UNIVERSITY OF CALIFORNIA
SANTA CRUZ

REAL-WORLD INSIGHTS INTO SCADA TRAFFIC: A CROSS-INFRASTRUCTURE NETWORK MEASUREMENT ANALYSIS

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

DOCTOR OF PHILOSOPHY

in

COMPUTER SCIENCE

by

Neil Anderson Ortiz Silva

December 2023

The Dissertation of Neil Anderson Ortiz Silva
is approved:

Professor Alvaro Cardenas, Chair

Professor Chen Qian

Professor Avishai Wool

Peter Biehl
Vice Provost and Dean of Graduate Studies

Table of Contents

List of Figures

List of Tables

Abstract

Real-World Insights into SCADA Traffic: A Cross-Infrastructure Network Measurement Analysis

Neil Anderson Ortiz Silva

In recent years, an increasing number of attacks have targeted Industrial Control Systems (ICS) worldwide, exposing the fragility of these systems. Understanding the SCADA networks that govern critical infrastructures is increasingly vital to protect this system. However, the confidential nature of ICS data typically restricts access to the real world, limiting efforts of academic research for more realistic studies. While previous studies have focused on some isolated network characteristics in a single infrastructure, none have taken a comparative approach across multiple critical infrastructures and multiple industrial protocols.

Aiming to fill this gap, our research dissects operational SCADA networks across multiple ICS based on real-world data. This study focused on network measurement of SCADA traffic in two ways: (1) between distinct ICS such as power, gas, and water, and (2) among the subsystems in the power grid from generation to end-customer grids.

Our analysis reveals distinct and shared behaviors of these networks, providing insight into their network behavior and configuration. It also reveals non-standard configurations, protocol operation characteristics, topology configurations, and considerable variations in periodic traffic patterns, high packet transmission rates, and message types. These insights contribute to a more realistic understanding of SCADA networks, challenging previous assumptions and emphasizing the existence of substantial diversity in SCADA traffic within these infrastructures. Our findings underscore the need for a specialized approach tailored to each critical

infrastructure and open the door for better network characterization for cybersecurity measures and more accurate designs in intrusion detection systems.

Chapter 1

Introduction

Supervisory Control and Data Acquisition (SCADA) systems represent the technology used to monitor and control remote large-scale physical processes such as the power grid, gas distribution, and water treatment. These systems manage several Industrial Control Systems (ICS) so vital to society that their incapacitation or malfunction could have a debilitating effect on national security, the economy, and public health.

Despite their criticality to our modern way of life, these networks have received limited attention from the academic measurement community. SCADA systems have migrated from serial communications to IP-based and Ethernet networks in the past two decades, and they can be analyzed with the same tools we have developed for measuring other modern networks.

Previous network measurements of SCADA networks focus on a single network using a single industrial protocol. As a result, previous work studies real-world SCADA networks in isolation, one at a time, rather than as a unified whole. Consequently, results and broad generalizations about SCADA networks in these previous studies may not represent the networks in other infrastructures. To address these concerns, in this study, (1) we define a taxonomy of the SCADA

networks and industrial protocols. (2) We examine data captured from three different operational facilities in different physical infrastructures: the power grid, gas distribution, and water treatment. Then, (3) we zoom in the SCADA networks in power grid by examing data capture from a System Operator, Transmission, Distribution and End-Consumer grids. We aim to systematically study the diversity of SCADA networks and identify similarities and differences in various infrastructures.

This thesis is based on these three publications in which I am the main author:

- ***"A Taxonomy of Industrial Control Protocols and Networks in the Power Grid"***, IEEE Communications Magazine (Volume: 61, Issue: 6), **Neil Ortiz**, Alvaro Cardenas, Avishai Wool,. June 2023. `https://ieeexplore.ieee.org/document/10155642`

- ***"From Power to Water: Dissecting SCADA Networks Across Different Critical Infrastructures"***, The Passive and Active Measurement (PAM) conference, Virtual Event, **Neil Ortiz**, Martin Rosso, Emmanuele Zambon-Mazzocato, Jerry den Hartog, Alvaro Cardenas,. March 2024. ***Paper submitted***.

- ***"SCADA World: An Exploration of the Diversity in Power Grid Networks"***, ACM SIGMETRICS / IFIP Performance conference, Venice-Italy, **Neil Ortiz**, Alvaro Cardenas, Avishai Wool,. June 2024. ***Paper submitted***.

In addition to these, I have contributed to the following publications.

SCADA traffic publications

- ***"Cybersecurity and Resilience for the Power Grid"***, Book: Resilient Control Architectures and Power Systems (IEEE Press Series on Power and Energy System). Xi Qin, Kelvin Mai, **Neil Ortiz**, Keerthi Korenu, Alvaro Cardenas,. Dec 2021. `https://ieeexplore.ieee.org/abstract/document/9646372`.

- ***"Towards a High-Fidelity Network Emulation of IEC 104 SCADA Systems."*** In Proceedings of the 2020 Joint Workshop on CPS&IoT Security and Privacy, 3–12. Virtual Event USA: ACM, 2020. Salazar Luis, **Neil Ortiz**, Xi Qin, and Alvaro A. Cardenas. `https://doi.org/10.1145/3411498.3419969`.

- ***"Uncharted Networks: A First Measurement Study of the Bulk Power System."*** In Proceedings of the ACM Internet Measurement Conference, 201–13. Virtual Event USA: ACM, 2020. Mai, Kelvin, Xi Qin, **Neil Ortiz**, Jason Molina, and Alvaro A. Cardenas. `https://doi.org/10.1145/3419394.3423630`.

- ***"IEC 60870-5-104 Network Characterization of a Large-Scale Operational Power Grid".*** In 2019 IEEE Security and Privacy Workshops (SPW). IEEE, 236–241. Kelvin Mai, Xi Qin, **Neil Ortiz**, and Alvaro A Cardenas. `https://doi.org/10.1109/SPW.2019.00051`

Tower attacks publications

- ***"Using Hotspot Analysis to Prioritize Security Efforts in Colombian Critical Infrastructure, a Focus on the Power Grid"*** Security Journal, June 7, 2021. Mendizabal, Agustin Palao, Jennifer S. Holmes, Mer-

cedez Callenes, **Neil Ortiz**, and Alvaro Cardenas. https://doi.org/10.1057/s41284-021-00300-7.

- ***"A Hotspot Analysis of Critical Hydrocarbons Infrastructure in Colombia: ELN (Ejército de Liberación Nacional) and FARC (Fuerzas Armadas Revolucionarias de Colombia) Attacks on Colombian Pipelines"*** Applied Geography 126 (January 2021): 102376. Mendizabal, Agustin Palao, Jennifer S. Holmes, **Neil Ortiz**, Mercedez Callenes, and Alvaro Cardenas. https://doi.org/10.1016/j.apgeog.2020.102376.

Chapter 2

Related Works

SCADA systems provide a human operator with real-time updates on the current state of the monitored remote process and give operators the capacity to control it remotely. To do this, SCADA systems require a wide-area telecommunications infrastructure and rely on industrial protocols to establish a predetermined message format and set of messages and responses.

Although modern SCADA systems use Internet-compatible protocols, the network measurement research community has largely overlooked these networks. One of the reasons is that companies that manage critical infrastructures, such as power grids, are very cautious about allowing outsiders to access their internal networks. Consequently, most of the published research related to SCADA networks has been based on simulations and testbeds.

Due to the difficulty of obtaining real-world data from operational ICS, analyzing emulated or simulated data is the most popular research approach. This line of work includes a testbed at SLN [13], simulated Modbus networks, or emulated C37.118 networks through hardware-in-the-loop simulation [50], or simulated GOOSE network [11]. However, simulations or testbeds do not represent real behaviors in industrial control network: as Lian and Nadjim-Terani's power grid

study showed [35], emulated datasets are prone to simple and regular patterns.

Some papers study operational ICS, but do not provide details of the system under study, such as protocol, type of supervised process, or type of infrastructure. Without this context information is not posible to know the specific challenges and nuances inherent to different SCADA environments. This gap in detailed, real-world data limits the broader applicability and depth of understanding that can be derived from these studies, underscoring the need for more comprehensive and research in operational ICS.

For example, Yang *et al.* [54] captured network traffic data from a real-world IEC 104 system without adding details of the type of system they analyzed, that is, whether they are from a transmission or distribution system. Likewise, Hoyos *et al.* [28] and Wressnegger *et al.* [53] indicate that their dataset network comes from a power plant, but they do not specify which network protocols are used or add any details of the network topology. Similarly, Jung *et al.* [32] analyzes the TCP connections of a distribution network without specifying protocols.

The works most closely related to ours are the 2022 PAM publication by Mehner *et al.* [44], the 2020 IMC publication by Mai *et al.* [41], and the Sigmetrics 2017 publication by Formby *et al.* [21].

Mehner *et al.* conducted a network characterization study in a Distributed Control System in a power utility. They examined packet traffic at the network layer, focusing on the field, control, and HMI levels. At the field and control levels, most traffic was from a proprietary protocol, while at the supervisory level, there was no legacy ICS protocol. This is distinct from our work, which analyzed IEC 104 at the HMI level (Power), IEC 104 at the control level (Gas), and Modbus/TCP at the field level (water).

Mai *et al.* conducted an analysis of IEC 104 traffic from a real-world bulk

power grid. They characterized traffic at the network, transport, and application level, including topology configuration, TCP flows, IEC 104 message types, and measurement and control commands. Their findings showed topological changes from one year to the next, with 90% of TCP connections lasting less than one second, as well as non-standard IEC 104 packet configurations. This research focused only on one protocol in one part of the power grid, while our study covers a wide range of protocols in all parts of the power grid.

All of these works study a relatively small component of a ICS and most of them focus on a singular industrial protocol. In contrast, our study (as illustrated in Table 2.1) aims to comprehensively study from different ICS.

In conclusion, previous research has often been constrained either by the type of data available (simulated, emulated, or lacking in detail) or by focusing on a narrow aspect of the system, such as a single protocol. This has resulted in a fragmented understanding of SCADA networks, particularly in how they operate across multiple ICS and with various industrial protocols. Our study addresses this gap by providing a comprehensive network measurement analysis of SCADA networks in a wide spectrum of ICS. It encompasses three different types of industrial process (water, gas, and power), multiple networks in the power grid (from generation to end-customer grids), and five industrial protocols. By using real-world traffic data, our research offers a more holistic and detailed view of SCADA system operations, contributing to a deeper understanding of these critical infrastructure networks.

Table 2.1: Related works Table

	[44]	[9]	[13]	[28]	[32]	[19]	[21]	[53]	[12]	[41]	[35]	[50]	[47]	**Our work**
Implementation														
Simulation	–	–	–	◐	–	–	–	–	●	–	●	–	–	–
Testbed	–	–	●	–	–	–	–	–	–	–	–	●	–	–
Real-World	●	●	–	–	●	●	●	●	–	–	●	–	●	●
Protocol														
IEC 104	–	–	–	–	–	–	–	–	–	●	●	–	●	●
MODBUS	–	–	●	–	–	●	–	–	–	–	–	–	–	●
C37.118	–	–	–	–	–	–	–	–	–	–	–	●	–	●
ICCP	–	–	–	–	–	–	–	–	–	–	–	–	–	●
GOOSE	–	–	–	●	–	–	–	–	●	–	–	–	–	●
DNP3	–	–	–	–	–	–	●	–	–	–	–	–	–	–
Unspecified	●	●	–	–	◐	–	●	◐	–	–	–	–	–	–
Infrastructure														
Power (Generation)	●	–	–	◐	–	–	–	●	–	●	–	–	–	●
Power (Transmission)	–	–	–	–	–	–	–	–	–	●	–	●	–	●
Power (Distribution)	–	◐	–	◐	◐	–	●	–	◐	–	●	–	–	●
Power (Consumer)	–	–	◐	–	–	●	–	–	–	–	–	–	–	●
Gas	–	–	–	–	–	–	–	–	–	–	–	–	●	●
Water	–	●	–	–	–	–	–	–	–	–	–	–	–	●

Legend: ●: considered by authors, ◐: not explicitly stated or exhibits ambiguity, –: not considered by authors.

Chapter 3

Background

3.1 Industrial Control Systems

Industrial Control Systems (ICS) are integral components in managing and automating industrial processes across various sectors. These systems encompass a range of control mechanisms, including Supervisory Control and Data Acquisition (SCADA) systems, Distributed Control Systems (DCS), and Programmable Logic Controllers (PLC), among others. ICS are primarily used in industries such as power grids, water treatment, oil and gas refining, and manufacturing. They are designed to ensure efficiency, reliability, and safety in industrial operations, often operating in real-time to monitor and control physical processes.

3.1.1 Power Grid

Electricity grids are the foundation for generating, transmitting, distributing, and providing electricity to end-users. These systems are divided into four interconnected grids: generation, transmission, distribution, and end-consumer. Generation plants are connected to the transmission grid through high-voltage

substations and transmission lines (usually rated at 220 kV and above). The combination of generation plants and the transmission grid makes up the Bulk Power System. This grid has a redundant configuration to ensure resilience against unexpected events. The Bulk Power System typically covers a large geographical area, such as an entire country. This paper focuses on the Bulk Power grid, and in particular, how a central control room monitors different substations spread geographically hundreds of miles apart to get the big picture of the operation of the power grid and then decides whether or not to change the setpoints of generators.

3.1.2 Water Treatment

Unlike Power and Gas, Water treatment plants are typically small-scale facilities, ranging from a few thousand square feet for a few hundred thousand gallons per day in small communities to several acres for millions of gallons per day for cities or industrial complexes. The purpose of these plants is to remove contaminants and pathogens from water to make it safe for drinking or use in industrial processes. The treatment process varies depending on the source of the water (natural sources like reservoirs or wastewater) and its intended use. It can be drinking water, wastewater, or a water recycling facility.

Three main types of processes are used to treat water: physical methods, biological methods, and chemical methods. (1) Physical methods involve separating pollutants by physical characteristics such as weight (sedimentation) or size (filtration), including advanced filtration techniques like microfiltration and reverse osmosis. (2) Biological methods allow microorganisms to metabolize pollutants and convert them into biomass that can be physically removed by settling and filtration. Coagulants aid in forming solid clumps in water (coagulation), which settle as sludge (sedimentation) and are filtered out. (3) Chemical methods pu-

rify water by adding specific substances, such as chlorine or ozone, to inactive pathogens (disinfection). Ozone precedes filtration, while chlorine follows to ensure the elimination of any lingering pathogens. Ultraviolet disinfection is also used, which introduces specific frequencies of light to break down cellular structures.

In terms of operation technology, water treatment plants are equipped with a range of sensors and actuators that are connected to programmable logic controllers (PLCs), which are then connected to Human-Machine Interface (HMI) stations. These components work together to regulate the various stages of the water treatment process, such as coagulation, sedimentation, filtration, and disinfection. Sensors measure different parameters, such as water flow rate, turbidity, and chemical levels, and send this information to the SCADA system. The system then processes the data and can automatically adjust the actuators to change valve positions, control pump speeds, or add chemicals as needed.

3.1.3 Gas Distribution plant

The gas transport and distribution grid follows a similar structure as the power grid. At specific locations, ingest stations receive and de-pressurize gas from the high-pressure nationwide transport grid (usually $>= 50\,bar$) and inject it into medium-pressure regional transport networks and, finally, the local distribution grid (around $10\,bar$). Local distribution grids operate at the level of cities or metropolitan areas. In regular intervals, small gas distribution closets can be found within streets and neighborhoods. These stations contain a mechanical pressure regulator that decreases pressure (often to $>= 1\,bar$) for the last mile. Usually, consumers are connected to more than one distribution station.

Even though gas distribution is considered part of a nation's critical infrastruc-

ture, gas distribution does not require consistent supervision or operator control, as gas 'just' flows. As a result, there is less need for redundancy of digital control equipment. Nonetheless, all stations are equipped with remote connections to allow the operator to monitor their system.

3.2 SCADA System

One of the essential tasks of the System Operator is to coordinate the power balance across multiple geographical regions to supply the demand and ensure the stability of the system. To achieve this, the systems operator relies on Supervisory Control and Data Acquisition systems (SCADA) to monitor and control the part of the grid they are responsible for in a Wide Area Network (WAN).

The SCADA system is designed to allow communication between networks that are geographically distributed in a Wide Area Network (WAN). In the case of power grids, SCADA systems are used to perform communications between (1) control room - substation, and (2) control room – control room. The first one allows the control room to perform the task of data collection from sensor devices (such as power meters) and to send command control to the control devices such as relays and switches. The second permits control rooms to share data among themselves so as to have visibility of the other parts of the network with which their system has interacted.

In order to provide continuous, reliable, and efficient communication, the SCADA uses a wide spectrum of industrial protocols. These protocols are designed to be compact, to run on top of the TCP/IP stack and most of them are defined in open standards. In addition, they have different types and communication models and were designed for diverse purposes. There are protocols for synchronous and unsynchronous, balanced and un-balanced communication, and

protocols that use client/server or publisher/subscriber models. There are also protocols for fast response to system events, or for high resolution in variable monitoring, and others for carrying a considerable amount of data.

In this thesis, we study real-world SCADA traffic from a bulk power grid, a distribution network, and an end-customer network, a water treatment facility and a gas distribution plant. We will look at SCADA protocols such as IEC 60870-5-104, ICCP, Modbus, Goose, Synchrophasor. To the best of our knowledge, SCADA traffic, such as the one in our study, hasn't been studied before.

3.2.1 Industrial Protocols

IEC 60870-5 (101 and 104)

When SCADA was first deployed, serial communication like Modbus and IEC 101 emerged as the communication standards. They served as a communication solution for exchanging data between equipment from different manufacturers. IEC 60970-5-101 (a.k.a IEC 101) [1] enables telecontrol messages between CR and substations. This point-to-point serial communication uses a low bandwidth bit-serial communication to transmit data objects and services over geographically wide areas. It also supports multi-drop communication (several devices connected to a single serial channel) in a client/server model. In addition, it uses balanced and unbalanced communications (in balanced communications, any party can initiate data transfers) and data acquisition by polling, cyclic transmission, spontaneous event, and general interrogation.

Later, with Ethernet and TCP/IP-based networking, a new range of SCADA protocols appeared. Protocols such as Modbus TCP, IEC 104, and DNP3 facilitated remote operation, maintenance, machine configuration, and interoperability across vendors. IEC 104 [3] is an application layer protocol that transmits over

TCP/IP using a client/server model. This protocol permits synchronous and asynchronous messages (a.k.a spontaneous/periodic messages) for balanced/unbalanced communications. In addition, it allows timestamps and quality attributes in the messages.

ICCP/Tase.2

Inter-Control Center Communications Protocol (ICCP) [2] exchanges time-critical data over WANs among CRs. This data exchange includes real-time monitoring and control data, measurement data, accounting data, and operator messages. It is widely used to tie together groups of utility companies, typically a regional system operator with the transmission, and distribution utilities, and generators. For example, regional operators may coordinate the import and export of power between regions across major inter-ties. In addition, Operators can use ICCP to exchange information between applications within a single control center. i.e., data exchange between the control center's Energy Management System (EMS) and a historian or SCADA [45].

ICCP can operate over either an ISO-compliant transport layer or a TCP/IP transport layer (although TCP/IP over Ethernet is the most common). In addition, ICCP employs a client/server model and sits in the upper sub-layer of layer 7 in the OSI reference model. A CR can be both a client and a server. All data transfers originate with a request by a CR to another CR that owns and manages the data. Each ICCP server performs access control on all incoming client requests based on bilateral association agreements. ICCP uses another industrial protocol MMS (Manufacturing Message Specification), for the required messaging services.

GOOSE

The Generic Object Oriented Substation Events (GOOSE) is a communication protocol defined by the IEC 61850 standard [30]. The IEC 61850 is an international standard that defines communication protocols to provide interoperability between all types of IEDs in a substation. GOOSE exchanges protection-related events (commands, alarms, and status) across digital substation networks.

It is an event-based protocol. The main objective of GOOSE messaging is to provide a fast and reliable way to exchange data sets between two or more IEDs. GOOSE enables the user to group any data format (status, value) into a data set It works directly over the Ethernet layer and follows a multicast communication model (a non-routable protocol that does not use IP addresses). To exchange data, GOOSE uses a publish/subscribe model [29].

Modbus/TCP

Modbus is one of the most common industrial protocols. It is easy to implement and maintain and has an open specification [4]. There are several versions, including Modbus RTU for serial communication and Modbus TCP for TCP/IP communications. The version that we will reference in this work is Modbus TCP. Modbus TCP is a simple request/response protocol in a client/server model widely implemented in both WANs and LANs networks. Only the controller (client) can initiate communication with the remote unit (server). i.e., the controller device must routinely poll each RTU or PLC (agents) and look for changes in the data. This means that, since there is no way for an agent device to report an exception, an agent only sends a message if requested by its controller.

Finally, unlike IEC 104, Modbus does not have timestamp or quality attributes in its packets. Furthermore, its format packets do not include an attribute for data

object descriptions. e.g., whether a register value represents a voltage value or a power measurement.

IEC C37.118

The IEEE C37.118-2 [5] is a widely used standard for PMUs sending synchronized phasor measurements (synchrophasor). A PMU computes the synchrophasor data. It transform electrical measurements for the current/voltage waveform at a given instant to a magnitude and phase angle. PMU implementations use WAMS network technologies for transmitting sysnchrophasor across large geographical areas due to its low latency requirements suitable for real-time supervision. Synchrophasor data is typically timestamped using Global Positioning System (GPS) time as a universal time source for higher accuracy. It is used in applications for dynamic observability, such as islanding detection, voltage stability monitoring, oscillation monitoring, and detection and wide-area frequency monitoring [52].

3.2.2 Network Endpoints

The following list summarizes the most common **endpoints** we have encountered in our study of power grid communications:

Control Room (CR) The control room, as the center of operations, orchestrates physical processes in the system, (such as power flow, voltage level, and frequency). They are computers collecting data from remote devices, (often referred to as "the SCADA"). They usually interface with other computers such as databases (a.k.a Historian) and Human Machine Interfaces (HMI) that operators can use to visualize the state of the physical process being managed.

Intelligent Electronic Device (IED) An IED is an embedded computer that

receives data from sensors, i.e., measurement transformers, and sends commands to power equipment, e.g., circuit breakers directly. Relays and digital fault records are examples of IEDs. Engineers deploy IEDs within a substation and only communicate locally with other IEDs or computers inside the substation (e.g., an RTU or a local CR). Because of the safety-critical nature of their operation (e.g., automatically disconnecting an overloaded electric line before a fire starts), they need to operate over highly-reliable and low-latency local networks.

Remote Terminal Units (RTU) Similar to an IED a RTU is an embedded computer that collects data mainly from IEDs, and then delivers it to the CR. While IEDs are employed for communication within a substation, RTUs are used for external communication. RTUs have been the traditional endpoint for exchanging data between a CR and a substation, but some modern substations utilize **substation gateways** as endpoints.

Programmable Logic Controller (PLC) A PLC is another industrial computer used to automatically control a physical process. They are more widely used in industry, such as water, chemical, and manufacturing systems. While IEDs focus on protecting electrical equipment, PLCs focus on controlling power generation machines. They are also popular in commercial end-consumers applications.

Phasor Measurement Unit (PMU) A PMU is a sensor that collects voltage and current values and calculates their synchrophasor measurements. They operate at a very high-frequency and in a time-synchronized way. They can be a stand-alone device or incorporated into an IED. It is a relatively new technology that is much faster and more time accurate (to the order of one millisecond) compared with the traditional SCADA technologies. They are used in Wide-Area Monitoring Systems (WAMS) and in synchrophasor-base Wide-Area Monitoring

Protection and Control (WAMPAC) applications.

Phasor Data Concentrator (PDC): is a server machine that collects, sorts (according to timestamps), and stores PMU data.

HMI: They gather data from the controller and enable the human operator to interact with SCADA through applications that process traffic data into information that humans can understand and other applications.

3.2.3 Agent and Controllers

For simplicity, in the remainder of this book, any endpoint that reports data will be referred to as an **agent (A)**, while any endpoint that collects data will be referred to as a **controller (C)**. For example: we will consider any PLCs or RTUs as "agents" and HMI or SCADA server as "controllers". In this way, we can focus on the characteristics of the network to facilitate our comparisons and diagrams, and not dwell on the individual devices.

Chapter 4

Dataset and Metodology

4.1 Dataset

Our SCADA traffic dataset consists of network packet captures from three different infrastructures: (1) a power grid that include: a system operator (SO) which had networks using different industrial protocols: IEC 104, C37.118, and ICCP. A transmission owner (TO) which captured data from a large substation where IEDs use GOOSE. A distribution operator (DO) monitoring various substations using IEC 104. Finally, a university campus that used the MODBUS protocol to monitor and control electricity consumption in their campus. (2) A gas distribution plant that uses IEC 104 to monitor and control processes in a metropolitan region. (3) A drinking/wastewater treatment facility that employs the industrial protocol Modbus/TCP to monitor PLCs on a few acres of land. Table 4.1 summarizes these data captures.

The datasets are packet traces stored in pcap format. Each row in Table 4.1 represents one network. The columns show the type of the Industrial Control System with the data comes from **ICS**, the specific part where the data was captured (**Location**), protocol used (**Protocol**), the number of hosts (**# hosts**),

duration of the capture (**Duration**) in hours, and the number of packets (**# Packets**) that cointains the traffic capture. Finally, the **Name** that we will use to refer to that network in the rest of this work.

Table 4.1: Summary characteristics of the datasets.

ICS	Location	Protocol	# hosts	Duration (hours)	Name
Gas	Distribution plan	IEC 104	157	2037	IEC104_G
Water	Drinking Water Treatment	Modbus/TCP	100	24.5	MODBUS_W
Power	Generation	IEC 104	39	8.2	Power or IEC104_B
Power	Transmission	IEEE C37.118-2	14	1	C37.118
Power	Transmission	ICCP	14	8.2	ICCP
Power	500 kV Substation	GOOSE	28	14.3	GOOSE
Power	Distribution	IEC 104	34	3.7	IEC104_D
Power	Consumer (Campus University)	Modbus/TCP	6	111	MODBUS_P

4.2 Metodology

The main questions we are trying to address in this study are to figure out what these networks look like, if their traffic patterns are similar to each other, and if within a given network, the communication dynamics are stable. In particular, we formulate the following five questions:

RQ1: What are the topologies of these networks?

RQ2: How do these networks differ in their communication patterns?

RQ3: Are the traffic patterns within a network different, and if so, how?

RQ4: What type of information is handled by these protocols?

RQ5: How much monitoring vs. control is done in these networks, and what types of control commands are sent?

Fig. 4.1 illustrates how our research questions create a general framework for analyzing SCADA networks. We start by understanding the topology of each network (RQ1), and then we start zooming in to understand the traffic differences between networks (RQ2), then traffic differences within a network (RQ3), the data types handled by the network (RQ4), and finally, the types of measurement and control commands sent back and forth between a controller and the agents (RQ5).

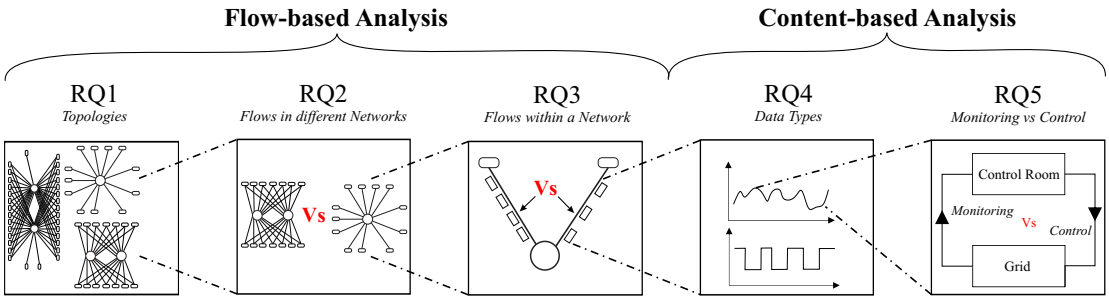


Figure 4.1: Framework for our research questions.

Chapter 5

A Taxonomy of Industrial Control Protocol and Networks in the Power Grid

Before beginning our analysis of SCADA network measurements, we defined a taxonomy to demonstrate the variety of configurations and protocols present in each SCADA network. This taxonomy was based on our dataset from the power grid.

Power grids are complex systems composed of multiple networks and a variety of industrial protocols. Each network is configured to meet the particular needs of the system, and protocols are used for different communication purposes (synchronous and asynchronous, request/response and unsolicited communications), and models (client/server or publisher/subscriber). Some protocols are designed to provide a rapid response to events, while others are intended to transmit large amounts of data. Some networks are used to transmit data over long distances, while others are used to keep data localized. Additionally, some protocols are used

to monitor stable-state events, while others are used to monitor transient events. In this chapter, we use our knowledge of datasets obtained from the various operators in a power grid to give a concise overview of the variability of industrial protocols in an ICS. We propose a taxonomy and examine their similarities and differences.

5.1 Taxonomy

We will now discuss our proposed taxonomy features. We examine the communication model used in the grids and their characteristics.

5.1.1 Communications

We identify three main models of communication which differ according to the specific needs of the network and the relationship between peers. The three models are client/server, publisher-subscriber, and peer-to-peer.

A client/server model is used mainly for supervision purposes. The client is the centralized SCADA system located in the CR, and the servers are the supervising devices in the field (such as RTUs, PMUs, and PLCs). The CR is always the initiator of communications and the field devices, answer only to the main server, so their relationship is hierarchical.

All the data is gathered at a particular point (CR), which means that the architecture of the network is centralized.

The publisher-subscriber is commonly used at the substation level, mainly for protection purposes. The publisher is the element that transmits data, and the subscribers are the consumers of that data. In this model, data is broadcast among all the devices and accepted only by those that need it. This is a many-to-many

communication type. Unlike the client/server model in which the devices that request data differ from the field devices, the devices in the publisher-subscriber model are typically the same and perform similar functions (usually IEDs), i.e., there is no hierarchical relationship. Since data is broadcast, the publisher does not need to know who is using the information and, the subscriber does not require the origin of the data. In this way, the electrical substation reduces delivery latency, and the connection complexity is decreased to a single point. As a result, protection schemes in a substation can be operated without delay. When an IED detects a fault, it broadcasts the alarm without concern about the destination address or connection problems.

The last communication model is the peer-to-peer model, where two or more peers pool their data in a decentralized system. Peers are CRs that communicate with each other directly without any intermediary and share the status (measurement data) of their substations. Usually, there is no hierarchical relationship. Each CR possesses the same capabilities: it can initiate communication and function as a client or a server (via a request-response message). Since the primary purpose is data monitoring between CRs, data speed and reliability are not the priority, unlike in the publish-subscribe model.

5.1.2 Connectivity, E2E ID, and Port

The industrial protocols we have analyzed have a variety of communication links. They range from Local Area Networks (LANs) to monitor and control devices relatively close to each other (including serial communication in IEC 101) to IP-based WANs used to monitor and control remote units. Therefore, we define the following features: (1) Connectivity denotes the type of connections the devices have; for example, IEC 101 is a serial link, IEC 104 uses WAN and GOOSE

uses LAN. (2) E2E ID defines the End-to-End identifier in the communication link. For example, IEC 104 communicates via IP addresses, while GOOSE via MAC addresses. Finally, the Port number identifies the port and the associated transport protocol used by the standard. TCP/IP is the most common protocol found in the Network layer, noteworthy for its reliable data transmission. Less common but also important, UDP is ideal for fast communication over long distances (WAN) while Ethernet is ideal for fast communication over short distances LAN. This is especially important for control and protection purposes. For example, PMU can use UDP for high data transmission between substations, and an IED can utilize GOOSE to transmit over Ethernet within the substations for rapid event responses.

5.1.3 Monitoring

We identified three ways that the control server monitors the status of devices and how the devices receive information from peers:

Spontaneous: These are events that the agent can send without receiving any previous request from the controller. The time report depends on when a value exceeds a pre-configured threshold or, in the case of status values, when they change. e.g., the networks that use IEC 104 widely use this type of transmission to reduce strain on the network.

Periodic: In this case, an agent reports value data at a fixed interval of time according to the configuration. The agents do not require an acknowledgment from the recipient, and the flow of communication can be in one direction. For example, C37.118 devices are mostly periodic, and only the controller communicates with the agent for configuration.

Request-Response: For these, only the controller can initiate communica-

tion and it must routinely poll each agent to look for changes. Agents do not report exceptions and never send a message unless their controller requests it. This is the only transmission mode used by Modbus. It has equal traffic flow in both directions, see Sections 6.5 and 7.6.

5.1.4 Object-Oriented

The design of object-oriented protocols is comparatively new. Their purpose is to address the complexity and interoperability of network devices by creating objects described as data attributes and operational services. Each object is an independent entity that can be replaced without affecting the whole system. For instance, a substation has measurement, control, and protection devices, each with its data type, functionalities, and services. Without a simplified means of communication, there would be no easy method of interoperability between devices.

5.1.5 Endpoints

Finally, we come to the endpoints of each network. They are labeled starting with the endpoint that sends out most of the information. For example, RTU-CR means that the endpoints of this network are RTUs and CRs and that RTU sends out the bulk of its data to the CR (although the CR can also send data to the RTU).

5.2 Analysis

Table 5.1 shows our proposed taxonomy. As we can see, there are no two protocols alike. First, among our datasets, IEC 101 represents the only serial

Protocol	Communications	Connectivity	E2E ID	Port	Monitoring	Object-oriented	Traffic Endpoints
IEC 101	C/S	Serial	Link address	-	P, SP		RTU-CR
IEC 104	C/S	WAN	IP	TCP/2404	P, SP		RTU-CR
ICCP	P2P	WAN	IP	TCP/102	R/R, P	✓	CR-CR
MODBUS	C/S	LAN	IP	TCP/502	R/R		PLC-CR
C37.118	C/S	WAN	IP	TCP/4712 UDP/4713	R/R, P		PMU-CR
GOOSE	PubSub	LAN	MAC	-	P, SP	✓	IED-IED

Table 5.1: Communications: Client/Server (C/S), Publish/Subscribe (PubSub), peer-to-peer (P2P). Mode: Request/Respond (R/R), Periodic (P), Spontaneous (SP).

protocol in the list. It is representative of legacy SCADA systems that use modems or other serial connections to communicate remotely.

Among the protocols that use communication networks, the client/server model is predominant. This means that most of the communication within the grid is end-to-end. With one exception, TCP/IP is the preferred transport protocol because packets can be re-transmitted if lost en-route. This is especially important for WAN, where the reception of packets is critical for control commands. The one time TCP is not utilized it is substituted with IEEE C37.118. Since this is a protocol predominantly designed for high-granularity monitoring purposes, it can use UDP for fast transmission without the need to acknowledge packets. IEEE C37.118 is a data hose over UDP, and the server receives as much of it as possible.

GOOSE is a different protocol from the others on the table. It is the only one that uses the Publisher/Subscriber model; this makes it more efficient for protection purposes. It is also the only protocol not using TCP (it runs directly on top of Ethernet).

As regards monitoring, most communications can be either periodic (asynchronous) or request/response (asynchronous).

Finally, we can see that protocols that use object-oriented paradigms are not widely used in the power grid; they are instead chosen for a particular section of the

grid. In our case, they are used for the short but fast packet transmission inside substations (GOOSE) and the long distances but heavy data volume between control rooms (ICCP).

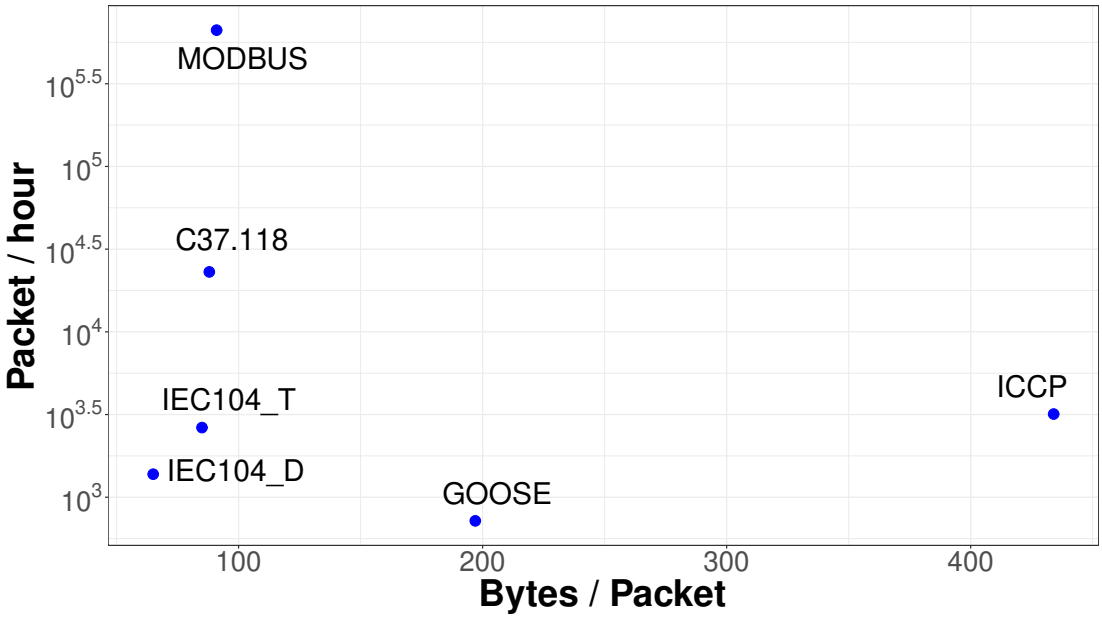


Figure 5.1: Packet/hour vs Packet size

To illustrate some of these differences, we obtained real-world data from operating power grids in different infrastructures, including a System Operator, a Transmission Owner, a Distribution Owner, and a University Campus (Consumer). They were primarily collected in each facility's control network (SCADA network of the CR) with the exception of GOOSE, which was captured in a substation. The datasets are packet traces stored in a pcap format. They contain unencrypted data from steady-stable operations (no system disturbance/events or attacks registered) including control and measurement data.

Fig. 5.1 shows our preliminary analysis of our datasets, showing clear clusters of activities. In trying to understand these clusters, we argue that the physical distance between the endpoints affects the amount of data that needs to be trans-

mitted. The Modbus TCP protocol used in a university campus and the GOOSE protocol used in the substation are LANs where all devices sit within a few dozen meters of each other. In contrast, the rest of the protocols operate over WANs, spanning hundreds of kilometers between devices. GOOSE shows relatively lower data transmission, however, and this is because, during our packet capture, there was no emergency event (e.g., line overload). Therefore, there was no need to send packets at higher rates.

We also see small packets (less than 100 bytes) but with a rapid transmission rate (C37.118). This reflects the high-frequency data collection from PMUs. On the other extreme, there are protocols with large packet sizes and moderate transmission rates (e.g., ICCP). This reflects that CRs have much data to share (they collect data from a wide area, while substations only collect data from one point in the network).

Conversely, IEC 104_T (IEC 104 used in the transmission system) and IE104_D (IEC 104 used in the distribution system) carry information from a single substation (per connection). Modbus has the highest transmission rate, nearly two orders of magnitude greater than the rest. This rate of transfer was a configuration decision of the university campus.

5.3 Conclusion

As this initial discussion shows, industrial control protocols differ considerably from each other. They have distinct properties, communicate with a variety of endpoints through different types of networks. This has implications for network design as well as security deployments.

We have demonstrated the distinctions between SCADA networks in terms of their configuration and features of the protocols. Now, we are ready to analyze our

actual data. We will begin by looking at the differences between ICS in power, gas and water. Afterwards, we will delve into the distinctions between the networks within the power grid.

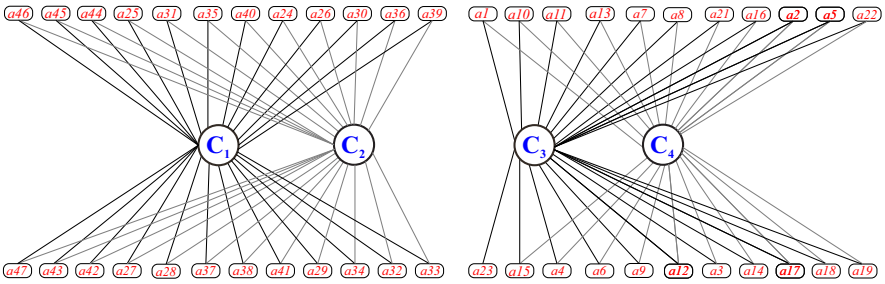
Chapter 6

From Power to Water: Dissecting SCADA Networks Across Different Critical Infrastructures

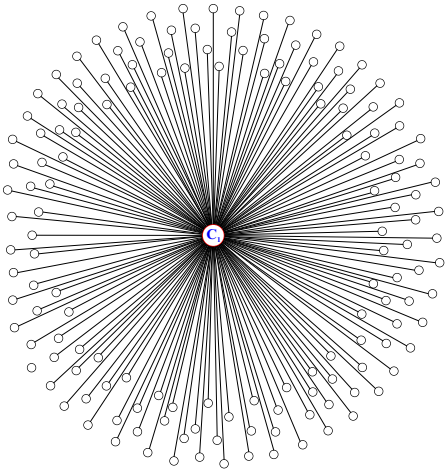
In Chapter 5, we discussed the distinguishing features of various industrial protocols and created a comprehensive taxonomy that highlights the variety of industrial networks. This chapter focuses on a thorough examination of SCADA networks in three Industrial Control Systems: power, water, and gas. Our aim is to analyze their network traffic patterns and how they relate to their particular operational processes. This analysis looks into the complexities of network topology, packet sizes, inter-arrival times, flow directions, and data types, providing a multi-dimensional view of the SCADA systems that control these essential infrastructures.

6.1 RQ1: Network Topology

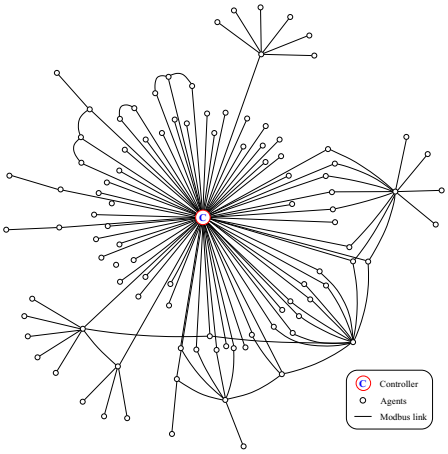
In this section, we examine the distinctive network topologies used across the distint industrial system of our dataset: power, gas and water. Each topology is tailored to the unique operational demands and physical charateristic of its repective industrial environment. This comparative analysis sheds light on the diverse approaches to network design in SCADA systems, highlighting their functionality and industrial requirements.



(a) Power



(b) Gas



(c) Water

Figure 6.1: Type of topology structure: a Complete bipartite, b Star, c Star-Hybrid: a star topology with other structures.

6.1.1 Power:

We observe that the power grid topology from Fig. 6.1a forms a Complete Bipartite Graph. A complete bipartite graph $K_{p,q}$ consists of a set of p vertex and a set of q vertex (in our case, $p = 2$) and pq edges joining the vertex of different types [27]. This type of topology is known as Spine Leaf topology [48, 25] in cloud data centers. The difference is that the Spine Leaf topology is used to forward packets through the Spine (the central nodes), while in SCADA networks, the central nodes consume data (they do not forward it).

In our Power network, each agent is connected to two controllers. This dual-purpose setup offers fault tolerance and load balance, which are essential for a network that focuses on the operational status of the process (see Section 6.3). This reduces the risk of the operators losing visibility in the event of a controller or link failure. If one controller fails, the other will take over, allowing control applications such as the Automatic Generation Control (AGC) algorithm to access the input data needed for its control operation. Furthermore, the operator can still monitor the grid from their HMI [8, 7].

We examine redundant connections from agents (RTUs) to controllers (SCADA servers). In Figure 6.1a, the active connections (i.e., those sending data to the controller) are represented in black, while the standby connections (i.e., those sending heart-beats to the controller to indicate that they are connected and ready to receive active connections) are shown in gray. Upon looking at the traffic of each connection, we find that from the two connections to a pair of controllers, one of these connections is used to send process data to one of the controllers (active link), while the other is used to keep the connection with the other controller alive, serving as a backup. The heartbeat signal consists of U-Format messages (TESTFR) described in the IEC 104 protocol.

This type of load balancing makes sense as we only require one active connection between a controller and an agent while the other connection is on standby, ready to be used in case of a failure. We further confirm with the ISO that all four controllers are physically in the same control room, so this network represents a control room with four servers arranged in pairs, each monitoring a different part of the grid.

In summary, the power grid network is composed of two bipartite graphs, $K_{2,18}$ and $K_{2,14}$, which differ from the star topology of the gas and water networks (discussed in Sections 6.1.2 and 6.1.3). These $K_{2,q}$ graphs are especially important for control operations, as they require standby connections to controllers, redundancy in controller servers, and links to ensure reliable communication for control purposes.

6.1.2 Gas:

The Gas network has a star topology, with a single controller connected to 155 agents, as shown in Fig. 6.1b. This is the largest network in our dataset.

It is noteworthy that the controller is a point of failure, meaning that if the controller fails, the entire network will become inoperable for an operator. This makes the controller a potential bottleneck for the supervision of the system. Most SCADA systems require a backup system and redundancy for the controller, but it is not always clear if such measures are in place.

We hypothesize that the lack of redundant and standby connections (like in the power grid) implies that it is not essential for the operator to receive updates from different monitoring points, and if necessary, automation systems can act locally without the direct involvement of the central controller.

6.1.3 Water:

Finally, the water network is characterized by a Star-Hybrid Topology, as seen in Fig. 6.1c.

We see that this network has the same single point of failure as the gas network; however, we also notice that the endpoints collaborate and exchange data at the edge of the network.

One major difference between our Water network and the Power and Gas networks is that the Water network is a LAN, and the Gas and Power networks are WANs. Therefore, we cannot see the MAC addresses of remote devices in the power and gas networks; however, in the water network, we can see them, and we can identify them as Programmable Logic Controllers (PLCs). In addition, a LAN network suggests that the water treatment facility is not spread over a large physical area. A simple LAN setup could mean that all devices are in close proximity. It can be assumed that the water treatment plant prioritizes operational reliability and simplicity.

In summary, the analysis of the SCADA netowrks for gas, power and water facilities reveals distinct topoligcal configurations, each suited to the specfic operational needs of the respective systems. The power network demostrates a unique Complete Bipartite Grahp topology. This configuration, provides fault tolerance and load balancing through dual connections to controllers for the network's control operations, ensuring uninterrupted communication and redundancy. In contrast, the gas network exhibits a star topology with a central controller , thus a single point of failure at the controller. The absence of redundant connections, unlike in the power grid, suggests a different operational emphasis, potentially allowing for local automation without central coordination. Finally, the water network, with its Star-Hybrid Topology, shares the single-point-of-failure charac-

teristic with the gas network but introduces collaborative data exchange between agents. Unique among the three, this network operates as a LAN, indicating a more compact physical setup focused on operational simplicity and reliability.

These topological differences underline the varied operational priorities and physical constraints of each network. While the power grid prioritizes fault tolerance and load balancing for control operations, the gas network appears to lean towards a centralized yet simpler structure. Water, with its LAN configuration and collaborative agents, appears to balance centralized control with operational simplicity.

6.2 RQ2: Traffic Pattern Differences Between Networks

We now focus on the traffic in these networks and examine how they differ in their communication patterns between networks. To do this, we use the traditional traffic analysis metrics of packet sizes and transmission timing (or inter-arrival times).

6.2.1 Packet Size Distribution

By analyzing the packet size distribution of each network and comparing them to one another (Fig. 6.2), we can observe that all three networks have a prevalence of small packets (0-100 bytes). Additionally, there is a contrast between the IEC 104 and Modbus protocols, with the former mainly composed of packets under 200 bytes, and the latter having a larger range of packet sizes.

- ⋆ **Common packet sizes:** Though these are three different networks using different protocols, most packets in the three networks are smaller than or

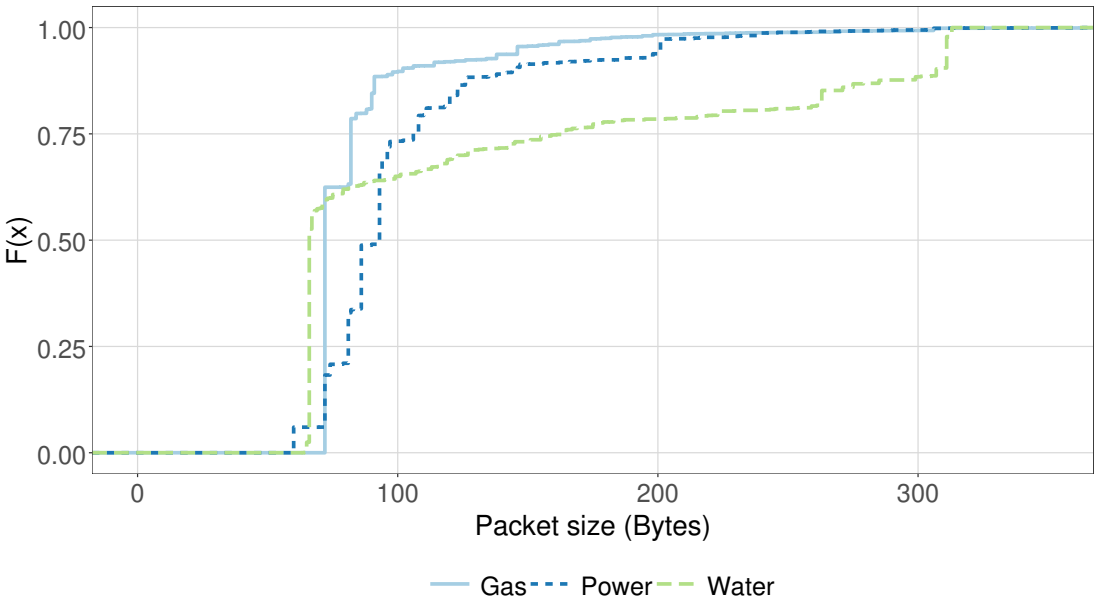


Figure 6.2: CDF Packet size

equal to 100 bytes. Fig. 6.2 shows that more than 50% of the packet sizes are in the range of 0-100 bytes: water (green dotted line) 60%, power (blue dotted line) 75% and gas (solid line) 80%. This is evidence of a common operation behavior between ICS networks.

- ⋆ **Most packets in IEC 104 networks are below 200 bytes:** The Power and Gas networks (which use the IEC 104 protocol) have 99.9% of their packets below the 200 bytes mark. The largest packet in an IEC 104 network was one gas packet of 744 bytes, and the largest packet in the power network was 1378 bytes.

- ⋆ **Larger Packet Sizes with Modbus:** In contrast, we can see a significant amount of packets in the water network larger than 200 bytes. This is due to Modbus' ability to encapsulate a large number of registers in a single packet, unlike IEC 104-based networks such as Gas, which divide their responses

into multiple messages. For example, a Gas network agent might respond to an Interrogation command with 105 status data points spread across five packets, while Modbus consolidates the same amount of data into a single reply, or even 2000 status data points in a single packet as seen in the biggest packet size (313 bytes). This efficiency might be based on the simpler data structures used in Modbus, when compared to the typed data from IEC 104.

In summary, the packet size distribution reveals similar and distinct communication patterns between networks. Although all networks use primarily small packets (0-100 bytes), significant differences emerge between the IEC 104 and Modbus protocols. IEC 104 networks, used in Power and Gas, predominantly feature packets under 200 bytes, highlighting a tendency toward smaller, more frequent transmissions. In contrast, the Water network, using Modbus, exhibits a broader range of packet sizes, often exceeding 200 bytes. This difference is due to Modbus' ability to encapsulate larger data volumes in single packets, unlike IEC 104, which opts for multiple, smaller messages. This divergence underscores the varied approaches to handling data transmission and encapsulation across these industrial protocols.

6.2.2 Packet Transmission Rate:

We now focus our attention on analyzing the intervals between packets. Our findings indicate that the gas network is the least active, with transmissions occurring mostly at regular intervals of minutes. Water has the highest rate of transmission, with almost half of the agents sending data in times less than a second. Power has a more dynamic pattern, event-driven nature, with varying transmission rates.

Fig. 6.3 shows the inter-arrival time of each data point per network. We

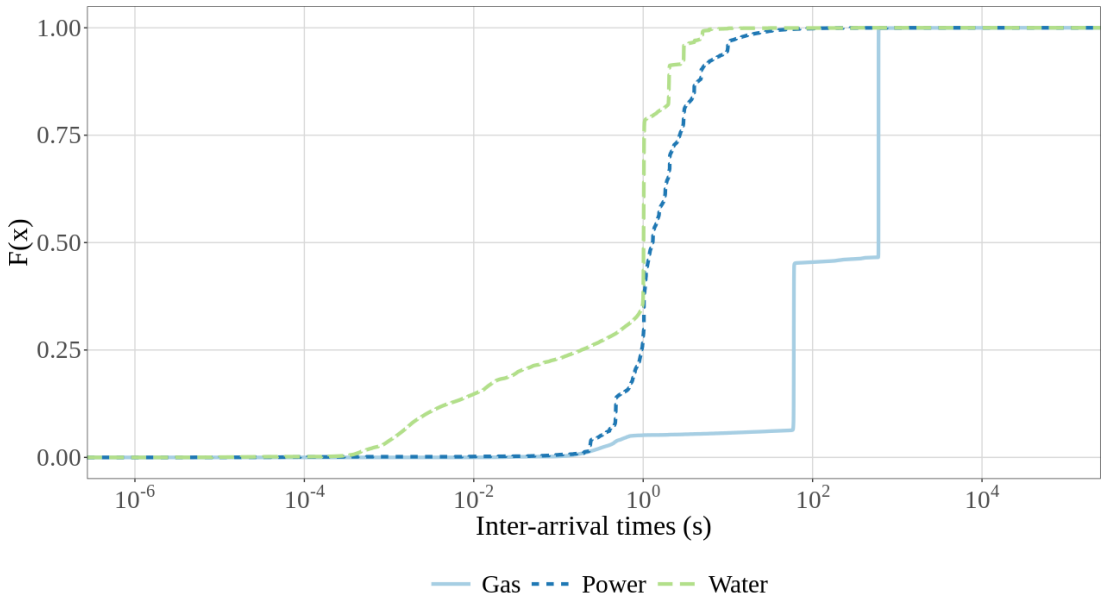


Figure 6.3: CDF of the Inter-arrival time for each ICS network

observe the following differences between networks:

- ⋆ **Gas:**

  The gas network is the least active of the three networks: Gas agents have a transmission rate much higher than power and water, with a difference of approximately one order of magnitude. For example, almost all agents in water and power transmit every 10 seconds, while only less than 6% of the data points in the Gas network are updated at that rate; in fact, 94% transmit in the order of minutes. This presumably means that the central controller in a gas network does not need to take any time-critical actions, and most of the control actions (if any) are done automatically by the substation without reporting them to the central controller. This means that while the transmission rates in water and power are in the range of seconds, in gas, it is in the range of minutes, making the gas network the

least active of the three networks.

Gas networks transmit at fixed intervals: The gas curve shows several distinct and clearly defined step changes: less than 1 second (5.5%), 1 minute (40%), and 10 minutes (54.5%). These steps are evidence of share set-up among agents that update their data points at the same time intervals. In other words, the data transmission depends on the set-up configuration of the devices than on the dynamics of the systems.

- **Water:**

Fast transmission rates: We can observe from the curve that water interarrival times start at one millisecond rates. However, this is present only for a small fraction of data points, less than 1%. Furthermore, a quarter of the network devices have a transmission rate of less than a second. Power and Gas have in common a minimal transmission rate of 100 ms.

Almost half of the PLCs are queried at 1s intervals: The controller interrogates 43% of the agents every second. (There are no spontaneous messages in the Modbus protocol, so all message exchanges are the response to a query, also known as an interrogation). There is a degree of consistency in the configuration of the devices in the Modbus protocol.

- **Power:**

Dynamic conditions: Unlike gas and water, the power curve has no distinct steps or plateaus. This implies that the agents have varying reporting patterns, which is a reflection of the majority of the spontaneous messages of this network. The continuous yet varied slope implies that the messages are spontaneous or event-driven. This is a sign that the network can respond to the dynamic conditions of the grid as a process. Thus, the smooth shape

of the power curve demonstrates the event-driven nature of this network.

In summary, we can see some similarities and differences among these networks. As discussed before, when a Modbus agent sends data to a controller, this is only in response to an interrogation. Agents do not have the ability to send data in Modbus spontaneously. In contrast, IEC 104 has the ability to configure agents so that they report whenever a value exceeds a threshold. As we discuss later, the power grid network takes advantage of this, and therefore, the transmission patterns are more diverse. The Gas network also uses IEC 104 and can configure devices to send spontaneous measurements. Surprisingly, the gas network does not take advantage of this, and instead, it is configured so that the controller interrogates the agents sporadically.

This difference may also explain the packet size analysis. Since half of the packets in both the Gas network and the Water network are interrogation queries, most of these packets will have the same size. In addition, they will be smaller than the responses with various measurements. This explains the big step early on in Fig. 6.2 for Gas and Water. More than half of the packets are small and of the same size because they are the same repeated query. Whereas for the power network, since there is no repeated central query, all the spontaneous messages have different packet sizes.

Finally, we can also see that the Power network and the Water network have fast transmission times. So, they are presumably operating a more time-critical process. Having said that, since both IEC 104 networks operate in a WAN, the minimum interarrival time in our datasets for any of them is 100ms. This may identify a time constraint for these networks managing assets in large geographical areas.

6.3 RQ3: Diversity Within Networks

We now look at the diversity within networks. Entropy is a helpful metric for understanding the diversity or randomness of data, and we can use it to evaluate the randomness of packet sizes or timings within a network.

In the context of packet sizes from network traffic, a higher entropy value (values close to 1) suggests a wide range of packet sizes being sent/received by an agent, while a lower entropy (closer to 0) implies that most packets have similar sizes. The spread and range of entropy values for each network provide insights into how varied the packet sizes are within each system. By comparing the entropy distribution using a cumulative distribution function (CDF), we can determine which network has the most predictable packet size distribution.

We use the Shannon entropy formula:

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log(p(x_i))$$

Where $p(x_i)$ is the probability of occurrence of a particular packet size for that agent. For each agent, we count the occurrence of each unique packet size. Then, we divide each count by the total number of packets for that agent to get the probability distribution. Then, we add each packet size a particular agent has sent at least once.

Fig. 6.4 shows the distribution of entropy values for the three networks. In the context of this plot, higher entropy values represent greater packet size diversity. The network with more agents (y-axis) having higher entropy values (x-axis) and more variabilility in its curve is likely the network that is less predictable in terms of packet sizes.

⋆ **A quarter of agents have uniform packet sizes:** The three networks

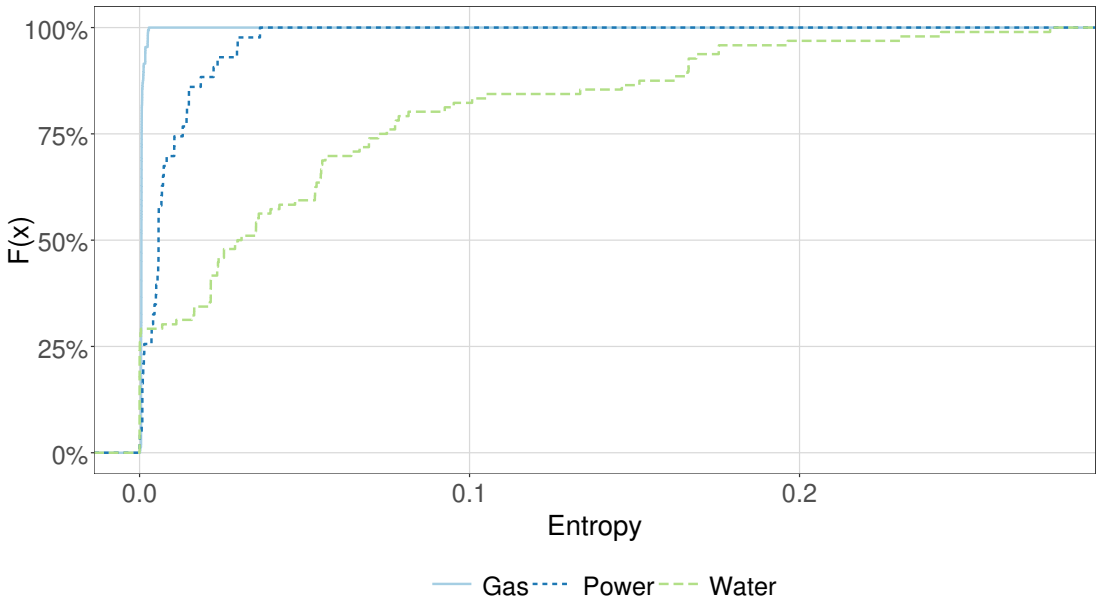


Figure 6.4: CDF of Entropy

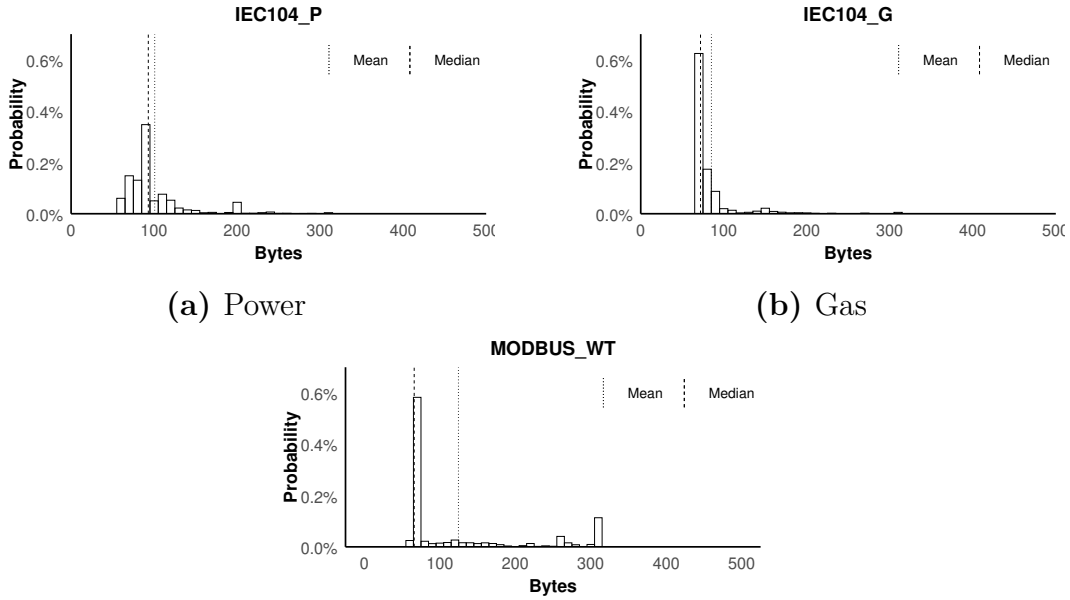
show a sharp increase at first until 25% (y-axis). This suggests that a large proportion of agents in the three networks have low entropy values. This implies that one-fourth of their agents have a consistent packet size or lack of diversity in packet sizes. For example, the agent with the least entropy in Power, sends repeatedly the same packet size of 72 bytes.

- ⋆ **Uniformity in Gas:** Most of the agents in the Gas network exhibit low entropy values, implying a lack of diversity or uniformity in packet sizes for these agents. Therefore, most of the packets have a fixed length.

- ⋆ **More diversity in Water:** The Water network has the highest entropy, which means that there is more uncertainty about the packet size that the controller will receive.

We can see a different version of these metrics in Fig. 6.5. From our Power network, we identify that packet sizes are primarily concentrated around 50-100

bytes. There are two modes: a primary mode in the 100-byte bin (90-100 bytes), which is the highest peak in the distribution. A secondary or minor peak at 200 bytes (exactly at 198 bytes). The former are spontaneous, and the latter are periodic (1-second) packets containing information (I-format in IEC 104). This serves as proof that the Power network is event-driven.



(a) Power **(b)** Gas

(c) Water

Figure 6.5: Packet size distribution. Binwidth 10.

Gas: From Fig. 6.5b, it is clear that the majority of packets are small. 70% of them are S-format and U-format messages, which are fixed-size packets defined in IEC 104 to acknowledge the receipt of data and to check the status of a connection. The default rate for acknowledging packets is higher in IEC 104 (w = 8), however, the gas network needs to send an S-format acknowledgment for almost every packet received (w = 1) due to the long interval of time (in the order of minutes) between transmissions. This explains why the packet sizes in the network are so uniform, as the majority of data types are S-format and U-format, both of which

are fixed-size packets.

Water: From Fig. 6.5c, we see a heavy tail of large packets. We identify that most of the small packets correspond to 'Read Coil Status ' (function code 1). These are binary values and usually represent the status of an actuator. Therefore, they are small packet sizes. In addition, any request messages are in this range of small packet size. In contrast, we see that the large packet sizes correspond to 'Read Holding Register' (function code 3) that are 16-bit register, much larger than coil status. While small packet sizes are predominant, the larger 'Read Holding Register' packets are notorious in the distribution, especially in the range 200-300 bytes. This evidences the diversity of packets in water indicated by the highest entropy.

In summary: From the Entropy analysis, we can see that a quarter of the agent in each network handles consistent packet sizes. The Gas network stands out for its pronounced uniformity, implying uniformity in its packets and communication patterns which is a reflection of its polling-drive nature. On the contrary, the power network presents a higher level of variability in packet size, reflecting a more event-driven network, especially because of spontaneous messages. Similarly, Water exhibits a greater variability of packet sizes between its agents, indicating that its operations and communications are more standardized or consistent given its request/respond model (polling-driven).

6.4 RQ4: Information Types

We now turn our attention to the information contained within the packets themselves. Each protocol has its own specific standard with clearly defined types of data that are included in the packets. We start with a general overview in

Table 6.1: Power ASDU types and their description

Type	Reference	Description	%
1	M_SP_NA_1	Single-point information.	<0.001
3	M_DP_NA_1	Double-point information.	0.08
5	M_ST_NA_1	Step position information.	0.05
7	M_BO_NA_1	Bitstring of 32 bit.	<0.001
9	M_ME_NA_1	Measured value, normalized value.	1.97
13	M_ME_NC_1	Measured value, short floating point number.	39.71
30	M_SP_TB_1	Single-point information with time tag CP56Time2a.	<0.001
31	M_DP_TB_1	Double-point information with time tag CP56Time2a	<0.001
34	M_ME_TD_1	Measured value, normalized value with time tag CP56Time2a.	0.51
36	M_ME_TF_1	Measured value, short floating point number with time tag CP56Time2a.	57.21
50	C_SE_NC_1	Set point command, short floating point number.	0.41
70	M_EI_NA_1	End of initialization.	<0.001
100	C_IC_NA_1	Interrogation command.	0.019
103	C_CS_NA_1	Clock synchronization command.	0.001

Table 6.2: Gas ASDU types and their description

Type	Reference	Description	%
1	M_SP_NA_1	Single-point information.	27.85
3	M_DP_NA_1	Double-point information.	0.03
9	M_ME_NA_1	Measured value, normalized value.	17.92
30	M_SP_TB_1	Single-point information with time tag CP56Time2a.	1.13
34	M_ME_TD_1	Measured value, normalized value with time tag CP56Time2a.	11.38
45	C_SC_NA_1	Single command.	0.01
48	C_SE_NA_1	Set point command, normalized value.	0.01
58	C_SC_TA_1	Single command with time tag CP56Time2a.	<0.001
70	M_EI_NA_1	End of initialization.	<0.001
100	C_IC_NA_1	Interrogation command.	36.38
101	C_CI_NA_1	Counter interrogation command.	<0.001
102	C_RD_NA_1	Read command.	3.79
103	C_CS_NA_1	Clock synchronization command.	1.47

Fig. 6.6. On the x-axis, we see the number of data types defined by the standard, and on the y-axis, we see the number of types present in our capture. There are two clusters: On the top right corner, we can see the IEC 104 networks, which have over 100 data types defined in the standard but only use around 10% of them in the capture. On the bottom left-hand corner is Modbus. Its standard do not contain many data types so, the variety of data types contained in the capture are relatively few.

IEC 104 Cluster: An IEC 104 packet can be either Information (I), Supervisory

Table 6.3: MODBUS types and their description

Type	Description	%
1	Read Coils	9.52
2	Read Discrete Inputs	0.39
3	Read Holding Register	89.1
4	Read Input Register	0.66
15	Write Multiple Coils	0.34
16	Write multiple register	<0.001
22	Mask write register	<0.001
23	Read/write multple registers	<0.001

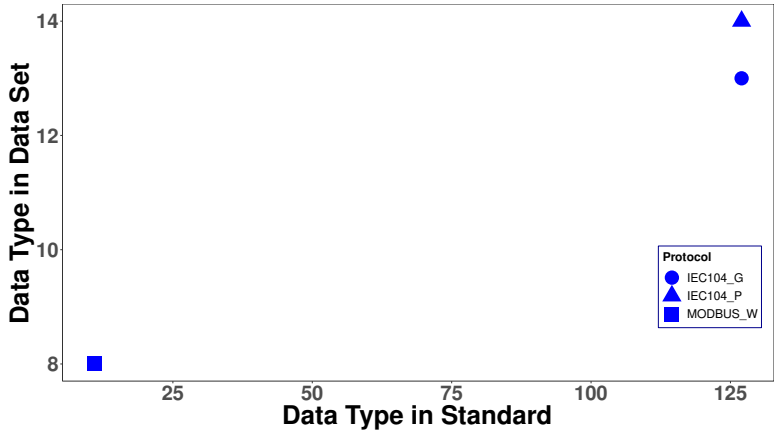


Figure 6.6: Data Types. IEC 104 gives a lot of flexibility for types, and they are used by operators

(S), or Unnumbered (U) APCI format. I-format packets are used to exchange sensor and control data, while S and U-format packets are used only for network signaling (acknowledgments and heartbeats, respectively).

For I-format packets, the standard [3] defines 127 different types of Information that can be exchanged. However, our Power network uses only 14 types, and Gas uses only 13 types, as seen in Tables 6.1 and 6.2 respectively. It is possible that we have not observed all types of traffic from these networks due to the brief duration of our traffic capture. Nevertheless, since our traffic is consistent with steady-state conditions, these data reflect the most frequent data message used during the operating stage of our power grid captures. In Table 6.1, we can see

that 99% of the information exchanged corresponds to only two types of messages: Type 36 (Measured value, short floating point value with time tag) and Type 13 (Measured value, short floating point). These values are power, voltage, and current measurements. The next most popular message (at only 0.845%) is perhaps the most critical message sent in this network: setpoint control commands to change the behavior of large power generators. Through these setpoint control messages, the SCADA system controls power generation in the grid. It is important to note that in this network less than 1% correspond to the binary values that report the status of actuators which indicate that power network is focused on the dynamic of the process.

On the other hand, gas networks give equal importance to monitoring the dynamics and status of the process (changes in actuators). Table 6.2 shows that status data such as Types 1 and 3 (28%) is the same proportion as measurement data such as Type 9 and 11 (around 29%). Furthermore, it is noteworthy that Interrogation commands, Type 100, make up a large portion, 36% of the traffic. This is the data type used by the controller to request data from agents, which is different from power network that do not require extra traffic to collect data since agents report data using spontaneous and periodic cause of transmission.

Nevertheless, both IEC 104 networks employ a limited number of data types in their process supervision. Although the IEC 104 standard offers a wide range of data types, the majority of them are not utilized in practice, regardless of the distinctions in network monitoring.

Water: In contrast to the different types available in IEC 104, other standards do not have the same diversity. Modbus is one of the oldest and simplest protocols used in SCADA systems. It only has three types of data: (Coils) bit access for binary values such as ON/OFF, (Registers) 16-bit access for continuous values,

and file record access. Modbus provides 11 function codes to interact with these variables. From these 11 function codes, our network uses eight functions, as seen in Table 6.3. Like power network, we can see that reading analog values makes up most of the traffic (89.76%) in the water network. 9.91% are reading binary values (the status of switches), while a very small percentage (0.34%) is a control command that changes the status of one of the binary values.

In summary, while the gas balance monitors the dynamic (analog measurements) and status (binary status) of the process, the power and water network monitors focus on the dynamic of the process. Therefore, the analog value data types are the predominant data type in the three networks in general.

6.5 RQ5: Flow, Monitoring and Control traffic

6.5.1 Flow direction:

We now look at our last research question, which relates to the direction of the flows. Traditionally, we expect SCADA networks to send more data to the controller than what the controller sends to the field devices. However, we do not see this pattern in most of our networks, mostly because of the interrogation commands.

We define a2C as the flow direction from the agent to the controller and C2a as the controller to the agent.

- ⋆ $a2C > C2a$ - In Power $a2C$ = 83.7% and $C2a$ = 16.3%. That is, most of the packets come from the remote substations (agents) to the controller. Based on that, we also make the following observation:

  Event Driven: Four-fifths of the traffic is in the monitoring direction ($a2C$). From that, the vast majority of the traffic (97.06%) are I-format

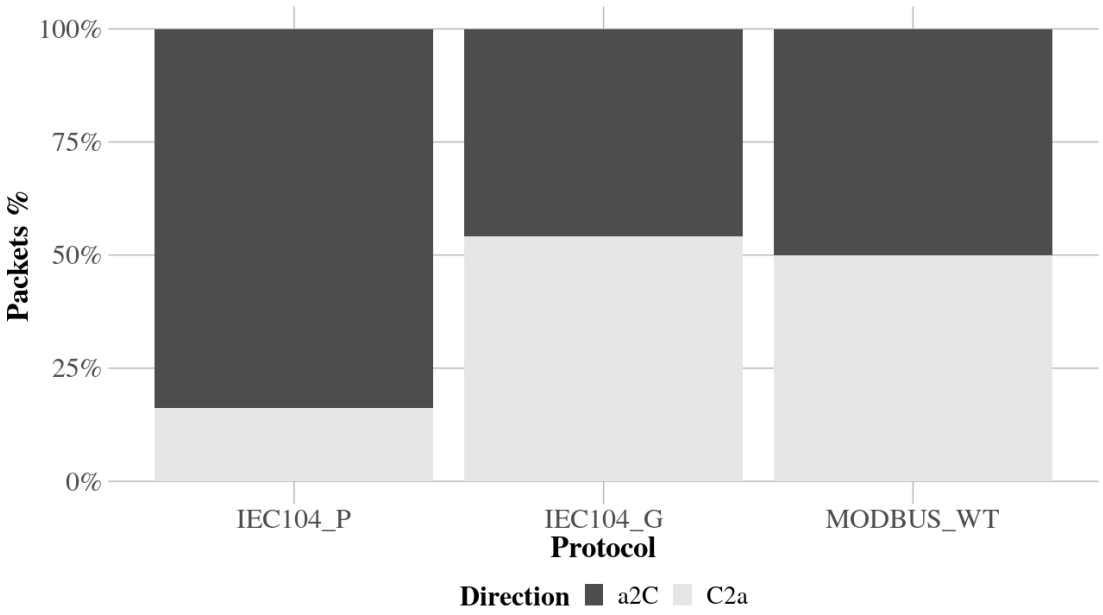


Figure 6.7: Flow directions: from Controller to agent (C2a) vs from agent to Controller (a2C)

messages. 90% of those are spontaneous packets (cause of transmission (COT) code '< 3 > spontaneous'). That means that 88% of the traffic from agents to controllers is generated by the occurrence of a particular event. Thus, most of Power's agents report changed data to the controller rather than sending static data (cyclic/periodic). For example, a change in the state of a binary point (e.g., a switch that passes from off to on), or in the case of analog points, when the values exceed a certain threshold (e.g., a frequency passes the 60.2 Hz threshold). In addition, as shown in the Table. 6.1, around 60% of the packets are time-stamp data. This is important for logging events, forensic analysis, and real-time control which is evidence that Power focus on the monitoring of the dynamic changes of the system.

Therefore, this indicates a network that prioritizes real-time monitoring and rapid response to changes in order to be able to react more quickly to real-

time changes, making it an event-driven network. This is crucial to the stability of the power grid.

Minimal Overhead: In an event-driven architecture like Power, resources are utilized more efficiently, given that data transmission is primarily triggered by significant events. This minimizes the amount of 'noise' in the system by reducing the transmission of redundant or unnecessary data. The approach also ensures that the network bandwidth is optimally used, making it easier to scale the system in the future or allocate bandwidth for other critical applications. Moreover, by focusing on real-time, event-triggered data, the system is better equipped to quickly identify and respond to abnormal conditions, thereby enhancing the overall reliability and security of the power grid.

Only one-fifth of traffic is generated by the controller, mainly for flow control: 80.7% for message control (*S-Format*[1] packets), and connection control (16.7% *U-frame*[2]). This means that most of the C2a data (80.3%) is dedicated to message acknowledgment, which is a small percentage (16%) of total traffic (a2C + C2a).

By looking the traffic data, we deduce that the controller has a larger acknowledgment window (w)[3] equal to 8. This means more data packets can be in flight before requiring an acknowledgment, resulting in better throughput. Additionally, an agent can send multiple packets before waiting for an

[1] *S-format* is a control field packet used for controlling the transport of information (ASDU packets). This protects against loss and duplication of I-format messages.

[2] *U-format* control field used to control the connection between stations. It is used as a start-stop mechanism for information flow. As a heartbeat to check connection. Also, as a mechanism for changeover between connections without loss of data when there are multiple connections available between stations.

[3] w specifies the maximum number of received I-format APDUs that the receiver should ACK at the latest. e.g., a w = 8 means that the controller will send to the agent an S-format message to ACK the last 8 I-format messages it receives.

acknowledgment, thus reducing round-trip time and improving latency.

★ $a2C < C2a$ - In the Gas network, more traffic is sent out from the controller than what is sent by the agent. There is a noticeable difference in the distribution of packets between controller-to-agent ($C2a$) and agent-to-controller ($a2c$): $a2C$ = 45.8% and $C2a$ = 54.2%. This imbalance can be attributed to two factors:

(1) **Polling Mechanism:** The controller employs Interrogation Commands to solicit data from the agents. These commands are sent as I-Frame packets, increasing the packet count in the $C2a$ direction.

(2) **Acknowledgment Scheme:** Unlike the controller, which acknowledges the receipt of each I-frame from the agent with an S-frame (w = 1), agents do not reciprocate. When the agent receives an I-frame (Interrogation command) from the controller, it sends back the requested data in an I-frame but does not acknowledge it with an S-frame. This unidirectional acknowledgment contributes to the imbalance in packet distribution.

In essence, for each cycle of data exchange initiated by a polling command, the controller sends two types of frames (first an I-frame to request data and then an S-frame to acknowledge receipt) while the agent only sends one I-frame in response. This results in a higher packet count in the $C2a$ direction.

We add the following observations:

- **One-to-One Acknowledgment:** Unlike the Power network, which waits until it receives 8 I-frames before it sends back an acknowledgment (w = 8), the Gas network operates with a smaller window size of just 1. This is because the interarrival times between I-frames is

so large (order of minutes, see Section 6.2.2), that they need to send acknowledgments for every packet.

- **A non-standard use of the IEC 104 protocol in an ICS:** in the Gas network, the controller utilizes station interrogation (interrogation commands) instead of Cyclic data transmission to synchronize the process data of the agents. The difference is that Interrogation commands acquire a full set of data, while polling only gets the data that is of interest. Interrogation commands are used to update the controller after initialization or after data loss or corruption of data [15].

  On the other hand, cyclic data is used to provide periodic updating of the process data to current values. Interrogation commands are event-based (loss of communication) or manually initiated (start a communication). Another difference between data acquisition by the Interrogation command and cyclic is that the former requires a request, while the latter does not. Interrogation commands are used to poll data from the agent, while cyclic does not require any commands; it is generated automatically by the agent (less traffic). Polling data by using interrogation commands is like a request/response; however, the agent can send the response in several messages, unlike Modbus, which sends the response in one message. The Gas network does not use cyclic data transmission, only general interrogation for polling data from agents. This is a non-standard use of the IEC 104 protocol in an ICS. It appears to be using a legacy approach like the one used in Modbus, but it implements it in IEC 104, without taking advantage of the new transmission mechanisms that modern protocols provide.

★ $a2C = C2a$ - In the final case, our Water network has an equal amount of

packets being sent by the controller to the agent, as well as from the agent to the controller. Like the Gas network, our Water network operates on a polling mechanism. Given its request-response protocol architecture, Water exhibits an equal traffic flow in both directions ($a2C = C2a$). In essence, for every data report the agent sends, the controller initiates the communication by sending a request. This implies that the controller frequently queries the agent to retrieve the latest state information or execute specific commands.

Response Granularity: Both Water and Gas utilize a polling-driven mechanism, but they diverge in how responses are sent by agents. For instance, in Gas, an agent might respond to an Interrogation command with 5 packets, each containing 21 IOA of ASDU Type 9. This results in 7 packets for the entire transaction: 3 for an Act, ActCon, ActTerm packet, and 4 for the actual data points. This increases the total number of packets in the transaction for one request. In contrast, Water adheres to a one-to-one request-response model, with each request from the controller receiving a single packet response from the agent. Consequently, a complete transaction in Water consists of just 2 packets: one for the request and one for the response. This streamlined approach minimizes packet loss and reduces network latency, making it more suitable for the time-sensitive operations in a water treatment facility.

6.5.2 Monitoring vs Control:

While the analysis above shows the diversity of the direction of the flows in SCADA networks, this flow-based (network layer) analysis without content analysis can be misleading if we want to quantify how many commands control centers send to their endpoints. We now categorize all of the traffic as either command

or measurement. Command data refers to instructions given by the Controller to request data or to set parameters in the network or the grid. Measurement data is the collection of information about sensor data values (analog, binary values) and the status of the network (clock synchronization, etc.). For example, some protocols like IEC 104 make this distinction easy by labeling each type as a "command type" (e.g., interrogation command) vs. a measurement type (e.g., measured value, short floating point). For the other protocols, we must look at the types and infer which ones are related to command actions (e.g., read vs. write in Modbus).

When we looked at the relationship between command data and measurement data, **we identified that more than 95% of the traffic is measurement data for all the protocols**. As shown in Fig. 6.8a, only the Gas network contained a significant amount of command data. However, for Power and Water, the percentage of command data is very low, 0.43% and 0.32%, respectively.

Looking in more detail at the types of commands, we define two types: **(1) control commands** and **(2) configuration commands**. Control commands make changes to the physical world. Configuration commands keep the devices and network configured correctly. Only one (Water) uses control commands. The Gas uses only configuration commands, as seen in Fig. 6.8b. In the case of Power, there are control commands for ramping up or down the generation of power plants. These commands are part of the AGC that the CR uses to maintain the power balance in a system. Even so, these commands are only 0.8% of all the traffic for this network. While Gas has a significant portion (40.17%) of configuration commands in our data (two types: interrogation and read command, Table 6.2).

In summary, we infer these networks run largely on "auto-pilot" with little interference by control commands. As far as we can see from our data, **control**

commands seem to be the exception rather than the rule and are not used throughout the whole network but rather confined to certain parts. Out of the small percentage of control commands that are sent, most of them (95%) are just monitoring commands.

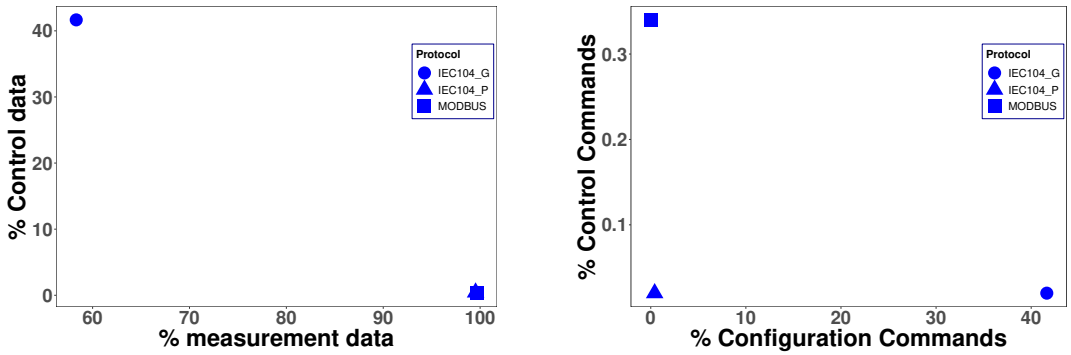


Figure 6.8: (a) Control vs Configuration commands, (b) Control vs Measurement data.

6.6 Conclusion

In this chapter, we uncovered revealing patterns and operational behaviors across different Industrial Control Systems. Through detailed analysis, it became evident that while Power grid and Water treatment networks often adhere to conventional protocol applications, the Gas distribution network deviates, reflecting intriguing operational choices—particularly in its data collection mechanism.

Notably, consistency was observed in the traffic across all networks: approximately a quarter of their traffic exhibited uniform packet sizes. Furthermore, a predilection for small packet sizes, falling within the 0-100 bytes range, was dominant in all three networks, accounting for more than half of their communications. When exploring transmission timings in ICS, durations typically spanned

from seconds (Power and Water) to minutes (Gas), with millisecond-order transmissions being rare exceptions and relatively inactive network exemplified by Gas.

IEC 104 networks, a pivotal focus of our study, revealed two consistent operational tendencies: a minimal IAT that hovers around one second and a maximum packet size capped at 200 bytes. Additionally, our findings highlighted the predictability of packet sizes for polling-oriented networks (such as Gas), while event-driven (like Power) networks presented a richer tapestry of diversity and fluctuation.

This exploration emphasized the heterogeneity within SCADA networks and the importance of customized and specialized approaches for each infrastructure. Our findings challenge generalized views on SCADA networks, advocating for more nuanced in the studies of these industrial systems networks.

In this chapter, we explore the diversity of SCADA networks in three different industrial control systems: Power, Gas and Water. We have observed their particular traffic characteristics and operational behaviors. We now move on to a more concentrated study in the chapter. 7, where we will zoom into the SCADA networks within a specific industrial system: the power system. We analyze the diversity of SCADA traffic across the entire supply chain, from generation to end-customer. By closely examining SCADA traffic within this particular industrial sector, we aim to gain a better understanding of its nuances in the power grid. This focused analysis will not only help us to gain a better understanding of SCADA networks in the context of power, but also provide insight into the theme of our study: the diversity and complexity of SCADA networks in various industrial settings.

Chapter 7

SCADA World: An Exploration of the Diversity in Power Grid Networks

In Chapter 6, we conducted a thorough examination of SCADA systems in various ICS, uncovering unique operational and traffic patterns. In this chapter, we focus our research on SCADA networks in the power grid. Our goal is to analyze, differentiate, and comprehend the SCADA networks in the energy supply chain, from generation to the end-user, in order to gain a comprehensive understanding of the complexity and heterogeneity of SCADA networks in critical infrastructure contexts.

Power grids are an essential part of today's world. They comprise a set of interconnected electrical grids that generate, transport, and deliver electrical power to consumers. Any disruption to the system can have significant societal, economic, and political consequences. Given that the reliable operation of modern power grids relies on computer networks, the scientific measurement community

needs to start discussing and analyzing these networks.

SCADA networks form the backbone of computer technology in the power grid. While the academic community has started to show interest in SCADA systems, obtaining operational data from power grid companies is difficult. As a result, most researchers are restricted to examining a single industrial protocol in datasets gathered from a single small part of the power grid. This means that prior research has studied networks in isolation from small parts of the power system. As a result, research papers portray SCADA networks as monolithic network infrastructures with periodic traffic and a fixed topology.

Through years of outreach, we have gathered operational data from the most extensive and diverse set of SCADA networks in the power grid. We use this unprecedented access to characterize the classes and diversity of SCADA networks, as well as to test if previous assumptions about their behavior hold in different networks.

Our aim is to compare and contrast the similarities and differences between these protocols and to gain insight into the diversity of SCADA networks.

This chapter presents an analysis of real-world data from a bulk power grid, a transmission substation, a distribution grid, and an end-customer facility. We focus on five industrial protocols: IEC 104, GOOSE, ICCP, C37.118, and Modbus/TCP (MODBUS).

7.1 Grid Operators

In addition to the diversity of endpoints (Section. 3.2.2), we have data from diverse companies, each with a different role in the power grid. In general, multiple companies oversee the bulk power system and the distribution system, each using their own SCADA systems that run one or more industrial protocols to

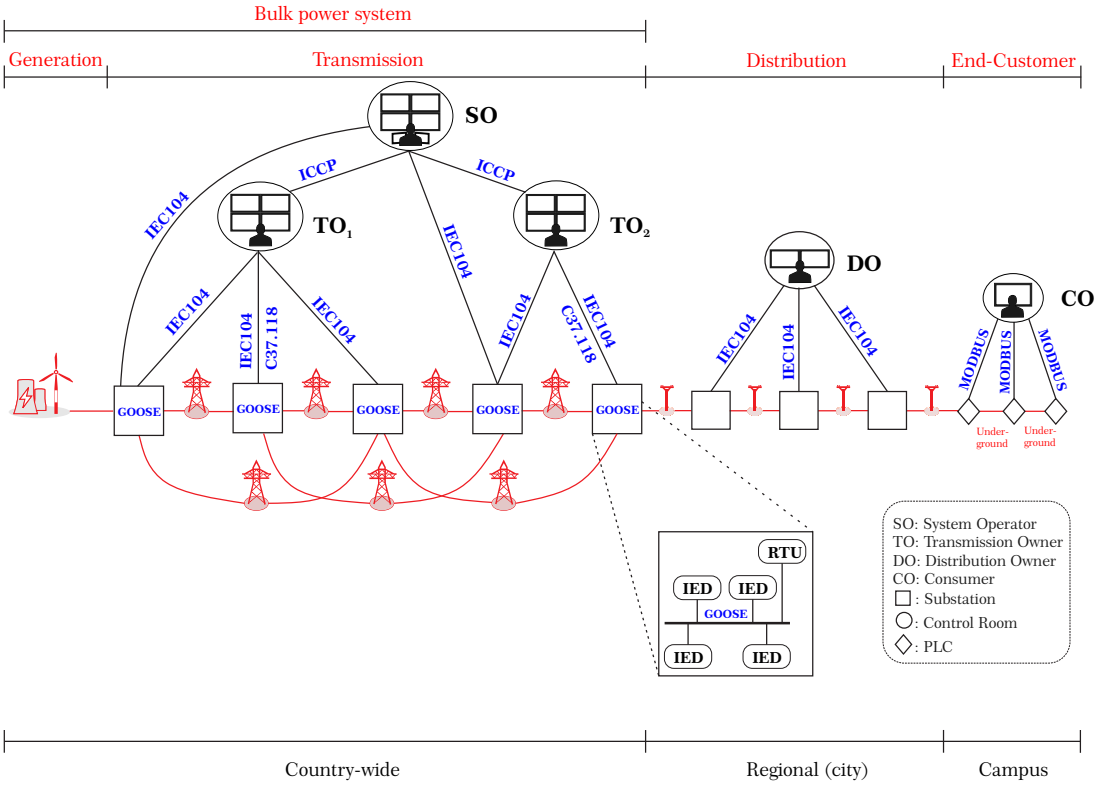


Figure 7.1: Representation of the industrial protocols contained in our dataset (e.g., IEC 104), the electrical and computer networks of the power grid from generation to end-customers. The operation of the power grid requires the coordination of multiple entities (SO, TO, DO, CO), protocols (IEC 104, GOOSE, Modbus, etc.), and devices (IEC, RTU, PMU, PLC).

monitor their part of the system. The main players (grid operators) in the bulk power grid are the system operators and the transmission owners. There are also distributor owners and consumers, such as small industries or facilities (such as a water treatment plant or a university campus) that supervise their portion of the grid. Figure 7.1 illustrates the different companies (SO, TO, DO, CO) and their respective control rooms.

System Operators (SO) orchestrate the operation of all power companies. In the U.S., system operators are called either Regional Transmission Operators (RTO) or Independent System Operators (ISO), depending on whether they ad-

minister the power grid among several states (RTO) or if they operate the grid in one state only (ISO). For example, the California Independent System Operator (CAISO[1]) operates the power grid for the entire state of California.

The SO needs a Wide Area Network (WAN) to exchange information with multiple power system operators. In addition, the operator is in charge of running the Automatic Generation Control (AGC) algorithm to control the power output of electric generators within an area in response to the system frequency or tieline loading. Therefore, the operator will need a network to communicate control commands from the CR to generation substations. Ultimately, they may also receive PMU data (synchrophasor) to monitor the power system's dynamics and, as a result, will need a Wide-Area Monitoring System (WAMS) infrastructure to communicate with distant substations. Due to the high transmission rate of PMU technology and the time accuracy requirements, this network can utilize a different communication infrastructure from the typical SCADA network.

Transmission Owners (TO) own assets in the transmission system, such as electrical towers and substations, along with the associated equipment, such as transformers and circuit breakers. They need industrial networks to connect their central control room(s) to remote substations.

Distribution Owners (DO) (including electric utilities) own assets at the distribution level and manage several distribution substations, distribution lines, and delivery to end consumers. They need industrial networks to receive data from local substations and send control commands to them.

Consumers (CO) receive electric power from distribution utilities. While many consumers are residential, others are commercial or institutional consumers (e.g., a university campus), which supervise their electricity consumption among different buildings. They need a network connecting the CR to different PLCs in buildings.

[1] `https://www.caiso.com`

7.2 RQ1: Network Topology

To start the comparison of these networks, study their network structure. By looking at the IP or MAC address endpoints, we create graphs and then compute the **average degree** $\bar{k}$ and the **largest eigenvalue** of the adjacency matrix, since these are metrics that measure the robustness of the networks [51, 14].

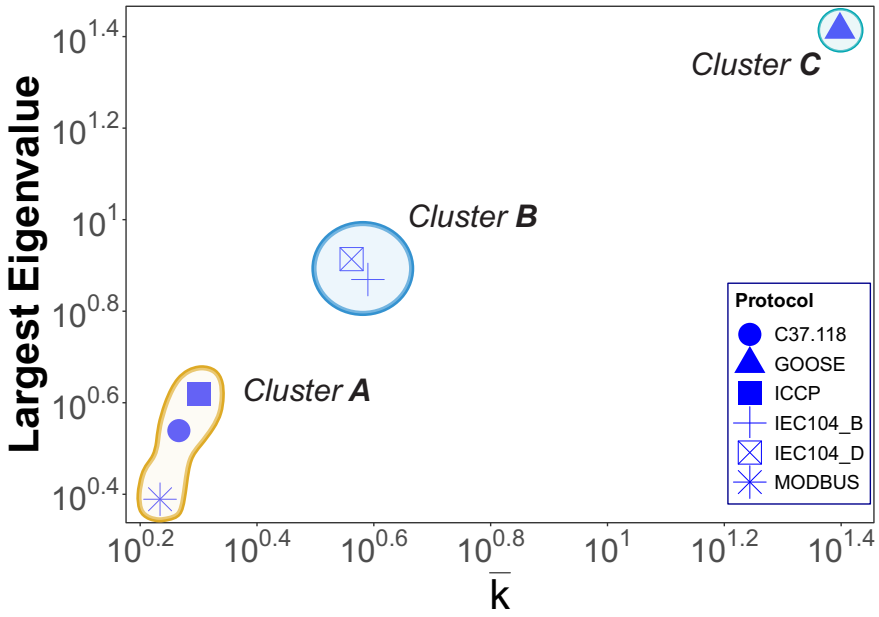
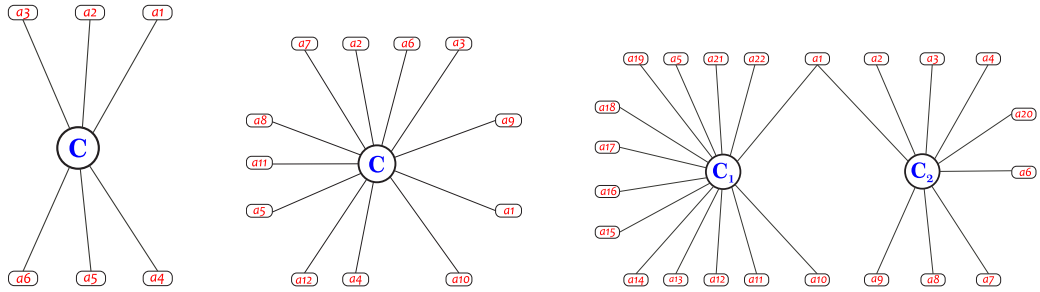


Figure 7.2: Network Robustness Metrics

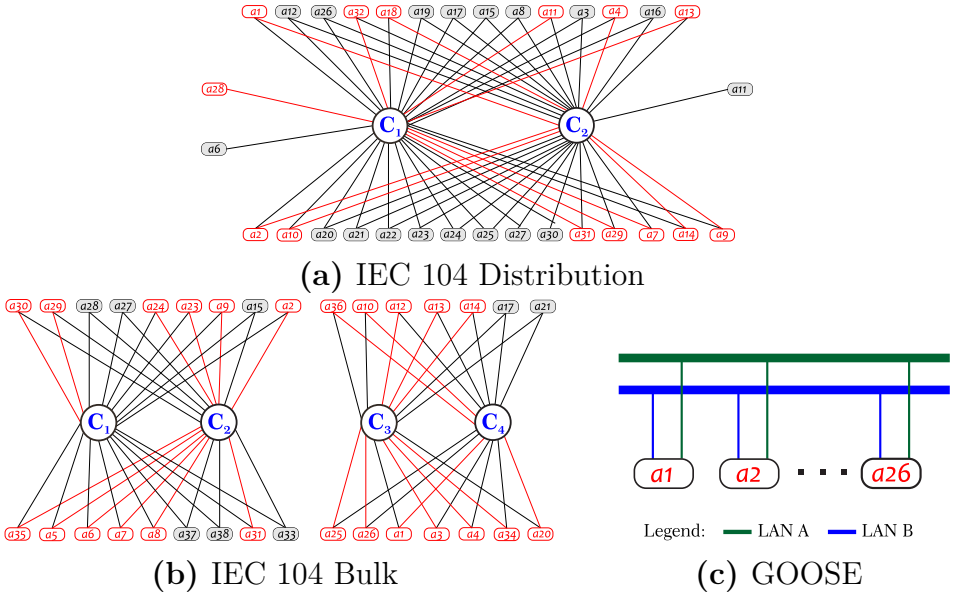
Fig. 7.2 illustrates these metrics in our data captures. Cluster C has the highest indicators of network connectivity, and the cluster only has the substation network using the GOOSE protocol (it is a LAN network with a redundant topology supported by a Link Layer protocol called HSRP). This makes sense as GOOSE is used for *protection* (i.e., to maintain safety in the system), and therefore these networks are robust and redundant so that safety messages can be shared effectively. Cluster B shows that the networks used to monitor substations (in transmission–IEC 104-B–and in distribution–IEC 104-D) tend to be more redundant and robust, because they have one or two control centers and each control

center has one or two control servers. Finally, the networks in Cluster A have a simple star topology where one central server connects to remote peers. A single cut in the network makes it disconnected.



(a) MODBUS **(b)** C37.118 **(c)** ICCP

Figure 7.3: Cluster A. *C*: Controller, *a:* Agent.



(a) IEC 104 Distribution

(b) IEC 104 Bulk **(c)** GOOSE

Figure 7.4: Cluster B (a),(b) and Cluster C (c).

We now look at these clusters and recognize three distinct topology configurations: star, complete bipartite, and complete graph.

Star Graphs: Cluster A consists of three networks: MODBUS, C37.118, and ICCP. This star configuration has each agent connected to a single controller

node. This setup is not fault-tolerant, as there is no backup for the central node, i.e., if the controller fails, the supervision is lost. In the case of MODBUS, this is not a major concern as it is a small network on a university campus and is not essential for the power grid. The C37.118 network transmits data to SCADA for monitoring purposes only. The operator has informed us that the PMU data are not being used for real-time control operations, so a lack of measurements is not a major issue.

On the other hand, the ICCP network consists of two (almost) separate star networks. Note that agent (*a1*) links the two-star graphs, which is a special case. The ISO company told us that $a1$ corresponds to the control room of the grid operator that owns the largest number of assets in the transmission system. If its data is lost due to a controller or connection failure, a large portion of the system will be without supervision. Also, we notice that *a1* has two simultaneous connections to each controller, where one-third of the data goes to *C1* and two-thirds go to *C2*.

Complete Bipartite Graphs make up cluster B. The only networks in this cluster are those using the IEC 104 protocol.

A complete bipartite graph $K_{p,q}$ consists of a set of p vertex and a set of q vertex (in our case, $p = 2$) and pq edges joining the vertex of different types [27]. This type of topology is known as Spine Leaf topology [48, 25] in cloud data centers. The difference is that the Spine Leaf topology is used to forward packets through the Spine (the central nodes), while in SCADA networks, the central nodes consume data (they do not forward it).

In our IEC 104 networks, each agent is connected to two controllers, except for three agents in IEC104_D (*a28,a6, and a11*) that registered only one connection. This dual-purpose configuration provides (1) fault tolerance and (2) load

balance. It reduces the risk of the CR losing control and supervision in the event of a controller or link fault. If one controller fails, the other will run, allowing control applications such as the AGC algorithm to access the input data necessary to perform its control operation. In addition, the CR operator maintains the supervision of the grid from its HMI [8, 7].

We look at redundant connections from the agents (RTUs) to the controllers (SCADA serves). We show in red (in Figs. 7.4a and 7.4b) links with an active connection (i.e., a connection sending data to the controller) and, in black, connections on standby (i.e., connections that send heartbeats to the controller to tell them they are connected and ready to receive any active connection). This type of load balancing makes sense as we only require one active connection between a controller and an agent, while the other connection is on standby ready to be used in case of a failure.

In IEC104_B, we see two bipartite graphs: $K_{2,18}$ and $K_{2,14}$. We confirm with the ISO that all four controllers are physically in the same control room, so this network represents a control room with four servers arranged in pairs, each monitoring a different part of the grid.

In summary, these $K_{2,q}$ graphs represent networks that are more critical, such as networks for control operations, than the simple star graphs from before and, therefore, need standby connections to the controllers.

Complete Graph: The GOOSE network is significantly different from any other network in our datasets for three reasons. First, it is a publish-subscribe network; therefore, sources of information send data as broadcast messages (as opposed to all other networks where there are clear end-to-end connections). Second, the network does not use IP addresses; instead, all the communication happens in the local network as Ethernet frames. Third, all devices of the GOOSE network have

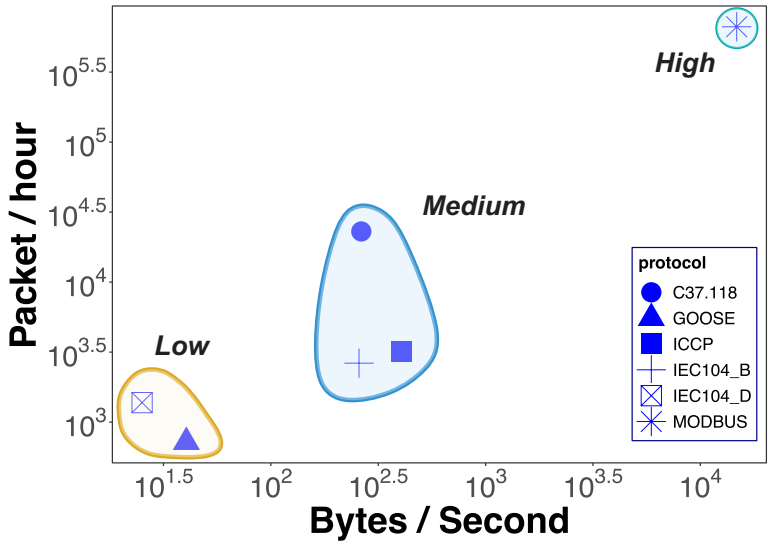


Figure 7.5: Packet/hour vs Bytes/second

two Ethernet adapters, and they send data redundantly (i.e., duplicate packets) in each interface; therefore, the physical connection of the network corresponds to two Ethernet networks in parallel to each other, each serving as a backup to the other network. This Ethernet redundancy is managed by the parallel redundancy protocol [34], which was designed for the recovery of highly available industrial networks.

While the physical topology corresponds to two parallel Ethernet networks, the logical topology is hard to determine. All we see in the pcap files are broadcasts of information. To learn which devices subscribe to each message (i.e., if the device will actually read the received Ethernet broadcast), we need extra information, such as the Substation Configuration Description (SCD) files [40], which denote which devices will read which messages. Unfortunately, we requested but did not receive such files. Therefore we mention that this is a fully connected graph in capability only–each agent has the ability to read the messages that any other agent sends in the network.

This architecture shows how GOOSE was designed for highly critical situa-

tions, where we need to ensure that messages are received in a very short time (3 ms for the most critical events [12]) and that they can be read by as many agents as needed. In fact, GOOSE networks are not considered monitoring networks (measurement acquisition); instead, they are considered events (alarms, status change) and control exchange [31] networks for **protection**. In contrast to monitoring, protection focuses on automatic responses to electrical disturbances.

Summary: We find three different patterns in our networks: star topologies for low-criticality monitoring-only networks, $K_{2,q}$ topologies for critical control networks, and a fully connected network to exchange highly critical event messages for protection. Outside the two-star graphs (MODBUS and C37.118), no other network in our dataset has the same topology.

7.3 RQ2: Traffic Pattern Differences Between Networks

We now focus on whether or not we can assume that all of them will have similar traffic flows. For example, are all flows periodic messages from agents to the controller? If not, how are these flows different from other networks? and what are the root causes for these differences?

To characterize the flows between networks we take the amount of packets sent between end devices (agents and controllers). Fig. 7.5 illustrates the amount of data being exchanged per protocol. Each axis shows the median data rate (of the per-connection median data rates) between an agent and the controller. We again identify three clusters, corresponding to three different data rates, (1) Low, (2) Medium, and (3) High.

Low data rates are represented by two networks: GOOSE and IEC 104 Distri-

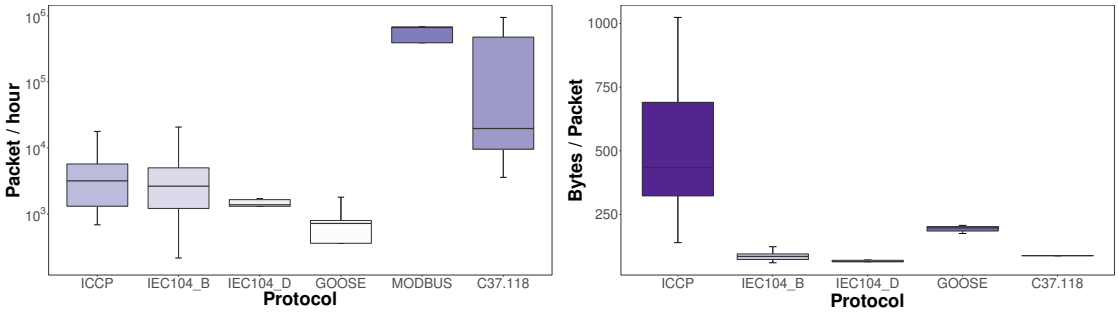
bution (IEC104_D). The reason for the low data rate in the GOOSE network is that during the 14 hours of our capture, there was no emergency, and therefore, there was no need for IEDs to report and change their configuration. So what we see in the 14 hours of our data capture, are simply regular heartbeats by the devices stating that nothing has changed since the last heartbeat. A similar argument can be made for the IEC104_D network, as the control center appears to be monitoring sporadically the status of devices in different substations.

Medium Data Rate: This cluster consists of all the wide area networks in the bulk power grid. As expected, monitoring and controlling the bulk power grid requires more information sent (per connection) because the bulk power grid is more complex than the distribution grid, and therefore substations, CRs, and PMUs have more information to send to control centers than the distribution grid.

High Data Rate: The high data rate network in our dataset appears to be an anomaly. This is a local area network that was configured by University operators to send information at a very high frequency. Furthermore, we also identify that while most of the data in the other networks is sent "spontaneously" (i.e., without a request from the controller), the MODBUS protocol does not have this capability. In order for the controller to receive information from agents (PLCs in this case), the controller needs to first send a request, which is then answered by the agent. This essentially doubles the amount of data being sent back and forth in the network.

We now look in more detail at the way each network sends its data by looking at the box plots of the packet rates and the packet sizes in Fig. 7.6.

Packet Rates: Fig. 7.6a illustrates the significant difference between the number of packets per hour of the C37.118 and MODBUS networks compared to the others. The number of packets per hour exchanged in the C37.118 and MOD-

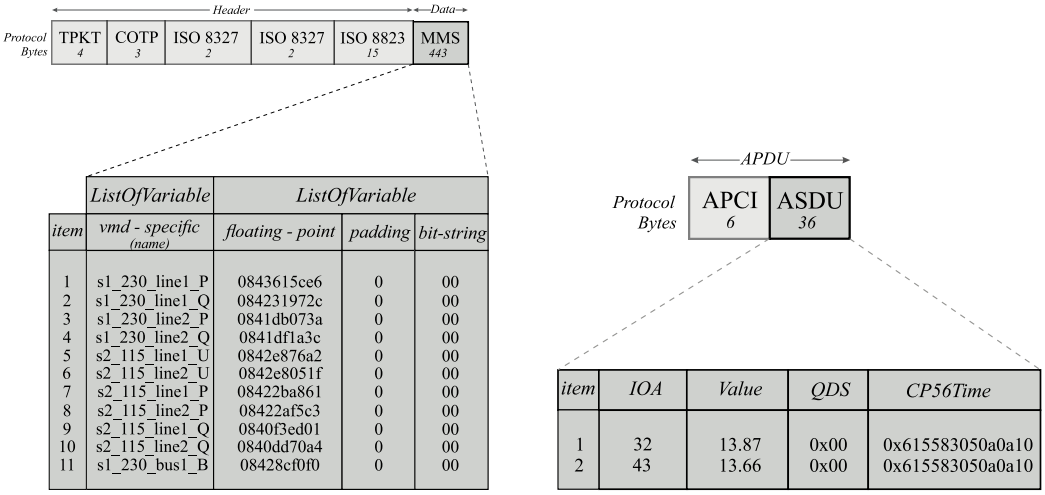


(a) Packets per hour **(b)** Packet size

Figure 7.6: Diversity between network flows.

BUS networks is one or two orders of magnitude larger than the other networks. This makes sense because PMU is a new technology that takes high reporting rates (maximum 60 or 120 frames per second [6]) voltage, currents that are time-stamped with high-precision clocks. These new real-time measurements can reveal system changes that are undetectable by traditional monitoring systems such as oscillation detection; therefore, we expect them to be the most active WAN in our datasets. With MODBUS, on the other hand, it is less clear why there is such a high rate of packets. One possible explanation is that there is no way for a MODBUS agent to report any change in a data point when it occurs without a query. The controller needs to keep requesting new data at a low frequency in order to identify any change in the status of the grid. Finally, Fig. 7.6a also shows that the networks that have the largest variance are again those in the "Medium Data Rate" Cluster: the SCADA networks operating in the bulk power grid: ICCP, IEC104_B, and C37.118. These networks have more diverse devices and tend to be more active.

Packet Sizes: From Fig. 7.6b, it is evident that ICCP is the main outlier. Not only does it have the largest packets, but it also has the greatest variance in packet sizes. The smallest packet sent in the ICCP network is larger than the largest packets sent by all other networks (except GOOSE), and the median packet



(a) Typical ICCP packet **(b)** Typical IEC104_B packet.

Figure 7.7: Payloads of ICCP vs IEC104_B (11 vs 2 measurement points).

size in the ICCP network is more than double the median of a GOOSE frame. Upon looking at this outlier we found that this is because ICCP shares data between control rooms, meaning *agents exchange data from multiple substations*. On the other hand, each endpoint in any non-ICCP network shares data from a single device or substation, thus ICCP networks transmit much more information (data points) per packet. For example, Fig. 7.7a and Fig. 7.7b compare the typical length of an ICCP and an IEC104_B packet. While the IEC104_B packet contains 2 data points from a single substation with a payload size of 36 bytes, the ICCP packet contains 11 data points from 2 substations (s1_230 and s2_115) with a payload size of 443 bytes. Lastly, when analyzing packet sizes, we found that MODBUS packets were truncated by our University, so we exclude MODBUS from the packet size analysis.

Summary: We find that the networks with the highest variability are those in the bulk power grid because these are the ones that cover the longest distances and diversity of devices. In contrast, the campus, distribution, and substation networks have low variability. We also found two outliers: ICCP is an outlier in

packet sizes because of the large amount of data that control rooms need to exchange to synchronize their state. The other outlier we found was the MODBUS network and its unusually high frequency of packet transmissions. We did not find a good reason for this behavior, but we note that these networks are highly configurable and in this case, the operator configured the network to report, perhaps more often than needed, the status of all PLCs.

7.4 RQ3: Diversity Within Networks

The variance of the flows in some networks in the last section tells us that there are some networks with high internal diversity. We now turn to examine the diversity of traffic activity within networks.

Fig. 7.8 and Fig. 7.9 illustrate how some networks have periodic traffic patterns which means most agents send similar-sized packets at regular intervals. Other networks have more agent activity diversity, sending packets of various sizes at different rates.

Fig. 7.8 suggests that traffic rates can be divided into two categories: periodic and aperiodic packet transmission. Periodic packet transmission refers to agents configured to send repeatedly the same packets over and over at the same intervals of time. They represent highly predictable behavior. We can see in that figure that our GOOSE, MODBUS, and C37.118 networks have several agents transmitting periodically and also forming very clear clusters of periods.

The periodicity of our GOOSE, MODBUS, and C37.118 networks can be attributed to the fact that their agents are highly homogeneous, sharing the same configurations. On the other hand, agents of ICCP and IEC 104 networks are heterogeneous, with different configurations, resulting in a higher diversity of packet characteristics. GOOSE agents are IEDs located in substations, C37.118 agents

are PMUs in substations, and MODBUS agents are PLCs, configured by one administrator. On the contrary, ICCP agents are CRs managed by different companies, and IEC104_B endpoints are RTUs in substations with varying equipment and managed by different entities. This makes uniformity in the configuration of the agents impossible.

For example, the GOOSE network is fairly predictable; Fig. 7.8a. The traffic rate reveals constant communications patterns with two characteristics: (1) periodic traffic, and (2) clusters with identical periods. We argue that this periodicity can be attributed to the network configuration among agents. Looking at the packet capture (and the standard parameters) we can infer Table 7.1 to explain the four levels of periodicity seen in Fig. 7.8a. For example, one cluster of agents is represented by sending 12 packets/min. Each agent transmits a packet per dataset every ten seconds. Since there are two datasets (GOOSE data unit) per agent, every agent has to send two packets every ten seconds. On the other hand, *a8* is the outlier because it is the only agent that must transmit three *datasets*. We can do a similar analysis for other periodic networks such as C37.118, and MODBUS.

Table 7.1: GOOSE network agent configuration

Pkt/min	**Cluster**	**Time_max [s]**	**#Datasets**
30	*a5, a1, a12, a20, a13, a3*	2	1
18	*a8*	10	3
12	*a21, a4, a9, a11, a18 a22, a27, a15, a23, a24*	10	2
6	*a10, a28, a14, a19, a2, a17 a26, a7, a16, a25, a6*	10	1

In contrast, we cannot obtain a similar simple table for the configuration of agents for IEC 104 networks, or for the control rooms in ICCP. One of the key

differences is that ICCP and IEC 104 transmit spontaneous packets (i.e., the transmission of the packet depends on the state of the grid, and not on any preconfigured timer). The other main reason for the diversity in these networks is that the endpoints are not homogeneous. Like ICCP, the agents of IEC 104 networks are not homogeneous, they represent RTUs in substations with different sizes, different equipment, and different functionalities, so even if there were no spontaneous packets, the configuration of each endpoint will be different.

A similar analysis can be made relating to packet sizes. Endpoints with the same packet size indicate devices with the same configuration.

Summary: The similarity of traffic within a network can be explained by the fact that the agents of our GOOSE, MODBUS, and C37.118 networks are very similar in configuration; in contrast, the agents of ICCP and the IEC 104 networks are not, thus leading to more diversity within those networks. For example, the endpoints of GOOSE are IEDs, the endpionts of C37.118 are PMUs, and the endpoints of MODBUS are PLCs. In contrast, the endpoints of ICCP are control centers managed by different companies. Similarly, the endpoints of IEC104_B are RTUs reporting all the data from substations with diverse equipment and managed by different companies. So having uniformity in the configuration of the agents is impossible. Another big difference is the fact that the networks with more diversity send packets spontaneously (i.e., depending on the state of the power grid, while the networks with less diversity send packets at pre-configured intervals and containing the same datasets).

7.5 RQ4: Information Types

We now turn our attention to the information contained within the packets themselves. Each protocol has its own specific standard with clearly defined types

of data that are included in the packets. We start with a general overview in Fig. 7.10. On the x-axis, we see the number of data types defined by the standard, and on the y-axis, we see the number of types present in our capture. There are two clusters: On the top right corner, we can see the IEC 104 networks which have over 100 data types defined in the standard but only use around 10% of them in the capture. On the bottom left-hand corner are the rest of the protocols. Their standards do not contain many data types so, obviously, the variety of data types contained in the capture are relatively few.

Now we look at the two clusters in more detail:

IEC 104 Cluster: An IEC 104 packet can be either in Information (I), Supervisory (S), or Unnumbered (U) APCI format. I-format packets are used to exchange sensor and control data, while S and U-format packets are used only for network signaling (acknowledgments and heartbeats respectively).

For I-format packets the standard [3] defines 127 different types of Information that can be exchanged. However, our IEC104_B network uses only eleven types and IEC104_D uses only seven types, as seen in Tables 7.2 and 7.3 respectively. It is conceivable that we have not observed all types of traffic from these networks, due to the brief duration of our traffic capture. Nevertheless, since our traffic is consistent with steady-state conditions, these data reflect the most frequent data message used during the operating stage of our power grid captures. In Table 7.2 we can see that 99% of the information exchanged corresponds to only two types of messages: Type 36 (Measured value, short floating point value with time tag) and Type 13 (Measured value, short floating point). These values are power, voltage, and current measurements. The next most popular message (at only 0.845%) is perhaps the most critical message sent in this network: setpoint control commands to change the behavior of large power generators. It is through

these setpoint control messages that the SCADA system controls power generation in the grid.

Table 7.2: IEC104_B ASDU types and their description

Type	Reference	Description	%
36	M_ME_TF_1	Measured value, short floating point value with time tag CP56Time2a.	72.4
13	M_ME_NC_1	Measured value, short floating point value.	26.5
50	C_SE_NC_1	Setpoint command, short floating point value.	0.845
3	M_DP_NA_1	Double point information.	0.183
100	C_IC_NA_1	Interrogation command.	0.003 53
9	M_ME_NA_1	Measured value, normalized value.	0.002 52
31	M_DP_TB_1	Double point information with time tag CP56Time2a.	0.002 52
103	C_CS_NA_1	Clock synchronization command.	0.001 51
30	M_SP_TB_1	Single point information with time tag CP56Time2a.	<0.001
1	M_SP_NA_1	Single point information.	<0.001
70	M_EI_NA_1	End of initialization.	<0.001

Table 7.3 illustrates the data types for IEC014_D. 93% of the values sent to the control center correspond to floating point values either in absolute values or normalized in the interval [-1,1]. These continuous variables represent typical power grid measurements such as power, voltage, and current. The third most common type in this network is noteworthy because it was the data type for IEC 104 that the Industroyer malware used to open circuit breakers in Ukraine [24]. Single point values usually report the state of binary variables, such as the status of circuit breakers. In our network, type 30 starts its reference description with an "M" which stands for "measurement," so this means that this type was sent from

the RTU to the control center. The specific type used by Industroyer must start with a "C" (command) in their reference.

In short, while the IEC 104 standard provides a large amount of data types, most of them are not used in practice, and the only reason we get a different cluster in Fig. 7.10 is because of low probability events.

Table 7.3: IEC104_D ASDU types and their description

Type	Reference	Description	%
9	M_ME_NA_1	Measured value, normalized value.	50.8
36	M_ME_TF_1	Measured value, short floating point value with time tag CP56Time2a.	42.3
30	M_SP_TB_1	Single point information with time tag CP56Time2a.	4.39
100	C_IC_NA_1	Interrogation command.	2.19
103	C_CS_NA_1	Clock synchronization command.	0.246
37	M_IT_TB_1	Integrated totals with time tag CP56Time2a.	0.140
31	M_DP_TB_1	Double point information with time tag CP56Time2a.	0.002 43

Non IEC 104 Cluster: In contrast to the different types available in IEC 104, other standards do not have the same diversity. To illustrate these restrictions, we use MODBUS. This protocol is one of the oldest and simplest used in SCADA systems. It only has three types of data access: (Coils) bit access for binary values such as ON/OFF, (Registers) 16-bit access for continuous values, and file record access. MODBUS provides 11 function codes to interact with these variables. Of those 11 function codes, our MODBUS network uses only three functions (1, 3, 5), as seen in Table 7.4. As with our IEC 104 networks, we can see again that measuring continuous variables makes up most of the traffic (76.9%). 23.09%

measures binary values (the status of switches), while a very small percentage is a control command changing the status of one of the binary values.

Table 7.4: MODBUS types and their description

Type	Description	%
3	Read Holding Register	76.9
1	Read Coils	23.09
5	Write Single Coils	<0.001

C37.118 contain the less diverse types with just one type of data in our dataset. The other networks in this cluster, while more expressive (e.g., GOOSE has 4 types: structure, boolean, bit-string, utc-time, and ICCP has 7 types: structure, floating-point, unsigned, bit-string, string, integer and boolean), they still represent mostly monitoring of floating point values.

Summary: The IEC 104 protocol provides a rich set of data types, however, more than 90% of the data sent in IEC104_B and IEC104_D is simply a floating point measurement. There are several other commands seen in our IEC 104 networks, but most of them represent less than 1% of the traffic. While the other networks provide less flexibility for the types of data they can send, they mostly follow the same pattern of reporting floating point (continuous) values to other endpoints.

7.6 RQ5: Monitoring vs. Control

The last section suggests that most of the information exchanged in SCADA networks focuses on measurement values. However, the "C" in SCADA stands for *control*, i.e., sending control commands to an agent to change the operation of the physical process. In this section, we analyze the traffic flowing from agents to controllers and the types of commands sent to agents.

A preliminary analysis of the direction of data flows of packets that goes from Controller to agent ($C2a$) and from agent to Controller ($a2C$) is illustrated in Fig. 7.11, showing the number of packets (in percentage) flow data percentage in each direction for each network. We identify four classes of networks based on the flow of data:

- $a2C$ = $C2a$ – The amount of packets sent from the agent is the exact amount sent from controller to agent. MODBUS network presents a $a2C$ = $C2a$ behavior. Given that it is a request/response protocol, it has equal traffic flow in both directions. i.e., for every set of values received from an agent, the Controller must first request those values.

- $a2C$ > $C2a$ – Traffic coming from the agent is greater than the traffic coming from the Controller. Both IEC 104 networks have similar behaviors; the majority of the traffic is from the agent to the controller.

- $a2C$ >> $C2a$ – The traffic coming from the controller is so small as to be insignificant, i.e., most of the data flow comes from the agent. A controller for C37.118 sends a packet to the PMUs, therefore, the controller receives data from them non-stop ($a2C$ >> $C2a$).

- $C2a$ > $a2C$ – More traffic is sent out from the controller than what is sent by the agent. The ICCP network does not monitor all the agents. Instead, the CR of the ISO acts as a peer to the other CRs of the companies operating the bulk power grid. The SO not only collects data from agents, but it also sends data to agents (other power companies in the system) through ICCP connections ($C2a$ > $a2C$).

While the analysis above shows the diversity of the direction of the flows in SCADA networks, this flow-based (network layer) analysis without content analy-

sis can be misleading if we want to quantify how many commands control centers send to their endpoints. We now categorize all of the traffic as either command or measurement. Command data refers to instructions given by the Controller to request data or to set parameters in the network or the grid. Measurement data is the collection of information about data values (voltage, current, etc.) and the status of the network (clock synchronization, etc.). For example, some protocols like IEC 104 make this distinction easy by labeling each type as a "command type" (e.g., interrogation command) vs. a measurement type (e.g., measured value, short floating point). For the other protocols, we must look at the types and infer which ones are related to command actions (e.g., read vs. write in MODBUS).

When we looked at the relationship between command data and measurement data, we identify that more than 95% percent of the traffic is measurement data for all the protocols. As we can see from Fig. 7.12, only three protocols (IEC104_B, IEC104_D, and ICCP) out of six contained command data. Even for these three protocols, the percentage of command data is very low, the highest (IEC104_D) being only 2.5%. On the other hand, we have C37.118, Goose, and MODBUS, which are almost 100% measurement data.

Looking in more detail at the types of commands, we define two types: **(1) control commands** and **(2) configuration commands**. Control commands make changes to the physical world. Configuration commands keep the devices and network configured correctly. Out of the three protocols that use command data, only one (IEC104_B) uses control commands. The other two protocols (IEC104_D and ICCP) use only configuration commands as seen is Fig. 7.12. In the case of IEC104_B, there are control commands for ramping up or down the generation of power plants. These commands are part of the AGC that the CR uses to maintain the power balance in a system. Even so, these commands are

only 0.8% of all the traffic for this network. While IEC104_D has configuration commands only in our data (two types: interrogation and clock synchronization, Table 7.3).

In conclusion, we infer that these networks run very largely on "auto-pilot" with little interference by control commands. As far as we can see from our data, control commands seem to be the exception rather than the rule and are not used throughout the whole network but rather confined to certain parts. Out of the small percentage of control commands that are sent, most of them are just monitoring commands.

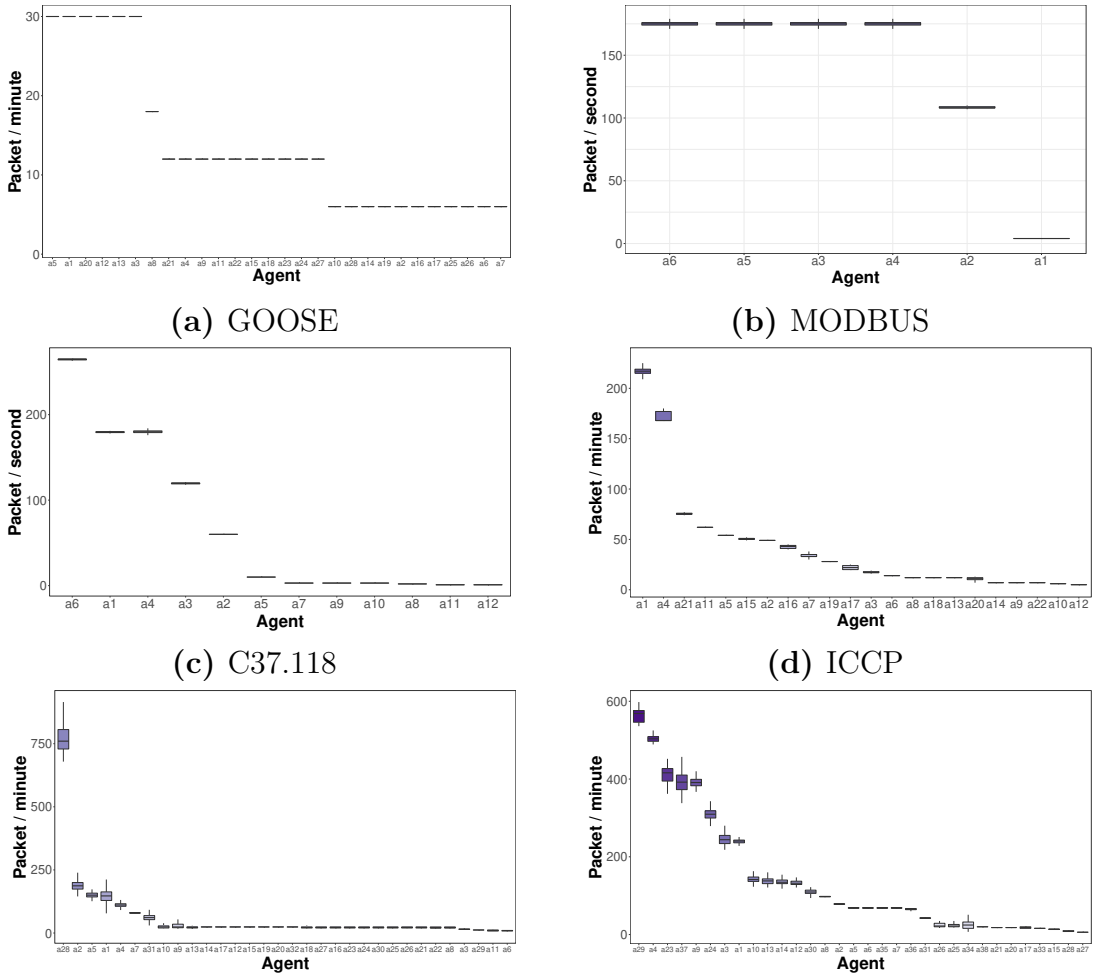
Summary: SCADA networks are predominantly monitoring networks, where CR or other nodes (GOOSE) consume data from agents. The only network sending control commands that change the physical world is IEC104_B (the ISO sends new setpoints for generators in different power plants to maintain the power balance in the grid). Even then, control commands consist only 0.8% of all the traffic in this network.

7.7 Conclusion

One of the contributions of this chapter is the use and analysis of datasets captured in an operational power grid infrastructure. This was possible through years of collaboration with industry and academic partners. Using this unique dataset, we show the variety and use of industrial control protocols in the power grid. We show that there are protocols for synchronous and asynchronous communications, request/response or unsolicited communications, and protocols that use client/server or publisher/subscriber models. Some protocols are designed for power emergencies (e.g., GOOSE) and others for carrying a considerable amount of information (e.g., ICCP). Some networks are used to transmit data through

hundreds of km (WANs) and others are located in a small area (LANs). Some of them are used to send remote control commands (IEC 104) and others for monitoring transient events (C37.118). Since each protocol has a specific application, SCADA traffic is diverse and impacts the network differently.

We found that more than half of the traffic in the distribution network is related only to keeping alive messages, dispelled commonly asserted assertions of SCADA systems, presented a new taxonomy, and discussed and analyzed industrial networks. We hope that our research motivates more inquiries about real-world SCADA networks as these networks migrate completely from serial to TCP/IP or Ethernet networks, and become more integrated into our information infrastructures we need to make sure we understand them and also know how to create security technologies applicable to the diverse ecosystem of SCADA networks.

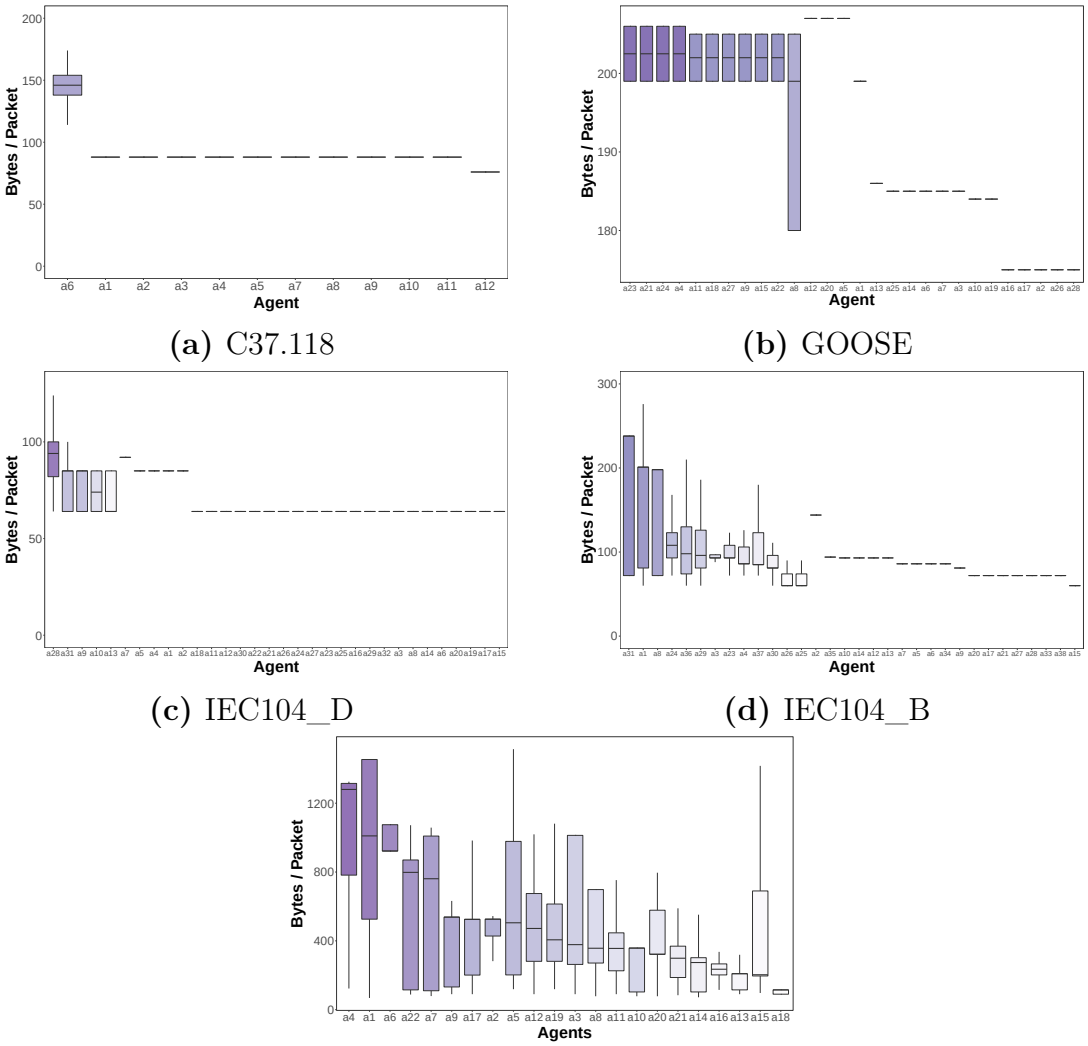


(a) GOOSE **(b)** MODBUS

(c) C37.118 **(d)** ICCP

(e) IEC104_D **(f)** IEC104_B

Figure 7.8: Packets per minute per endpoint. Some endpoints have very periodic transmission patterns (e.g., endpoints in GOOSE) while other endpoints have more variance in their transmission patterns (endpoints in IEC104_B).



(a) C37.118

(b) GOOSE

(c) IEC104_D

(d) IEC104_B

(e) ICCP

Figure 7.9: Packet sizes per endpoint. Several endpoints send the same packets over and over again (IEC104_D), while in other networks endpoints send diverse packets (ICCP).

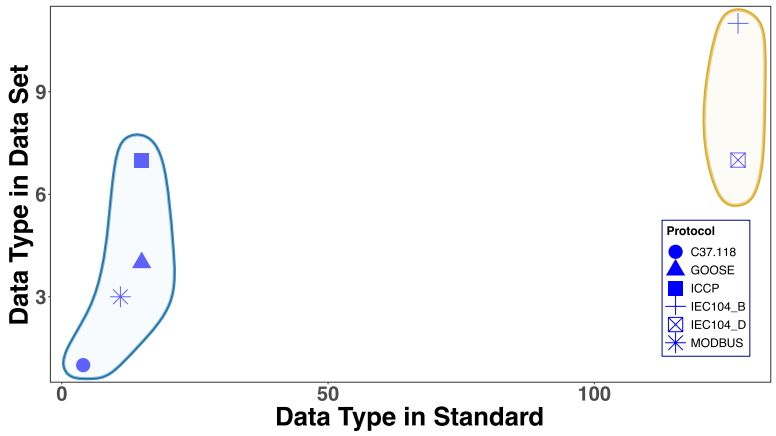


Figure 7.10: Data Types. IEC 104 gives a lot of flexibility for types, and they are used by operators

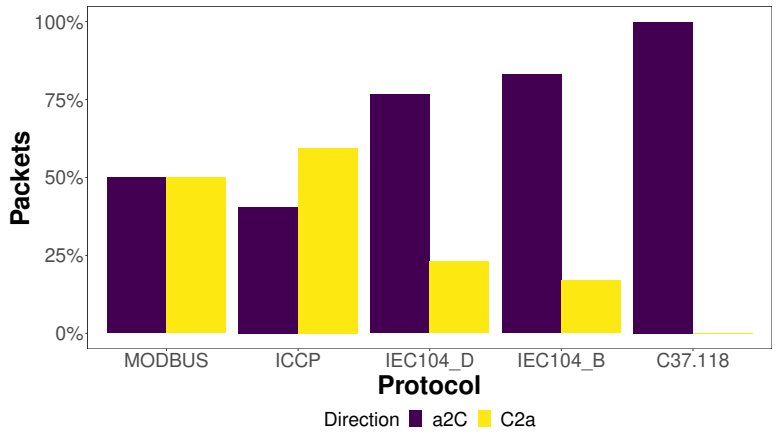


Figure 7.11: The number of packets sent from an agent to a controller (*a2C*) and from a controller to an agent (*C2a*). While MODBUS has equal traffic flow in both directions, C37.118 traffic only flows in the controller direction. Moreover, ICCP presents a larger traffic from C2a than a2C.

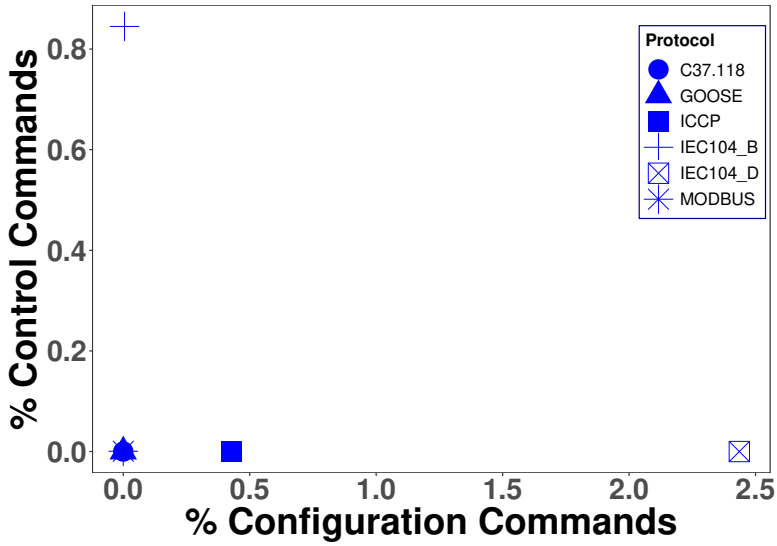


Figure 7.12: Control vs Configuration Commands. While some networks send commands to devices (e.g., IEC104_D sends clock synchronization commands), the only network that sends control commands to change the operation of the grid is IEC104_B.

Chapter 8

Discussion

In the previous chapters, we showed that SCADA networks are not monolithic entities. The individual protocols, their topology, the traffic characteristics, and the individual data types used in all networks. They show a diverse ecosystem for monitoring and controlling different industrial control systems. These results can help us dispel misconceptions about SCADA networks.

8.1 Dispelling Misconceptions

SCADA communications have been analyzed by the community for some 20 years. However, previous research has focused on results from testbeds or only one real-world network. Therefore, the conclusions of previous research are sometimes inaccurate or not representative of various real-world operational systems.

The common wisdom we have seen repeated in the literature is that SCADA networks are similar, and they tend to be painted under the same broad strokes. We bring a few examples below, contrasted with our observations.

Polling: "Due to the polling mechanisms typically used to retrieve data from field devices, industrial control network traffic exhibits strong periodic patterns"

(Barbosa *et al.* 2012) [9]. "most of the SCADA traffic is expected to be generated periodically due to the polling mechanism used to gather data. " (Barbosa *et al.* 2016) [10]. "Due to the use of request-response communication in polling, SCADA traffic exhibits stable and predictable communication patterns.". (Lin *et al.* 2018) [39]. **Reality:** Our data set reveals that Request/Respond is not always the mode of communication employed. Fig. 6.7 indicates that only MODBUS has a flow that reflects a polling flow pattern with an equal percentage of traffic in both directions.

Flow Direction: "the bulk of the traffic is generated from field devices regularly reporting data to the master and the master occasionally sending commands as needed" (Formby *et al.* 2017) [21]. **Reality:** We saw in the last Section 6.5 that for Gas, this relationship can be reversed. In this case, the master sends more data to the other endpoints of the connections (C2a > a2C).

Simplicity of topology, traffic, and protocols: "control systems tend to have static *topology, regular traffic, and simple protocols.*"(Cheung *et al.* 2006) [13]. **Reality:** The *topology* of SCADA networks is dynamic. Mai *et al.*, 2020 [41] shows topological differences in a bulk power grid over two consecutive years. They found that processes such as energy dispatch can affect the topology daily by adding or removing generation nodes regularly according to the demand needs. In addition, the frequency of maintenance in electrical elements, such as generation machines and transformers, removes nodes temporally. Furthermore, expansion projects can add new nodes to the network.

Traffic: Fig 6.3 shows that several protocols do not have regular traffic patterns. Protocols that use spontaneous transmission, such as IEC 104, present high variability in traffic because the data report depends on the status of the physical world. Furthermore, some SCADA networks are composed of heteroge-

neous devices configured by other companies, and therefore, they have different configurations, resulting in diverse traffic patterns.

Protocol: Finally, as we discussed in our analysis, the first SCADA protocols, such as MODBUS, were fairly simple; however, modern protocols have a more complex structure. We also saw how protocols like IEC 104 provide hundreds of data types that amplify the range of data types the user can choose. So, overall, we argue that industrial protocols are becoming more complex than what some researchers expect.

Network topology: "SCADA systems typically use primary-backup approaches to provide disaster recovery capabilities. Specifically, a hot backup of the central control server (the SCADA master) can take over immediately if the primary SCADA master fails, and in many SCADA systems, a cold-backup control center can be activated within a couple of hours if the primary control center fails." (Babay *et al.* 2018) [8]. **Reality:** As we showed in this paper, there are several SCADA topologies, and most do not satisfy this primary backup assumption. In fact, as we can see in our datasets, the backup server in our $K_{2,q}$ networks is not stand-by or inactive; it is a secondary server helping us load-balance the network and taking an active part in the monitoring of the system.

Timing and periodicity of traffic: "SCADA systems for the power grid must deliver device status updates and supervisory commands within 100-200ms." (Babay *et al.* 2018) [8]. **Reality:** As we saw in our analysis, data reporting can change significantly, not only among networks but even among different endpoints in the same network. Most of the status updates in our networks took more than 200ms.

We contend that the prevailing academic perspective on SCADA protocol usage in real systems is analogous to observing just a fraction of a larger puzzle.

Often, researchers draw conclusions based on an isolated SCADA network, overlooking the broader context because they don't have access to other operational networks. Our paper aims to shed light on the multifaceted and evolving nature of SCADA systems within the power grid, striving for a more comprehensive understanding.

8.2 Limitations

While the data captures we received were relatively brief, this suggests we might not have observed all the devices in the network. Some devices could have been offline or not generating traffic during our capture period, thus not registering in our dataset.

Nevertheless, even if our data might not provide a complete perspective of the systems, these captures still offer valuable insights into the diversity of SCADA traffic within an Industrial Control System. From this data, we drew definitive conclusions about variations in SCADA transmission rates, packet sizes, and topologies. Furthermore, our dataset is the most diverse ever reported in an academic context, boasting broad coverage across different companies, devices, and protocols within ICS.

Chapter 9

Conclusions

This study aimed to investigate the variations in SCADA traffic behavior across three critical infrastructures: power, water, and gas. Also, we examined the subsystems in the power grid, from generation to end-customer. The primary research question focused on identifying their differences and similarities in network traffic and understanding traffic patterns based on their protocol characteristics and operational processes.

Our analysis identified several key differences and commonalities in network traffic patterns, contributing significantly to our understanding of SCADA systems in ICS. These include:

- **Predominance of Small Packet Sizes:**A preference for small packet sizes, ranging from 0 to 100 bytes, was evident in all three ICSs, making up more than half of their communications. This observation that the majority of communications involve small packets is significant for data management and security strategies.

- **Uniform Packet Sizes Across Networks:** Notably, consistency was observed in the traffic across all networks: approximately a quarter of their

traffic exhibited uniform packet sizes. The discovery of a consistent pattern in packet sizes across all networks challenges our previous assumptions that different industries would have vastly different communication patterns. This insight could simplify certain aspects of network monitoring and anomaly detection across various ICS.

- **Variation in Transmission Times:** Transmission times spanned from seconds (Power and Water) to minutes (Gas). Milliseconds-order transmission are rare exceptions. Extraordinary, Gas network is the less active network with long transmission periods in the order of minutes. The identified transmission times, ranging from seconds in power and water to minutes in gas, offer valuable insights into the operational tempo of these systems. This information is good for designing network infrastructure and security protocols that can accommodate these differing requirements.

- **IEC 104 Protocol Behavior:** IEC 104 reveals two consistent operational tendencies: a minimal IAT that hovers around one second and a maximum packet size capped at 200 bytes. This provides a more refine understanding of how this protocol operates in real-world settings. This could influence how network traffic is analyzed and secured in systems using IEC 104.

- **Directional Flow Variability:** The flow direction of the packets vary in SCADA networks. They are going to depend on the protocol characteristics such as Modbus, or the configuration of the network, such as Gas network. This is an unforeseen aspect that has not been document before. This finding could be important for designing network monitoring tools and security systems that consider flow patterns.

- **Gas Network's Unique Configuration:** Gas networks reveleas a non-

standard configuration of the IEC 104 protocol, increasing the traffic and normal flow traffic behavior of IEC 104 networks. The unexpected revelation about the non-standard configuration of the IEC 104 protocol in gas networks, which increases traffic and alters normal flow behavior, is particularly intriguing. This could point to customizations in the gas sector that may not be well understood or documented, presenting unique security challenges.

This study was limited by its reliance on available data, which may not capture all nuances of SCADA network traffic. Some devices could have been offline or not generating traffic during our capture period, thus not registering in our dataset. Nevertheless, even if our data might not provide a complete perspective of the systems, these captures still offer valuable insights into the diversity of SCADA traffic within an Industrial Control System.

This study has highlighted differences and found similarities in SCADA network behaviors across various infrastructures, offering a wide perspective of these operational technologies in ICS. The insights gained not only challenge conventional notions but also provide a better understanding of SCADA networks. We hope that our research motivates more inquiries about real-world SCADA networks as these networks migrate completely from serial to TCP/IP or Ethernet networks, and become more integrated into our information infrastructures we need to make sure we understand them and also know how to create security technologies applicable to the diverse ecosystem of SCADA networks.

Bibliography

[1] IEC 60870-5-101:2003, June 2003.

[2] IEC tr 60870-6-505:2002, June 2005.

[3] IEC 60870-5-104:2006, June 2006.

[4] Modbus application protocol specification v1.1b3. apr 2012.

[5] IEEE Standards Association. IEEE Std C37.118.2-2011 IEEE Standard for Synchrophasor Data Transfer for Power Systems. page 53, December 2011.

[6] IEEE Standards Association. Ieee std c37.118.2-2011 ieee standard for synchrophasor data transfer for power systems. page 53, dec 2011.

[7] Amy Babay, John Schultz, Thomas Tantillo, Samuel Beckley, Eamon Jordan, Kevin Ruddell, Kevin Jordan, and Yair Amir. Deploying intrusion-tolerant scada for the power grid. In *2019 49th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*, pages 328–335, 2019.

[8] Amy Babay, Thomas Tantillo, Trevor Aron, Marco Platania, and Yair Amir. Network-attack-resilient intrusion-tolerant scada for the power grid. In *2018 48th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*, pages 255–266, 2018.

[9] Rafael R. R. Barbosa, Ramin Sadre, and Aiko Pras. Difficulties in modeling scada traffic: A comparative analysis. In Nina Taft and Fabio Ricciato, editors, *Passive and Active Measurement*, pages 126–135, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.

[10] Rafael Ramos Regis Barbosa, Ramin Sadre, and Aiko Pras. Exploiting traffic periodicity in industrial control networks. *International Journal of Critical Infrastructure Protection*, 13:52–62, June 2016.

[11] Atul Bohara, Jordi Ros-Giralt, Ghada Elbez, Alfonso Valdes, Klara Nahrstedt, and William H. Sanders. ED4GAP: Efficient Detection for GOOSE-Based Poisoning Attacks on IEC 61850 Substations. In *2020 IEEE International Conference on Communications, Control, and Computing Technologies*

for Smart Grids (SmartGridComm), pages 1–7, Tempe, AZ, USA, November 2020. IEEE.

[12] Atul Bohara, Jordi Ros-Giralt, Ghada Elbez, Alfonso Valdes, Klara Nahrstedt, and William H. Sanders. Ed4gap: Efficient detection for goose-based poisoning attacks on iec 61850 substations. In *2020 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm)*, pages 1–7, 2020.

[13] Steven Cheung, Bruno Dutertre, Martin Fong, Ulf Lindqvist, Alfonso Valdes, and Keith Skinner. Using model-based intrusion detection for scada networks. *Proceeding of the SCADA Security Scientifc Symposium*, page 12, 2007.

[14] Fan Chung. *Spectral Graph Theory*, volume 92 of *CBMS Regional Conference Series in Mathematics*. American Mathematical Society, 1 edition, December 1996.

[15] G. Clarke, D. Reynders, and E. Wright. *Practical Modern SCADA Protocols: DNP3, 60870.5 and Related Systems*. 01 2004.

[16] C.A.S. da Cunha, O. Rein, J.A. Jardini, and L.C. Magrini. Electrical utilities control center data exchange with ICCP and CIM/XML. In *2004 IEEE/PES Transmision and Distribution Conference and Exposition: Latin America (IEEE Cat. No. 04EX956)*, pages 260–265, November 2004.

[17] J. Donghui P, Summers and M. Walstrom. Cyberattack on critical infrastructure: Russia and the ukrainian power grid attacks, 2017.

[18] EPRI. Inter-Control Center Communications Protocol ICCP. USer's Guide Rev. 1. 1999.

[19] Noam Erez and Avishai Wool. Control variable classification, modeling and anomaly detection in modbus/tcp scada systems. *International Journal of Critical Infrastructure Protection*, 10:59–70, 2015.

[20] Noam Erez and Avishai Wool. Control variable classification, modeling and anomaly detection in Modbus/TCP SCADA systems. *International Journal of Critical Infrastructure Protection*, 10:59–70, September 2015.

[21] David Formby, Anwar Walid, and Raheem Beyah. A case study in power substation network dynamics. *Proc. ACM Meas. Anal. Comput. Syst.*, 1(1), jun 2017.

[22] David Formby, Anwar Walid, and Raheem Beyah. A Case Study in Power Substation Network Dynamics. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 1(1):1–24, June 2017.

[23] Niv Goldenberg and Avishai Wool. Accurate modeling of Modbus/TCP for intrusion detection in SCADA systems. *International Journal of Critical Infrastructure Protection*, 6(2):63–75, June 2013.

[24] Andy Greenberg. *Sandworm: A new era of cyberwar and the hunt for the Kremlin's most dangerous hackers*. Anchor, 2019.

[25] Vipul Harsh, Sangeetha Abdu Jyothi, and P. Brighten Godfrey. Spineless data centers. HotNets '20, page 67–73, New York, NY, USA, 2020. Association for Computing Machinery.

[26] 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, New York, NY, USA, 2017. Association for Computing Machinery.

[27] A. J. Hoffman. On the Line Graph of the Complete Bipartite Graph. *The Annals of Mathematical Statistics*, 35(2):883 – 885, 1964.

[28] Juan Hoyos, Mark Dehus, and Timthy X Brown. Exploiting the goose protocol: A practical attack on cyber-infrastructure. In *2012 IEEE Globecom Workshops*, pages 1508–1513, 2012.

[29] Juan Hoyos, Mark Dehus, and Timthy X Brown. Exploiting the GOOSE protocol: A practical attack on cyber-infrastructure. In *2012 IEEE Globecom Workshops*, pages 1508–1513, Anaheim, CA, USA, December 2012. IEEE.

[30] International Electrotechnical Commission, Technical Committee 57, and International Electrotechnical Commission. *Communication networks and systems for power utility automation. Part 8-1,*. 2011. OCLC: 914242803.

[31] Devika Jay, Himanshu Goyel, Umayal Manickam, and Gaurav Khare. Unsupervised learning based intrusion detection for goose messages in digital substation. In *2022 22nd National Power Systems Conference (NPSC)*, pages 242–247, 2022.

[32] Sang Shin Jung, David Formby, Carson Day, and Raheem Beyah. A first look at machine-to-machine power grid network traffic. In *2014 IEEE International Conference on Smart Grid Communications (SmartGridComm)*, pages 884–889, 2014.

[33] Rafiullah Khan, Kieran McLaughlin, David Laverty, and Sakir Sezer. IEEE C37.118-2 Synchrophasor Communication Framework - Overview, Cyber Vulnerabilities Analysis and Performance Evaluation:. In *Proceedings of the*

2nd International Conference on Information Systems Security and Privacy, pages 167–178, Rome, Italy, 2016. SCITEPRESS - Science and and Technology Publications.

[34] Hubert Kirrmann, Mats Hansson, and Peter Muri. Iec 62439 prp: Bumpless recovery for highly available, hard real-time industrial networks. In *2007 IEEE Conference on Emerging Technologies and Factory Automation (EFTA 2007)*, pages 1396–1399. IEEE, 2007.

[35] C.-Y. Lin and Simin Nadjm-Tehrani. A comparative analysis of emulated and real iec-104 spontaneous traffic in power system networks. In Habtamu Abie, Silvio Ranise, Luca Verderame, Enrico Cambiaso, Rita Ugarelli, Gabriele Giunta, Isabel Praça, and Federica Battisti, editors, *Cyber-Physical Security for Critical Infrastructures Protection*, pages 207–223, Cham, 2021. Springer International Publishing.

[36] C.-Y. Lin and Simin Nadjm-Tehrani. A Comparative Analysis of Emulated and Real IEC-104 Spontaneous Traffic in Power System Networks. In Habtamu Abie, Silvio Ranise, Luca Verderame, Enrico Cambiaso, Rita Ugarelli, Gabriele Giunta, Isabel Praça, and Federica Battisti, editors, *Cyber-Physical Security for Critical Infrastructures Protection*, volume 12618, pages 207–223. Springer International Publishing, Cham, 2021. Series Title: Lecture Notes in Computer Science.

[37] 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*, pages 51–60, Incheon Republic of Korea, May 2018. ACM.

[38] Chih-Yuan Lin and Simin Nadjm-Tehrani. Protocol study and anomaly detection for server-driven traffic in scada networks. *International Journal of Critical Infrastructure Protection*, page 100612, 2023.

[39] Chih-Yuan Lin, Simin Nadjm-Tehrani, and Mikael Asplund. Timing-based anomaly detection in scada networks. In *Critical Information Infrastructures Security*, pages 48–59, Cham, 2018. Springer International Publishing.

[40] Juan C. Lozano, Keerthi Koneru, Neil Ortiz, and Alvaro A. Cardenas. Digital substations and iec 61850: A primer. *IEEE Communications Magazine*, 61(6):28–34, 2023.

[41] Kelvin Mai, Xi Qin, Neil Ortiz, Jason Molina, and Alvaro A. Cardenas. Uncharted networks: A first measurement study of the bulk power system. In *Proceedings of the ACM Internet Measurement Conference*, pages 201–213, Virtual Event USA, oct 2020. ACM.

[42] 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, 2019.

[43] Peter Maynard, Kieran McLaughlin, and Berthold Haberler. Towards Understanding Man-In-The-Middle Attacks on IEC 60870-5-104 SCADA Networks. In *2nd International Symposium for ICS & SCADA Cyber Security Research 2014*. BCS Learning & Development, September 2014.

[44] Stefan Mehner, Franka Schuster, and Oliver Hohlfeld. Lights on power plant control networks. In Oliver Hohlfeld, Giovane Moura, and Cristel Pelsser, editors, *Passive and Active Measurement*, pages 470–484, Cham, 2022. Springer International Publishing.

[45] S. Mohagheghi, J. Stoupis, and Z. Wang. Communication protocols and networks for power systems-current status and future trends. In *2009 IEEE/PES Power Systems Conference and Exposition*, pages 1–9, March 2009.

[46] A Muir and J Lopatto. Final Report on the August 14, 2003 Blackout in the United States and Canada: Causes and Recommendations, April 2004.

[47] 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. In *2022 IEEE Security and Privacy Workshops (SPW)*, pages 243–250, San Francisco, CA, USA, May 2022. IEEE.

[48] Pedro Juan Roig, Salvador Alcaraz, Katja Gilly, and Carlos Juiz. Modelling a leaf and spine topology for vm migration in fog computing. In *2020 24th International Conference Electronics*, pages 1–6, 2020.

[49] C. A. Ruiz, N. J. Orrego, and J. F. Gutierrez. The Colombian 2007 black out. In *2008 IEEE/PES Transmission and Distribution Conference and Exposition: Latin America*, pages 1–5, Bogota, Colombia, August 2008. IEEE.

[50] Leonidas Stylianou, Lenos Hadjidemetriou, Markos Asprou, Lazaros Zacharia, and Maria K. Michael. A behavioral model to detect data manipulation attacks of synchrophasor measurements. In *2021 IEEE PES Innovative Smart Grid Technologies Europe (ISGT Europe)*, pages 1–6, Espoo, Finland, October 2021. IEEE.

[51] Hongsuda Tangmunarunkit, Ramesh Govindan, and Sugih Jamin. Network topology generators: Degree-based vs. structural. *ACM SIGCOOM*, pages 147–159, 2002.

[52] Muhammad Usama Usman and M. Omar Faruque. Applications of synchrophasor technologies in power systems. *Journal of Modern Power Systems and Clean Energy*, 7(2):211–226, March 2019.

[53] Christian Wressnegger, Ansgar Kellner, and Konrad 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, 2018.

[54] 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, 2013.