comparison, we do not intend to make an exemplary model with each system among its own type. Instead our main goal is to show the security community that it is necessary and favorable to study the operational network over testbeds. Each real-world industrial control system has its own uniqueness in the monitoring and control patterns although they comply to the same communication protocol. These differences introduce the complexity to building system behavior baseline in the anomaly detection design.

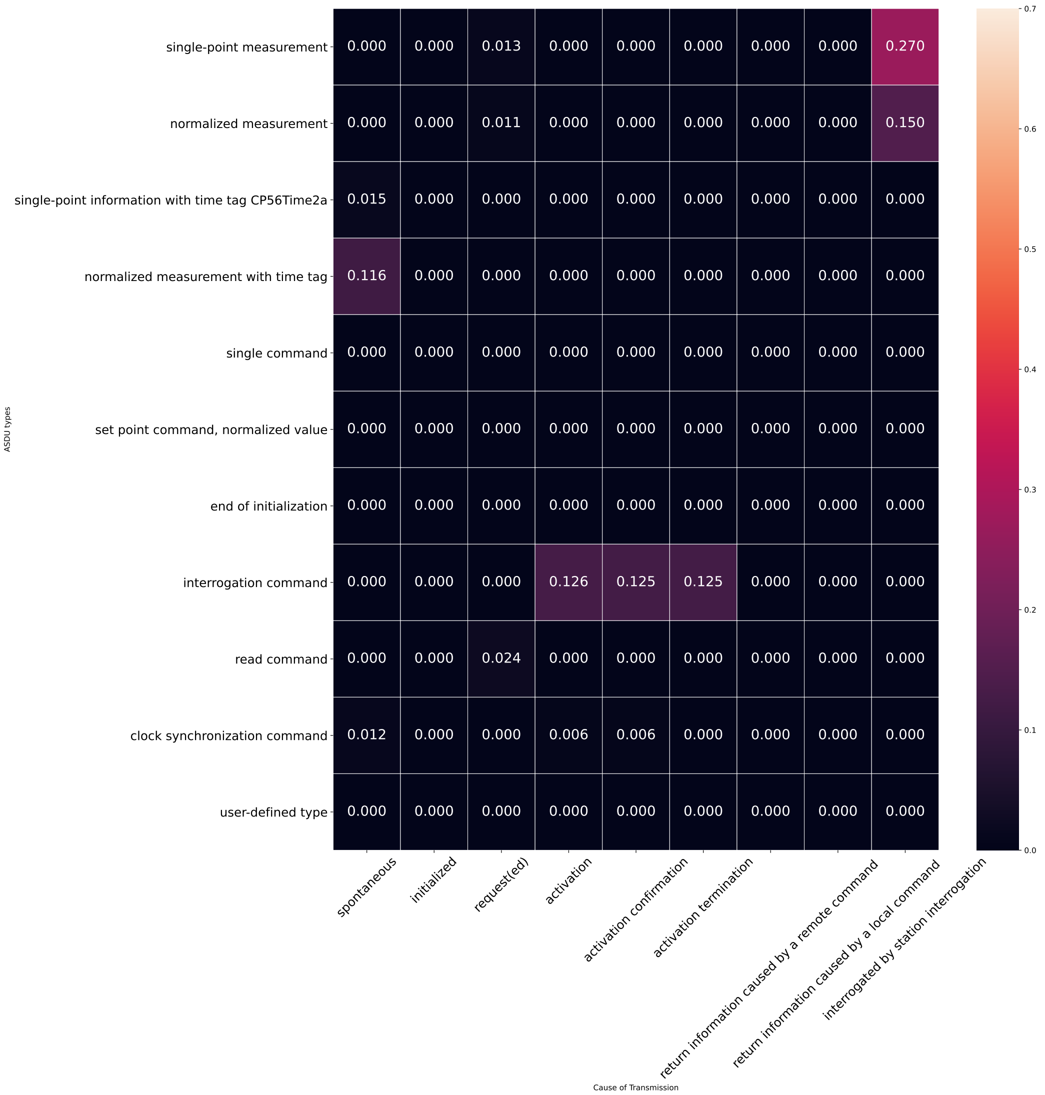


Figure 6.3: Distributions of transmission causes for I-format messages, the natural gas distribution network

Chapter 7

Anomaly Detection in ICS SCADA Network

As introduced in chapter 2, the most effective anomaly detectors for ICS testbeds are based on the statistical modeling of sensor reading values. They usually have two approaches to build the baseline, one is to identify the control system invariants by a state estimation, or to fit a regression/Bayesian/neural network model to the time series.

7.1 Preliminary Defense Design

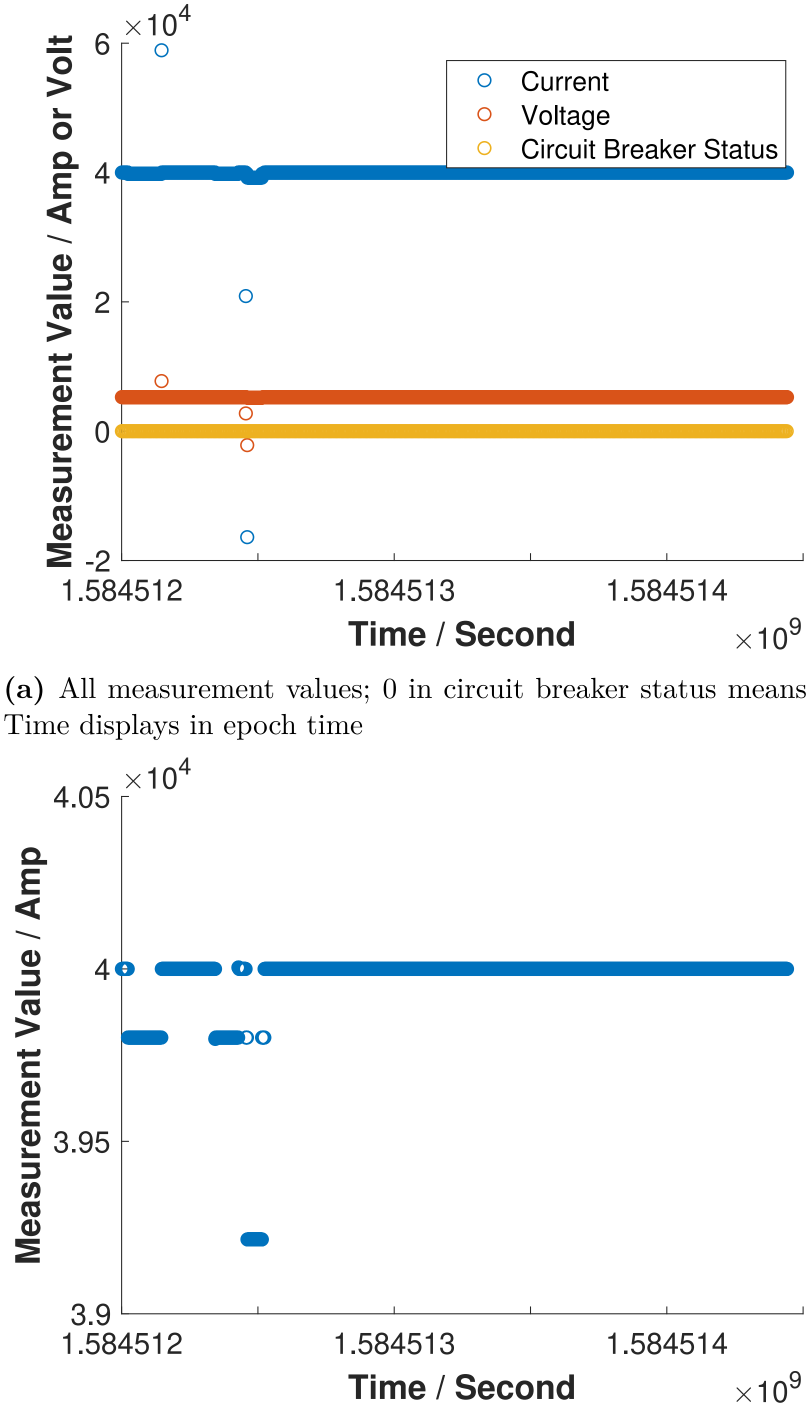
With the software testbed designed for a smart grid in compliance with IEC 104[68], We construct our first defense design with the measurement values in the process variables.

In simulation, we analyze two scenarios and collect network traffic captures for each. As a starting point, we place the initial capture during normal system operation. The SCADA system takes measurements and status from the RTUs and sends commands issued by the operator in normal operation. Then, when the command injection described in [68] assaults the system during normal operation, we perform the second capture. As a result, the traffic traces in the second capture are intermingled. We may use these two datasets to first measure and construct a baseline of normal operations, and then examine the effects of the attack on physical device measurements.

The majority of network traffic in SCADA systems is about passive monitoring the status of variables inside an RTU. Fewer data formats are designed for the exchange of active control commands, compared to the formats for reporting variables under monitoring. Under the protocol of IEC 104, payload data units from type 45 single command to type 50 set-point command, are designed for commands. We identify that power grid operators use type 50 set-point command for power modulation in the automatic generation control (AGC) process of load balancing. In the attack simulation we creatively use type 50 to transmit the injected malicious commands under type 45 in the attack, to open circuit breakers and to cause a blackout.

7.1.1 Visualization of Time Series

In Figure 7.1, After preprocessing, we show the visualization of measurement values. Current, voltage, and circuit breaker status are all physical factors that we model based on their real-world behavior. Most of the time, the voltage and circuit breaker status are relatively stable. In a real-world context, current is more erratic than the other two. There are several (less than 10) scattered points for all three types of measurements, assuming from the unknown simulation noise.



(a) All measurement values; 0 in circuit breaker status means close; Time displays in epoch time

(b) Current measurement values

Figure 7.1: Real-time measurement of all three types of physical variables of one RTU to the control station under normal operation.

7.1.2 Feature Selection

For each APDU in the network traffic, we pick the payload data under IEC 104 ASDU types as categorical features and measurement values as a numeric feature. We use one-hot encoding to create the feature vectors, as shown in Table 7.1.

Table 7.1: Examples of Feature Vectors

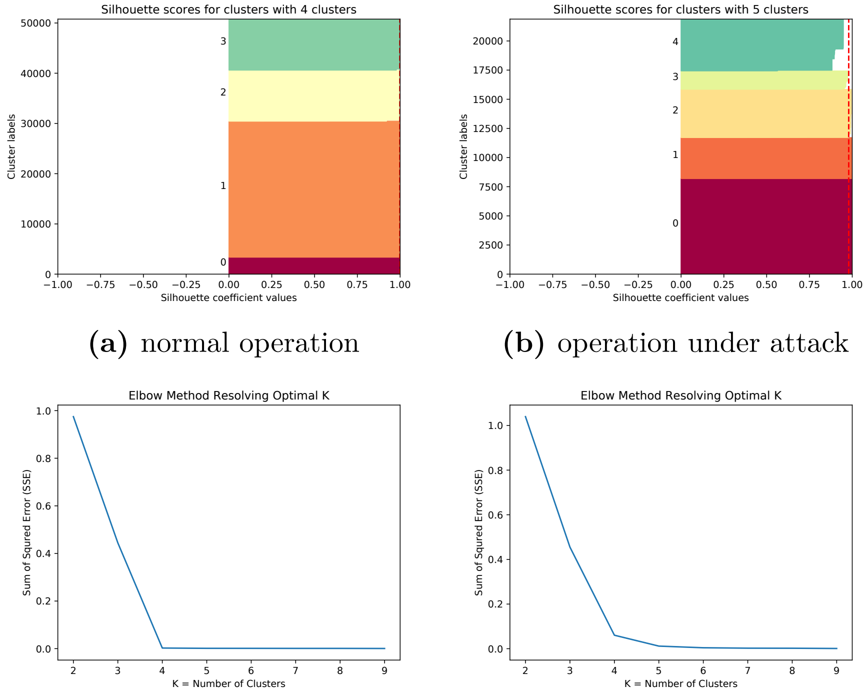
	type3	type36	type50	measurement
command data	0	0	1	0
voltage data	0	1	0	5200

7.1.3 Anomaly Detection

We apply clustering algorithms to detect abnormal traffic from traffic capture. First, we train K-Means [47] and DBSCAN algorithms on the traffic captured without attacks. Since the clustering result from both algorithms are similar, we present K-Means results only in this section. First, we use the Silhouette score [67] and Elbow method [74] to decide the number of clusters when using K-Means, as shown in Figure 7.2c, Figure 7.2a, Figure 7.2d and Figure 7.2b. If the Silhouette coefficient on the x-axis value is greater than 0.5, then it indicates that the data point is well matched to its cluster and poorly matched to neighboring clusters. If most objects have a high value, then the clustering configuration is appropriate. If the majority of points have a low or negative value, then the clustering configuration may have too many or too few clusters. Elbow method can help us pick the cluster number at the elbow position with the least imprecise cluster assignments. Therefore, we conclude that 4 clusters are the best choice. In Figure 7.3a, we use compressed feature components to visualize the clustering results with principal component analysis (PCA). PCA compresses four feature components into two principal components and preserves 86% variance of the original feature vector. Each cluster composition is listed in Table 7.2.

Table 7.2: Cluster composition under regular operation

cluster 0	circuit breaker status
cluster 1	current measurement around 40000 amps
cluster 2	control command confirms the close of breakers
cluster 3	voltage measurement around 5200 volts



(a) normal operation **(b)** operation under attack

(c) normal operation **(d)** operation under attack

Figure 7.2: Choosing the proper number of clusters with K-Means.

We then apply the same clustering algorithm on the traffic capture when the testbed is in the same operation configuration but under attack. From Fig. 7.2d, we have a cluster 4 successfully grouping all the abnormal values of current and voltage when they suddenly change to negligible values, in Figure 7.3b. Comparing the cluster groups in Figure 7.3a and Figure 7.3b, there is a new fifth cluster in Figure 7.3b. This new cluster successfully aggregates all the network traces containing the abrupt drop of the currents and voltages to negligible values. This phenomenon happens because the attacker opens all the circuit breakers in the transmission line. So the measurements of the field devices all drop to close to zero if the devices locate in the same transmission line as those circuit breakers. The operators can further investigate all the data points in cluster 4 to confirm the lines affected by the attack.

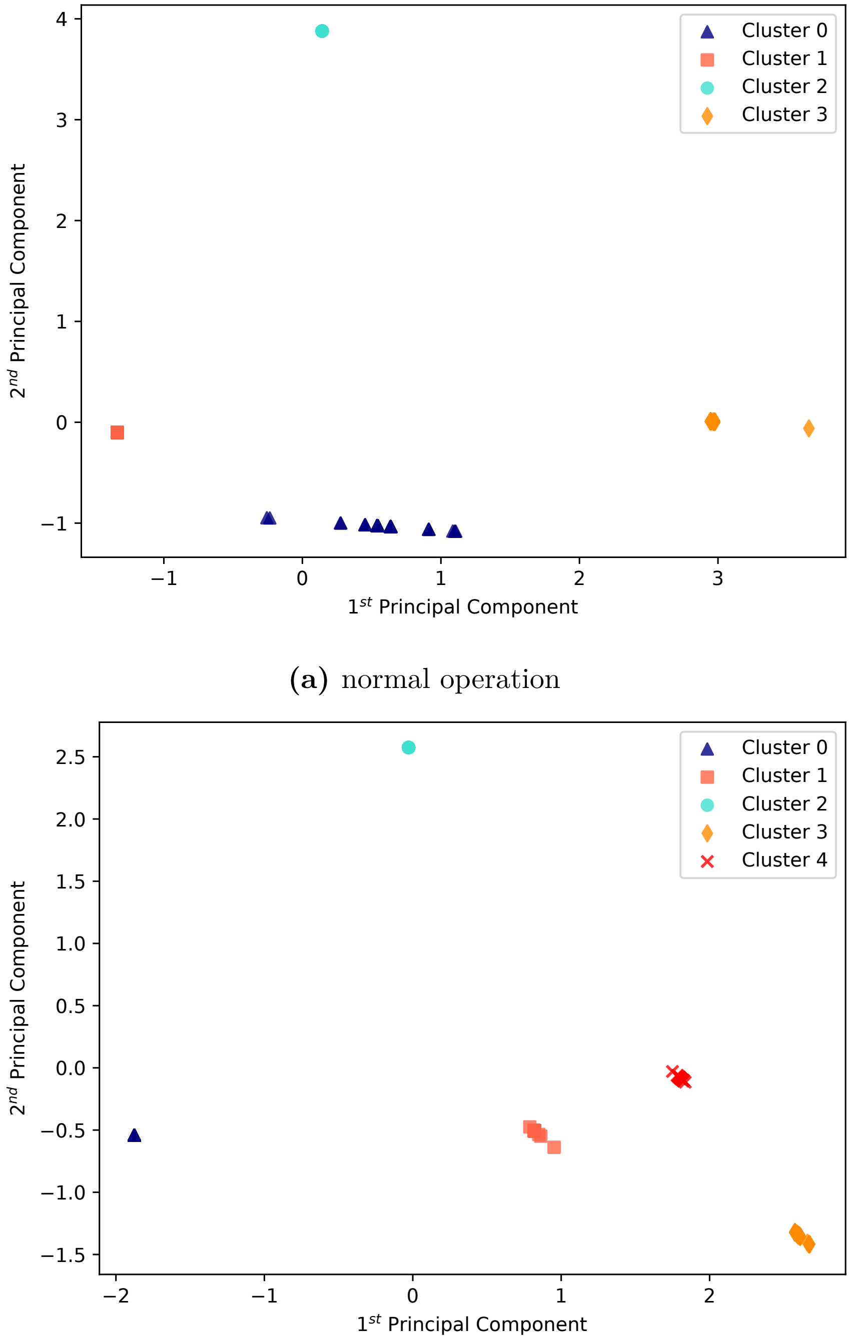
7.1.4 Limitations

We have a cluster successfully grouping all the abnormal values of current and voltage when they suddenly changed to negligible values. In real scenarios, the attackers could launch multiple delicate trials, especially in the reconnaissance stage. And our unsupervised learning method is expected to catch the nuance changes in the measurement values if the attack affects the process variables.

However, because of the primitive physics modeling in this testbed, the attack scale was limited to one RTU. In the real world, the SCADA network has a broad attack contamination surface, and all RTUs with circuit breakers can be effectively shut down by issuing a circuit breaker open command. It may be difficult for clustering to discover the outlier group if there is a global spike of negative or zero values. A different time series modeling should be investigated instead.

7.2 Advanced Defense Design: Device Profiling

By applying the deep-packet inspection of IEC 104 traffic, we can build a dataset from the time series of the measurements in point variables. Taking each measurement value at a particular timestamp as each row in the data table, we



(a) normal operation

(b) operation under attack

Figure 7.3: K-Means results in two situation with the identical testbed configuration

compute the aggregated statistical characteristics over a specific rolling window (e.g., hourly/daily) per point variable and use them as features. With the plain-text description retrieved from Telnet traffic, we label the dataset with the physical semantics of actuators or sensors connected to the points. Specifically, we focus on the classification of analog variables, because the classification of digital variables is a trivial problem when they are mostly about open/close gas valves in this dataset. One can recognize the function by a number one as opening and zero as closing. Through this modeling, we build a normal-behavior profile for each category of points so that this model can detect an unexpected activity and raise the alert.

This detection model is potential for defending against false data injection (FDI) attack by maintaining the operational states of the gas network through monitoring the point variables' time series. [54] first proposed the attack design of FDI for the power system under a strong threat model, where the adversary has access to alter the sensor values with the full knowledge of the network topology and parameters. This line of research evolves with weaker threat models, where the most recent work in [50] assumes the attacker may have full system knowledge but have limited access to change meter readings due to the improved meter protections by the operators, and the attack method in [84] where the adversary can use Principle Component Analysis to generate a stealthy attack vector. With all these potential attack vectors, our detection model is threat-model-agnostic and simply focuses on the classification of the point variable based on the collected time series.

7.2.1 Dataset Generation and Description

Raw Dataset

We are fortunate to access two real-world SCADA network traffic datasets in PCAP files. One is for *the bulk* power system (defined by Kelvin et al. in [57]) and another for the distribution network in a natural gas system. *The bulk* power system consists of the generation/production plant and transmission network, out of the three main components of most industrial control systems with the distribution network as the third. As introduced in **some previous section here**, we extract the time series of the signals in actuators and sensors after performing deep packet inspection of the IEC 104 packets. Each data sample in the original intact packets is properly timestamped with the packet arrival time, with the polled value in the payload of an Application Service Data Unit at this specific timestamp and tagged by the point variable type if known (we only know partial variable labels for both datasets). There are digital and analog values for a variety of point variables, with the analog ones being real number values or normalized values in the range of $[0, 1]$. The choice of normalization totally depends on the operator's choice of the encapsulation format, i.e. ASDU types in IEC 104. During the In Table 7.3, we list the units of different variable classes in these two systems after a label clean-up procedure for both datasets (will introduce in the next part of "Physical semantics and labels"). Considering the computation efficiency, out of all 154 RTUs we choose a representative subset of six RTUs, covering all four gas station types with two Gas Distribution Stations (GDS) RTU 22 and 172, Gas Leak Test Station (GBS) RTU 26, Gas Expansion Station (GET) RTU 228, Gas Measurement Station (GMS) RTU 230, and Biogas Generation Station RTU 11027. After comparisons of several different subset combinations with different GDS data, it shows $< 0.1\%$ variation in classification

Table 7.3: Present analog measurement units

The number of occurrences of a unit is higher than the actual number of variables that are actually measured in the respective unit, i.e., due to alarm thresholds using the same unit of measurement. Various spellings of the same unit were aggregated, e.g., various ways of abbreviating “Millibar”.

Dataset	Unit	Concept	Count
Gas	mBar	pressure	341
	%	relative	111
	Bar	pressure	67
	Cts	valve position	61
	m^3/h	flow	11
	Sec	time (duration)	10
	°C	temperature	21
	ppm	parts per million	5
The bulk	kW	power	190
	kV	voltage	101
	A	current	81
	Hz	frequency	23
	kW	AGC set-point	2
	-	motor position	1

performance which indicates almost no degradation in choosing a specific group.

Physical Semantics and Labels

In the bulk, the classes of the variables are relatively neater, i.e. the standard physical variables in a power system. While the variable class labels of the gas network come from the Telnet traffic with rather specific descriptions, and some details are trivial to dissect. We identify the physical semantics of 1048 point variables (705 digital ones and 343 analog ones) from the Telnet traffic. It is unreasonable to apply hundreds of distinctive point types as labels. Instead, we distill a representative taxonomy of these types. For example, *PT1 Gas Pressure High High Limit* and *Gas Measurement Station K Pressure 100mB High High Limit* are both alarm configuration signals. Therefore, we perform a cleanup for these labels by pruning the redundant details and merging into the core types. Eventually we obtain five meaningful variable class labels, alarm_config (i.e. alarm configuration signals), flow (i.e. flow rate), position, pressure and temperature. In the bulk power system, we have access to the labels of 471 (73 digital ones and 398 analog ones) out of 856 point variables.

Feature Engineering

First, we would like to compute the statistical characteristics from the data flows of different variables. We first form groups of time series for each point variable over the time length of certain rolling window (per minute/hour/day). With the Python package *TSFRESH* [18], we apply the statistical computation over the data samples within each aggregated group. Initially, we start with the complete set of 78 features provided by the package. Each individual feature may have customizable parameters so the final set of features can be over 1000. Then we reduce the feature space by excluding the invalid features that generate null values, introduced from a divide-by-zero situation. We can further prune the features by limiting only the top features from the feature importance analysis with tree-based supervised learning algorithms. Then, we add protocol-specific features to the instances in the aggregated groups resulted from the previous step. Each point variable matches one information object in terms of IEC 104 packets. According to the information object belonging to what ASDU and CauseTx types, we conduct one-hot encoding for the ASDU types and cause-of-transmission (CauseTx) types to introduce more domain knowledge. For example, assuming the dataset has observed ASDU IDs 1, 34, and 45, one instance is ASDU ID 34. Then we name features asdu1, asdu34, asdu45, and this instance will have value one in asdu34 and value zeros in asdu1 and asdu45, resulting in a feature vector of 010. Similarly for CauseTx, the features are like causetx2, causetx20 and so on.

The complete process to construct this well-structured time series dataset is visualized in Figure 7.4. Eventually we have the two datasets as summarized in Table 7.4. The aggregated instances are the actual data input for our following machine learning experiments.

Table 7.4: This table explains the dataset in a rule-of-thumb way for time series

Dataset	#Samples	#Aggregated Instances	#Features	#Classes
Gas	15.1 millions	61354	201	5
The Bulk	5.5 millions	3862	202	6

7.2.2 Modeling for Point Variable Type Profiling

For the two SCADA networks, we would like to develop a monitoring model that can be trained offline with the historical measurement time series data of the variables, and then perform the inference of device type (e.g. gas pressure or electricity voltage) with the fresh data. Therefore, our task is to build a supervised learning model for time series, with the aggregated statistical features and protocol-specific features and point variable types as tags.

Experimental Setup

Train-Test Split: The way we split the train and test set is to use the earlier 75% of time series for training and the later 25% of time series for testing, with all classes stratified sampled. One thing to note about the train-test-split method is that, we don't run cross validation across various combinations of train-test splits. This way we can isolate the model from two implausible situations, peeking future information during training phase and forecasting occurred events during testing phase. Then the model's performance can testify its predictive power in this

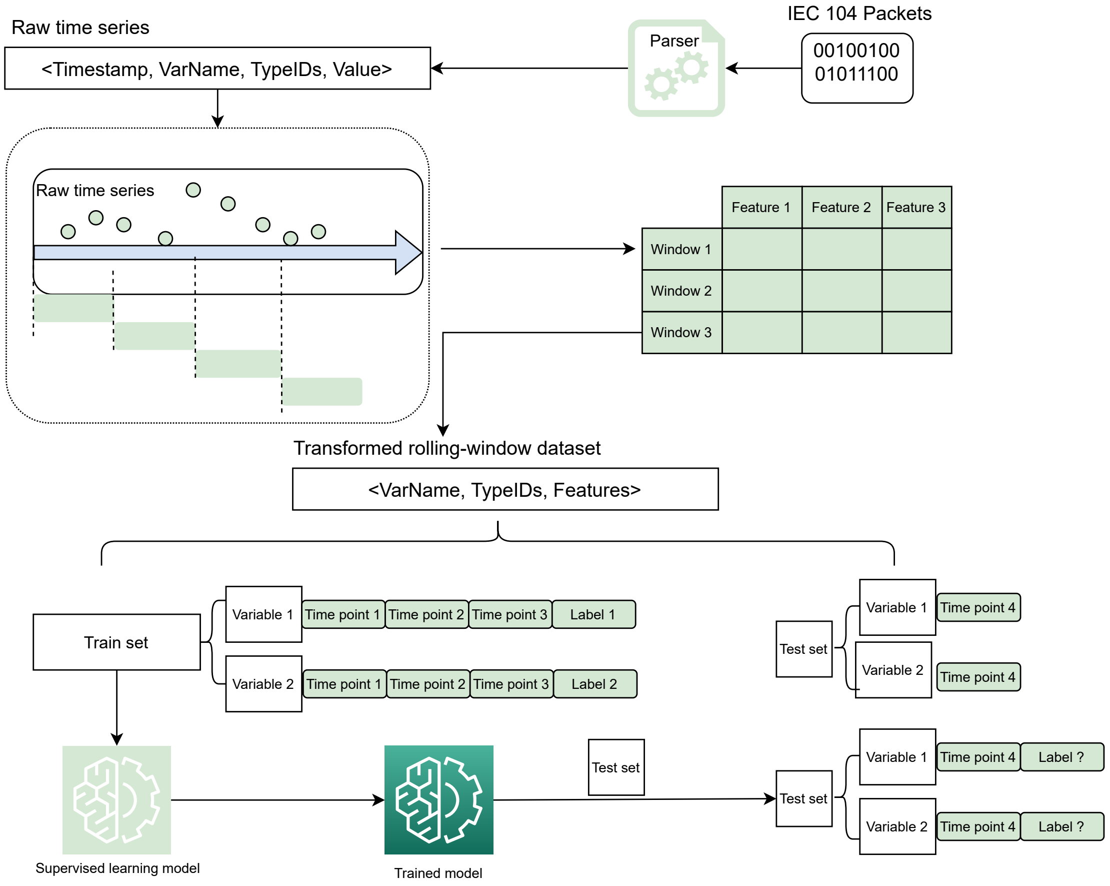


Figure 7.4: The flow of this time series experiment, including the processes of label construction, rolling window for sample aggregation and seasonality feature computation, and training/testing experiments.

specific application scenario, where time series models are designed for anomaly detection in future series.

Algorithm choices: Since we've defined this problem as a supervised learning problem, we evaluate the performance of multiple popular supervised learning algorithms, including the most performant gradient boosting algorithm XGBoost, Random Forest, Support Vector Machine with a linear kernel, K-Nearest Neighbors, and Logistic Regression. Other than XGBoost with its own toolkit, we use *scikit-learn* package to apply the rest algorithms to our datasets.

Hyperparameter Tuning: We will present the eventual model performance of the tested algorithms in next section section 7.2.2. XGBoost outperforms all the rest ones with a more acceptable computation efficiency when all the algorithms tested have been equally treated in fine-tuning. Since it is a tree-based gradient boosting algorithm, the hyperparameters most effective for tuning are mainly about adjusting the tree branches, height and regularization items. Specifically, we tune the learning parameters *learning_rate*, two regularization items *reg_alpha* and *reg_lambda*, and the boosting tree based parameters *n_estimator* the number of trees, *max_depth* the tree height limit, *subsample* the random sampling rate for each tree construction, to achieve a properly fitted learnt model. The final sets of parameters for both datasets are as follows:

Table 7.5: Fine-tuned hyper-parameters of XGBoost algorithm

Dataset	learning_rate	max_depth	n_estimators	subsample
Gas	0.08	3	150	0.8
The Bulk	0.1	3	100	0.8

Results and Discussions

We present the most performant result in the format of confusion matrix in Figure 7.6, and precision-recall curves in Figure 7.5. In both cases, gradient boosting tree algorithm with XGBoost package [17] performs the best.

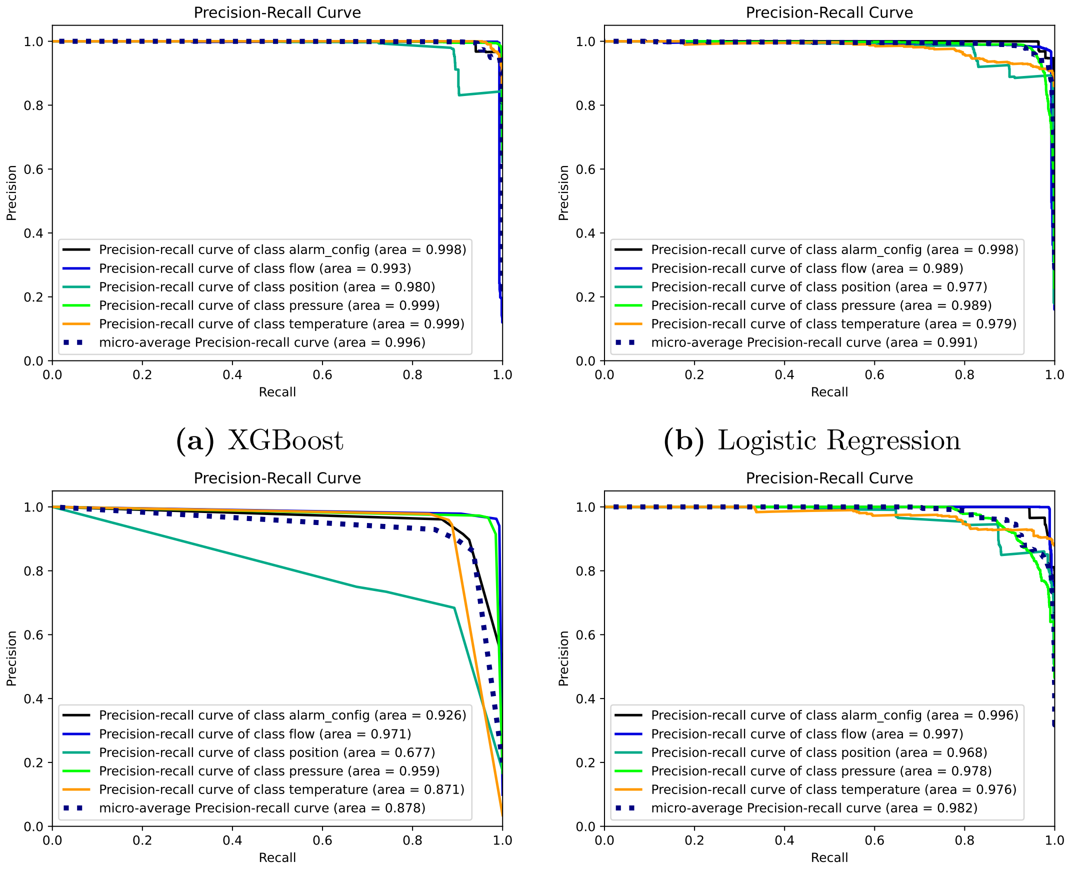
Table 7.6: Algorithm performance comparison

Dataset	Algorithm	F1	Accuracy
Gas	Logistic Regression	0.95	0.95
	K-Nearest Neighbors	0.90	0.89
	SVM (linear kernel)	0.96	0.96
	Random Forest	0.93	0.93
	XGBoost	0.97	0.97
The Bulk	Logistic Regression	0.45	0.45
	K-Nearest Neighbors	0.42	0.42
	SVM (linear kernel)	0.45	0.45
	Random Forest	0.65	0.65
	XGBoost	0.92	0.92

Table 7.7: Top 10 features ranked by gain

Rank	Feature Meaning	Importance Score
1	the variance value over a corridor given by the quantiles 0 and 0.8 of the sample distribution	0.2294471
2	root mean square	0.14340197
3	A complexity estimate based on the Lempel-Ziv compression algorithm (bin width = 10)	0.05593252
4	IEC 104 ASDU type 34 normalized values with timetag	0.039654512
5	percentage of reoccurring datapoints	0.03806225
6	first location of minimum	0.037393752
7	the variance value over a corridor given by the quantiles [0, 0.4]	0.024982385
8	the size of data samples	0.021845581
9	the maximum value	0.021015033
10	the variance value over a corridor given by the quantiles [0.8, 1]	0.020797458

Explainable Learning with Feature Importance Analysis in XGBoost and SHAP First, we conduct the feature importance analysis by computing the average classification correctness gain during tree branch splits using a specific feature. This analysis result doesn't indicate a direct causal relationship between features and inference results, instead it shows the quantified impact of



(a) XGBoost

(b) Logistic Regression

(c) K-nearest neighbor

(d) Random forest

Figure 7.5: Precision-recall curves for device profiling in gas dataset

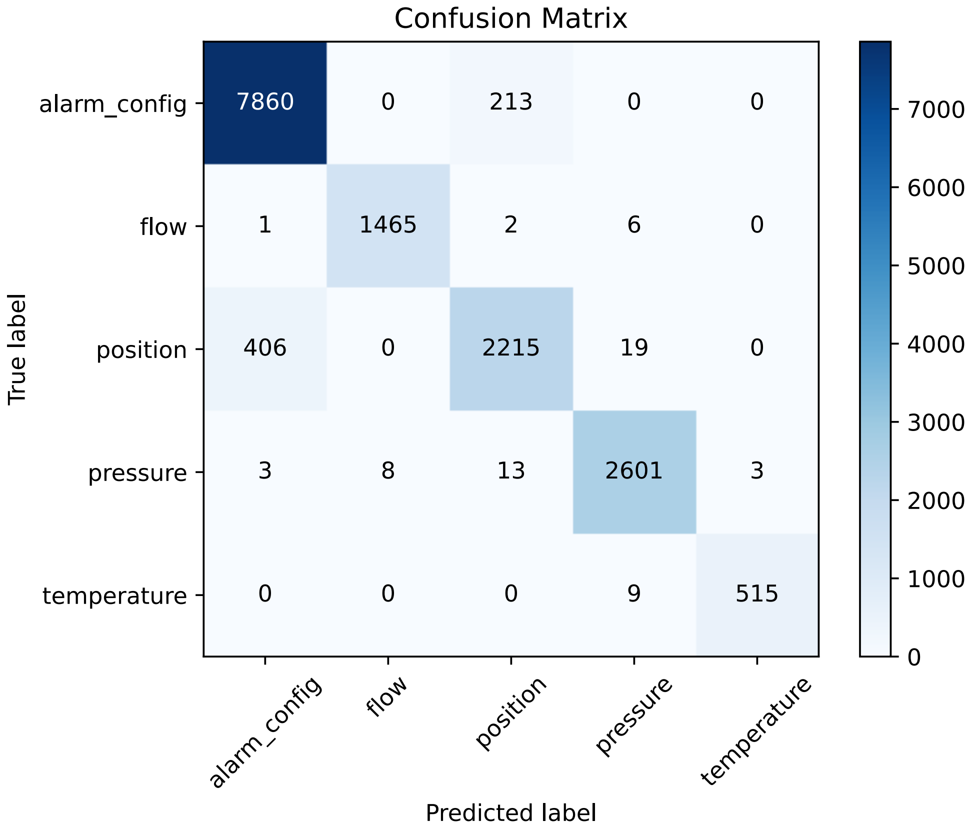


Figure 7.6: The most performant classification results in confusion matrix: the majority false detection results come from alarm configuration and position signals.

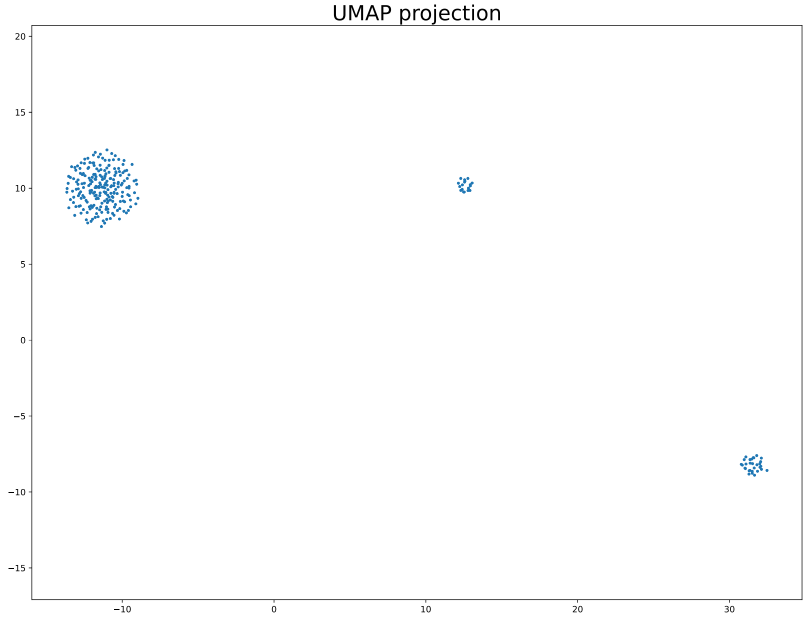


Figure 7.7: Clusters of the wrong classification results

each feature over the classification result. In Table 7.7, it gives the top 10 features. We can observe that during the decision process of the XGBoost algorithm, the variance over a certain range of quantiles and the root mean square error (RMSE) are more important (22.9% and 14.3%) for the classifier to make a correct judgement than all the rest features (each is $< 5.6\%$). These top two features together demonstrate that the classifier reckons with the volatility information of the input signal. To be noted, although no protocol-specific feature appears in the top 10 list, there is the top 17 feature ASDU type 34 (a format for normalized values with a special timetag). This is reasonable because this feature separates flow and temperature samples (both in type 34) from pressure, alarm configuration signals and position (all in type 9).

SHAP, the SHapley Additive exPlanations method explains the marginal contributions from each feature to the classification output of machine learning models based on game theory. It trains the model with each possible combination, i.e. the power set, of each feature across with other features and compares the performance difference under different impacts. Precisely for a tree-based boosting algorithm, we claim a dataset N with features M and prediction function as f_x. We define S representing any subset of features that exclude the i-th feature, and $|S|$ is the cardinality of the subset. Then, the contribution of each feature ϕ_i on the prediction/classification output is computed through each marginal contribution. Officially it is determined through the following formula:

$$\phi_i = \sum_{S\subseteq N\setminus\{i\}} \frac{|S|!(M-|S|-1)!}{M!}[f_x(S\bigcup\{i\} - f_x(S)]$$

First, we run SHAP analysis for all the features in the gas dataset. As shown in 7.8, we give the top 20 contributing features. Specifically, the variation in the sum of absolute value changes mostly impacts the class decision, especially for the alarm_config class. This observation is reasonable because being the largest group among all classes, the signals of configuration signals are fixed with little or no changes over the whole timeline. While the second top contributing feature of the mean value over the quantiles between 0 and 0.6 mainly impacts the recognition of pressure readings. This indicates the pressure's local average within such long quantiles are more volatile than signals in the rest classes. As a matter of fact, we observe that eight of top 20 features (2nd, 7th, 8th, 9th, 11th, 14th, 16th, 18th) are related to the analysis of the time series' volatility within quantiles. As for one protocol-relevant features, we only observe the ASDU type nine (a type for the normalized measurement values) as one top contributing feature. It leads the model to classify as pressures, alarm configuration signals, or positions when it mostly directs the recognition of configuration signals due to more samples in alarm configuration class. Other classes are in ASDU type 34, which is the normalized measurement value with a special formatted timestamp unit as result of clock synchronization.

7.3 Process-Aware Anomaly Detection in Gas Processes

7.3.1 Automated Process-Aware Detection Design

For the unsupervised anomaly detection, it is crucial for the model to learn the normal patterns of SCADA network, specifically the physical dynamics in the process variables. To introduce more insights of the correlated process variables, we plan to develop multi-variate modeling, i.e. allow the model to learn not only the seasonality in individual variables, but also the dependencies among multiple correlated variables. Intuitively, this multi-variate modeling can also detect the anomaly type where the dependency is violated. Moreover, the correlation among variables facilitates the interpretation of the detected anomalies.

The main architecture is the Long Short-Term Memory (LSTM) autoencoder decoder, which has been shown to be state-of-the-art in anomaly detection for numeric time series[60][49]. To begin, we create a new dataset using the time series of five signals, one pressure signal, and four pertinent safety alarm configuration signals representing various alerting levels. In big steps, we divide this dataset at one time point into the subsets of *normal session* and *contaminated session*, assuming that only the latter subset contains the abnormal process values. When a known process anomaly occurs in our dataset, we have information of the rough timestamp. In real-world applications, historical data of no process anomaly occurring yet can be used for training. Then in the training phase, the *normal session* subset is the input to the LSTM autoencoder-decoder. After the

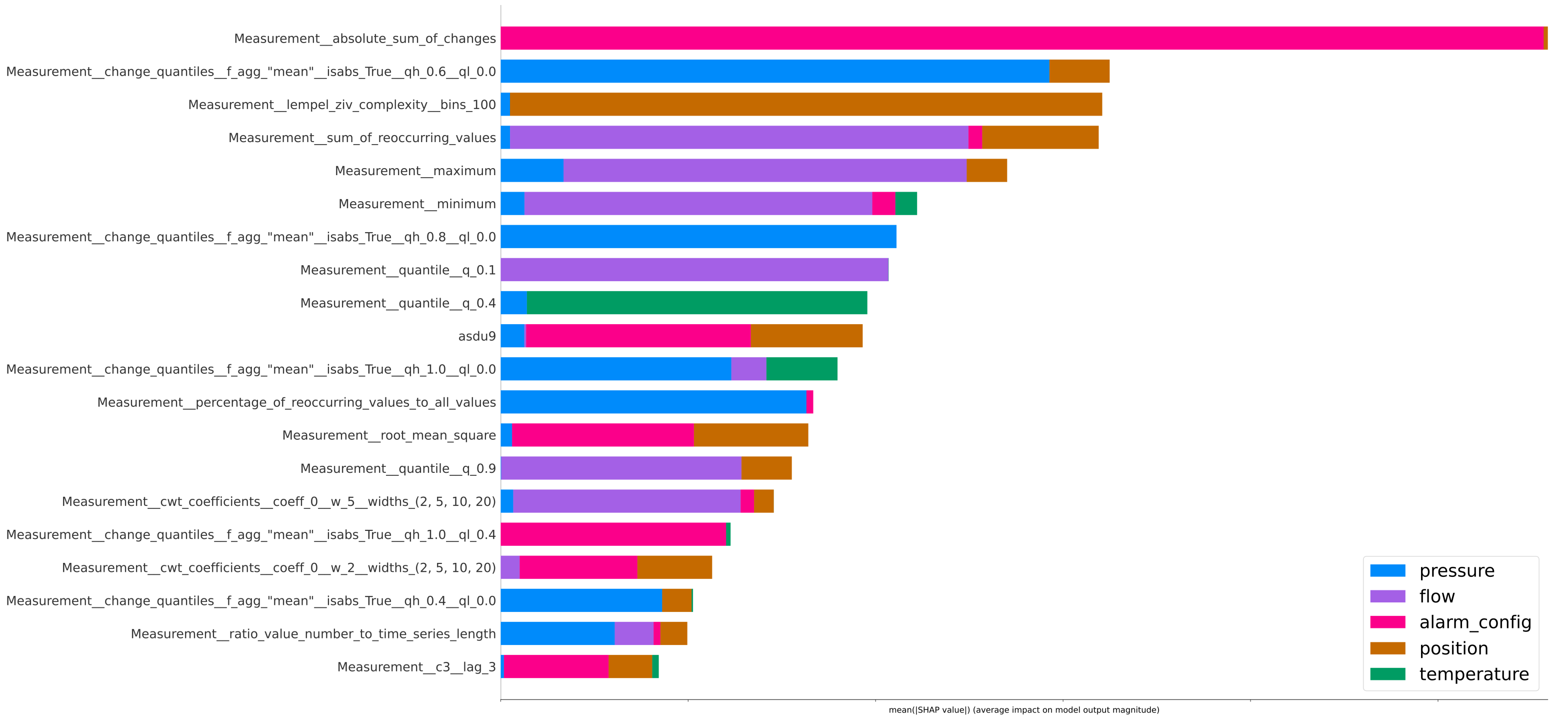


Figure 7.8: The marginal contribution of the top 20 features to the classification results

model converges in the training, we apply the well-trained neural network to the test dataset. In both training and testing, the neural network produces the reconstruction of both the *normal session* and *contaminated session*. By calculating the difference between input samples and reconstructed samples, i.e. the reconstruction loss, we get the maximum loss in all the reconstructed training data and define this max loss as the threshold. Later we call any testing data having a reconstruction loss higher than this threshold as anomaly. The overall framework is shown in Figure 7.9 and Figure 7.10. After a series of fine-tuning, our final LSTM network structure is in Figure 7.11, with the Adam optimizer and mean absolute error (MAE) loss optimization. In the following case study, we study the gas pressure signal and the correlated safety alarm configuration signals.

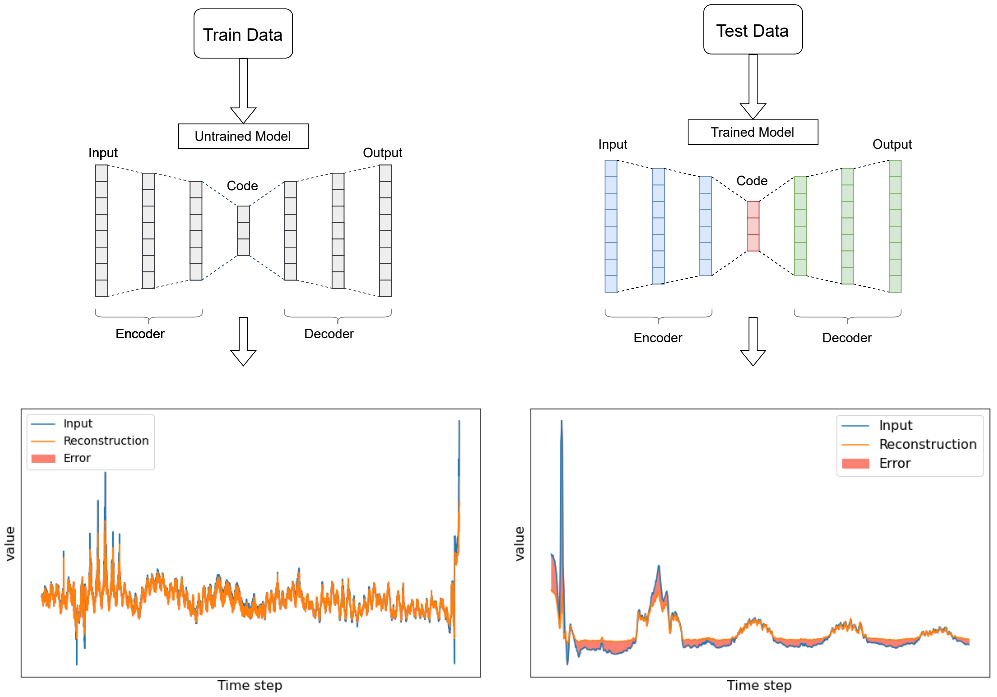


Figure 7.9: Overall framework of the automated process anomaly detection

Finally, we show the detection result in Figure 7.12. Our model applies to a the time series of a pressure sensor in a distribution station. The detector identifies

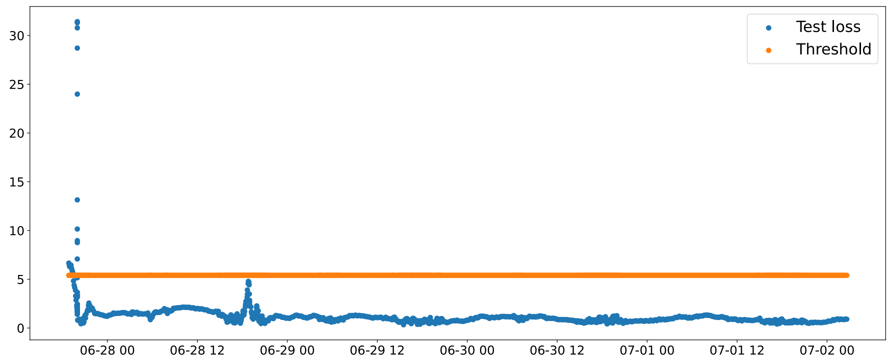


Figure 7.10: The trained LSTM neural network learns a threshold of the reconstruction loss from training data, and determine any sample with the loss higher than this threshold as anomaly

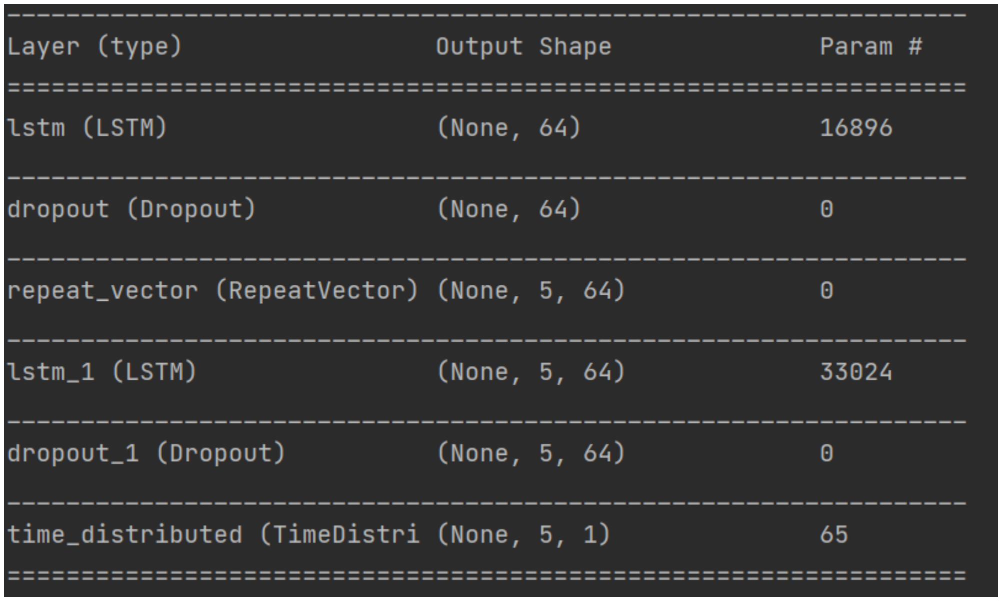
the unexpected rising pressure at least five minutes before the value exceeding the safety threshold. Therefore, the hazard of pressure explosion can be avoided with the detection alert. It allows the engineers to have some time window to react and confine the situation.

We will provide the anomaly interpretation in the next section.

7.3.2 Process Anomalies in Detailed Views

Here we describe a typical process control action issued by operators to Gas Distribution System (GDS) RTU 172, used in the application scenario for automated anomaly detection. We observe the same behavior in many other RTUs. To carry out our analysis, we leverage information regarding the mapping of IOAs and their relation with the physical world from our previous analysis.

As depicted by the timeline in Figure 7.14, the process control operation starts with the SCADA server sending an activation command (CoT: Act) of type single-command to RTU 172 to switch the binary value at IOA 5162 to “on”. The RTU immediately confirms receiving the message (CoT: ActCon) and confirms finishing



```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
lstm (LSTM)                  (None, 64)                16896

_________________________________________________________________
dropout (Dropout)            (None, 64)                0

_________________________________________________________________
repeat_vector (RepeatVector) (None, 5, 64)             0

_________________________________________________________________
lstm_1 (LSTM)                (None, 5, 64)             33024

_________________________________________________________________
dropout_1 (Dropout)          (None, 5, 64)             0

_________________________________________________________________
time_distributed (TimeDistri (None, 5, 1)              65
=================================================================
```

Figure 7.11: Neural network layer structure of the fine-tuned LSTM autoencoder decoder model

the transaction (CoT: ActTerm). We visualize how the single command works in Figure 7.13.

Switching the value to "on", results in a message from the RTU informing the SCADA server that IOA 1066 changed its value to "on" due to a remote control operation (CoT: Retrem). As can be seen in Table 7.8, on the RTU, IOA 5162 and 1066 both reference the same low-pressure-valve regulation indicator, one in IEC 104 command and one in monitoring direction. This is expected as IEC 104

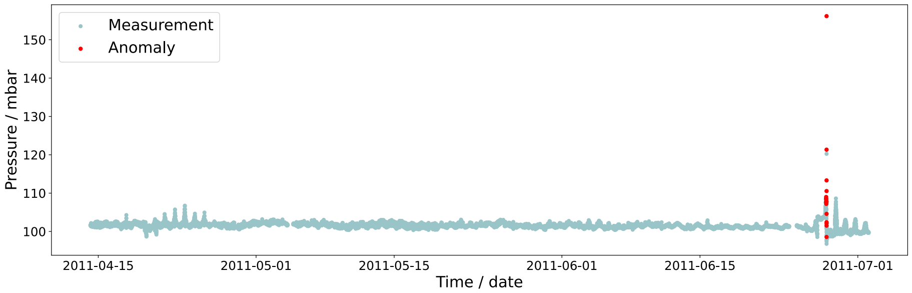


Figure 7.12: Detection result in a pressure sensor

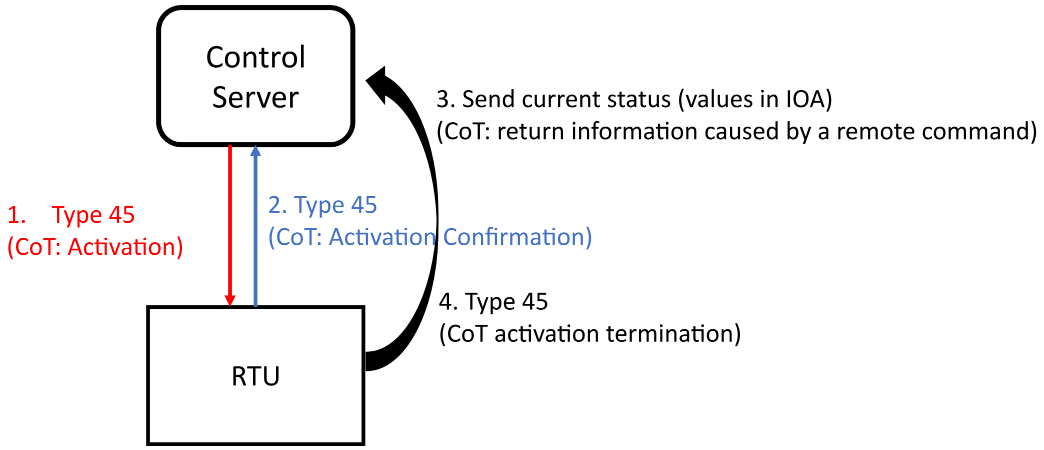

Control Server
3. Send current status (values in IOA)
(CoT: return information caused by a remote command)
1. Type 45
(CoT: Activation)
2. Type 45
(CoT: Activation Confirmation)
4. Type 45
(CoT activation termination)
RTU

Figure 7.13: Legit operation sequences of single command

strictly distinguishes between IOs (and thus IOAs) used in command and monitoring directions. Interestingly, the local control logic of the RTU (CoT: Retloc) switches the low-pressure-valve regulator to "off" immediately after. This is a typical pulse-signal behavior. While the IEC 104 protocol would allow one to specify a pulse behavior directly in a single command, the feature is not used here, and the pulse behavior is implemented with additional, custom control logic, written in the RTU. In response to the initial command, the control motor for the valve is powered up. The motor status is tracked by the RTU, which sends a spontaneous (CoT: Spont) single-point information message to the SCADA Server indicating that IOA 81 (i.e., the motor, see Table 7.8) is now "on". In a second single-point information message, the RTU informs the SCADA Server that IOA 1029, which resembles the binary (on/off) status of the low-pressure-valve regulator is now "on" as well. From the IEC 104 message we know that this status change is caused by a remote control operation (CoT: Retrem), i.e., a direct consequence of the initial command. We then observe IOA 1066 resetting its value to "off" due to a local control operation triggered by the control logic on the RTU (CoT: Retloc). After a few seconds, the sequence ends with the RTU informing the SCADA server

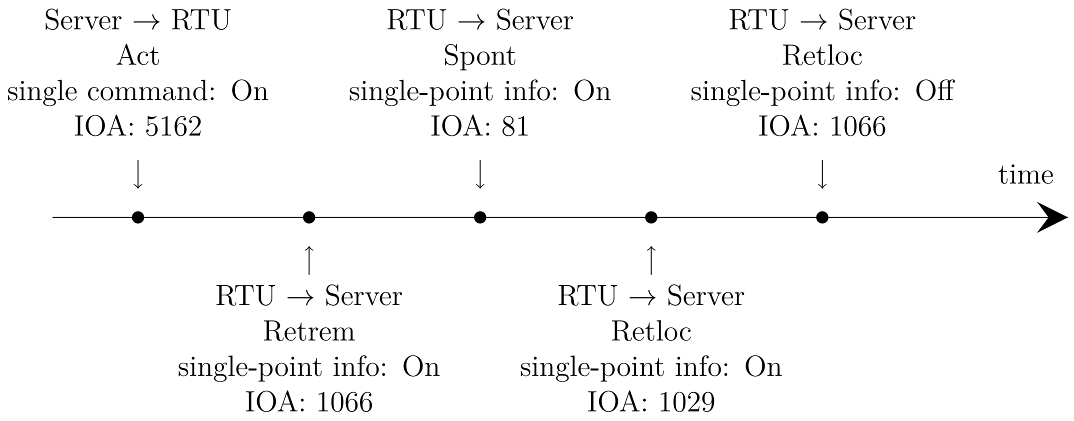


Figure 7.14: Exemplary (and simplified) IEC 104 control message exchange (ignoring ActCon and ActTerm) for RTU 172

Table 7.8: Simplified IOA to local I/O port mapping for RTU 172

Monitoring IOA	Control IOA	Local I/O port	Description*
81	N/A	Di-081	Motor Status
1066	5162	Do-042	Operate Low-Pressure Regulator
8213	N/A	Ai-021	Valve Position Status
1029	5125	Do-005	Low-Pressure Regulator Status

that low-pressure-valve regulation indicator (IOA 1029) and motor (IOA 81) are now “off” due to a local control operation (CoT: Retloc).

These control operations are relatively rare. On RTU 172, for example, this sequence/operation repeats irregularly, with a few exceptions, roughly once a day with gaps in between. We visualize the time sequence of events triggered by the initial control command in Figure 7.15. The time series shows that the immediate response of the RTU (in fact some of the IEC 104 messages are transmitted in the same TCP segment), including the custom implementation of the pulse-signal. A few seconds later (13 seconds in the example in the chart), the motor switches off and the control operation completes.

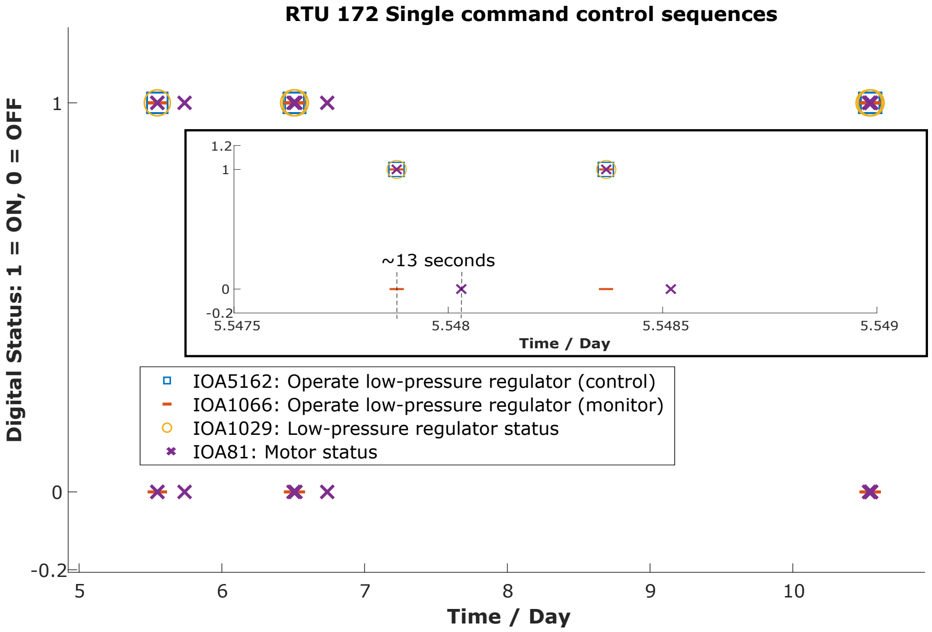


Figure 7.15: The control and monitor sequences between the SCADA server and RTU 172

Normal Behaviors of Gas System Components

After the natural gas is produced from the decomposition of rock formations, it is either compressed or enlarged to fit different volume needs for storage, transmission and distribution.

Lower to Higher Pressure Depending on the the input pressure range, gas regulators have the types of low-pressure ones and the high-pressure ones. The gas compressors accomplish the transformation from lower pressure to higher pressure through the electronic regulation of the gas flow with the low-pressure gas regulators. During the compression process, excessive heat is generated due to the Conservation of energy, and there is a cooler system to dissipate this heat [25].

Higher to Lower Pressure: High-range to Mid-range At the gas expansion turbine station (GET), the high-pressure regulator takes in the national high-pressure gas of tens or hundreds bars (40 bars in our dataset), through heating the gas eventually outputs the regional mid-range pressure (8 bars in our dataset). This process intrinsically avoids the energy loss by collecting the extra kinetic energy and transforming into another form of useful energy, e.g. the energy source for power generators to produce electricity.

Higher to Lower Pressure: Mid-range to Low-range The local distribution station, i.e. a GDS, compresses 1 to 8 Bar mid-pressure gas to the < 1 Bar low-pressure with the high-pressure gas regulators.

All kinds of regulators follow the same basic principle, which is changing the medium's pressure in the regulator cylinder through the interactive forces between an internal spring and a diaphragm. The common components are a spring, a metal or rubber thin plate called the diaphragm, and a valve.

Position sensors apply broadly in oil and gas systems. They are the monitoring tool to ensure the system in a safe and reliable status even in a harsh environment. Direct monitoring objects are valves, actuators, and motors.

Safety Control Ideally, every gas station has an emergency shutdown system connected to automation processes. The operator schedules it to be activated when abnormal events occur, such as the gas having an unanticipated out-of-limit pressure change or a leakage. After detection, depending on which control unit is connected, the emergency system will either shutdown the corresponding unit such as a gas regulator or a motor, or open the vent to release excessive gas or heat.

Anomaly Interpretation

Abnormal Event 1: Pressure Peak in the Distribution Station RTU 172 monitors a distribution station (GDS). Among all the point variables this RTU manages, there is a set of point variables connecting to a mid-pressure gas regulator. The function of this regulator is to convert the 8-bar pressure to the 100-milibar pressure. We identify several crucial points from the Telnet traffic representing the outlet 100-mbar pressure (with IOA 8193) and the 100-mbar corresponding safety thresholds HH/H/L/LL (IOA 9217 to IOA 9220) and the alarm (IOA 88), position sensor for the valve status (IOA 8213) and the corresponding safety variables (IOA 9221 to 9224 for thresholds, IOA 45 for the alarm).

The outlet pressure has a burst reading on the day of June 28th, and the peak was 156.9 mbar comparing to the normal range of 109.4 to 120.4 mbar. With the 100-mbar pressure reading marked as blue crosses, Figure 7.16 shows that this reading has been stable and stayed within the four thresholds until one day in late June (June 28th). During the burst, the pressure first exceeded the H threshold of 120 mbar and then the HH threshold of 130 mbar. Figure 7.17 zooms into the time frame when this occurred as an hourly view. Both digital variables of the valve closed-or-not status and the safety alarm should be a 0/1 value. To avoid overlapping in the plot, we add number two offset to the valve status value, i.e. value two is actually zero and value three is actually two. Until the local time midnight of June 28th, the valve remained open all the time. Starting from 2 am, the valve got closed for around 4 hours 20 minutes (to 6:20 am) from some unknown maintenance, and the closing caused visibly larger fluctuation of the pressure reading. In the middle of this closing, the valve got a brief opening at around 4:59 am, and immediately closed again. Reacting to this brief opening, the pressure reading burst and the safety alarm alerted within 3 seconds. After the valve opening gave the pressure a relief, we can observe that the pressure reading peak dropped to the normal range right away. At the time of this abnormal event, the position reading of the corresponding regulator also had a burst of 100%, exceeding the H threshold of 85% and the HH threshold of 90% in Figure 7.18.

The polling modes of measurement values from the pressure sensor, valve and alarm status are at a frequency of 10-minute by interrogation command.

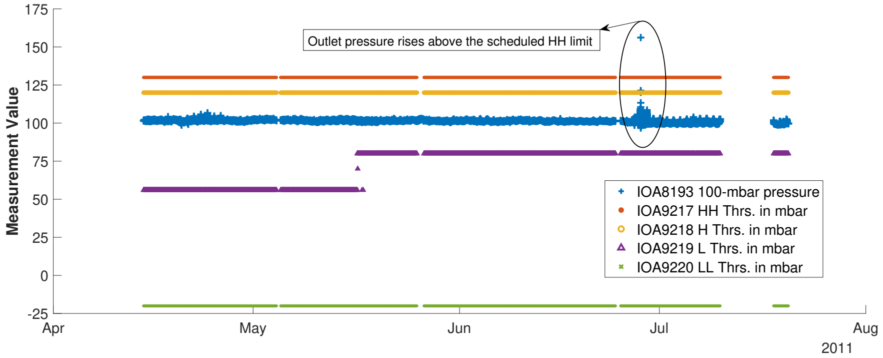


Figure 7.16: Measurement time series of the relevant points in RTU 172 over the whole capture

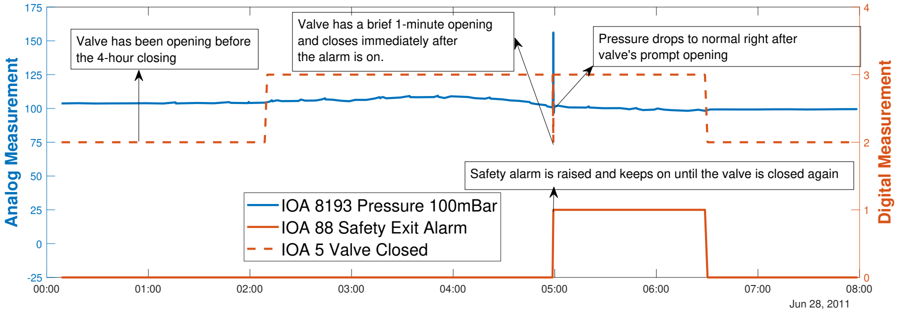


Figure 7.17: Safety exit alarm triggering event

Abnormal Event 2: Pressure Reading Myths in the Expansion Station

RTU 228 monitors a expansion station (GET) where the main task is to convert high pressure of 40 bar to mid-range pressure 8 bar. Therefore, the key point variables are the relevant pressure sensors and the corresponding safety control

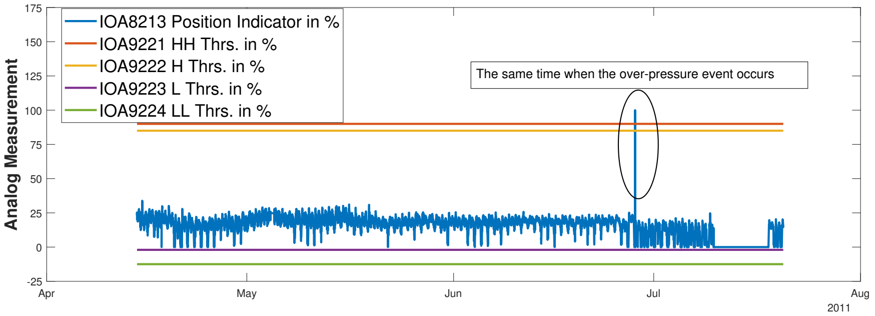


Figure 7.18: Position sensor readings correlate with the over-pressure anomaly

variables. During the whole capture, there are two abnormal events worthy of a deep dive (Figure 7.19.

The first abnormal event on the day of May 5th (Figure 7.20), involves the high-pressure point IOA 8195 and the emergency-stop safety control. The key takeaway is that during this event the system has issued emergency stop out of a series of suspicious faulty reading. Similarly, we compensate an offset of 40 to the digital readings of IOA 1025 for presentation purpose. The signal of emergency stop (the green line) stayed zero during the whole capture except at the moment of sudden increase in the 40-bar variable IOA 8195 (the yellow triangles). Normally, the polling mode of IOA 1025 is periodic at the frequency of 10 minutes by the station's interrogation schedule and the polled value is always zero. However, during this event, two readings of ones have been returned within 15 seconds at 9:15 am as the answer to a remote control command, which is indicated through the cause of transmission (type 11, return information caused by a remote command). Based on our experience, we suspect that before the morning of May 5th, the readings in IOA 8195 of the 40-bar pressure sensor are faulty negative values out of unknown reasons. The system makes the judgement that this is a potential safety threat after a while the pressure reading is still too low, and issues the emergency stop. For example the system can assume that there is a gas leakage going on. Then the emergency stop can interrupt the gas transmission at the moment, and allow the operators to inspect the specific pipelines. After the high-pressure readings come back to normal at around 40 bars, the system assigns no more emergency stop commands.

The second abnormal event in Figure 7.21 is still relevant to the 40-bar pressure readings on June 27th. During this event, the high pressure drops again gradually from 42 bars to 10 bars over 8 hours, and then within one hour it climbs up back to normal. The system hasn't issued any emergency stop during these 9 hours.

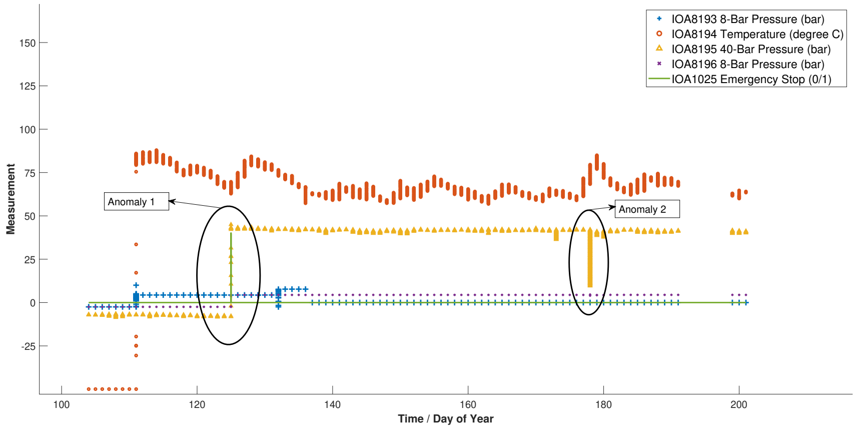


Figure 7.19: Measurement time series of the relevant points in RTU 228 over the whole capture

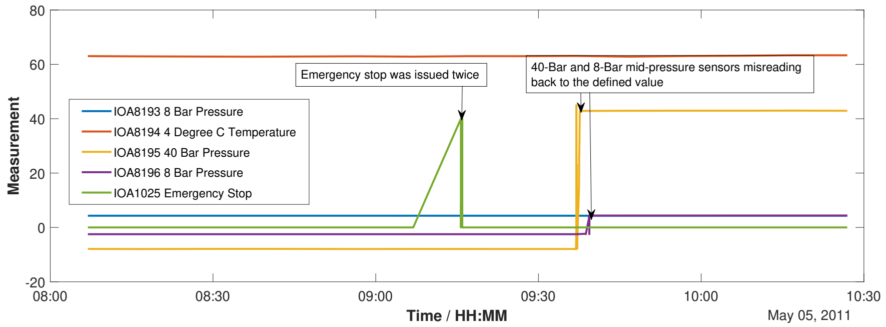


Figure 7.20: System issues emergency stop at the detection of 40-bar pressure sensor reporting negative values

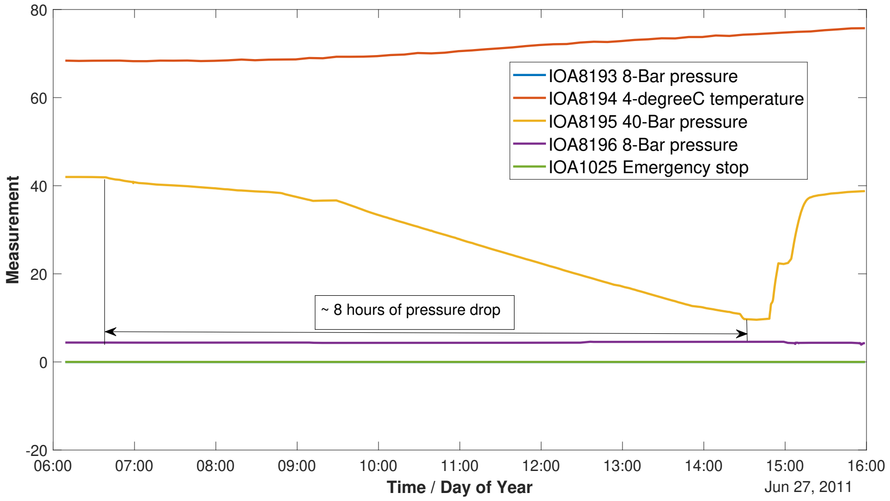


Figure 7.21: High pressure of 40-bar gradually degrades to 10 bars

7.4 Discussion

For critical infrastructures providing the essential utilities, the availability and integrity of the services mostly prevail over the other security principles. Therefore, it is unfeasible to interrupt the physical processes and perform penetration testings. As a result, firewalls and the intrusion/anomaly detection systems are the main defense mechanism deployed in ICS. Specifically, the network intrusion/anomaly detection is the focus of this dissertation. According to the taxonomy given by [21][51] and [34], there are the designs based on predefined attack scenarios (i.e. misuses), signatures, system and protocol specifications and process physics.

With the mechanisms of misuse and signature based, the SCADA network may still be prone to any attacks outside of the predefined detection rules. With the specification-based detector, stealthy attacks that follow the specification rules may still evade. For instance, the adversary can inject a false time series to the sensor in compatible with the protocol standards, still the false data can lead the system to a hazardous state. Therefore, we think the anomaly detection must have the awareness of the physical processes, and the essence of the design should be about how we build profiles from different perspectives, how the detection system identifies a suspicious deviation, how we understand the detection results in the downstream tasks.

Chapter 8

Discussions, Conclusions and Future Work

8.1 Discussions

8.1.1 Experience and Lessons

We are the first to analyze the SCADA traffic in a bulk power grid and gas distribution network, revealing a thorough picture of how operators supervise the physical operations of power grid transmission, distribution, and gas delivery. We'd like to offer our approach in this chapter, which has been validated by our own datasets and has led to fruitful discoveries of unanticipated network misconfigurations, maintenance events, and real-world process anomalies. We'd also like to highlight some of the key lessons we've learned from projects, which may help researchers avoid pitfalls in the future.

First, we want to share our methodology when working with such network captures, and summarize our methodology in Figure 8.1. With this workflow, one can avoid being overwhelmed by all the network services and focus on the Operational Technology network activities.

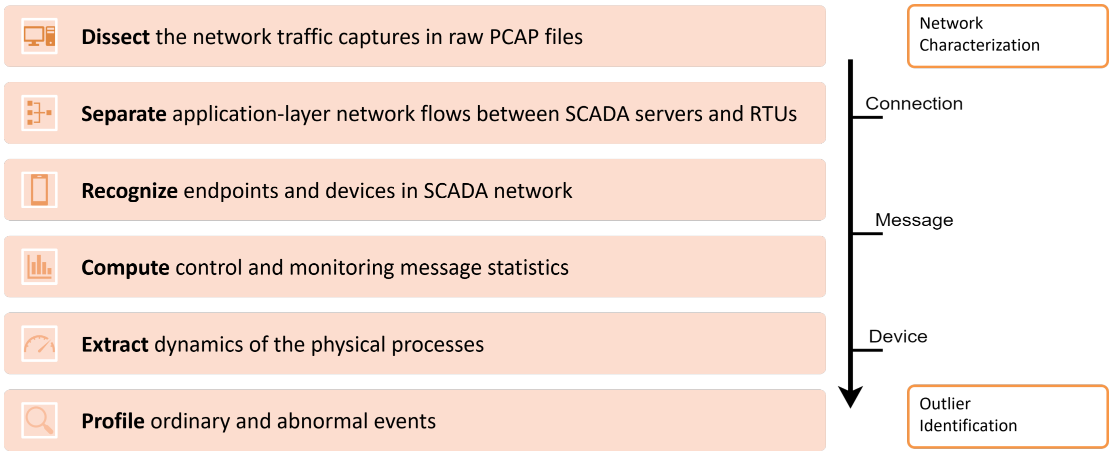

Dissect the network traffic captures in raw PCAP files
Separate application-layer network flows between SCADA servers and RTUs
Recognize endpoints and devices in SCADA network
Compute control and monitoring message statistics
Extract dynamics of the physical processes
Profile ordinary and abnormal events
Network Characterization
Connection
Message
Device
Outlier Identification

Figure 8.1: Our attested approach when working with operational SCADA network traffic datasets

Secondly, do not rely on the network endpoint and device fingerprints based on previous IT network expertise. When we understand that the gas stations are not fingerprinted by IP addresses, but rather by an IEC104 feature, the gas dataset in our study demands us to be attentive to unexpected IP changes.

Thirdly, *READ* the protocol manual. It may appear to be a nuance for network security researchers. Our experience with IEC 104 protocol standards ensures that we will be able to spot interesting anomalies and deduce their origins. An thorough investigation of the protocol itself is a must for understanding the network's normal behavior when designing a monitoring system.

Last but not the least, a ubiquitous metric system is needed for detection evaluation. Supervised learning algorithms can use standard metrics, such as accuracy, precision, recall, etc. However, anomaly detection for these network monitoring datasets mostly does not have ground truth of real attacks, i.e. the labels of 0/1. It is an essential step for the researchers first identify suspicious irregular activities, and then define specific anomalies from there. This is also a limitation in our current research, in which we try to mitigate through defining process anomaly categories based on specific physical processes, and detection performance is measured by whether or not these anomaly categories are discovered. However, generalizing all process abnormalities is difficult, and considering all anomalies at the same level of severity and stealthiness is unfair.

8.1.2 Security, Safety, and Ethics

The dataset was collected as part of a research project aimed to increase cyber security and resiliency of critical infrastructures. Confidentiality was practised when handling the data. We explicitly do not disclose the network operator, the exact geographical location of individual devices, public IP addresses, or any other properties that may put the operator or the infrastructure at risk.

Our analysis shows a trusted insider assumption is used in this SCADA network. The network we study is not part of the Internet (it is a private network on leased lines). Anyone with access to this private network will have open access to all devices in the network. Traffic is not authenticated (or encrypted), and attackers can potentially spoof any device.

The usage of clear text protocols (i.e. IEC 104) is (still) common practise in SCADA networks. The International Electrotechnical Commission (IEC) has published a security specification for IEC 104 in 2013, to provide sender authentication and to ensure the integrity of data units (i.e. APDUs). However, operators tend to be reluctant to upgrade the security features of IEC 104 channels with this release, probably under the consideration for the expenses and interruption to routine operations. Besides most SCADA protocols predating modern best practises in protocol design, according to [27] the introduction of encryption may decrease compatibility, introspection, and monitorability of the network, as well as introduce additional complexity and latency in the control process [27]. In this context we do not consider the use of plain text protocols in a private and controlled network a security issue by itself. Any Man-in-the-Middle attacker could most likely, with or without the credentials gained from parsing Telnet traffic, cause major interruptions on the gas delivery process.

Given this situation, accurate network security monitoring is needed. Any such network monitoring solution must take into account that IEC 104 has a monitoring and a control direction and thus two IOAs that are observed on the network may refer to the same local variable on the RTU. we show several correlations between variables and messages. This type of network profile can help in creating a list of acceptable or expected behavior of the network. In addition to protecting the network, the local control logic on the RTU can actually be a security and safety feature. With a strong control logic on the RTU, an Attack like Industroyer[8] would have less impact as the RTU can locally perform safety-checks to not allow entering invalid or unsafe states regardless of the incoming commands.

8.2 Conclusion

In this dissertation, we conduct the first network measurement and anomaly detection of the IEC 104-based SCADA system controlling a bulk power grid and the natural gas distribution network, both as the large-scale state-level networks spanning over multiple geographic areas. We list and explain several interesting observations with respect to IEC 104 usage and RTU configurations in this gas network.

By combining the information obtained from engineering and IEC 104 network traffic, we reconstruct the bulk and gas distribution systems' layouts, including the type and purpose of the substations and the physical properties of the gas that enters the SCADA system. Our analysis shows that it is possible to extract this information, essential for security monitoring, purely from the raw network traffic and without background knowledge provided by the control system engineers. We also note that configuration changes in SCADA environments, although probably less frequent than in IT environments, are not as rare and exceptional as the research community assumed.

Comparison with other IEC 104 networks and further exploiting context information, such as communication using other protocols within the SCADA control center, are planned to be explored further in follow-up studies.

We observed several differences in usage of the IEC 104 protocol between the gas and the bulk power grid described. Most notably the operators of the gas distribution network do not rely on IEC 104 periodic (e.g., an RTU transmits the current variable in regular time intervals) or spontaneous reports (i.e., an RTU automatically notifies the SCADA server when a value exceeds a defined threshold). Instead, they interrogate the RTUs in constant time intervals (see Figure 5.8). This is in contrast to the IEC 104 observations in the power grid. We believe that the physical process of gas delivery is much more straightforward, with devices mostly being valves/regulators. Because gas "just flows" through the network as long as constant pressure is maintained.

8.3 Future Research Directions

- **Investigation of differences between the use of IEC 104 in different sectors:** While we understand the intuition behind the differences between the power grid and the gas network, it is important to further investigate more system types. The reasons may be the operator's design choices or result from differences in the underlying physical processes. This will benefit the generalization in the SCADA network monitoring.

- **Exploration of the interactions inside of the control room:** Besides the SCADA and Telnet traffic between the RTUs and the SCADA server, gas dataset contains network traffic exchanged within the SCADA control center as well. While the control center traffic is not in the scope of this dissertation, we believe that understanding and describing an operational real-world SCADA control center may be of use to the academic community.

- **Downstream tasks of anomaly detection:** It is tricky to identify anomalies from a dataset without ground truth adversarial event labeled. In our case, we have the remote management traffic in Telnet to facilitate the unraveling of the myth. So one downstream task of anomaly detection can be the parsing of the management and diagnostic traffic.

- **Strengthening the motives for both researchers and industrial operators:** We can already notice an increase in interest in defending critical infrastructures since more critical facilities have been targeted in recent decades. However, both parties' collaboration and motivations are still insufficient. The operators can provide real-world viewpoints on various technical jobs, assisting researchers in developing more accurate threat models and detection models, as well as improving the monitoring system's scalability and real-time running difficulties. On the other hand, after earning the trust of the operator and gaining access to datasets, the researchers construct more robust and resilient intrusion/anomaly detection systems. In gas dataset, for example, we see messages exchanged via proprietary protocols between the two SCADA servers and an HMI.

Table .1: ASDU I-Format Type Identification Codes and Semantics

Type ID Code	Acronym	Description
1	`M_SP_NA_1`	Single-point information
3	`M_DP_NA_1`	Double-point information
5	`M_ST_NA_1`	Step position information
7	`M_BO_NA_1`	Bitstring of 32 bits
9	`M_ME_NA_1`	Measured value, normalized value
11	`M_ME_NB_1`	Measured value, scaled value
13	`M_ME_NC_1`	Measured value, short floating point number
15	`M_IT_NA_1`	Integrated totals
20	`M_PS_NA_1`	Packed single-point information with status change detection
21	`M_ME_ND_1`	Measured value, normalized value without quality descriptor
30	`M_SP_TB_1`	Single-point information with time tag CP56Time2a
31	`M_DP_TB_1`	Double-point information with time tag CP56Time2a
32	`M_ST_TB_1`	Step position information with time tag CP56Time2a
33	`M_BO_TB_1`	Bitstring of 32 bit with time tag CP56Time2a
34	`M_ME_TD_1`	Measured value, normalized value with time tag CP56Time2a
35	`M_ME_TE_1`	Measured value, scaled value with time tag CP56Time2a
36	`M_ME_TF_1`	Measured value, short floating point number with time tag CP56Time2a
37	`M_IT_TB_1`	Integrated totals with time tag CP56Time2a
38	`M_EP_TD_1`	Event of protection equipment with time tag CP56Time2a
39	`M_EP_TE_1`	Packed start events of protection equipment with time tag CP56Time2a
40	`M_EP_TF_1`	Packed output circuit information of protection equipment with time tag CP56Time2a
45	`C_SC_NA_1`	Single command
46	`C_DC_NA_1`	Double command
47	`C_RC_NA_1`	Regulating step command
48	`C_SE_NA_1`	Set point command, normalized value
49	`C_SE_NB_1`	Set point command, scaled value
50	`C_SE_NC_1`	Set point command, short floating point number
51	`C_BO_NA_1`	Bitstring of 32 bits
58	`C_SC_TA_1`	Single command with time tag CP56Time2a
59	`C_DC_TA_1`	Double command with time tag CP56Time2a
60	`C_RC_TA_1`	Regulating step command with time tag CP56Time2a
61	`C_SE_TA_1`	Set point command, normalized value with time tag CP56Time2a
62	`C_SE_TB_1`	Set point command, scaled value with time tag CP56Time2a
63	`C_SE_TC_1`	Set point command, short floating-point number with time tag CP56Time2a
64	`C_BO_TA_1`	Bitstring of 32 bits with time tag CP56Time2a
70	`M_EI_NA_1`	End of initialization
100	`C_IC_NA_1`	Interrogation command
101	`C_CI_NA_1`	Counter interrogation command
102	`C_RD_NA_1`	Read command
103	`C_CS_NA_1`	Clock synchronization command
105	`C_RP_NA_1`	Reset process command
107	`C_TS_TA_1`	Test command with time tag CP56Time2a
110	`P_ME_NA_1`	Parameter of measured value, normalized value
111	`P_ME_NB_1`	Parameter of measured value, scaled value
112	`P_ME_NC_1`	Parameter of measured value, short floating-point number
113	`P_AC_NA_1`	Parameter activation
120	`F_FR_NA_1`	File ready
121	`F_SR_NA_1`	Section ready
122	`F_SC_NA_1`	Call directory, select file, call file, call section
123	`F_LS_NA_1`	Last section, last segment
124	`F_AF_NA_1`	Ack file, ack section
125	`F_SG_NA_1`	Segment
126	`F_DR_TA_1`	Directory
127	`F_SC_NB_1`	Query Log, Request archive file

Bibliography

[1] How much carbon dioxide is produced when different fuels are burned? https://www.eia.gov/tools/faqs/faq.php?id=73&t=11. Accessed: 2021-10-27.

[2] Ieee standard for electric power systems communications-distributed network protocol (dnp3). *IEEE Std 1815-2012 (Revision of IEEE Std 1815-2010)*, pages 1–821, 2012.

[3] U.S. Energy Information Administration. Natural gas explained natural gas pipelines, 2021.

[4] Hermine N Akouemo and Richard J Povinelli. Probabilistic anomaly detection in natural gas time series data. *International Journal of Forecasting*, 32(3):948–956, 2016.

[5] Ehab Al-Shaer, Qi Duan, and Jafar Haadi Jafarian. Random host mutation for moving target defense. In Angelos D. Keromytis and Roberto Di Pietro, editors, *Security and Privacy in Communication Networks*, pages 310–327, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.

[6] American Gas Association. How Does the Natural Gas Delivery System Work?, 2021.

[7] Sajjad Amini, Fabio Pasqualetti, and Hamed Mohsenian-Rad. Dynamic load altering attacks against power system stability: Attack models and protection schemes. *IEEE Transactions on Smart Grid*, 9(4):2862–2872, 2016.

[8] ESET Anton Cherepanov. Win32/industroyer a new threat for industrial control systems, 2017.

[9] Spyros Antonatos, Periklis Akritidis, Evangelos P Markatos, and Kostas G Anagnostakis. Defending against hitlist worms using network address space randomization. *Computer Networks*, 51(12):3471–3490, 2007.

[10] M. Atighetchi, P. Pal, F. Webber, and C. Jones. Adaptive use of network-centric mechanisms in cyber-defense. In *Sixth IEEE International Symposium*

on Object-Oriented Real-Time Distributed Computing, 2003., pages 183–192, 2003.

[11] A. Baiocco and S. D. Wolthusen. Indirect Synchronisation Vulnerabilities in the IEC 60870-5-104 Standard. In *2018 IEEE PES Innovative Smart Grid Technologies Conference Europe (ISGT-Europe)*, pages 1–6. IEEE, 2018.

[12] Rafael Barbosa, Ramin Sadre, and Aiko Pras. Flow whitelisting in scada networks. *International Journal of Critical Infrastructure Protection*, 6:150–158, 12 2013.

[13] Rafael Ramos Regis Barbosa, Ramin Sadre, and Aiko Pras. A first look into scada network traffic. In *Proceedings of the 2012 IEEE Network Operations and Management Symposium*, pages 518–521, Maui, HI, USA, 2012. IEEE.

[14] Shameek Bhattacharjee, Aditya Thakur, and Sajal K. Das. Towards fast and semi-supervised identification of smart meters launching data falsification attacks. In *Proceedings of the 2018 on Asia Conference on Computer and Communications Security*, ASIACCS '18, page 173–185, New York, NY, USA, 2018. Association for Computing Machinery.

[15] Christopher Bing and Stephanie Kelly. Cyber attack shuts down u.s. fuel pipeline 'jugular,' biden briefed, 2021.

[16] BP. Statistical Review of World Energy 2020. Technical Report 69, BP plc, 2020.

[17] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794, New York, NY, USA, 2016. Association for Computing Machinery.

[18] Maximilian Christ, Nils Braun, Julius Neuffer, and Andreas W. Kempa-Liehr. Time series feature extraction on basis of scalable hypothesis tests (tsfresh – a python package). *Neurocomputing*, 307:72–77, 2018.

[19] George Constable and Bob Somerville, editors. *A Century of Innovation: Twenty Engineering Achievements that Transformed our Lives*. The National Academies Press, Washington, DC, 2003.

[20] Council of European Energy Regulators. 6th CEER benchmarking report on the quality of electricity and gas supply. Technical Report 6, Council of European Energy Regulators, 2016.

[21] Hervé Debar, Marc Dacier, and Andreas Wespi. Towards a taxonomy of intrusion-detection systems. *Computer networks*, 31(8):805–822, 1999.

[22] Jan Diettrich, Wided Medjroubi, and Adam Pluta. SciGRID_gas IGGI.

[23] EIA. U.s. energy-related carbon dioxide emissions, 2019, 9 2020.

[24] Sandro Etalle. From Intrusion Detection to Software Design. In Simon N. Foley, Dieter Gollmann, and Einar Snekkenes, editors, *Computer Security – ESORICS 2017*, volume 10492 of *Lecture Notes in Computer Science*, pages 1–10. Springer International Publishing.

[25] Penn State Extension. Understanding natural gas compressor stations, 2015.

[26] Nicolas Falliere, Liam O Murchu, and Eric Chien. W32.stuxnet dossier, 02 2011.

[27] Davide Fauri, Bart de Wijs, Jerry den Hartog, Elisa Costante, Emmanuele Zambon, and Sandro Etalle. Encryption in ics networks: A blessing or a curse? In *2017 IEEE International Conference on Smart Grid Communications (SmartGridComm)*, pages 289–294, 2017.

[28] Cheng Feng, Tingting Li, and Deeph Chana. Multi-Level Anomaly Detection in Industrial Control Systems via Package Signatures and LSTM Networks. In *2017 47th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*, pages 261–272. IEEE, 06 2017.

[29] David Formby, Sang Shin Jung, John Copeland, and Raheem Beyah. An empirical study of tcp vulnerabilities in critical power system devices. In *Proceedings of the 2Nd Workshop on Smart Energy Grid Security*, SEGS '14, pages 39–44, New York, NY, USA, 2014. ACM.

[30] David Formby, Preethi Srinivasan, Andrew Leonard, Jonathan Rogers, and Raheem Beyah. Who's in Control of Your Control System? Device Fingerprinting for Cyber-Physical Systems. In *Proceedings 2016 Network and Distributed System Security Symposium*, San Diego, CA, 2016. Internet Society.

[31] David Formby, Anwar Walid, and Raheem Beyah. A case study in power substation network dynamics. *Proc. ACM Meas. Anal. Comput. Syst.*, 1(1):19:1–19:24, June 2017.

[32] Karl Pearson F.R.S. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.

[33] Song Fu, Shisheng Zhong, Lin Lin, and Minghang Zhao. A re-optimized deep auto-encoder for gas turbine unsupervised anomaly detection. *Engineering Applications of Artificial Intelligence*, 101:104199, 2021.

[34] Jairo Giraldo, David Urbina, Alvaro Cardenas, Junia Valente, Mustafa Faisal, Justin Ruths, Nils Ole Tippenhauer, Henrik Sandberg, and Richard Candell. A survey of physics-based attack detection in cyber-physical systems. *ACM Comput. Surv.*, 51(4), jul 2018.

[35] Cyril Goutte, Peter Toft, Egill Rostrup, Finn Å. Nielsen, and Lars Kai Hansen. On clustering fmri time series. *NeuroImage*, 9(3):298 – 310, 1999.

[36] Andy Greenberg. *Sandworm: A New Era of Cyberwar and the Hunt for the Kremlin's Most Dangerous Hackers*. Doubleday, 2019.

[37] Andy Greenberg. The colonial pipeline hack is a new extreme for ransomware, 2021.

[38] Brendan D. Gregg. chaosreader, 2014.

[39] Ersi Hodo, Stepan Grebeniuk, Henri Ruotsalainen, and Paul Tavolato. Anomaly Detection for Simulated IEC-60870-5-104 Trafiic. In *Proceedings of the 12th International Conference on Availability, Reliability and Security*, ARES '17, pages 100:1–100:7, New York, NY, USA, 2017. ACM.

[40] International Electrotechnical Commission (IEC). Power systems management and associated information exchange - Data and communications security - Part 5: Security for IEC 60870-5 and derivatives.

[41] International Electrotechnical Commission (IEC). Telecontrol equipment and systems - Part 5-101: Transmission protocols - Companion standard for basic telecontrol tasks. Technical report, International Electrotechnical Commission (IEC), 02 2003.

[42] International Electrotechnical Commission (IEC). Telecontrol equipment and systems - Part 5-104: Transmission protocols - Network access for IEC 60870-5-101 using standard transport profiles. Technical report, International Electrotechnical Commission (IEC), 06 2006.

[43] International Electrotechnical Commission (IEC). Telecontrol equipment and systems - Part 5-7: Transmission protocols - Security extensions to IEC 60870-5-101 and IEC 60870-5-104 protocols (applying IEC 62351). Technical report, International Electrotechnical Commission (IEC), 07 2013.

[44] International Energy Agency (IEA). Gas – Fuels & Technologies, 2021.

[45] Celine Irvene, Tohid Shekari, David Formby, and Raheem Beyah. If i knew then what i know now: On reevaluating dnp3 security using power substation traffic. In *Proceedings of the Fifth Annual Industrial Control System Security (ICSS) Workshop*, pages 48–59, 2019.

[46] Celine Irvene, Tohid Shekari, David Formby, and Raheem Beyah. If I Knew Then What I Know Now: On Reevaluating DNP3 Security using Power Substation Traffic. In *Proceedings of the Fifth Annual Industrial Control System Security (ICSS) Workshop*, pages 48–59, San Juan PR USA, December 2019. ACM.

[47] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: A review. *ACM Comput. Surv.*, 31(3):264–323, September 1999.

[48] Dorene Kewley, Russ Fink, John Lowry, and Mike Dean. Dynamic approaches to thwart adversary intelligence gathering. In *Proceedings DARPA Information Survivability Conference and Exposition II. DISCEX'01*, volume 1, pages 176–185. IEEE, 2001.

[49] Tung Kieu, Bin Yang, Chenjuan Guo, and Christian S Jensen. Outlier detection for time series with recurrent autoencoder ensembles. In *IJCAI*, pages 2725–2732, 2019.

[50] Gaoqi Liang, Junhua Zhao, Fengji Luo, Steven R Weller, and Zhao Yang Dong. A review of false data injection attacks against modern power systems. *IEEE Transactions on Smart Grid*, 8(4):1630–1638, 2016.

[51] Hung-Jen Liao, Chun-Hung Richard Lin, Ying-Chih Lin, and Kuang-Yuan Tung. Intrusion detection system: A comprehensive review. *Journal of Network and Computer Applications*, 36(1):16–24, 2013.

[52] Chih-Yuan Lin and Simin Nadjm-Tehrani. Understanding iec-60870-5-104 traffic patterns in scada networks. In *Proceedings of the 4th ACM Workshop on Cyber-Physical System Security*, CPSS '18, pages 51–60, New York, NY, USA, 2018. ACM.

[53] Hui Lin, Jia-Ning Zhuang, and Yih-Chun Hu. Defrec: Establishing physical function virtualization to disrupt reconnaissance of power grids' cyber-physical infrastructures. 2020.

[54] Yao Liu, Peng Ning, and Michael K. Reiter. 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, page 21–32, New York, NY, USA, 2009. Association for Computing Machinery.

[55] Yao Liu, Peng Ning, and Michael K. Reiter. False data injection attacks against state estimation in electric power grids. *ACM Trans. Inf. Syst. Secur.*, 14(1):13:1–13:33, June 2011.

[56] Kelvin Mai, Xi Qin, Neil Ortiz Silva, and Alvaro A Cardenas. Iec 60870-5-104 network characterization of a large-scale operational power grid. In *2019 IEEE Security and Privacy Workshops (SPW)*, pages 236–241. IEEE, 2019.

[57] Kelvin Mai, Xi Qin, Neil Ortiz Silva, Jason Molina, and Alvaro A. Cárdenas. Uncharted Networks: A First Measurement Study of the Bulk Power System. In *IMC '20: ACM Internet Measurement Conference, Virtual Event, USA, October 27-29, 2020*, pages 201–213. ACM, 2020.

[58] Petr Matoušek. Description and analysis of iec 104 protocol. Technical report, 12 2017.

[59] Peter Maynard, Kieran McLaughlin, and Berthold Haberler. Towards understanding man-in-the-middle attacks on iec 60870-5-104 scada networks. In *ICS-CSR*, 2014.

[60] Ali H. Mirza and Selin Cosan. Computer network intrusion detection using sequential lstm neural networks autoencoders. In *2018 26th Signal Processing and Communications Applications Conference (SIU)*, pages 1–4, 2018.

[61] Thomas Morris and Wei Gao. Industrial Control System Traffic Data Sets for Intrusion Detection Research. In Eduardo Bayro-Corrochano and Edwin Hancock, editors, *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, volume 8827 of *Lecture Notes in Computer Science*, pages 65–78. Springer International Publishing.

[62] Netbeheer Nederland. Basisinformatie over energie-infrastructuur. Technical report, Netbeheer Nederland, 2019.

[63] Department of Energy. Building a better grid initiative to upgrade and expand the nation's electric transmission grid to support resilience, reliability, and decarbonization, 2022.

[64] Modbus Organization. *MODBUS Messaging on TCP/IP Implementation Guide: V1. 0b*. Modbus Organization, 2006.

[65] Xi Qin, Martin Rosso, Alvaro A Cardenas, Sandro Etalle, Jerry den Hartog, and Emmanuele Zambon. You can't protect what you don't understand: Characterizing an operational gas scada network.

[66] ESET Research. Industroyer2: Industroyer reloaded, 2022.

[67] Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53 – 65, 1987.

[68] Luis Salazar, Neil Ortiz, Xi Qin, and Alvaro A Cardenas. Towards a high-fidelity network emulation of iec 104 scada systems. In *Proceedings of the 2020 Joint Workshop on CPS&IoT Security and Privacy*, pages 3–12, 2020.

[69] Congressional Research Service. Pipeline cybersecurity: Federal programs, 2021.

[70] Tohid Shekari, Celine Irvene, Alvaro A. Cardenas, and Raheem Beyah. Mamiot: Manipulation of energy market leveraging high wattage iot botnets. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, CCS '21, page 1338–1356, New York, NY, USA, 2021. Association for Computing Machinery.

[71] Saleh Soltan, Prateek Mittal, and H. Vincent Poor. BlackIoT: IoT botnet of high wattage devices can disrupt the power grid. In *27th USENIX Security Symposium (USENIX Security 18)*, pages 15–32, Baltimore, MD, August 2018. USENIX Association.

[72] Rui Tan, Varun Badrinath Krishna, David KY Yau, and Zbigniew Kalbarczyk. Impact of integrity attacks on real-time pricing in smart grids. In *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security*, pages 439–450, 2013.

[73] Marcio Andrey Teixeira, Tara Salman, Maede Zolanvari, Raj Jain, Nader Meskin, and Mohammed Samaka. SCADA system testbed for cybersecurity research using machine learning approach. *Future Internet*, 10(8):76, 2018.

[74] Robert L. Thorndike. Who belongs in the family? *Psychometrika*, 18(4):267–276, Dec 1953.

[75] Wonkee Donkee Tools. How does a gas regulator work?, 2021.

[76] Kris Villez, Venkat Venkatasubramanian, Humberto Garcia, Craig Rieger, Tim Spinner, and Raghunathan Rengaswamy. Achieving resilience in critical infrastructures: A case study for a nuclear power plant cooling loop. In *2010 3rd International Symposium on Resilient Control Systems*, pages 49–52, 2010.

[77] Zisheng Wang and Rick S. Blum. Topology attack detection in natural gas delivery networks. In *2019 53rd Annual Conference on Information Sciences and Systems (CISS)*, pages 1–6, 2019.

[78] Joe H Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244, 1963.

[79] Joe Weiss. Aurora generator test. *Handbook of SCADA/Control Systems Security*, page 107, 2016.

[80] J. E. White. A User TELNET Description of an Initial Implementation. Internet Requests for Comments.

[81] C. Wressnegger, A. Kellner, and K. Rieck. Zoe: Content-based anomaly detection for industrial control systems. In *2018 48th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*, pages 127–138, June 2018.

[82] Y. Yang, K. McLaughlin, T. Littler, S. Sezer, B. Pranggono, and H. F. Wang. Intrusion detection system for iec 60870-5-104 based scada networks. In *2013 IEEE Power Energy Society General Meeting*, pages 1–5, July 2013.

[83] Y. Yang, K. McLaughlin, S. Sezer, Y. B. Yuan, and W. Huang. Stateful intrusion detection for iec 60870-5-104 scada security. In *2014 IEEE PES General Meeting | Conference Exposition*, pages 1–5, July 2014.

[84] Zong-Han Yu and Wen-Long Chin. Blind false data injection attack using pca approximation method in smart grid. *IEEE Transactions on Smart Grid*, 6(3):1219–1226, 2015.

[85] Fareed Zakaria. We're headed for a global energy crisis. what we need is a transition strategy, 10 2021.