Integrity and Privacy Protection for Cyber-physical Systems (CPS)

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in Computer Science and Engineering

by

Arpan Bhattacharjee

Dr. Shahriar Badsha – Thesis Advisor
August, 2021
THE GRADUATE SCHOOL

We recommend that the thesis prepared under our supervision by

Arpan Bhattacharjee

entitled

Integrity and Privacy Protection for Cyber-physical Systems (CPS)

be accepted in partial fulfillment of the requirements for the degree of

Master of Science

Shahriar Badsha, PhD
Advisor

Shamik Sengupta, PhD
Committee Member

Mohammed Ben-Idris, PhD
Graduate School Representative

David W. Zeh, Ph.D., Dean
Graduate School

August, 2021
Abstract

The present-day interoperable and interconnected cyber-physical systems (CPS) provides significant value in our daily lives with the incorporation of advanced technologies. Still, it also increases the exposure to many security privacy risks like (1) maliciously manipulating the CPS data and sensors to compromise the integrity of the system (2) launching internal/external cyber-physical attacks on the central controller dependent CPS systems to cause a single point of failure issues (3) running malicious data and query analytics on the CPS data to identify internal insights and use it for achieving financial incentive. Moreover, (CPS) data privacy protection during sharing, aggregating, and publishing has also become challenging nowadays because most of the existing CPS security and privacy solutions have drawbacks, like (a) lack of a proper vulnerability characterization model to accurately identify where privacy is needed, (b) ignoring data providers privacy preference, (c) using uniform privacy protection which may create inadequate privacy for some provider while overprotecting others.

Therefore, to address these issues, the primary purpose of this thesis is to orchestrate the development of a decentralized, p2p connected data privacy preservation model to improve the CPS system’s integrity against malicious attacks. In that regard, we adopt blockchain to facilitate a decentralized and highly secured system model for CPS with self-defensive capabilities. This proposed model will mitigate data manipulation attacks from malicious entities by introducing bloom filter-based fast CPS device identity validation and Merkle tree-based fast data verification. Finally, the blockchain consensus will help to keep consistency and eliminate malicious entities from the protection framework.
Furthermore, to address the data privacy issues in CPS, we propose a personalized data privacy model by introducing a standard vulnerability profiling library (SVPL) to characterize and quantify the CPS vulnerabilities and identify the necessary privacy requirements. Based on this model, we present our personalized privacy framework (PDP) in which Laplace noise is added based on the individual node’s selected privacy preferences. Finally, combining these two proposed methods, we demonstrate that the blockchain-based system model is scalable and fast enough for CPS data’s integrity verification. Also, the proposed PDP model can attain better data privacy by eliminating the trade-off between privacy, utility, and risk of losing information.
Letter of Declaration

I, Arpan Bhattacharjee, student of Department of Computer Science & Engineering, University of Nevada, Reno, hereby declare that the work presented in this thesis paper titled “Integrity and Privacy Protection for Cyber-physical Systems” has been carried out by me and has not been previously submitted to any other organization or for any academic purpose except where due acknowledgment has been made. The work I have presented does not breach any existing copyright; therein is based on materials collected and developed by myself.

Arpan Bhattacharjee
Student ID No: 8001295308
Masters of Computer Science & Engineering
Department of Computer Science & Engineering,
University of Nevada, Reno
Acknowledgments

All the praises and thanks go to God, who allowed me to reach this endeavor. I wish to express my deep gratitude to my honorable thesis supervisor Dr. Shahriar Badsha, my co-advisor, Dr. Shamik Sengupta, and thesis committee member Dr. Mohammed Ben-Idris, for their persistent guidance, keen interest, and continuous encouragement. I am also grateful to all Department of Computer Science & Engineering, UNR faculty members for their kind guidance and direction. I am using this opportunity to express my gratitude to everyone who supported me throughout this M.S thesis paper. I am thankful for their aspiring guidance, invaluably constructive criticism, and friendly advice during the thesis work. I am sincerely grateful to them for sharing their truthful and illuminating views on several issues related to the project.

I would also like to thank my family, my father Debadideb Bhattacharjee, my mother Arpita Bhattacharjee, my wife Adity Biswas, and my brothers, who always provided my mental strength and courage throughout this masters journey. I am grateful for their prayers and the sacrifice they have made for me. Last but not least, I would like to express my gratitude to all my friends in and outside of UNR who directly or indirectly supported me throughout this time.
Credits

Portions of the materials used in this thesis have previously appeared or are under consideration in the following scientific publications:

Conferences


Book Chapter

- S. Islam, A. Bhattacharjee, S. Badsha, “Towards Secure Cyber Infrastructure for SmartCities: Learning-based Intelligent Solutions,” Accepted as Springer Book Chapter
# Table of Contents

1 Introduction 1
   1.1 Motivation ................................................. 3
   1.2 Research Contribution ...................................... 4
   1.3 Thesis Organization ...................................... 5

2 Background 7
   2.1 Preliminaries ................................................ 7
      2.1.1 Cyber-physical Systems (CPSs) ....................... 7
      2.1.2 Synchrophasor Network ................................ 7
      2.1.3 Advanced Manufacturing System ....................... 8
      2.1.4 Blockchain Overview .................................. 9
      2.1.5 Differential Privacy (DP) ............................. 9
      2.1.6 Data Perturbation Mechanisms ......................... 11
   2.2 Related Work ............................................... 12
      2.2.1 Critical Security Challenges in Synchrophasor .......... 12
      2.2.2 Blockchain for Synchrophasor and Smart Grid .......... 12
      2.2.3 Analysis of Existing Threats and Risks in Advanced Manufacturing System and their Countermeasures .......... 13
      2.2.4 Critical Privacy Challenges for CPS Data .............. 14
      2.2.5 Importance of Data Privacy-Preservation .............. 15
2.2.6 Privacy Preservation in CPS ............................................ 15
2.2.7 Advancement of this Thesis over Existing Works .......... 16

3 Achieving Integrity for Cyber-physical Systems (CPS) ........ 18

3.1 Case Study 1: Developing Block-Phasor: A Decentralized Blockchain Framework to Enhance Security of Synchrophasor ........ 18
3.1.1 Motivation and Contribution ...................................... 18

3.2 Proposed Blockchain based Synchrophasor Wide Area Monitoring System ............................................................. 20
3.2.1 Bloom Filter based PMU Device ID Validation .......... 21
3.2.2 ECDSA based PMUs Measurement Authentication ...... 22
3.2.3 Merkle Tree based Verification ................................. 22
3.2.4 Mining and Generation of Distributed Ledger .......... 23

3.3 Experimental Analysis .................................................... 24
3.4 Summary ................................................................. 26

3.5 Case Study 2: Blockchain-based Secure and Reliable Manufacturing System ............................................................ 27
3.5.1 Motivation ............................................................ 27
3.5.2 Threat Model .......................................................... 27
3.5.3 Contribution .......................................................... 28

3.6 Proposed Blockchain Based Solution ............................... 28
3.6.1 Clustering Mechanism ............................................... 30
3.6.2 Authentication Mechanism ....................................... 30
3.6.3 Raft-based Distributed Consensus and Ledger Generation . 32

3.7 Experimental Analysis .................................................... 33
3.8 Summary ................................................................. 34
4 Vulnerability Characterization and Privacy Protection Model for Cyber-physical Systems

4.1 Motivation and Contribution ........................................ 36
4.2 Proposed Vulnerability Characterization Model .................... 37
  4.2.1 Smart Grid Topological Relationship .......................... 37
  4.2.2 Vulnerability Characterization Framework ...................... 38
  4.2.3 Vulnerability Assessment by Building an Attack Profile ....... 39
  4.2.4 Standard Vulnerability Profile Generation ...................... 40
4.3 Proposed Personalized Differential Privacy Model for CPS ....... 41
  4.3.1 Personalized Differential Privacy Modelling .................... 44
  4.3.2 Calculating Distance between Nodes of the Graph ............ 45
  4.3.3 Laplace Distribution and \( \varepsilon \)-Personalized Privacy .... 45
  4.3.4 Mechanism of Achieving PDP .................................. 45
  4.3.5 Criteria Mapping to Personalized Privacy Preference .......... 46
  4.3.6 Proof: Personalized Differential Privacy (PDP) ............... 46
4.4 Experimental Analysis .............................................. 47
  4.4.1 Utility Analysis ............................................. 47
  4.4.2 Privacy Analysis .......................................... 47
  4.4.3 Risk Factor Analysis ......................................... 48
4.5 Summary .......................................................... 54

5 Conclusion .......................................................... 55

5.1 Open Issues, Challenges and Future Research Directions .......... 56
  5.1.1 Suitability Issues ........................................... 56
  5.1.2 Scalability, Latency and Throughput Issues .................... 56
  5.1.3 Consideration of Multiple Layers in Blockchain Design ....... 56
  5.1.4 Big Data .................................................. 57
5.1.5 Machine Learning ............................................. 57
5.1.6 Applicability of PDP in Adversarial Settings ............ 58
List of Tables

4.1 Criteria Mapping to Personalized Privacy Preservation . . . . . . . . 46
# List of Figures

2.1 Manufacturing Control Information Exchange Mechanism. \hspace{1cm} 8  
2.2 Differential Privacy Distinguishes between Private and General Information. \hspace{1cm} 10  
3.1 Blockchain based Proposed system Model for Synchrophasor WAMS. \hspace{1cm} 20  
3.2 Bloom Filter based PMUs Authentication. \hspace{1cm} 21  
3.3 Merkle tree generation time for different number of PMUs on a single PDC. \hspace{1cm} 24  
3.4 Relation between number of PMUs and number of required hashes for verification in a Merkle tree. \hspace{1cm} 25  
3.5 Relation between number of PMUs and hash verification time. \hspace{1cm} 26  
3.6 System Model of Blockchain-based Manufacturing overlay network. \hspace{1cm} 29  
3.7 Transaction Processing. \hspace{1cm} 30  
3.8 State Transition Model of RAFT. \hspace{1cm} 32  
3.9 Throughput by Transaction Rate. \hspace{1cm} 34  
4.1 Vulnerability Characterization Framework. \hspace{1cm} 38  
4.2 Trust Distance based Personalized Privacy Model for CPS. \hspace{1cm} 41
4.3 Evaluation of varying privacy loss(ε) in grid data. (a) Hourly Consumption (KWh) of Meter 1 for different Privacy loss(ε) (b) Comparison between original vs. sanitized data with fixed Privacy loss(ε) (c) Comparison between original vs. differentially private aggregated query result. ............................................................... 49

4.4 Relation between Trust distance ($T^D$) and Privacy Loss (ε) ......... 50

4.5 Evaluation on personalized privacy protection where (a) to (f) shows as the trust distance decreases from highest to lowest the privacy preference sequentially decreases thereby system generates less noisy output and ensure better data utility. ............................... 51

4.6 Utility Analysis. ................................................................. 52

4.7 Distribution of losses over uniform and personalized privacy ......... 53

4.8 Comparative analysis between privacy loss and disclosure risk of private information for uniform and personalized privacy .............. 54
Chapter 1

Introduction

Cyber-physical systems like smart grids and advanced manufacturing with the incorporation of modern information and communication technologies like industrial internet of things (IIoT), machine intelligence, and advanced data analytics are making CPS an intelligent, interconnected, and automated networking of business process [1]. However, these technological revolutions aren’t ready for existing CPS platforms that suffer from security and reliability issues mainly because of their dependency on the trust of the central controller. The current CPS systems are becoming highly susceptible to single point of failures issues because of their interdependencies. Attackers regularly launch cyber and physical attacks such as hardware tampering, side-channel attack, reverse engineering, IoT sensor identity spoofing, malicious data injection, and tampering data to compromise the integrity of the CPS systems [2] [3] [4].

However, the security and privacy issues in CPS are not only caused by external attacks but also about 60 percent of the attacks are done by internal employees and contractors to steal intellectual properties of the CPS system [5]. One example of a cyberattack against the industrial control system is the Stuxnet attack against the Iranian Nuclear facility where attackers conducted four-zero day exploits to compromise the centrifuges and eventually compromise the control system [6] [7]. So, the inter-operable and interconnected cyber-physical systems provide significant value in our day-to-day lives and also increases the exposure to many security risks, with critical and financial impacts.

To counter these issues, we argue that there must be a novel and decentralized way of mitigating the cyberattacks at the edge of the CPS system before it hits the control center and causes a single-point-of-failure. The cyberattack on the Ukrainian power grid that caused a significant blackout for three hours underlines the necessity
of the decentralization of the CPS system [8]. Besides the single point of failure issue, different sensing and monitoring devices of the CPS are susceptible to data manipulation attacks; as a result, they question the system’s integrity. Interestingly, blockchain, the core technology developed to revolutionize the traditional payment system, has the ability and necessary mechanisms to counter most of the significant vulnerabilities of the CPS network. Blockchain can offer decentralization, security, immutability, and integrity by design. That can mitigate the notable and potential vulnerabilities of the CPS system, such as data manipulating cyber-attacks and other insider and outsider attacks.

Moreover, CPS systems generate, transmit, and collect a tremendous amount of privacy-sensitive data. This data ensures a wide range of advantages to all involved systems parties but also raises a privacy risk to data providers whose shared data can be statistically correlated to identify the private information of the system and its users [9]. Topological investigations by researchers have shown that energy-CPS smart grids and advanced manufacturing systems have inter-dependencies between their different nodes [10]. Some of these nodes have relatively higher importance as they generate essential data for the CPS operation. So, identifying the vulnerabilities of these critical nodes and ensuring a robust privacy-preservation mechanism can improve the CPS systems resiliency against privacy-compromising attacks [11].

However, currently, uniform data privacy is provided to all the data providers of the CPS systems, leading to inadequate privacy preservation for some data providers while over-protecting others [12]. One of the major disadvantages of this uniform privacy protection is that it introduces an unacceptable amount of noise because the same level of privacy loss (noise) is added for all data providers, ultimately degrading the data utility and system performance.

To address these privacy problems in CPS, this thesis focuses on designing a comprehensive privacy characterization and quantification model to identify the privacy issues during data sharing, publishing, or aggregating between the associated parties. This privacy quantification model will be subsequently integrated with a Personalized Differential Privacy (PDP) solution in which data providers specify their personal data’s privacy requirements. This integrated privacy characterization and protection model ensures data privacy and utility by eliminating additional noise from the data. As a result, adversaries who run malicious analytics on these privacy-preserved data will not identify sensitive private information.
1.1 Motivation

One of the cyber-physical systems, for example, smart grid’s significant aspects is to aggregate the power consumption and generation data from multiple levels (e.g., regional, central) and use it for grid’s network planning, balancing loads, regulating power generation, monitoring consumption and predicting prices based on the demand [13]. However, privacy has become a challenging issue for the smart grid. Mainly, the data collected from the grid’s end devices contain a significant amount of sensitive private information [14]. Adversaries use malicious techniques to access this data and launch serious security and privacy-compromising attacks like stealing meter data and eavesdropping on the communication channel to hack the central controlling system of the grid [14]. Malicious analysts can further use this fine-grained power/consumption data to infer the renewable energy sources’ operational mechanism and analyze the appliance usage to determine consumers’ daily lifestyle [15].

Similarly, the incorporation of industrial IoT and advanced data analytics in the manufacturing paradigm provide scalability, efficiency, interoperability, and automation to improve enterprise productivity. Still, it raises some challenges to secure its infrastructure, components, and data. To ensure IoT’s security and privacy, it is essential to assure the integrity of the underlying components, particularly protection of sensor, code, and data against malicious manipulations. Moreover, the technological revolutions aren’t ready for existing manufacturing platforms that suffer from security and reliability issues mainly because of their dependency on the trust of the central controller. Therefore the primary purpose of this study is to orchestrate the development of a distributed decentralized peer-to-peer connected system architecture that will improve the reliability of the CPS process by providing a secure and immutable control information exchange platform.

To eliminate the limitations of modern cyber-physical systems by enhancing the self-defensive capability of CPS against cyberattacks, we aim to ensure reliability, security, integrity, decentralization, data privacy-preservation, and analytics at low latency. For that, we have combined two computing paradigms in our work. We are introducing Blockchain technology in our CPS system model, secured by a new differential privacy scheme called "Personalized Differential Privacy (PDP)" for CPS data privacy preservation. This PDP scheme is integrated into the system to provide user-specific privacy requirements and ensure resiliency against internal and external privacy-compromising data disclosure, correlation, or collusion attacks. The high-
level idea is to achieve trustlessness, integrity, and decentralization (Via Blockchain) and privacy-preservation (Via personalized differential privacy), which adds user-specific explicit noise to the aggregate data to protect data’s confidentiality and privacy. Finally, the goal of using these two techniques as mentioned above (i.e., Blockchain and personalized differential privacy) is to achieve privacy bound tighter than the state-of-the-art encryption-based techniques.

The primary motivations of this work have been summarized as follows:

- Traditional cyber-physical systems (e.g., Smart Grid/Manufacturing Paradigm) are centralized controller dependent. This makes them untrusted and unreliable and also prone to a single point of failure issues.

- The introduction of advanced technologies like IoT sensors and data analytics in the CPS paradigm without providing necessary system security mechanisms and data privacy leads to new security, reliability, and privacy challenges. For example, malicious manipulation of the CPS data and sensors to interrupt the system’s optimal operation or running advanced analytics to identify the internal confidential insights of the system.

- Lack of a proper CPS vulnerability characterization model to accurately identify where security and privacy are needed.

- Existing privacy-preservation models mainly use uniform privacy protection, resulting in inadequate privacy for some nodes while overprotecting others. Based on the type of CPS data, the need to maximize data privacy/security/utility may vary. For instance, information about distributed energy resource (DER) owners, grid configuration, product demand and supply, etc., requires high data security and privacy, whereas data utility is not a high priority. In contrast, voltage and frequency measurements require high data utility.

### 1.2 Research Contribution

To mitigate the security issues in the CPS network while collecting and forwarding measurement data through different nodes, we propose a blockchain-based and decentralized security framework for CPS to mitigate data tampering attacks in the system at the edge. Furthermore, we introduce a personalized data privacy model by
characterizing and quantifying the vulnerabilities and privacy requirements of CPS. The main contributions can be summarized as follows:

- We develop a blockchain-based decentralized CPS framework to share synchrophasor data securely and eliminate the system’s single point of failure issues.
- We incorporate bloom filter for fast CPS device identity validation and authentication and Merkle tree for fast data verification to achieve integrity by mitigating data tampering attacks.
- To demonstrate blockchain’s applicability in resource-constrained CPS systems, we incorporate a clustering algorithm and distributed trust management strategy to form a blockchain overlay network that helps to minimize CPS overheads and provide end-to-end security.
- We furthermore introduce a Standard Vulnerability Profiling Library (SVPL) using an attack graph and networks centrality metrics to combine the risk sources, data privacy weaknesses, feared events, and associated harms to arrange nodes in a rank from minimum to maximum vulnerable based on their privacy loss scenario.
- Finally, we develop a Personalized Differential Privacy (PDP) model that guarantees data privacy based on individual data providers’ selected privacy preferences. Our experimental analysis demonstrates that this PDP model ensures better data utility and privacy in CPS over traditional Uniform Differential Privacy (UDP).

1.3 Thesis Organization

The rest of the thesis is organized as follows:

- **Chapter 2: Background and Related Work:** The background information on cyber-physical smart grid and advanced manufacturing systems critical security and privacy-related challenges as well as state of the art privacy preservation mechanisms related information are presented in this chapter. We also describe the related work separately in terms of research questions and research contributions in this chapter.
• **Chapter 3: Achieving Integrity for Cyber-physical Systems:** The main purpose of this chapter is to orchestrate the development of a distributed decentralized peer-to-peer connected system architecture for cyber-physical systems that will ensure the integrity of the system by providing a secure and immutable data and control information exchange platform. The contribution of this chapter appear in [16] [17].

• **Chapter 4: Vulnerability Characterization and Privacy Protection Model for Cyber-physical Systems:** The objective of this chapter is to focus on building a new methodology for vulnerability analysis by modeling the system’s topology and understanding the relationship between cyber events and creating an attribute-based attack profile according to individual nodes’ vulnerabilities to achieve personalized privacy preservation. Furthermore, this chapter demonstrates the building of a reliable and robust data privacy preservation model to address the limitations of uniform data privacy preservation. The high-level idea is to achieve a user-specific privacy guarantee to increase data utility by eliminating excess privacy loss. The contribution of this chapter appear in [18,19].

• **Chapter 6: Conclusion and Future Work:** This chapter concludes our thesis by summarizing the contributions as well as key findings of our work. Furthermore, the significance of our research and potential future directions, and some limitations of our proposed methods are also discussed here.
Chapter 2

Background

2.1 Preliminaries

2.1.1 Cyber-physical Systems (CPSs)

Cyber-physical systems (CPSs) are ubiquitous and serve as the backbone of numerous critical infrastructure applications. The applications range from smart grids, advanced manufacturing, intelligent transportation to internet of things (IoT) installations [20]. These applications provide multiple functionalities in our lives and society, which imparts a positive influence and convenience. However, it might be hindered by novel threats that undermine the safety of CPSs. Some of such threats are malware injection, network intrusion, data manipulation, etc. Cyber-physical attacks such as Stuxnet, Black Energy, Industoyer, Triton, etc., against CPS systems have increased in the last couple of years. [9]. Although CPSs have been highly equipped with time-synchronized high-resolution measurement data for advanced monitoring and diagnostic of the cyber-physical infrastructures, they have also consistently become an attractive target for advanced attacks [3]. Therefore, we require novel methods to safeguard CPSs by ensuring the system’s integrity and privacy of the data.

2.1.2 Synchrophasor Network

The synchrophasor technology enables real-time system monitoring, control, and operation using the phasor measurement units (PMUs). The real-time phasor measurement is facilitated by using the GPS satellite time signal as the time reference for calculating the phase angles of the voltage, current, and frequency at different nodes.
Phasor data concentrators (PDCs) collect all the PMU data and send them to the central PDC or super PDC. The collected real-time PMU data at the control center can then be used for monitoring applications, such as state estimation during steady-state conditions or post-disturbance analysis and control during grid contingencies. Transmission of PMU data to PDCs and super PDCs is done by TCP/IP or UDP/IP network protocol [21].

### 2.1.3 Advanced Manufacturing System

The intelligent manufacturing system is facilitated with the increasing integration of Internet of Things (IoT) technologies in the cyber-physical systems of the factories to provide better interoperability between the layers of the manufacturing line. The product manufacturing line is a hierarchical layout, such as in Fig. 2.1 where there are at least five layers, namely: Resource layer, Perception layer, Manufacturing Service Provider layer, Storage layer, and Terminals [22]. The basic communication primitive between the layers of the manufacturing system is to exchange control information where the resource layer mainly receives commands from the service providers layer and completes manufacturing tasks based on that command. The perception layer is responsible for sensing the physical resources with the help of sensors and actuators, gathering data, and exchanging control commands between the physical resources.
and the service providers. The service provider layer monitors the control systems to manage the production line and communicate with end-users and different service providers via the cloud. This layer also stores data in the cloud and sends service responses to the terminals. The goal of the cloud includes knowledge management, provide cloud services, and efficient data storage. At last, Terminal provides remote end-interaction components, application interfaces, and intelligent devices to transmit specific commands and operate the whole manufacturing system.

2.1.4 Blockchain Overview

Blockchain is a shared ledger of transactions that records and stores all transactions between participators of the network in chronological order. The transactions are verified and authenticated by the participating nodes, which provide better audibility as well as eliminates the dependency on central authority [23]. These transactions are stored in the ledger in the form of a block of data containing the transactions’ hash values. In the blockchain, blocks can be further classified into Block Header and Transactions List. The block header contains the cryptographic hash of previous blocks, timestamps, and transaction list stores the hash of transactions as follows: $\text{Hash}(\text{Tran}_ID \parallel P - \text{Tran}_ID \parallel PU_k \parallel \text{Sign} \parallel \text{Data})$.

Mining and consensus are two of the most important aspects of blockchain technology. Before adding blocks to the ledger, the transactions must go through a mining process. The nodes are the miners in the network, and they perform verification to check the validity of the transactions. Once the miners finish their verification, they ask the participating nodes of the network to provide their consensus to add the new block in the ledger. Different consensus mechanisms exist in blockchain like Proof of Work, Proof of Stake, Proof of Authority, Proof of Elapsed Time, Proof of Burn, etc., and these mechanisms help the network to agree on changes to the ledger [24]. The integration of blockchain to achieve security and integrity has been found effective in various types of applications such as smart grid [16], smart city [25, 26], ride-sharing application [27], location privacy preservation [28], cyber insurance [29], dynamic task allocation [30] as well as cybersecurity information exchange [31].

2.1.5 Differential Privacy (DP)

Dwork [32] first introduced differential privacy that describes as “DP makes sure that the query output is indistinguishable regardless of the absence or presence of specific
Figure 2.2: Differential Privacy Distinguishes between Private and General Information.

data points in the dataset $D$. This means that the malicious analysts will not be able to identify the specific private information of a particular dataset with absolute confidence. The formal definition of differential privacy (DP) can be expressed based on adjacent databases $D$ and $D'$ where both $D$ and $D'$ differ from each other with at least one data point as shown in Fig. 2.2. We will further discuss the definition mathematically in this section:

**Definition 1 (Adjacent Datasets):** An arbitrary function $F$ fulfills $\varepsilon$-differential privacy condition $P_F$ if for any two neighbouring datasets $D$ and $D'$, and for any possible end result $\xi \in \text{Range}(F)$, we get:

$$P_F[F(D) \in \xi] \leq \exp(\varepsilon) \times P_F[F(D') \in \xi]$$ (2.1)

where, $\text{Range}(F)$ = output of resultant function $F$ and $\varepsilon$ = Privacy parameter that establishes the required level of privacy in the system and $\varepsilon$ is preferred to have lower value which provides stronger privacy.

**Definition 2 (Global Sensitivity):** The sensitivity value mainly decides the appropriate amount of perturbation necessary for the differential private data streams. Moreover, global sensitivity refers to the maximum amount of noise perturbation required to differentiate between the adjacent datasets by a single data point. For an arbitrary query $q : D \rightarrow F$, the global sensitivity $\Delta q_{gs}$ can be discovered by using this equation:

$$\Delta q_{gs} = \max_{D, D'} \| q(D) - q(D') \|$$ (2.2)
2.1.6 Data Perturbation Mechanisms

In differential privacy, three noise addition mechanisms such as Laplace mechanism [33], Gaussian mechanism [34], and Exponential mechanism [35] are used as the preferred mechanisms to add noise to the data and protect the dataset from privacy-compromising disclosure, linking, correlation and differencing attacks. We will further investigate these noise addition mechanisms here:

**Laplace Mechanism**

In the Laplace mechanism for noise addition, the Laplacian function adds the noise and perturbs the dataset where the amount of noise added is computed from the LM distribution. The differential privacy sensitivity function regulates the amount of noise added to the dataset. For example, in a given dataset $D$, the function $F$, and global sensitivity value $\Delta q_{gs}$, the arbitrary equation satisfies -differential privacy parameter, when, noise $\sim Lap(\Delta q_{gs}/\epsilon)$ and the equation is:

$$E = F(D) + Lap(\Delta q_{gs}/\epsilon)$$

(2.3)

**Exponential Mechanism**

Exponential mechanism is another noise addition method which is mainly introduced for conditions where the best response is needed to be picked. This is also used for non-numerical outputs. In a given dataset $D$, $a \in A$ presents the possible query answer, in a score function $f : D \times A \rightarrow A$; and a randomized algorithm chooses a probability dependent answer where the randomized algorithm will prove the $\epsilon$-differential privacy.

$$(D, u) = a : P_{a}[a \in A] \propto \exp(\epsilon u(D, a)/2\Delta u)$$

(2.4)

where $\Delta u$ is the sensitivity of the exponential function.

**Gaussian Mechanism**

Analogous to the Laplace mechanism, Gaussian is another crucial building block of differential privacy-based systems where noise is added from a gaussian distribution.
In a query function $f$, the Gaussian perturbation $\sigma$ will be:

$$\sigma = \frac{\Delta_2 f}{c} \sqrt{2 \log(1.25/\varepsilon)}$$ (2.5)

where value of $\varepsilon$ falls between 0 and 1.

## 2.2 Related Work

### 2.2.1 Critical Security Challenges in Synchrophasor

Synchrophasor provides reliable real-time system monitoring over a large geographical area, but malicious attackers attack the network connections and exploit the system because of the vulnerable communication technologies. Cyber-attacks such as Denial of service, Man-In the middle, Packet analysis, Malicious code injection, and Data spoofing happen on the network layer. This causes four general classes of attacks in the PMU data: data Interruption, fabrication, modification, and interception. In DDoS attacks, the attacker can gain control of the whole streaming synchrophasor data by using the Internet Control Message Protocol (ICMP) attack [36]. DoS / DDoS attacks on the transport layer and local area network of synchrophasor network are done by analyzing the vulnerabilities of the system network protocol [37]. If proper data encryption is not used in transmission, attackers can analyze the non-encrypted data and harvest valuable information such as passwords by using Wireshark, which is a packet analyzing tool [38]. Researchers of [39] describe two types of data modification attacks. Malicious code injection and false data injection attacks cause data inconsistency and inaccuracy, leading to system monitoring failure.

Therefore, to mitigate the issues in the synchrophasor network, our goal of this work is to develop a decentralized protection framework with proper cryptographic mechanisms (e.g., Bloom Filter, ECDSA, and Merkle Tree) to tackle the vulnerabilities in PMU data transmission and provide integrity and resiliency against synchrophasor data manipulation.

### 2.2.2 Blockchain for Synchrophasor and Smart Grid

Blockchain technology holds the answer to improve the security issues and standards of the synchrophasor monitoring system. A cyber-secured monitoring device must have the following properties: it has to be authentic, available, and confidential in
its connectivity with the other devices and sensors. Blockchain technology with its distributed ledger, smart contracts, system interoperability, peer-to-peer connection, immutability, trustless consensus, identity management, and tamper-proof transactions can provide all those solutions in the energy sector [40]. Researchers of [41] have surveyed blockchain importance in the smart grid from both industrial and research perspectives, where they have demonstrated blockchain-based frameworks for energy grid protection and security. Blockchain’s applicability has so far gained much attraction in energy trading. Researchers have proposed blockchain as a platform for decentralized energy trading where customers have anonymity and privacy to directly negotiate about energy prices and perform energy trading with the service providers [42]. More recently some researcher has focused their research on the applicability of blockchain in the phasor measurement system. They proposed a collective decision algorithm to eliminate the possibility of system failure in the case of nodes become malicious [43]. Another group of researchers used blockchain in smart inverters to develop a cost-effective micro-PMU-based behind-the-meter system to enhance the visibility and situational awareness of the smart grid [44].

Despite all the efforts towards smart grid security and privacy, including the IEEE C37.118-2 standard, very little research focused on ensuring the integrity and security of phasor data during transmission and protecting the PMUs and PDCs identity from malicious entities. Hence, we enhance the integrity and confidentiality of synchrophasor communication standard in this work through blockchain and essential cryptographic tools.

2.2.3 Analysis of Existing Threats and Risks in Advanced Manufacturing System and their Countermeasures

The smart factories use sensors and actuators equipped with computing and communication capabilities to collect and exchange data between machine-to-machine and machine-to-human. This collected information is analyzed with the help of modern data analysis tools and transforms into a decision. This significantly increases the industry’s ability to improve productivity by uncovering essential insights from the manufacturing process. Still, the internet connectivity and the use of communication networks to transfer and store data into the data center make the adversaries actively looking to compromise this whole manufacturing process [45]. Security and privacy threats such as device identity spoofing, data tampering, unauthorized infor-
mation disclosure, as well as cyber-attacks such as DoS, MIM, Malicious code, and data tampering attacks [2] cause critical security issues related to network, data, and information security in the industries. Researchers with cutting-edge technologies and research have proposed different solutions against smart manufacturing systems’ security and privacy issues. Researchers of [46] proposed a SPAKE protocol-based lightweight authentication mechanism between the IoT sensors. Other researchers used an attribute-based signature (ABS) scheme to anonymously authenticate sensors and operators, MAC to authenticate system gateway, and Multi-receivers encryption (MRE) to build secure communication between entities [46] [47]. Blockchain technology has recently become an efficient and effective platform for secured data sharing. Researchers have introduced a blockchain-based framework to enforce privacy and trustworthiness on IIoT data by amalgamating differential privacy, federated ML, and smart contract [2].

To counter the issues mentioned above, we have introduced the concept of blockchain technology in the context of an advanced manufacturing system that is compatible with adapting the technological changes and creates a trustable decentralized environment without the need for third-party authorities. The proposed system framework is then described based on its architecture. The clustering mechanism helps maintain the participators of the system, authentication and certification mechanism helps establish a fine-grained access control for remote users and consensus to manage the distributed ledger.

2.2.4 Critical Privacy Challenges for CPS Data

Integration of computing and communications intelligence effectively improves the quality of monitoring and control of CPS. However, the reliance on information technology also raises vulnerability to malicious attacks. For example, False Data Injection (FDI), which attacks data integrity, is emerging as a severe threat to the SCADA system [48]. Furthermore, attackers are launching sophisticated attacks by analyzing the sensitive grid data and identifying the breakers’ locations to trip it off, which causes large-scale blackouts by interrupting the grid’s optimal operation [49]. Recently, to defend against these attacks, researchers already took some steps like real-time state measurement data from the SCADA system that have been leveraged to detect the attacks that cause the blackout [50]. However, these current solutions are limited because they don’t provide any protection for privacy-sensitive valuable
data. Thus, to protect the confidential data of CPS systems, our focus is to ensure privacy protection and security assurance against adversarial privacy-compromising attacks without degrading the utility of the CPS data.

2.2.5 Importance of Data Privacy-Preservation

Privacy preservation is the process of protecting private information against being published or revealed by unauthorized as well as authorized users [51]. Aggrawal and Srikan [52] first introduced it as a new research field in 2008 to mitigate privacy-compromising access and analytics to industrial systems' private data and identify valuable insights. Privacy-preserving techniques aim to modify, transform, distribute, and hide information to prevent exposing the original data during processing and analyzing for intrusion or anomaly detection. Several privacy-preserving approaches such as encryption-based, perturbation-based, authentication-based, differential privacy (DP) based have been proposed so far to effectively protect sensitive information and identify attacks from the cyber-physical networks [53] [54].

However, these solutions lack a proper privacy characterization and measurement technique to identify the expected privacy requirements of individual data providers. As a result, traditional data privacy-preservation mechanisms provide the same level of privacy for all data providers that ensure the confidentiality of the sensitive data but compromise the utility and quality of valuable data.

2.2.6 Privacy Preservation in CPS

Lu [55] proposed a privacy-preserving method for power networks based on the computing protocol of cosine similarity to protect big data of CPS systems. Gope and Sikdar [56] propose a privacy-preserving authentication scheme based on one-way non-collision hash functions for securely analyzing the energy consumption of end-users. Gai [57] introduced a blockchain-based technique to address privacy leakage challenges in smart grids in his research work. Recently blockchain and homomorphic encryption-based privacy-preserving query processing has gained prominence to provide security and privacy assurance in CPS paradigm [58] [59] [60] [61]. Layer wise perturbation and DP-based deep belief network techniques have been also developed recently to examine perturbation points and accomplish data privacy [62].

There are many existing studies on DP and Laplace noise addition mechanisms in the current literature to accomplish effective data privacy protection [62]. However,
not many researchers have focused on personalized differential privacy (PDP), least of PDP in CPS, to improve the data utility while preserving CPS data privacy. To counter these limitations of existing works, we have adopted PDP for the CPS system by characterizing the system’s vulnerabilities, which will help achieve optimal data privacy while maintaining utility and reducing the risk of data disclosure.

2.2.7 Advancement of this Thesis over Existing Works

The advancement of modern cyber-physical systems has increased untrusted third party’s integration in the CPSs. This change presents a sharp growth of unique cyber threats in number and sophistication. For instance, divided administration like distributed energy resources (DERs) are owned mainly by third-party vendors instead of utility can cause large-scale coordinated attacks because of DER’s remote accessible functions. Moreover, the increased cyber-physical interdependencies also impact the CPS systems by launching malicious commands to get unauthorized access into CPS controllers. As a result, untrusted CPS operators can easily access the remote sensor and private data and use it to interrupt the CPSs optimal operation to gain a financial incentive or compromise the reliability of the whole system. CPS service providers can also run advanced analytics on the aggregated sensitive data that will infer the accurate behavior profiles of consumers. CPSs multiple administrative boundaries force the service providers to do key exchange and revocation between various parties in cryptographic security solutions, weakening the security mechanism.

Researchers recently focused their attention on the importance of having a privacy-preserving data aggregation and analytics mechanism for both the CPS service providers and their consumers and introduced methods like K-anonymity, L-diversity, or different encryption-based schemes [63]. However, these researchers mainly rely on cryptographic techniques, which are computationally intensive for resource-constrained CPS devices, raises security vs. budget tradeoff [64]. To address these limitations, some researchers have proposed differential privacy to preserve the confidentiality of sensitive data while eliminating the computation burden. However, differential privacy because of its uniform noise distribution for all users in a dataset causes a serious challenge that is “Limited privacy for some users while overprotecting others” [65].

To counter these security and privacy issues and eliminate the limitations of existing works as well as to protect the cyber-physical systems, we argue that CPSs requires an integrated and automated cyber-physical defense tool that can provide them (a) a
secured, trustless, decentralized data and control information exchange platform(b) efficient privacy-preserving data aggregation and analytics ability while assuring data utility. To this end, we introduce a blockchain-assisted data privacy-preservation model for cyber-physical systems. This integrated defense tool will empower the CPS paradigm by enabling the ability to (i) use a blockchain-based network security layer to protect the highly confidential data, commands and controlling protocols from integrity compromising attacks (ii) personalized data privacy for the CPS systems by characterizing the system’s vulnerabilities, which will help achieve optimal data privacy while maintaining utility and reducing the risk of data disclosure.
Chapter 3

Achieving Integrity for Cyber-physical Systems (CPS)

This chapter addresses the security issues of central controller-dependent cyber-physical systems (CPS). We use blockchain, clustering algorithm, and cryptographic mechanism as a solution concept to improve the integrity of CPS against those security and privacy-related issues. We further present two case studies in this chapter, divided into two subsections. One is "Block-Phasor: A Decentralized Blockchain Framework to Enhance Security of Synchrophasor," and the other subsection is "Blockchain-based Secure and Reliable Manufacturing System" to demonstrate the applicability of our solution concept.

3.1 Case Study 1: Developing Block-Phasor: A Decentralized Blockchain Framework to Enhance Security of Synchrophasor

3.1.1 Motivation and Contribution

Traditional SCADA-based systems capture snapshots of the grid every 4-6 seconds, resulting in broader time gaps for monitoring steady-state conditions and dynamic disturbances and post-disturbance due to the lack of synchronization and continuity of data stream. Synchrophasor-based systems address this issue by enabling real-time GPS-synchronized phasor data capturing and analysis through the use of phasor measurement units (PMUs) and phasor data concentrators (PDCs) which are 100 times
faster than SCADA measurements [66]. Although PMU data has reshaped the power grid with superior monitoring and control tools, these tools can become vulnerable to a broad array of cyberattacks, including PMU data interruption, interception, modification, and fabrications. We argue that there must be a novel and decentralized way of mitigating a cyberattack at the edge of the grid before it hits the control center and causes a single-point-of-failure. The cyberattack on the Ukrainian power grid that caused a major blackout for three hours underlines the necessity of the decentralization of the smart grid’s security enforcement [8]. Besides the point of failure issue, different grid sensing and monitoring devices are susceptible to data manipulation attacks. Interestingly, blockchain, the core technology developed to revolutionize the traditional payment system, has the ability and necessary mechanisms to point to major vulnerabilities of the synchrophasor network. Blockchain can offer decentralization, security, immutability, and verifiability by design that can mitigate the notable and potential vulnerabilities of synchrophasor systems such as data manipulation and other insiders and outsider attacks.

**Our Contributions:** To mitigate the security issues in the synchrophasor network while collecting and forwarding measurement samples through PDCs, we propose a blockchain-based and decentralized security framework to mitigate data tampering attacks and achieve accountability of all devices in the system at the edge. Specifically, we propose:

- A blockchain-based and decentralized framework to overcome the system’s single point of failure issue. This will enable us to connect the PMUs with the PDCs through blockchain and share synchrophasor data securely.

- We propose bloom filter-based fast identity validation techniques for the PMUs to quickly ensure the PMU device’s authenticity before sending the measurement samples to PDC.

- Merkle tree-based scalable and fast measurement data verification to achieve accountability and mitigate data tampering attacks.

- Using this method, the PMUs and corresponding PDC can gather the proof of sent data from PMU without revealing the data to other PMUs. This process can be done within a few milliseconds to build the Merkle tree and achieve verifiability for practical scenarios.
3.2 Proposed Blockchain based Synchrophasor Wide Area Monitoring System

Synchrophasor network uses the IEEE C37.118-2 communication protocol to transmit measurements between PMU, PDC to the control center over the communication network. This communication may lack proper security measurements, which makes it a target for cyber-attacks. So, we have proposed a decentralized synchrophasor framework in this section with proper authentication and data verification mechanisms to mitigate the security issues of synchrophasor data transmission. This system also enables PDCs to act in a decentralized manner and eliminate security vulnerabilities at the edge of the substations. The system model is equipped with bloom filter [67] for PMU identity validation so that malicious or Sybil PMUs can not access the system and pretend like an honest PMU, Elliptic Curve Digital Signature Algorithm (ECDSA) [68] signature mechanism to ensure the authenticity of PMU and its sent measurement data to PDC, and finally, Merkle tree to store and verify the proof of PMU data to ensure integrity and accountability of the device. The detailed working mechanism of our proposed system model is shown in Fig. 3.1 and described below.
3.2.1 Bloom Filter based PMU Device ID Validation

PMU devices provide real-time monitoring of the phasor data but primarily are located at unsafe locations. There is a chance that an adversary may create a false identity and access the network without registering into the system properly. To mitigate this issue and considering the nature of the synchrophasor network, we need to ensure fast identity validation mechanisms. A Bloom filter is a fast query algorithm that can be used for immediate identity validation. It can quickly confirm without revealing true identity whether an ID is legal in the system.

Bloom filter initialization phase

All the PMUs in the private blockchain network will store their unique ID in a database. Bloom filter requires \( k \) hash functions to produce an output on the possible inputs. Let us denote the identity of a PMU as \( ID_i \) where \( i = 1, 2, ..., m \) and \( m \) is the total number of PMUs for a given region. For \( ID_i \), at first we choose \( k \) hash functions. These functions will be used to calculate the hash values of PMU \( ID_i \) and set the mapping values to the corresponding positions. The hash is calculated as \( H_j(ID_i) \mod n \) (according to Fig. 3.2), where \( H_j \) is one hash function of total \( k \) and \( n \) is the size of the array. Initially, the array value is set to be zero when an ID is stored in the array. The hash functions output as index values and set the mapping values of the corresponding positions in the array to be one. This way, all authorized PMUs at first will store their corresponding ID in the bit array for future verification. Fig. 3.2 shows the overall process of bloom filter-based PMU authentication.

Bloom filter based Validation Phase

To check if \( ID_i \) is authorized in the authentication phase, we first calculate the same hash values and check the bit values. If all \( k \) bits are 1, it means the record exists.
Even if a single bit is 0, the record does not exist. By using the bloom filter in Fig. 3.2, we can easily assume the legal and illegal IDs based on the fact that the mapping values contain zero or one.

### 3.2.2 ECDSA based PMUs Measurement Authentication

Although the bloom filter can provide fast identity validation, there is still a chance that attackers may forge the real identity of PMU. To solve this issue, we also use the ECDSA based fast PMU authentication to guarantee phasor measurement integrity and authenticity during transmission. ECDSA uses pre-computed tokens to generate signatures quickly and also has a small key size as well as low computation and communication overheads which make it a suitable candidate for the energy system. To achieve this authentication, while sending the data to PDC, PMUs can generate public and secret key pairs as $P_{ki}$ and $S_{ki}$ respectively and sign the hash of the data as $\text{sign}(S_{ki}, H(\text{msg}_i))$, where $\text{sign}()$ denotes the signature function and $\text{msg}_i$ denotes the measurement data sent by PMU ID$_i$. Then, the PMU ID$_i$ sends $\{P_{ki}, \text{sign}(S_{ki}, H(\text{msg}_i)), \text{msg}_i\}$. From this message, the PDC can verify if the PMU is authenticated or not by computing $H(\text{msg}_i)$ and then $\text{verify}(P_{ki}, \text{sign}(S_{ki}, H(\text{msg}_i)))$, where $\text{verify}()$ denotes verification function.

### 3.2.3 Merkle Tree based Verification

To achieve the integrity of phasor measurement data sent from PMUs to PDCs, we use a Merkle tree-based structure to store the data into the blockchain. Merkle tree is a binary tree-like structure that can help achieve fast verifiability without revealing any PMU data to each other. Since PMUs are primarily busy with sensing and sending the data to PDC and PDCs have more computation power than PMUs, we assume that PDCs build this Merkle tree and store the Merkle root into the blocks. Later these blocks are shared among the PDCs to store them in the blockchain. For instance, the data message coming from PMU ID$_i$ is converted as a transaction for a given time window. Then PDC converts the received data into a hash and stores them in a Merkle tree as shown in Fig. 3.1 (Merkle tree-based block structure). The PDC can fix a time window and build one Merkle tree and a Merkle root, gathering all the PMU data transactions together within the window. Then the PDC stores the Merkle root into the block. In case of any dispute or forensic request, the input data can be verified without disclosing any other PMU data. For instance, to verify
if PMU0's data exists, PDC can reveal Hash1, Hash2n, and the Root, and it can be verified of PMU0 data was tampered or not.

3.2.4 Mining and Generation of Distributed Ledger

In the proposed framework, as shown in Fig. 3.1, we have used a private blockchain model where PDCs are assumed to be the mining nodes. We also assume that there is one PDC for each region with multiple PMUs. The PDC nodes are connected in peer to peer fashion. Once the PDCs receive the transactions from PMUs and generate the block, including the Merkle root, it broadcasts the block with other PDCs for verification. All PDCs follow the same process and reach consensus through PBFT consensus protocol [69].

The primary goal of this consensus algorithm is to reach a final agreement to add blocks into the ledger based on the acceptance of the majority of the honest nodes. To commit a block and progress, the nodes in a PBFT network go through five phases: Request, Pre-Prepare, Prepare, Commit and Reply. Among the PDCs, they work in different roles such as client, leader, and replica. These roles are selected in a round-robin fashion.

The working mechanism is as follows. At first, the primary PDC node receives a request to add a block into the ledger from the client PDC node. The leader PDC node forwards this block request to the other three replica nodes. The nodes perform the requested service, and when the majority of the nodes agree on the validity of the block, they send a valid reply to the client node. But, when a PDC node is crashed or manipulated, this request goes through all five phases to reach a consensus. Finally, when the majority of the nodes agree on the consensus, the primary node commits the block into the ledger. We also consider that if the primary node goes rouge, the majority of the honest nodes can vote on the legitimacy of the current leading node and replace it with the next leading node in line. So, the criteria for valid leader node acceptance are: \( \frac{L}{P} > \tau \), where, \( P \) : PDC nodes, \( L \) : most yes votes to replace Leader Node and \( \tau \): threshold whose value must be greater than 50 percent). When the majority of the honest nodes are in agreement, then the primary node is replaced.
3.3 Experimental Analysis

In this section, we present the detail of our experiment. We conducted the experiment on an Intel(R) Xeon(R) CPU@2.30GHz PC using Python 2.7 (64 bit). The implementation of the proposed method on the grid edge devices with lower computational capability is left of our future work. Here, we present the evaluation of three of the critical relationships upon which the efficiency and efficacy of the proposed blockchain layer are largely depended\(^1\):

1. Impact of the number of PMUs on Merkle tree generation time.

2. Relation between the number of PMUs and number of required hashes for verification.

3. Relation between number of PMUs and hash verification time.

We run the experiment for the following PMUs: 10, 50, 100, 150, 200, 250, and 300. Another goal of this experiment is to understand and analyze different scalability issues associated with integrating the proposed blockchain framework into the synchrophasor network. Currently, a single PMU transmits 30 to 60 measurement samples per second to a PDC. If we consider each measurement as an individual transaction, then both communication and computation costs on PMU and PDC will increase rapidly. Hence, in the experiment, we consider a transaction happens every

\(^1\)The computation cost of blockchain creation is out of scope of this thesis as the proposed blockchain is an overlay network of synchrophasor and ensuring fast verifiability with merkle tree is important to cope with fast PMU measurement data transfer.

![Figure 3.3: Merkle tree generation time for different number of PMUs on a single PDC.](image-url)
second, and the measurements sent by a PMU in a second are packed into that single transaction. The experimentation with different time intervals is left of our future work.

Simulation and Discussion

The experimental result on the relation between the different number of PMUs and Merkle tree generation time is depicted in Fig. 3.3. While the figure shows a linear relationship between two variables, the most crucial point is the time needed to generate the tree. In a PDC with a few PMUs (e.g., 10), the required time is less than 0.2 milliseconds. If the number of PMUs increases by 400% to 50, the required time doubles (less than 0.4 milliseconds). If we further triple the PMU number (200% increase), the required time increases only by only 45%. Even for 300 PMUs, the required time is only around 1.5 milliseconds. Such a low number underlines that the proposed blockchain framework can be utilized for real-time security mechanisms in the smart grid.

Next, we evaluate the relationship between the number of PMUs and the required hashes to generate a Merkle tree root hash for verification. This verification process on a PMU should yield low computation and communication complexities. To verify whether a Merkle tree, generated by a PDC, is correct, a PMU needs $\log_2(n)$ number of hashes, where $n$ is the number of nodes in the tree. In other words, this process incurs an additional $O(\log_2(n))$ communication cost. The value of $n$ depends on the number of PMUs under a single PDC. Fig. 3.4 shows the relationship between the number of PMUs and the number of required hashes to generate a Merkle tree root,

![Figure 3.4: Relation between number of PMUs and number of required hashes for verification in a Merkle tree.](image-url)
Figure 3.5: Relation between number of PMUs and hash verification time.

where for a large number of PMUs, the required number of hashes are pretty low. It is important to note that, in the current implementation of the synchrophasor network, for hash verification, a PDC needs to send only at most 7 hashes.

Finally, in Fig. 3.5, we present the analysis of the relation between the number of PMUs and hash verification time. Unlike the Merkle tree generation operation, which a PDC must compute every second, hash verification is done only when necessary. Thus, the overhead yields by the hash verification are occasional. Similar to the curve in Fig. 3.4, the curve in Fig. 3.5 also shows a logarithmic relationship between hash verification time and the number of PMUs. Furthermore, the actual required time is quite low (e.g., for 50 and 300 PMUs, it is 0.0020 and 0.0040 milliseconds, respectively).

3.4 Summary

In this case study, we provide an overview of synchrophasor network cybersecurity challenges and introduce a decentralized blockchain-based technology as a protection framework to tackle the vulnerabilities in PMU data transmission. The proposed blockchain-based framework with a bloom filter and Merkle tree substantially improves the grid devices’ self-defensive capabilities at the edge against data manipulation by cyber attackers.
3.5 Case Study 2: Blockchain-based Secure and Reliable Manufacturing System

3.5.1 Motivation

The Industrial Internet of Things (IIoT) integration in the manufacturing system has shifted the traditional industries into a fully connected and automated smart industry. The product manufacturing system consists of $M$ machines, each equipped with $n$ IoT sensors such as $M(n)$. Because of the computing and communication capabilities, the IoT sensors are connected with each other $(n_1, n_2, ..., n_j)$ and with the manufacturing controllers $C$ through the internet, which makes the Industrial production system an attractive target for cyberattacks. Attackers regularly launch cyber and physical attacks such as hardware tampering, side-channel attack, reverse engineering, IoT sensor identity spoofing, malicious data injection as well as tampering product design and configuration data to compromise the production line [4]. However, the security and privacy issues in intelligent manufacturing are not only caused by external attacks but also about 60 percent of the attacks are done by internal employees and contractors to steal intellectual property [5]. So, the inter-operable and interconnected manufacturing systems provide significant value in the production process; it also increases the exposure to many security risks, with critical and financial impacts.

3.5.2 Threat Model

In this work, we focus on protecting IIoT sensor-based devices from unauthorized access and provide immutability and resiliency against data and command modification as soon as possible before the attack causes irreversible damages to the production line, such as compromising the whole production process. Therefore, we consider $p_i$ is the probability that the attacker compromises sensor $i$ and manipulates the product configuration data $m$. We assumed that the production line of the manufacturing system consists of a $n$ number of IIoT sensor devices. This sensor network controls the state of the production process, and they communicate with the controller and share operational measurements (e.g., product design, development, and configuration data) and control commands between them. The measurement vector $m(t) = \{m_1(t), ..., m_n(t)\}$, where $m_i(t)$ denotes the measurement sent by sensor $i$ at time $t$. Let, $\tilde{m}(t) \in \mathbb{R}^n$ denote the received measurements by the controller at time $t$. Based on these received measurements, the controller or the control system defines control
actions to maintain a certain production process. If some of the sensors are under attack or the measurement data or commands are tampered with then, $\tilde{m}(t)$ will be different from the real measurement $m(t)$. These general attacks on IIoT sensors and manipulation of sensor measurements will impact the manufacturing process’s confidentiality, integrity, and availability.

3.5.3 Contribution

To mitigate the security issues of the manufacturing line, we propose a decentralized blockchain-based secured and immutable control information (code, command, and data) exchange platform to protect the production control system from unauthorized access and provide resiliency against data and information manipulation attacks. The main contribution of this case study is as follows:

- A p2p connected system model is proposed using blockchain to migrate the central controller-dependent manufacturing system into an intelligent decentralized system and eliminate the need for trusted third parties.

- A clustering algorithm is used to group the participating nodes into clusters to ensure scalability and reduce processing and packet overhead on IoT devices. The cluster header is responsible for managing the blockchain.

- A digital signature-based certification mechanism is introduced to ensure the integrity of private information such as control commands and data and prevent them from being manipulation by unintended participators.

- A distributed consensus algorithm is used here to achieve the consistency of the proposed decentralized system. This helps the network participators to work collaboratively and store the transactions securely and effectively.

3.6 Proposed Blockchain Based Solution

The proposed blockchain framework for the advanced manufacturing system includes Service providers, IIoT equipped smart machines, Cloud storage, and Terminal to jointly form an overlay network. Fig. 3.6 represents the participators of the overlay network. Similar to other blockchain-based frameworks cryptography mechanism is used here to provide participators anonymity, identity authentication, and fine-grained access control between the participators of the overlay network to protect it
Figure 3.6: System Model of Blockchain-based Manufacturing overlay network.

from unauthorized access. The production line is primarily equipped with IIoT devices, and because of the resource constraint of these devices verifying each new block and transaction is challenging. To overcome this issue and ensure better scalability, low latency, and better throughput, we have used a clustering algorithm to group the overlay network participators into clusters. The participating nodes of the clusters also select a cluster header as well as a co-leader of the cluster in the case cluster header becomes malicious or leaves the network due to sudden loss of connectivity. These cluster headers are responsible for maintaining and managing the blockchain-based overlay network of the manufacturing system. Managing the blockchain means providing a secured platform between the manufacturing layers to generate, verify, and store each transaction in an immutable, decentralized, and trustless way to protect the transactions from malicious modifications.

The working mechanism of the system model in Fig. 3.7 is when participating nodes wish to ask for a request, consider that an operator sends the command to the service provider to change the production parameter to reduce the product manufacturing time; they at first need to publish a corresponding transaction in the blockchain overlay network. In our system, we adopt a private blockchain network that consists of some processing and consensus nodes (i.e., cluster header) to maintain the blockchain according to the consensus mechanism. Unlike transactions that are broadcasted in the network and stored on-chain, the data generated by the IIoT devices are stored off-chain, mainly in the cloud. This helps the system model ensure better scalability by reducing the packet overhead and memory requirement of the
blockchain network.

3.6.1 Clustering Mechanism

The industrial production line is equipped with randomly placed IIoT enabled machines and devices. This unstructured environment creates overlapping network topology, which is not suitable. So, we have used a clustering algorithm such as [70] to group the IIoT devices into clusters. This Algorithm 1 discovers network nodes by exploiting local network topology knowledge and forming clusters in random geometric graphs. The clusters of our blockchain overlay network consist of two types of nodes: Cluster header or also referred to as Blockchain Manager (BM), and Cluster Members. The cluster header is selected based on its maximum coverage ability in the network. This will facilitate covering the whole network with a minimum number of nodes.

3.6.2 Authentication Mechanism

The main goal of our approach is to create a secured control information exchange environment for the manufacturing system so that it will be resilient against IIoT device unauthorized modification and control data manipulation. The Blockchain managers of our proposed system are p2p connected and consider every non-group
Algorithm 1 Cluster Head Selection

Input: Topology (T)
Output: Cluster Head $CH_{id}$.

1: Each node has their topological characteristics; Such as Total number of neighboring Nodes;
2: Nodes Calculate their Degree ($ND_i$) from the number of edges incident to that node.
3: Each Node Advertises their own $ND_i$ to one hop away neighbors;
4: Each node collects its local and neighboring nodes degrees as $ND_i, j$;
5: Cluster Head Selection ($CH_{id}$): Node with the maximum degree ($ND_i, j, ND_i$)
6: If node i is a cluster head ($CH_{id}$) without any member
then
7: $CH_{id}$ joins the neighboring cluster as a member where cluster head $CH_k$ has the maximum degree $ND_k$;
8: End

node as malicious. So, when a participating node $A$ of the blockchain overlay network sends a transaction to another node $B$, then (1) $A$ sends the request to the cluster header, (2) Cluster header checks the authenticity of $A$, (3) If, $A$ is valid then cluster header broadcast the transaction in the network, Finally, (4) $B$ can read the transaction. The authentication mechanism of the proposed system model, which is based on the Elliptic Curve Digital Signature Algorithm (ECDSA), is described below.

Initialization Phase

The cluster header ($CH_{id}$) of the groups can be considered as certification authority. The group members in the overlay are known by their public and secret key pairs as $PU_i$ and $SE_i$ respectively. Then, when a cluster member request a transaction. At first, it signs the hash of the transaction as $\text{sign}(SE_i, H(\text{Tran}_i))$, where $\text{sign}()$ denotes the signature function and $\text{Tran}_i$ denotes the transaction sent by member $M_{id}$. Then, the $M_{id}$ sends the {$PU_i, \text{sign}(SE_i, H(\text{Tran}_i)), \text{Tran}_i$} to the $CH_{id}$.

Verification and Certification Phase

The $CH_{id}$ checks the authenticity of $M_{id}$ by computing $H(\text{Tran}_i)$ and then $\text{verify}(PU_i, \text{sign}(SE_i, H(\text{Tran}_i)))$, where $\text{verify}()$ denotes verification function. After verifying the authenticity of the $M_{id}$ and $\text{Tran}_i$; the cluster header ($CH_{id}$) uses a lightweight certification mechanism to generate a certificate which contains: (1) the cluster ID ($C_{id}$) of the blockchain overlay network and (2) the cluster member’s identifier
\{PU_i, \text{sign}(SE_i, H(\text{Tran}_i)), \text{Tran}_i\}. The \(CH_{id}\) then signs the certificate with its private key \(CH_{prk}\) and broadcast the transaction in the network. The receiving cluster nodes verify the integrity of the transaction with the cluster header’s \((CH_{id})\) public key \(CH_{puk}\). If, the certificate is valid, then the overlay nodes stores the certificate as well as its hash content \(H(C_{id}, M_{id}, PU_i, C_{puk}, Tran_i)\) in the ledger.

### 3.6.3 Raft-based Distributed Consensus and Ledger Generation

The raft is a distributed leader-based consensus algorithm that can play a crucial role in an advanced manufacturing system by providing fault-tolerance as well as maintaining safety, efficiency, low latency, and better throughput [71]. The primary goal of this consensus algorithm is to solve the problem of getting the majority of servers (nodes) to agree on a shared state of the system even in the time of node failures. This will ensure an immutable ledger generation and protect against malicious block and ledger generators. At any given time, the cluster heads of our proposed overlay network can be in any one of these three states: 1) leader, 2) follower, or 3) candidate. The leader node is responsible for managing the replicated ledger. It accepts transaction requests from member nodes and replicates and forwards these newly received transaction requests to its followers. At this state, the requests are still uncommitted and stay in a volatile situation. Once the leader node receives the valid response of the majority of the followers, it updates the request in the replicated ledger and notifies the participators that the transaction is recorded. Fig. 3.8 represents the state transition mechanism of the members of our system. The algorithm (2) for leader election and ledger generation for our proposed system is as follows:

![Figure 3.8: State Transition Model of RAFT](image_url)
Algorithm 2 Leader Election and Ledger Generation
Raft decomposes consensus into three sub-problems: Leader Election, Ledger Replication & Ledger Safety.

**Input:** Remote procedure calls (RPCs).
**Output:** Received or NotReceived

1. Leaders send periodic heartbeats = *AppendEntries* RPCs to all followers to maintain their authority.
   
   If
   
   Followers notreceived *AppendEntries* RPCs
   
   then
   
   Followers → *RequestVote* RPC to all the nodes
   
2. *AppendEntries* RPC with requested transactions.
   
   If successful
   
   then
   
   update requested transactions in the ledger.
   
   else
   
   leader retries *AppendEntries* RPCs indefinitely until all followers eventually store all the log transactions.

3. **End**

3.7 Experimental Analysis

For the suitability and applicability of blockchain in an advanced manufacturing system to improve the performance, security, and reliability, we performed a numerical analysis of our framework in the context of hyperledger fabric 2.0. Hyperledger is a private blockchain platform for enterprise-level application development. It is suitable for our system because while most blockchain platform employs a transaction, order-execute model. The transaction flow of our system model follows the transaction execution-validation-consensus-update model similar to hyperledger fabric.

To evaluate the performance of the proposed system model, we have conducted a numerical analysis based on the benchmark experiments that have been conducted in [72] [73]. The resulting analysis of the system model was done against the blockchain framework’s throughput metrics, which refers to successful transactions per second in the system.

In Fig. 3.9 the analysis is done to observe the impact of different transaction rates and their impact on the blockchain framework. Transactions are classified as ‘open’ and ‘query’ transactions, where ‘open’ transactions perform both read and write in the system, and ‘query’ is simply a read-only transaction.
We analyze Fig. 3.9 for different transaction rates, varying from 20, 40, 60, 80, and 100 tps; how the transactions impact the throughput of the system. For ‘query’ transactions, the blockchain system can handle higher transaction rates without any significant delay as only read operation performs here. We also observed in the case of ‘open’ transactions, as the transaction rate increases, the system overhead also increases. This overhead impact the throughput of the system. The throughput increases linearly as expected, but it gradually starts decreasing around 60 tps. This delay occurs because of the specific system environment. So if we can scale up the system environment to support higher transaction rates, then the throughput will also increase. We can also assume from this that the system’s latency will also depend on the different workloads where ‘query’ transactions will produce minimal latencies.

From the analysis, we can generalize the scalability of the proposed blockchain system model depends on different parameters, i.e., the total number of cluster members, no. of cluster headers, the peer number, total number of transactions, verification time, and the consensus protocol. If the system’s hardware configuration has a higher spec, it will significantly improve the system’s performance. The goal of this numerical analysis is to initially evaluate the feasibility of our reference architecture.

3.8 Summary

In this case study, we have introduced the concept of blockchain technology in the context of an advanced manufacturing system that is compatible to adapt the tech-
nological changes and create a trustable decentralized environment without the need of third-party authorities. The proposed system framework is then described based on its architecture. The clustering mechanism helps maintain the participators of the system, authentication and certification mechanism helps establish a fine-grained access control for remote users and consensus to manage the distributed ledger. Initial performance analysis regarding the benchmark platform shows that blockchain is suitable for the advanced manufacturing system.
Chapter 4

Vulnerability Characterization and Privacy Protection Model for Cyber-physical Systems

The primary objective of this chapter is to ensure CPS-based system’s sensitive data’s privacy preservation while supporting high data utility. For that, we have developed a novel privacy preservation mechanism, "Personalized Differential Privacy (PDP)," to determine the required privacy protection level. This PDP is created by characterizing and identifying individual CPS node’s privacy preferences to maximize the utility of the data while ensuring individual node’s data privacy preservation.

4.1 Motivation and Contribution

Understanding vulnerability and characterizing weaknesses before an attack occurs is essential to minimize security, privacy, and performance issues. However, the lack of comprehensive vulnerability assessment of CPS systems and over-reliance on random fault analysis or quantification of privacy after an attack creates a significant gap in the security and privacy of CPSs [74]. Moreover, traditional data privacy preservation mechanisms (e.g., Differential privacy, k-anonymity, l-diversity, etc.) mainly provide the same level of data privacy to all the data providers, which ultimately leads to inadequate privacy preservation for some data providers while over-protecting others because not all data providers require the same level of privacy [12]. To counter this problem, our goal in this chapter is to build a data privacy preservation mechanism
where data providers will specify their private data’s privacy requirements.

**Our Contributions:** In this chapter, we focus on introducing a vulnerability characterization and privacy quantification model for CPS. The contributions can be summarized as follows:

- We at first characterize the privacy issues of the CPS entities through a Standard Vulnerability Profiling Library (SVPL) during CPSs data generation, sharing and aggregation. This characterization helps us propose a robust personalized privacy model for the CPS by obfuscating individual data providers’ attributes.

- This proposed PDP scheme is based on individual node’s privacy requirements. By personalizing the node’s privacy preference, we eliminate the issue of protecting privacy uniformly. This consequently meets the user-specific privacy guarantee, which increases the data utility by eliminating excess privacy loss.

- This proposed privacy model is proved to be efficient enough for data privacy-preservation in our experimental analysis, where we consider the whole environment is untrusted. Aggregators can only learn their desired statistics without inferring additional information from the system. This helps the CPS system to provide resiliency against data correlation, linking or disclosure attacks.

4.2 **Proposed Vulnerability Characterization Model**

The formulation of the proposed vulnerability characterization model is shown step by step in this section, and based on it; we formulate our proposed personalized data privacy model.

4.2.1 **Smart Grid Topological Relationship**

Electric grids can be represented with nodes (system buses) and their associated connectivity links (e.g., transmission and distribution lines). Such relationship can be modelled as a graph \( G(\mathcal{N}, \mathcal{L}) \) where \( \mathcal{N} \) and \( \mathcal{L} \) represents the number of nodes and communication links. Multiple communication links connected to the same node pair are considered as one link here. The interconnectivity of the graph \( G(\mathcal{N}, \mathcal{L}) \) is represented by a \( N \times N \) adjoint matrix \( \mathcal{A} \) where \( \mathcal{A}_{ik} = 1 \) if the nodes \( i \) to \( k \) are coupled by a direct link; otherwise \( \mathcal{A}_{ik} = 0 \). The \( N \times N \) adjacent matrix is presented as a Laplacian matrix \( L \) to present the properties of the grid topology in \( L = \Delta - \)
Figure 4.1: Vulnerability Characterization Framework

$A$, where $\Delta = \text{diag} (e_1, \ldots, e_N)$ is the diagonal edges of the matrix connecting the centers of opposite nodes with the edges and their associated elements $E_i = \sum_{k=1}^{N} a_{ik}$.

4.2.2 Vulnerability Characterization Framework

In this work, we use an attribute-based attack graph to assess the network’s vulnerabilities. The vulnerability characterization framework incorporates five parts as shown in Fig. 4.1. These are CPS network properties, attack graph creation, vulnerability profile generation, privacy loss assessment using SVPL and maximum privacy loss calculation, and at last, vulnerability characterization and quantification module. The evaluation includes the following steps:

- First, we generate an attack graph by analyzing existing information like the grid’s topological relationship to traverse and assess the previous attacked incidents to collect the vulnerabilities of the compromised nodes.

- This attack graph is then leveraged to build SVPL by calculating the vulnerability scores generated from risk sources, data privacy weaknesses, feared events, and associated harms.

- According to the computed vulnerability score, the maximum to minimum privacy loss is calculated by combining the privacy loss magnitude and the frequency of the successful privacy-compromising attacks.

- Finally, this maximum privacy loss assists in identifying the network’s key fragile points by creating the best attack profile that will help the grid utilities to reinforce their budgets appropriately and ensure defense before an attack occurs.
4.2.3 Vulnerability Assessment by Building an Attack Profile

To perform a complete privacy risk assessment and develop the vulnerability attack profile, we need at first to establish a common link by defining the relationship between the “feared events", “risk sources," and “privacy weaknesses" to model their impacts called “privacy harm."

Definition 1 (Risk Source:) Risk sources define an individual or organizational entity whose legitimate or illegitimate ownership of a dataset and actions causes intentional or unintentional privacy harm to the system.

Definition 2 (Privacy Weaknesses:) Privacy weakness is a limitation of the data protection mechanisms’ security measures, including inadequate security functionality or error-prone system design, which ultimately causes privacy harm in the system.

Definition 3 (Feared Event): Feared events are the malicious actions on the system that exploit the system’s privacy weaknesses, leading to privacy harm.

Definition 4 (Privacy Harm:) Privacy harm is the negative impact on datasets, individual data providers, or the network resulting from one or more feared events to compromise the system’s physical, cyber or financial stability.

Based on these four attributes and traditional attack graph model [75], we develop our vulnerability assessment attack graph (VAAG), which comprises eight tuples as follows: $VAAG = (N, E, PC, M, D, R, PLM, FPLE)$

- $N$ represents the nodes of the system which consists of start node $S_n$, attacked nodes $A_n$ as well as attacked path $A_p$, where, $N = S_n \cup A_n \cup A_p$. If $A_n \in A_p$, then $A_n$ is a compromised node on the attack path.

- $E$ represents the edge set of all $N$ nodes where $E \subseteq (S_n \times A_p) \to (A_p \times A_n)$, and where every privacy attack is denoted by an action $a \in E$.

- $PC$ denotes the physical connectivity of the nodes, where $PC \subseteq (N \times N_l)$. $N_l$ is the physical link between the nodes.

- $M$ is a mapping function that maps the edge sets $E$ to the compromised set $C$, which is $M : E \to C$. In the VAAG, every malicious attack ($a \in E$) is the result of a vulnerability $C(e)$ on that node according to two privacy risk factors: Privacy Loss Magnitude (PLM) and Frequency of Privacy Loss Events (FPLE).

- $D$ denotes the vulnerability dependencies of the nodes presented by $d_i : v_i \to v_j$. This means if adversaries try to compromise node $N$ then both $v_i$ and $v_j$
vulnerability set is needed.

- The privacy risk score is calculated by decomposing two risk factors: PLM, which denotes the privacy attacks on the data points, and FPLE, which presents the successful occurrences of an attack by a malicious actor. This PLM and FPLE are calculated from the existing vulnerability information of the CPS infrastructure. The overall risk is calculated by combining these risk factors: $R = PLM \times FPLE$.

### 4.2.4 Standard Vulnerability Profile Generation

The formulation of our proposed SVPL to identify and characterize the vulnerabilities of the CPS nodes mainly consists of three parts: traversing the compromised nodes, developing an attack graph, and using these sub-level attack graphs to generate the nodes’ global attack profile. In this process of traversing and identifying the compromised nodes, we start our process by going through the existing vulnerability information of the nodes $V$ and the attack incidents $AI$ and use it as a baseline profile to generate global vulnerability assessment attack graph "VAAG". The attack incident set $AI$ is the set of attack preconditions $AI_P$ and attack consequences $AI_C$. The vulnerability profile generation follows Algorithm 3, where $AI$ at first checks if there is an available attack incident by traversing and matching the current information

---

**Algorithm 3** Vulnerability Profile Generation

**Input:** Nodes existing vulnerability set $V$, attack incident set $AI$ and newly added nodes $N$.

**Output:** Newly Compromised nodes $CN$ and their edge sets $E$.

**Step 1:** if $AI \neq \emptyset$ then **return**

**Step 2:** for each $ai \in AI$

- if $N \in ai \cdot AI_p \land (ai \cdot AI_p - \{ N \}) \subset V$ then
  1. Start attack node creation $A_n$
  2. $CN \leftarrow CN \cup \{A_n\}$
  3. for each targeted nodes $T_n \in ai \cdot AI_p$
  4. $E \leftarrow E \cup \{<T_n, A_n>\}$
  5. for each $T_n \in ai \cdot AI_c$
  6. if $T_n \notin V$ then
  7. $V \leftarrow V \cup \{T_n\}$; **End**

**Step 3:** for each $A_n \leftarrow N$

8. $CN$ Match ($V$, $AI$, $AI_p$ & $AI_c$)

**End; return** $CN$, $E$
with the system’s existing information. If abnormalities are found in the new node \( N \) then it is considered as a compromised node. This compromised node \( N \) will be used to find the previously connected nodes by linking through the edges and put the whole compromised node-set in the \( CN \) and their edges in \( E \). If this newly vulnerable node is not included in the \( V \) set, it will be added to the existing vulnerable node set. Finally, every node of \( CN \) will be matched with the \( AI \) set to profile the different vulnerabilities.

### 4.3 Proposed Personalized Differential Privacy Model for CPS

In our proposed personalized differential privacy (PDP) scheme in the Fig. 4.2; the end devices \( S = M1..M(n); PMU(1)..PMU(n); DER(1)..DER(n) \) regularly transfer their ready-to-publish data to the adjacent fog (\( F \)) and cloud nodes (\( C \)). Analysts, when submitting their access request to the central control center, it is considered as a query in the aggregated database. However, before releasing the data to the analysts, each data provider first processes their data by their preferred \( \epsilon \)-differentially private mechanism and submits it to the neighboring fog nodes. PDP is introduced in our scheme by employing a Laplace mechanism where individuals generate their non-uniform noisy outputs \( \epsilon(0) \).

![Figure 4.2: Trust Distance based Personalized Privacy Model for CPS.](image-url)
This scheme is dependent on every source node’s privacy preferences that is shown in subsection 4.3.5 in Table 4.1. For example, based on the criteria of access to nodes, distance is crucial because longer distances (more hops) will force the transmitting data to travel through more malicious nodes and raises the likelihood of getting compromised. Trust distance ($T^D$) is calculated from their effective distance from the central cloud center. The longer the trust distance from the central control center, the higher the required privacy protection level. We assume end devices with significant hop distance are more prone to privacy attacks by malicious entities. However, before describing our proposed privacy scheme, we first need to characterize the privacy issues in CPS.

**Privacy Characterization and Quantification**

To introduce our personalized differential privacy scheme in the CPS, we need to characterize individual end nodes’ privacy requirements by considering the nodes’ combinational relationships using different attributes. In our privacy quantification, we consider any individual node $S$ owns at least one dataset $D$. We further assume that these datasets are regarded as a combination of multiple variables $v = \{v_1, v_2, \ldots, v_n\}$, where each $v$ has a set of attributes $A = \{a_1, a_2, \ldots, a_n\}$ in their corresponding datasets. To ensure the privacy-preservation of dataset $D$, the different attributes ($a$) of the dataset need to be obfuscated to avoid linking individuals in the aggregated database. For this, each dataset requires a sanitization to obscure its quasi-identifiers before publishing or disclosing it to the requesters. In short, we set a finite set $A = \{a_1, a_2, \ldots, a_n\}$ where $n$ shows the possible characteristics of the sensitive attributes in the dataset (e.g., individuals appliance usage data in a consumption dataset). The quasi-identifiers are presented as $Q_ID$ where two distinct nodes $s_i$ and $s_j$ are assumed quasi-equivalent if $s_i(Q_ID) = s_j(Q_ID)$, i.e., they share some of the equivalent attributes in their dataset. This privacy quantification aims to identify the attribute-based linkage in the dataset and use our privacy scheme to generalize individuals’ quasi-identifying attributes to protect sensitive information from attackers.

This sanitization is performed by mapping the datasets’ original values to a generalized value using individuals privacy preference ($\Phi$), which depends on a mapping function called “trust distance ($T^D$)”. This ($T^D$) denotes individuals with a larger hop distance from the central control center require greater privacy protection. Datasets are sanitized and the variables are presented as: $v' = \{v'_1, v'_2, \ldots, v'_n\}$, where $v'$ is the sanitization of $v$. The order of every generalized attributes are denoted as $a'_i$. 
Now, considering two datasets \((D, D')\), with their associated attributes \((a, a')\) and a mapping function \(f : D \rightarrow D'\) the sanitization is:

**Definition 3:** (Datasets Attribute Generalization): For \((D, D')\) their associated values and corresponding attributes \((v, v') \& (a, a')\), dataset generalization is a mapping function \((f : D \rightarrow D')\) that maps any dataset \(D, D'\) to a generalized form \(D^+\).

**Personalized Differential Privacy Scheme**

For our proposed PDP scheme, the concept of adjacent datasets and their associated source is very important. We assume two datasets are adjacent if the larger dataset is a proper subset of the smaller one where the larger set has exactly one more tuple in its set.

**Definition 4:** Two dataset \(D \& D' \subset D\) are referred to be adjacent when \(D \sim D'\), if \(D \subset D'\) and \(|D'| = |D| + 1\). Here, \(D \sim D'\) presents that \(D\) and \(D'\) are adjacent.

From the definition of adjacent datasets, we can infer the importance of using personalized protection of individual data sources before publishing the data to the neighboring nodes. Suppose privacy isn’t guaranteed before transferring the data. In that case, the system will be bombarded by data correlation, linking, or collusion attacks because malicious analysts will use the relation of adjacent datasets to identify individual sources private data.

Before describing our proposed scheme, we at first present the privacy specification of individual data providers. The privacy specification of our scheme is, each data provider \(S\) independently specifies their privacy requirements.

**Definition 5:** Individual data providers privacy preference can be mapped using: \(\Phi : S \rightarrow \mathbb{P}\), where stronger privacy preference is represented by a smaller \(\epsilon\) value. \(\Phi^s\) is used to present the preferred privacy requirement of corresponding end nodes \(s \in S\). Here, individuals privacy requirements are shown as, \(\Phi = \{(s_1, \epsilon_1), (s_2, \epsilon_2), ..., (s_n, \epsilon_i)\}\), where \(s_i \in S\) and \(\epsilon_i \in \mathbb{P}\).

Based on the adjacent datasets and individual data providers’ privacy specifications, our personalized differential privacy scheme is presented in definition: 6, derived from the definition: 3.

**Definition 6:** In a setting, a set of end nodes \(S\) and their privacy requirement is \(\Phi\), then the randomized mechanism \(M : D \rightarrow P\) meets \(\Phi\) personalized privacy protection for every adjacent datasets \(D, D' \subset D\), with \(D \sim D'\). The privacy-preserving output of end nodes is \(O \subseteq P\). So, \(\text{PDP}[M(D) \in O] \leq e^{\Phi^s} \times \text{PDP}[M(D') \in O]\)

where \(\Phi^s\) denotes \(S\) end nodes privacy preference. This helps our proposed end
nodes-fog-cloud framework attain strong privacy preservation of their associated data where privacy guarantees are assured based on each node’s personal privacy preference. The operational mechanism of the proposed PDP scheme follows Algorithm 4.

**Algorithm 4** Personalized Differential Privacy Procedure for our System

**Input:** Trust Distance $T^D$ ;  
**Output:** Aggregate individual nodes privacy-preserved Data $S\{\epsilon(D)\}$ based on their preferred privacy preference $\Phi$.

**End Nodes (S):**  
1: Central Cloud $C$ requests end nodes $S$’s data $D$;  
2: Map trust distance $T^D$ along with preferred privacy loss $\epsilon$ using Sigmoid function $S(\cdot)$;  
3: Privacy loss or noise are added using Laplace mechanism $\text{Lap}\left(\frac{\epsilon}{\epsilon}\right)$;  
6: Publish privacy-preserved data $S\{\epsilon(D)\}$ to neighbouring fog nodes $F$  

**Fog Nodes (F):**  
7: Aggregate end nodes $S$ noisy data.  
8: Fog nodes ($F$) publish its aggregated end nodes ($S$) data in the central cloud ($C$) by using the nodes preferred privacy $\Phi$ and create a proxy: $\hat{p}_{ij} = D + \text{Lap}\left(\frac{\epsilon}{\epsilon}\right)$;  
**IF** $\hat{p}_{ij}$ satisfy $\epsilon(T^D)$ differential privacy of its associated End nodes ($S$);  
**then** Cloud Level Aggregation: Aggregate the sum of the fog level noisy data in the central cloud as: $C = \sum_{j=1}^{F} \epsilon\{\hat{p}_{ij}\}$;  
**else** Repeat Stage $2 \rightarrow 8$;  

End

4.3.1 Personalized Differential Privacy Modelling

In this section, we modeled the nodes (source, destination) based on their privacy preference. We assume that a source node $S_n$ when transferring its private data $p_d \in P_D$, where $P_D$ presents the private data needs privacy protection before sending the data to the destination. The privacy protection mechanism starts with the source node $S_n$ when releasing the data to its adjacent fog nodes $F_n$; it at first calculates a approximate data $d'_s$ and release this $d'_s$ data instead of the sources original data $P_D$. This approximate data $d'_s$ needs to satisfy the source nodes privacy protection preference $\epsilon \in \left(\frac{1}{\Delta_{ij}}\right)$ where $\Delta_{ij}$ is the distance function $D : \Delta_{ij} : S_n \rightarrow D$ and $\epsilon : D^\kappa \rightarrow D^\kappa$ presents the mapping function that transforms the distance $\Delta_{ij}$ into $\epsilon \in \left(\frac{1}{\Delta_{ij}}\right)$ personalized differential privacy preference level.
4.3.2 Calculating Distance between Nodes of the Graph

In this layered grid architecture, attackers can compromise private data from any source/destination node. This means the more significant transmission distance between the source and destination node causes more nodes to be involved in each one-hop distance for receiving, storing, and sending the transmitted data. This causes more untrusted nodes to be involved, raising the risk of getting compromised by malicious adversarial attacks. However, to keep the privacy protection approach stable and efficient, we set a maximum threshold distance $TH_d$ beyond which we will provide a random privacy loss ($\varepsilon$). So, the nodes’ effective distance calculation in our weighted graph follows a least-cost approach using the “Dijkstra Algorithm” where the shortest path is calculated by finding the least cost from one node to every other connected node. So, the distance $\Delta_{ij}$ between the source and destination node is calculated as follows: $\Delta_{ij} = D(n) - S(n)$.

4.3.3 Laplace Distribution and $\varepsilon$-Personalized Privacy

The Laplace mechanism adds preferred noise to the private data from the Laplace distribution to fulfill each node’s personal differential privacy preference. The privacy loss is employed by using the Laplace mechanism $L$ to inject $\varepsilon$-personalized private noise $Lap(\frac{\Delta f}{\varepsilon})$ to the sensitive data $D$, that can be denoted as: $L = f(D) + Lap(\frac{\Delta f}{\varepsilon})$. This ensures $\varepsilon$-differential privacy to transform each nodes’ private data $P_D$ to a noisy data $d'_s$.

4.3.4 Mechanism of Achieving PDP

The formulation of PDP follows sequential composition mechanism which guarantees the composition of $\sum_i \varepsilon_i$. For example, there is a set of source nodes $S = \{S_1, S_2 \ldots S_n\}$, and their corresponding Laplacian mechanisms $L = \{L_1, L_2 \ldots L_m\}$, generates $\varepsilon_1, \varepsilon_2 \ldots \varepsilon_m$ privacy. For every $L_i$ needs ensuring $\varepsilon$-differential privacy which ultimately guarantees the composition of $\sum_i \varepsilon_i$-differential privacy. Here, $\sum_i \varepsilon_i$ refers upper privacy bound of sequential composition mechanism. So, when a dataset is released from a node it goes through randomized privacy loss ($\varepsilon$) based on individual nodes preference and the final privacy guarantee will be the summation of total privacy loss $\varepsilon$. 
Table 4.1: Criteria Mapping to Personalized Privacy Preservation

<table>
<thead>
<tr>
<th>Criterion/Nodes</th>
<th>Edge</th>
<th>Fog</th>
<th>Cloud</th>
</tr>
</thead>
<tbody>
<tr>
<td>Access to nodes</td>
<td>High</td>
<td>Medium</td>
<td>Low</td>
</tr>
<tr>
<td>Access frequency to Private Data</td>
<td>High</td>
<td>Medium</td>
<td>Low</td>
</tr>
<tr>
<td>Attackers background knowledge</td>
<td>Low</td>
<td>Medium</td>
<td>High</td>
</tr>
<tr>
<td>Communication medium</td>
<td>Low</td>
<td>Medium</td>
<td>High</td>
</tr>
<tr>
<td>Operational complexity</td>
<td>High</td>
<td>Medium</td>
<td>Low</td>
</tr>
</tbody>
</table>

4.3.5 Criteria Mapping to Personalized Privacy Preference

The traditional differential privacy follows a uniform privacy protection for all nodes; thus, it compromises the data utility. To overcome this limitation, we have adopted a privacy preference profile that is assigned to all of the nodes based on some criterion as shown in Table 4.1. The privacy preference of each node is mapped as high, medium, and low depending on the desired privacy and utility requirement. For example, it is comparatively easier to access the edge nodes than the fog and cloud nodes. Hence, privacy and security attacks (e.g., eavesdropping, membership inference, data poisoning, etc.) are more likely to occur in the edge nodes than the other nodes, which is why high data privacy is desired in the edge nodes. Here to find the appropriate $\varepsilon$-differential privacy, we use the following mapping function: $L(\varepsilon_i) = \frac{\exp(K_i x)}{\sum_{m=1}^{M} \exp(K_M)}$; where $i = 1$, $K$ presents a parameter, and $K_M$ is the mapping parameter.

4.3.6 Proof: Personalized Differential Privacy (PDP)

To guarantee personalized $\varepsilon$-privacy, we at first take a source node with two different privacy preference $\varepsilon_p$ and $\varepsilon_{p+1}$, where $\varepsilon_{p+1} > \varepsilon_p$ and similarly we take another node with two different privacy level $\varepsilon_q$ and $\varepsilon_{q+1}$, where $\varepsilon_{q+1} > \varepsilon_q$. The sequential composition mechanism of our system is built upon the upper bound which guarantees $\sum_i^n \varepsilon_i$ privacy. So, the PDP for the two nodes is as follows: $PDP(\varepsilon_p + \varepsilon_q) = PDP(\varepsilon_p + 1)$ or $PDP(\varepsilon_p + \varepsilon_q) = PDP(\varepsilon_q + 1)$. Upon observing this equation, we can see $\varepsilon_p < \varepsilon_{q+1}$ and $\varepsilon_q < \varepsilon_{p+1}$ which refers that the two nodes with noisy responses generate different level of privacy-preserved data.
4.4 Experimental Analysis

In this part, we present the suitability and applicability of using personalized differential privacy to ensure the CPS data’s privacy preservation while maintaining optimal data utility. The experimental analysis is conducted on a Core i5 CPU running at 2.7 GHz with 8 GB of RAM using Python 3.7.6 (64 bit). We use a real-world smart grid dataset [76] for our experiment where the data are collected from EPFL-Campus Medium-Voltage Grid.

For simplicity, we have extracted the noise from the Laplace distribution with zero means ($\theta = 0$) and variance $2\theta^2$. The reason for using Laplace distribution is that it has a fatter tail, which can provide better-randomized noise. The experimental evaluation is based on the following four criteria:

4.4.1 Utility Analysis

In the CPS infrastructure, the data is generated from the edge nodes and transmitted to the cloud node through the fog nodes. In edge nodes, the data utility is ‘1’ and data privacy is ‘0’ since edge nodes generate and preserve granular level data where no privacy is added to the data. Privacy is added only when edge nodes start sharing their data with the fog and cloud nodes. Any privacy-preserving mechanism (e.g., differential privacy) incurs utility loss over the original value, which can be measured through different methods (e.g., MSE, MAE, Max_Error, RMSE, MedAE, etc.). In our experiment, we interpret the utility loss of the fog and cloud nodes by Mean Absolute Error (MAE) which can be calculated using the following formula: $\text{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n}$ where, $\text{MAE} =$ mean absolute error; $y_i =$ privacy preserved value; $x_i =$ original value and $n =$ total number of data points. To measure the data utility loss on the scale of ‘0’ to ‘1’, we have calculated the percentage deviation (PD) using MAE and Mean (i.e., $PD = \text{MAE}/\text{Mean}$) and then subtracting the percentage deviation from ‘1’. Here, the data utility of ‘1’ reflects the highest possible utility of the nodes when no privacy preservation mechanism is applied and vice versa.

4.4.2 Privacy Analysis

Our privacy analysis is conducted by generating an interactive privacy preference profile based on some criteria as shown in Table 4.1 for every edge, fog, and cloud node. Based on the discussion on Subsection 4.3.5, edge nodes require high data
privacy as they are more frequently accessed than other nodes, and their resources are limited, which in turn restricts them to apply other privacy and security measures (e.g., encryption, firewall, etc.). However, if the attacker can compromise the fog or central cloud, he can exploit the aggregated data to conduct more devastating attacks. So, in this case, the fog and cloud nodes require better privacy than the edge node. Also, the number of physical communication mediums is lower in the fog and cloud nodes than in the edge node. So, they are more attack-prone than the edge nodes based on this communication medium criteria. The preference profile is assigned to all of the nodes based on some criterion as shown in Table 4.1. The privacy preference of each node is mapped as high, medium, and low depending on the desired privacy and utility requirement. For example, it is comparatively easier to access the edge nodes than the fog and cloud nodes. Hence, privacy and security attacks (e.g., eavesdropping, membership inference, data poisoning, etc.) are more likely to occur in the edge nodes than the other nodes, which is why high data privacy is desired in the edge nodes.

4.4.3 Risk Factor Analysis

In risk factor analysis, we consider the risk of disclosure and re-identification of individual data points. The risk factor is high when the data privacy level is low. More specifically, if the noise of the differential privacy mechanism is low, the attacker can easily re-identify individual data points by conducting multiple aggregation queries and averaging out the results. This correlation is also reflected in our empirical evaluation outlined in the discussion part of this thesis.

Simulation and Discussion

Here, we introduce the query model that our proposed system’s data analysts will use to analyze real-time and historical power data and identify its internal insights. For example, an analyst needs to learn the average power consumption pattern of consumers living in Reno whose usage falls in the range [80, 90] kWh per day to provide demand-based supply in that region. The central cloud of our proposed model (i) forwards this query to its associated Fog nodes (ii) Fog nodes further request all of its associated individual edge nodes to provide their hourly average consumption value (ii) Edge nodes receive this query and calculate their average hourly consumption using $\epsilon$ personalized differential privacy and send the data to the Fog node (iv) Fog
Figure 4.3: Evaluation of varying privacy loss ($\epsilon$) in grid data. (a) Hourly Consumption (KWh) of Meter 1 for different Privacy loss($\epsilon$) (b) Comparison between original vs. sanitized data with fixed Privacy loss($\epsilon$) (c) Comparison between original vs. differentially private aggregated query result.

We first use 6 meters sub-fraction (seconds) level granular data to simulate the actual consumption pattern of the meters for 24 hours. This pattern is further used to simulate our personalized privacy model, where a comparison between sensitive data blends with a large dataset as well as aggregates from individual selective data sources is shown.

**Analysis based on Varying Privacy Loss($\epsilon$)**

We take a single meter out of the 6 meters where we use different privacy loss ($\epsilon$) on the original consumption data of meter 1. Fig. 4.3(a) shows the hourly consumption of Meter 1 with and without the DP-mechanism. Using varying privacy loss, alternatively, the query result of Meter 1 becomes indistinguishable at different privacy levels. Fig. 4.3(a) also denotes with a small value of the privacy loss ($\epsilon$), the differentially private query result deviates further from the original query result. We can then infer when the value of the privacy loss is lower, the scale factor of the randomized noise distribution (in our case, the Laplace) becomes larger following this equation:

$$b = \frac{\Delta f}{\epsilon};$$

where $\Delta f$ denotes the data sensitivity & $\epsilon$ represents privacy loss.

Consequently, we can conclude when the tail of the PDF (Probability Density Function) of our Laplace distribution extract more noise and becomes fatter, it ensures better privacy in the system. Contrarily, suppose we continue increasing the value of the privacy loss ($\epsilon$). In that case, we can see that the differentially private query
result comes closer to the actual result. As a result, it becomes vulnerable to privacy-compromising data disclosure or linking attacks.

**Analysis based on Fixed Privacy Loss (\(\epsilon\))**

We also observe the impact of differential privacy mechanism for all the Meters with the same privacy loss (\(\epsilon = 0.1\)) in Fig. 4.3(b). This experiment is based on comparing the actual Vs. the sanitized data using fixed noise. The observation supports our experimental plot of Fig. 4.3(a) hourly consumption where meter 1 shows the highest consumption and the highest differentially private result among all the meters. The value of the differentially private result of meter 1 from Fig. 4.3(b) differs from Fig. 4.3(a). This happens because we have added randomized noise in each step of our experiment.

**Comparison between Original Vs. DP-based Aggregated Query Result**

In our final aggregated query result, we have observed the comparison between the original average vs. the differentially private average of the query result. We find that the differential privacy mechanism adds noise at a magnitude of 8 approximately in this scenario. If we run the same query again over the same dataset, the final aggregated average will remain same (around 16 as denoted by the ‘red’ colored curve in Fig. 4.3(c)). However, the differentially private average will vary each time due to the DP-mechanism’s \(\epsilon\)-indistinguishable property.

**Analysis on Personalized Privacy Protection**

After demonstrating the effectiveness and limitations of using uniform differential privacy, in Fig. 4.3, we represent the evaluation of using user-specific privacy protection

![Figure 4.4: Relation between Trust distance (\(T^D\)) and Privacy Loss (\(\epsilon\))](image.png)
Figure 4.5: Evaluation on personalized privacy protection where (a) to (f) shows as the trust distance decreases from highest to lowest the privacy preference sequentially decreases thereby system generates less noisy output and ensure better data utility. to model the effectiveness of using personalized privacy in an untrusted environment. This evaluation is based on using the exact hop distance as an index to map individual nodes’ personalized privacy protection level as shown in Fig. 4.4 whereas the distance increases from central control center privacy preference level increases.

Based on this, in Fig. 4.5, we demonstrate the personalized privacy protection of individual data sources in specific time slots. In Fig. 4.5 (a), PV inverters have the longest effective distance from the authority-owned central cloud server; thus, the privacy level is the highest, so that analysts get the largest noisy output when running queries on it. In Fig. 4.5 (b) to 4.5 (e), the effective trust distance of the nodes from the central control center continuously decreases; thus, the privacy protection level is also decreasing as nodes are situated in a more secured place. At last, in Fig. 4.5 (f), the end node has the shortest effective distance from the central authority, thereby requiring the lowest privacy protection and higher data utility. The average noises are 0.1, 0.25, 0.5, 0.75, 1.0 and 10, respectively. In addition, we can further infer that fog generates more accurate data than the adjacent end nodes in terms of privacy loss.
Impact on Data Utility

This analysis’s primary goal is to demonstrate that by taking personal privacy preferences into account, our proposed PDP mechanisms can often attain more accurate data utility compared to traditional differential privacy, which provides only a uniform privacy guarantee. To that end, we compare our proposed mechanism in Fig. 4.6, in terms of root mean squared error (RMSE) on real data, with varying privacy loss. From the analysis, we can see as the privacy loss increases, the root means squared error (RMSE) decreases, which means better utility assurance. However, due to randomized noise in the DP mechanism, sometimes negative bias can be added to the original data; it may cause the RMSE to be higher even with a higher privacy loss. For example, approximate privacy loss (0.70) shows a higher RMS value above the regression line. However, from the regression line, we can see the general trend follows the statement “higher privacy loss incurs lower RMS error and higher data utility”.

Comparative Analysis between Personalized and Uniform Privacy

To comprehensively understand the importance of applying PDP over UDP, we selected a set of two different privacy levels for both the fog and cloud aggregation, which is $\varepsilon = 0.6$ and $\varepsilon = 0.8$. Data aggregation is not required in the edge nodes; rather aggregation function is performed while transferring the data from the edge node to the fog and fog to the cloud. In UDP setting, two cases are demonstrated where both the fog and cloud aggregation adopts the same privacy levels at a time (i.e., case 1: $(\varepsilon_f, \varepsilon_c) = \{0.6, 0.6\}$; case 4: $(\varepsilon_f, \varepsilon_c) = \{0.8, 0.8\}$). However, in the PDP set up we considered the same privacy level but in an alternative manner (i.e., case 2:
Figure 4.7: Distribution of losses over uniform and personalized privacy

$(\varepsilon_f, \varepsilon_c) = \{0.6, 0.8\}$; case 3: $(\varepsilon_f, \varepsilon_c) = \{0.8, 0.6\})$. Fig. 4.7 (a) and (b) depict the loss distribution of these four cases for uniform and PDP accordingly, while Fig. 4.7 (c) presents the mean and standard deviation of these loss distributions. Generally, the higher standard deviation yields better privacy since the noise is distributed widely from the mean value. Also, the higher mean yields to higher loss (alternatively, lower data utility). So, for the case 1, where both the standard deviation and mean are high, privacy will be highest, but the data utility will be lowest. In contrast, for the case 4, the privacy will be the lowest, and data utility will be the highest. So, if we consider uniform privacy, it may offer excessive privacy control to a subset of nodes while applying insufficient protection to another subset.

However, if we consider the PDP setting, we can achieve optimized data privacy and security. For this, we need to consider cases 2 and 3. In the application, where data privacy is more desirable than the data utility case, 2 will be preferred over case 3 and vice versa. We have considered privacy loss 0.6 and 0.8 in our experiment, but different sets of privacy loss can also be opted. Nevertheless, the distribution follows the same trend as Fig. 4.7 for any set of privacy loss.

Comparative Analysis between Privacy Loss, Data Utility and Data Disclosure Risk

We outlined the data disclosure or re-identification risk for the same four cases under uniform and PDP settings as shown in Fig. 4.8. The correlation among the privacy loss, data utility, and data disclosure risk shows that risk and utility decrease along with the increment of privacy loss ($\varepsilon$) in the UDP settings. For instance, in case of house id 114 if we increase the privacy loss from 0.6 to 0.8, the risk decreases approximately from 0.81 (Fig. 4.8(a)) to 0.79 (Fig. 4.8(d)) for both the fog and cloud...
Figure 4.8: Comparative analysis between privacy loss and disclosure risk of private information for uniform and personalized privacy aggregation. However, if we consider the PDP for the same house, the minimum attainable risks can be 0.61 (for cloud aggregation, Fig. 4.8(b)) and 0.79 (for fog aggregation, Fig. 4.8(c))). If the CPS analysts or customers pull the data from the cloud node, then the case 2 (i.e., $\varepsilon_f, \varepsilon_c = \{0.6, 0.8\}$, Fig. 4.8(b)) is more preferable over the rest three cases since the cloud risk factor is the lowest for this case.

Moreover, a little increment in the privacy loss ($\varepsilon$) of fog aggregation ($\varepsilon_f$) yields to larger risk increment in the cloud aggregation. For example, for the same house id (i.e., 114), increasing $\varepsilon_f$ from 0.6 to 0.8 increases the cloud aggregation risk from 0.81 to 1.01 (Fig. 4.8(a) and Fig. 4.8(c)). Although the UDP decreases the disclosure risk factor by increasing the fog and cloud privacy losses ($\varepsilon_f, \varepsilon_c$), it does not consider the data utility in the process. In contrast, a PDP mechanism tunes both the fog and cloud privacy loss levels to the optimized values ($\varepsilon_f^*, \varepsilon_c^*$) to attain optimal data privacy and data utility with optimal disclosure risk.

4.5 Summary

This chapter proposes a privacy scheme, “Personalized Differential Privacy (PDP)” by characterizing the privacy issues during data sharing, aggregation, and publishing by CPS end nodes. This proposed PDP scheme with the added benefit of a user-specific privacy guarantee helps us eliminate the differential privacy model’s weakness by protecting the privacy of individual node’s sensitive data non-uniformly. Lastly, we have validated our proposed privacy scheme using data-driven experiments. The experimental analysis has shown that our proposed model can optimize privacy protection by masking original data and providing resiliency against privacy-compromising attacks.
Chapter 5

Conclusion

Cybersecurity has become a significant issue in modern cyber-physical systems because of the growing number of monitoring applications and controlling devices. Attackers can manipulate sensitive CPS data or compromise CPS devices that can lead to significant economic loss and reliability as well as security issues for CPS service providers and their customers. In this thesis work, we provide an overview of CPS network cybersecurity challenges and introduce a decentralized blockchain-based technology as a protection framework to tackle the vulnerabilities in CPS data transmission. The proposed blockchain-based framework with a bloom filter and Merkle tree substantially improves the self-defensive capabilities of the grid devices at the edge against data manipulation by cyber attackers. We also incorporate comprehensive vulnerability characterization and privacy quantification as well as protection model in this thesis work to deal against the security and privacy-related issues in CPS during data generation, transmission, and collection. The proposed SVPL model in work helps identify the CPS system’s security and privacy vulnerabilities accurately by building a standard vulnerability profile considering each data provider’s privacy loss scenario. Furthermore, a customized PDP model is also introduced to improve the data utility by reducing data disclosure risk while assuring high-level data privacy protection for the CPS. Finally, our experimental analysis compares proposed personalized differential privacy (PDP) with uniform differential privacy. It proves that this PDP model ensures better data privacy protection while maintaining the data utility and minimizing the risk of data disclosure.
5.1 Open Issues, Challenges and Future Research Directions

Blockchain is an emerging technology and is currently adopted in a wide area of the research domain. But, like any other emerging technology, blockchain also has its limitations. So, it is essential to address the potential limitations of blockchain before integrating it into cyber-physical systems. Moreover, due to the dynamic characteristics of CPS, the implementation of personalized differential privacy in the CPS physical network faces many challenges. This section will discuss the challenges, open issues, and future research directions to achieve personalized data privacy in CPS.

5.1.1 Suitability Issues

Blockchain lacks a data storing mechanism based on which current Synchrophasor monitoring largely depends. Blockchain only stores validated transactions into the ledger, which raises the question of whether the smart grid authority wants better security or insecure data storing.

5.1.2 Scalability, Latency and Throughput Issues

Blockchain’s integration into the smart grid raises the scalability issue because the power management system deals with a large volume of measurement transactions every second. But, currently, blockchain applications have a low transaction speed which causes low throughput and latency because they are two key scalability metrics. In our future work, we will look into the subdivision of the whole network into small networks (such as region-based [24]) and how such subdivisions will help us to achieve scalability, reduced latency, and improved throughput.

5.1.3 Consideration of Multiple Layers in Blockchain Design

We perceive a single layer for our proposed blockchain model, but the synchrophasor network can be considered a layered network. So, in the future, we will explore the viability of the layered blockchain [77] design and identify the best approach for inter-blockchain communications.
5.1.4 Big Data

Big data vision is not limited to CPS; it covers almost all aspects of human life. In this thesis, we have previously discussed personalized differential privacy in some big data scenarios. With increased data, the algorithm based on personalized differential privacy has become more advanced and more responsive. However, still has some challenges for big data that need to be solved.

Intuitive Privacy Definition

In personalized differential privacy itself, one of the biggest challenges is defining precise privacy. Even after establishing mathematical tests and strict privacy models, personalized differential privacy and DP lack an intuitive definition of privacy based on big data. Therefore, finding a more intuitive explanation of privacy based on extensive data analysis is still challenging for data scientists [78].

Composition Theorem

The composition theorem plays an active role in algorithm design and privacy budget allocation. However, the existing method of using the composition theorem to determine the privacy budget is not optimal [79]. Therefore, optimizing the composition of differential privacy in big data analysis is still an unresolved challenge. Similarly, in big data, maintaining privacy protection and the dimensionality problem with large data volume and high computational overhead is a big challenge for researchers.

5.1.5 Machine Learning

The actual objective of any machine learning algorithm is to identify helpful information from a given dataset. However, maintaining individual data privacy while using ML algorithms will be one of the most challenging tasks in the future. To address this topic, researchers have launched differential privacy-based data preservation mechanisms with machine Learning algorithms [80]. In the future, extensive research on personalized data preservation with complex machine learning algorithms needs to be conducted.
5.1.6 Applicability of PDP in Adversarial Settings

To identify the applicability of the personalized differential privacy model in CPS against adversarial attacks, we need first to identify the factors and constraints affecting the PDP model’s data privacy, security, and utility. Continuing this research line, we need to analyze the impact of different adversarial attacks on the proposed PDP-based model to assess the privacy-preservation model’s suitability and applicability to fight against malicious attacks.
Bibliography


[17] A. Bhattacharjee, S. Badsha, and S. Sengupta, “Blockchain-based secure and reliable manufacturing system,” in *2020 International Conferences on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData) and IEEE Congress on Cybermatics (Cybermatics)*, 2020, pp. 228–233.


[60] A. Bhattacharjee, S. Badsha, and S. Sengupta, “Blockchain-based secure and reliable manufacturing system,” in *2020 International Conferences on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData) and IEEE Congress on Cybermatics (Cybermatics)*. IEEE, 2020, pp. 228–233.


