Synthetic Network Generation and Vulnerability Assessment of Cyber-Physical Power Systems

Liu, Y.

DOI
10.4233/uuid:d096dc20-f758-4ed2-9890-dd8d7ba2625b

Publication date
2025

Document Version
Final published version

Citation (APA)
Liu, Y. (2025). *Synthetic Network Generation and Vulnerability Assessment of Cyber-Physical Power Systems*. [Dissertation (TU Delft), Intelligent Electrical Power Grids]. https://doi.org/10.4233/uuid:d096dc20-f758-4ed2-9890-dd8d7ba2625b

Synthetic Network Generation and Vulnerability Assessment of Cyber-Physical Power Systems

Synthetic Network Generation and Vulnerability Assessment of Cyber-Physical Power Systems

Dissertation

for the purpose of obtaining the degree of doctor

at Delft University of Technology

by the authority of the Rector Magnificus, prof. dr. ir. T.H.J.J. van der Hagen,

chair of the Board for Doctorates

to be defended publicly on

Wednesday 5 February 2025 at 17:30 o'clock

by

Yigu LIU

This dissertation has been approved by the promotors.

Promotor: Prof. dr. P. Palensky
Copromotor: Dr. A.I. Ştefanov

Composition of the doctoral committee:

Rector Magnificus,	chairperson
Prof. dr. P. Palensky,	Delft University of Technology, *promotor*
Dr. A.I. Ştefanov,	Delft University of Technology, *copromotor*
Independent members:	
Prof. dr. C.C. Liu,	Virginia Tech
Prof. dr. ir. P.H.A.J.M. van Gelder,	Delft University of Technology
Dr. S. Lakshminarayana,	University of Warwick
Dr. Z. Erkin,	Delft University of Technology
Dr. J. Hong,	University of Michigan-Dearborn
Prof. dr. ir. A.H.M. Smets,	Delft University of Technology, reserve member

This research was financially supported by the China Scholarship Council.

Keywords: Cyber-Phyiscal Power Systems, Synthetic Networks, Vulnerability Assessment, Smart Grids, Machine Learning

Cover by: Y. Liu

E-mail: liuyigu_a@126.com

ISBN (Paperback/softback): 978-94-6384-719-3

ISBN (E-book): 978-94-6518-003-8

An electronic copy of this dissertation is available at
https://repository.tudelft.nl/.

Dedicated to those precious moments

Yigu Liu

Contents

Summary

Power system operation is increasingly reliant on Information and Communication Technologies (ICTs), which are essential for enhancing the resilience, reliability, and security of the future electricity supply. The advancement of ICTs has tightly integrated power grids with communication networks, giving rise to Cyber-Physical power Systems (CPS). However, this growing digitalization also increases system complexity, heightens vulnerability to cyber attacks, and alters traditional operational patterns. Consequently, this trend underscores the critical need for continued exploration and innovation in CPS to address emerging challenges. In this context, the availability of reliable test cyber-physical systems is crucial. The test CPS models must enable realistic analyses without exposing sensitive information about critical infrastructures, allowing researchers to thoroughly investigate newly introduced vulnerabilities and ensure the reliability and cyber security of CPS. To this end, we are motivated to have the following research focus: the synthetic network generation and vulnerability assessment of cyber-physical power systems.

Synthetic Network Generation: The rising demand for advanced research on cyber-physical power systems compels the creation of realistic and reliable test systems. However, the national security concerns prevent the public sharing of real Critical Infrastructure (CI) data, including the real CPS models. As a solution, synthetic networks aim at generating realistic projections of real-world networks while concealing the actual, sensitive system models and data, e.g., CPS topologies, characteristics, and operational parameters, and maintaining similar overall characteristics as the real cyber-physical systems.

CPS Vulnerability Assessment: Due to the rapid integration of cyber and physical infrastructures, modern power systems are becoming more efficient while also exhibiting increased vulnerabilities. This emerging risk was starkly demonstrated by the three major cyber attacks on the Ukrainian power grid in 2015, 2016, and 2022, underscoring the critical need for enhanced security measures in this landscape. The evolving communication infrastructures have significantly altered the propagation mechanisms of cascading failures in CPS. These changes present novel challenges in ensuring a safe system operation. Consequently, it is imperative to thoroughly investigate the new cascading mechanisms and pinpoint the critical components of CPS, which will enable the implementation of timely mitigation strategies, thereby enhancing the overall security and resilience of CPS.

Based on the discussion above, this thesis proposes novel methods for synthetic network generation and vulnerability assessment of cyber-physical power systems using complex network theory, graph neural networks and data mining techniques. The generated synthetic CPS models and datasets preserve the complex network features and statistical properties of the real CPS. The proposed CPS vulnerability assessment method identifies the critical components to increase CPS resiliency. It is conducted on both real and synthetic CPS models. The results are statistically similar. The generated synthetic CPS models and datasets can be freely shared with the research community without disclosing the confidential CPS, while adversaries cannot reverse engineer it. The major contributions of this thesis are elaborated as follows.

From the perspective of synthetic networks, this thesis proposes three different generators to accurately generate synthetic CPS by utilizing real CPS models and data with varying levels of availability and completeness. First, a two-stage generative model is proposed to generate synthetic communication topologies of large-scale CPS based on the existing power grids. It reproduces the existing communication network design process and is capable of generating statistically realistic networks. The proposed method is implemented to create a realistic, large-scale synthetic CPS for the interconnected power grids in continental Europe. Then, to generate synthetic networks for both cyber and physical system layers, a scalable generative model, namely Graph-CPS, is proposed to generate a synthetic CPS topology. This method is capable of reflecting realistic network feature distributions while ensuring the confidentiality of the real CPS models and data. Graph-CPS can learn and reproduce various complex network parameters, not only across different network types but also across varying network sizes. This method paves the way for a deeper understanding of CPS characteristics, offering valuable insights into CPS structures and network parameter configurations. In the end, a hybrid generator, namely SibGen, is proposed to generate the digital sibling of the real CPS. The core idea behind digital siblings lies in balancing the fidelity of synthetic models with the need for data confidentiality, ensuring that overall system behaviours are accurately captured without compromising sensitive information. Moreover, SibGen not only learns the topological features of CPS but also effectively captures the operational characteristics. This dual capability allows research to be conducted on the generated digital sibling to closely mirror real-world scenarios, making the research findings more practical and convincing.

From the perspective of CPS vulnerability assessment, this thesis systematically evaluates and identifies the vulnerabilities of CPS considering time-varying operational states, with an emphasis on the correlation between CPS components. In this thesis, two types of correlations—manifest and latent—are defined to better reveal the cascading mechanism in CPS. These correlations are used to investigate both apparent and potential cascading relationships between CPS components. By jointly analysing manifest and latent correlations, this thesis introduces a critical components identification model, i.e., GraphCCI. This model effectively captures the cascading failure characteristics across various operational states and generates a weighted cascading graph database for graph data mining. Once frequent cascading sub-graphs are identified, the proposed Node Criticality Index (NC-Index) is used to accurately pinpoint critical CPS components, enhancing the overall system's security and resilience.

The ultimate goal of this thesis is to advance the development of CPS while safeguarding its security. This thesis offers realistic and reliable synthetic networks for research while ensuring the confidentiality of real system models and data. Additionally, the proposed vul-

nerability assessment methods effectively reveal cyber-physical cascading mechanisms and accurately identify critical CPS components. These findings provide valuable insights for the continued development of CPS and help ensure the safety of system operations.

SAMENVATTING

De werking van energiesystemen is steeds meer afhankelijk van Informatie- en Communicatietechnologieën (ICT), die essentieel zijn voor het verbeteren van de veerkracht, betrouwbaarheid en veiligheid van de toekomstige elektriciteitsvoorziening. De vooruitgang van ICT heeft stroomnetwerken nauw geïntegreerd met communicatienetwerken, wat heeft geleid tot Cyber-Physical Power Systems (CPS). Echter, deze toenemende digitalisering vergroot ook de systeemcomplexiteit, verhoogt de kwetsbaarheid voor cyberaanvallen, en verandert traditionele operationele patronen. Hierdoor wordt de dringende noodzaak benadrukt voor voortdurende verkenning en innovatie binnen CPS om nieuwe uitdagingen het hoofd te bieden. In dit kader is de beschikbaarheid van betrouwbare test cyber-fysische systemen cruciaal. De test-CPS-modellen moeten realistische analyses mogelijk maken zonder gevoelige informatie over kritieke infrastructuren bloot te geven, zodat onderzoekers nieuw geïntroduceerde kwetsbaarheden grondig kunnen onderzoeken en de betrouwbaarheid en cyberveiligheid van CPS kunnen waarborgen. Om deze reden zijn wij gemotiveerd om ons te richten op de volgende onderzoeksthema's: de synthetische netwerkopwekking en kwetsbaarheidsanalyse van cyber-fysische energiesystemen.

Synthetische Netwerkopwekking: De toenemende vraag naar geavanceerd onderzoek naar cyber-fysische energiesystemen vereist de creatie van realistische en betrouwbare testsystemen. Nationale veiligheidskwesties voorkomen echter dat echte gegevens van kritieke infrastructuren (CI), inclusief de echte CPS-modellen, openbaar worden gedeeld. Als oplossing richten synthetische netwerken zich op het creëren van realistische projecties van netwerken in de echte wereld, terwijl ze de werkelijke, gevoelige systeemmodellen en gegevens, zoals CPS-topologieën, kenmerken en operationele parameters, verhullen en vergelijkbare algemene kenmerken behouden als de echte cyber-fysische systemen.

CPS Kwetsbaarheidsanalyse: Door de snelle integratie van cyber- en fysieke infrastructuren worden moderne energiesystemen efficiënter maar ook kwetsbaarder. Dit opkomende risico werd duidelijk aangetoond door de drie grote cyberaanvallen op het Oekraïense elektriciteitsnet in 2015, 2016 en 2022, wat de kritische noodzaak van verbeterde veiligheidsmaatregelen in deze omgeving benadrukte. De evoluerende communicatie-infrastructuren hebben de verspreidingsmechanismen van cascade-uitval in CPS aanzienlijk veranderd. Deze veranderingen vormen nieuwe uitdagingen voor een veilige systeemwerking. Het is daarom noodzakelijk om de nieuwe cascade-mechanismen grondig te onderzoeken en de kritieke componenten van CPS te identificeren, waardoor tijdige mitigeringsstrategieën kunnen worden geïmplementeerd en de algehele beveiliging en veerkracht van CPS kunnen worden verbeterd.

Gebaseerd op bovenstaande discussie, stelt deze scriptie nieuwe methoden voor voor synthetische netwerkopwekking en kwetsbaarheidsanalyse van cyber-fysische energiesystemen met behulp van complexe netwerktheorie, grafneuronale netwerken en data miningtechnieken. De gegenereerde synthetische CPS-modellen en datasets behouden de kenmerken van complexe netwerken en statistische eigenschappen van de echte CPS. De voorgestelde CPS-kwetsbaarheidsanalyse-methode identificeert de kritieke componenten om de veerkracht van CPS te vergroten. Deze analyse wordt uitgevoerd op zowel echte als synthetische CPS-modellen. De resultaten zijn statistisch vergelijkbaar. De gegenereerde synthetische CPS-modellen en datasets kunnen vrij worden gedeeld met de onderzoeksgemeenschap zonder de vertrouwelijkheid van de CPS prijs te geven, en vijanden kunnen deze niet omgekeerd-engineeren. De belangrijkste bijdragen van deze scriptie worden hieronder nader toegelicht.

Vanuit het perspectief van synthetische netwerken stelt deze scriptie drie verschillende generatoren voor om nauwkeurig synthetische CPS te genereren door gebruik te maken van echte CPS-modellen en -gegevens met verschillende niveaus van beschikbaarheid en volledigheid. Ten eerste wordt een tweefasen-generatiemodel voorgesteld om synthetische communicatietopologieën van grootschalige CPS te genereren op basis van bestaande stroomnetwerken. Het reproduceert het bestaande ontwerpproces van communicatienetwerken en kan statistisch realistische netwerken genereren. De voorgestelde methode wordt toegepast om een realistisch, grootschalig synthetisch CPS te creëren voor de onderling verbonden stroomnetwerken in continentaal Europa. Vervolgens wordt een schaalbaar generatief model, Graph-CPS, voorgesteld om synthetische netwerken voor zowel de cyber- als fysische systeemlagen te genereren. Deze methode is in staat om realistische netwerkfunctiedistributies te weerspiegelen terwijl de vertrouwelijkheid van de echte CPS-modellen en -gegevens wordt gewaarborgd. Graph-CPS kan verschillende parameters van complexe netwerken leren en reproduceren, niet alleen over verschillende typen netwerken, maar ook over verschillende netwerkafmetingen. Deze methode biedt een dieper inzicht in CPS-kenmerken en netwerkparameterconfiguraties. Tot slot wordt een hybride generator, SibGen, voorgesteld om de digitale tegenhanger van de echte CPS te genereren. Het kernidee achter digitale tegenhangers is het evenwicht tussen de betrouwbaarheid van synthetische modellen en de noodzaak van gegevensvertrouwelijkheid, zodat het algehele systeemgedrag nauwkeurig wordt vastgelegd zonder gevoelige informatie prijs te geven. Bovendien leert SibGen niet alleen de topologische kenmerken van CPS, maar vangt het ook effectief de operationele kenmerken op. Deze dubbele capaciteit maakt het mogelijk om onderzoek uit te voeren op de gegenereerde digitale tegenhanger, zodat deze dicht bij realistische scenario's komt en de onderzoeksresultaten praktischer en overtuigender maakt.

Vanuit het perspectief van CPS-kwetsbaarheidsanalyse evalueert deze scriptie systematisch en identificeert zij de kwetsbaarheden van CPS, rekening houdend met tijdsafhankelijke operationele toestanden, met nadruk op de correlatie tussen CPS-componenten. In deze scriptie worden twee soorten correlaties—manifeste en latente—gedefinieerd om het cascade-mechanisme in CPS beter te onthullen. Deze correlaties worden gebruikt om zowel de duidelijke als de potentiële cascaderelaties tussen CPS-componenten te onderzoeken. Door manifeste en latente correlaties gezamenlijk te analyseren, introduceert deze scriptie een model voor kritieke componentenidentificatie, GraphCCI. Dit model legt effectief de kenmerken van cascade-uitval vast over verschillende operationele toestanden en genereert een gewogen cascade-grafiekdatabase voor grafiekdata mining. Zodra frequente cascade-subgrafieken zijn geïdentificeerd, wordt de voorgestelde Node Criticality Index (NC-Index) gebruikt om nauwkeurig kritieke CPS-componenten te identificeren, waardoor de algehele systeembeveiliging en veerkracht worden verbeterd.

Het uiteindelijke doel van deze scriptie is het bevorderen van de ontwikkeling van CPS met behoud van de veiligheid. Deze scriptie biedt realistische en betrouwbare synthetische netwerken voor onderzoek en waarborgt de vertrouwelijkheid van echte systeemmodellen en -gegevens. Bovendien onthullen de voorgestelde kwetsbaarheidsanalyse-methoden effectief cyber-fysische cascade-mechanismen en identificeren ze nauwkeurig kritieke CPS-componenten. Deze bevindingen bieden waardevolle inzichten voor de voortdurende ontwikkeling van CPS en helpen de veiligheid van systeemwerking te waarborgen.

List of Publications

Journals published:

Yigu Liu, Haiwei Xie, Alfan Presekal, Alexandru Ştefanov, Peter Palensky. "A GNN-Based Generative Model for Generating Synthetic Cyber-Physical Power System". *IEEE Transactions on Smart Grid*, vol. 14, no. 6, pp. 4968-4971, 2023.

Yigu Liu, Alexandru Ştefanov, Peter Palensky. "Generating Large-Scale Synthetic Communication Topologies for Cyber-Physical Power Systems". *IEEE Transactions on Industrial Informatics*. Early Access, DOI: 10.1109/TII.2024.3438232

Yigu Liu, Alexandru Ştefanov, Ioannis Semertzis, Peter Palensky. "GraphCCI: Critical Components Identification for Enhancing Security of Cyber-Physical Power Systems". *IEEE Transactions on Industrial Cyber-Physical Systems*, vol. 2, pp. 340-349, 2024.

Journals under review:

Yigu Liu, Alexandru Ştefanov, Alfan Presekal, Peter Palensky. "SibGen: A Hybrid Generator for Generating the Digital Sibling of Cyber-Physical Power System," *IEEE Transactions on Smart Grid*. (submitted, first round of review)

Conference Paper:

Yigu Liu, Ioannis Semertzis, Alexandru Ştefanov, Peter Palensky. "Critical components identification for cyber-physical power systems considering time-varying operational states". *9th Workshop on Modeling and Simulation of Cyber-Physical Energy Systems, MSCPES 2021*, Held as part of the Cyber-Physical Systems and Internet-of-Things Week, Proceedings, Nashville, TN, USA, 2021.

1

Introduction

1.1. Background and Motivation

The power system operation is increasingly dependent on Information and Communication Technologies (ICTs), which ensure the resilience, reliability, and security in electricity supply of the future power grid. It can be envisioned that on top of the power system infrastructure reside integrated layers of ICTs, which form an interdependent and complex Cyber-Physical power System (CPS). Such drastic transformation brings evolutionary changes and challenges to modern power systems. From an adversarial perspective, such integration introduces new vulnerabilities to the CPS, which can be exploited to conduct cyber-physical attacks. From the standpoint of power system operation, the integration of cyber and physical systems changes the operational mechanisms of the CPS. It not only alters the inherent topological structure of the CPS but also changes the patterns of how cascading failures propagate within the system. Consequently, the scale and depth of fault propagation are likely to be greatly aggravated. This situation has already manifested in the real world. CPS disruptions can cause equipment damage, financial loss and even a loss of lives.

Numerous cyber-physical events [1], [2], [3], [4], [5], [6], [7], [8], [9] have been reported in Critical Infrastructures (CIs). We summarize the major cyber-physical events with large social impact since 2000 as in Figure 1.1. One representative event highlighting the impact of power grid and communication network coupling is the blackout in Italy on September 28, 2003 [1], [2]. It started with the shutdown of a substation, which led to the failure of nodes in the communication network. Cascading failures were triggered in both the cyber and physical system layers of the Italian power grid, resulting in over €120 million in financial losses and affecting around 56 million people.

From the perspective of cyber security, in 2015 and 2016, the Ukraine power grid suffered two serious cyber attacks which led to large-scale power outages [6], [7]. In 2015, hackers intruded into the ICT systems of three distribution system operators. Seven 110 kV and twenty-three 35 kV substations were disconnected from the power grid for hours. The cyber attacks in Ukraine are the first publicly acknowledged incidents to result in power outages that affected 225,000 customers. The hackers shut down power by using phishing emails, BlackEnergy3, virtual private network and credential theft, network and host discovery, and Operational Technology (OT) hijack. Attackers opened circuit breakers to cause power outages and used KillDisk to damage the OT system, i.e., Supervisory Control and Data Acquisition (SCADA) system. Subsequently, in 2016, the attacks were focused on the SCADA system at transmission level targeting a single 330 kV substation. This led to a power outage in the distribution system where 200 MW of load was unsupplied. The real-world events highlight the urgent need to strengthen the systematic security of CPS. With evolving communication infrastructures, the operational mechanisms of CPS have changed, introducing new challenges to ensuring safe system functionality. As a result, it is critical to analyze these new characteristics and pinpoint critical CPS components to deepen our understanding to CPS and thereby enhancing system security. To this end, effective vulnerability assessment methods are needed to accurately identify newly emerging cyber-physical system vulnerabilities. It is essential

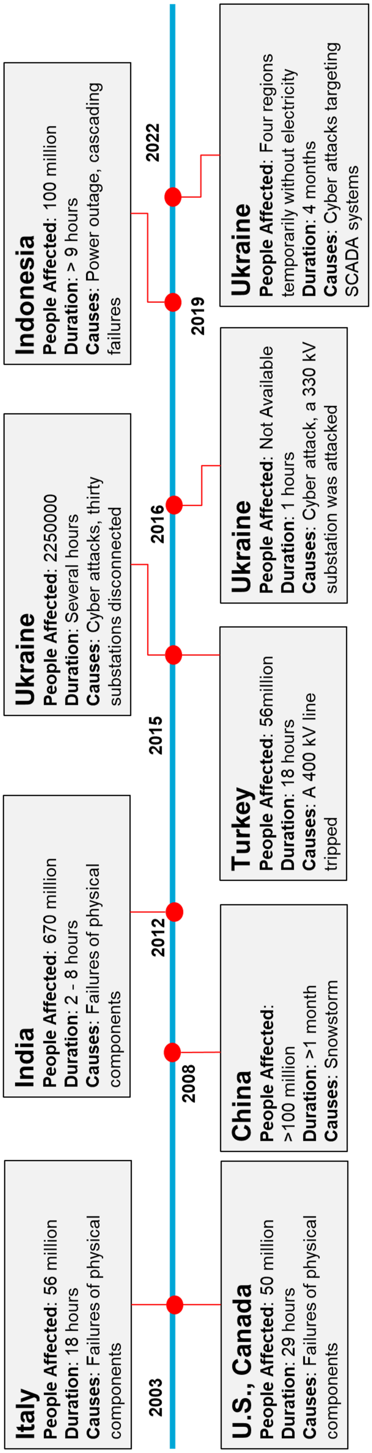

2003
Italy
People Affected: 56 million
Duration: 18 hours
Causes: Failures of physical components
U.S., Canada
People Affected: 50 million
Duration: 29 hours
Causes: Failures of physical components
2008
China
People Affected: >100 million
Duration: >1 month
Causes: Snowstorm
2012
India
People Affected: 670 million
Duration: 2 - 8 hours
Causes: Failures of physical components
Turkey
People Affected: 56million
Duration: 18 hours
Causes: A 400 kV line tripped
2015
Ukraine
People Affected: 2250000
Duration: Several hours
Causes: Cyber attacks, thirty substations disconnected
2016
Ukraine
People Affected: Not Available
Duration: 1 hours
Causes: Cyber attack, a 330 kV substation was attacked
2019
Indonesia
People Affected: 100 million
Duration: > 9 hours
Causes: Power outage, cascading failures
2022
Ukraine
People Affected: Four regions temporarily without electricity
Duration: 4 months
Causes: Cyber attacks targeting SCADA systems

Figure 1.1 The History of Cyber-Physical Events

to thoroughly analyze the operational behavior of CPS and evaluate system loopholes to achieve efficient pre-attack defense and post-attack mitigation.

Conducting research on cyber-physical systems requires realistic models and data. However, real CPS models and data are highly confidential due to national security concerns. If real CPS data were made publicly available, adversaries could initiate accurate and effective attacks, leading to catastrophic consequences. The current literature on developing synthetic networks is mainly focused on the physical power system [10], [11], [12], [13], [14], [15], while how to generate a realistic synthetic cyber-physical system is rarely investigated. Therefore, the effective synthetic CPS generative models are desirable. First, a qualified generative model should conceals all sensitive input information—such as system topology and parameter configurations—while maintaining similar topological characteristics and operational features as the real CPS. Moreover, the generated synthetic CPS should be able to providing realistic feedback to researchers while developing CPS-related methods. That is, the synthetic CPS should exhibit similar performances and behaviors as the real systems when these methods are implemented.

Motivated by the circumstances described above, the research objective of this thesis is as follows.

Generate synthetic CPS with realistic topology and operation models, and systematically evaluate CPS vulnerabilities under time-varying operational states.

The generated synthetic CPS models and datasets retain the complex network characteristics and statistical properties of real CPS. The proposed vulnerability assessment method effectively identifies critical components to enhance CPS resilience. This assessment is performed on both real and synthetic CPS models, yielding statistically similar results. The generated synthetic CPS models and datasets can be freely shared with the research community without disclosing the confidential CPS, while adversaries cannot reverse-engineer it. The key philosophy behind this dissertation lies in filling the gaps in the availability, completness and confidentiality of CPS models and data while enabling a comprehensive analysis of CPS behaviour, ultimately fostering deeper insights into CPS characteristics and mechanisms.

1.2. CHALLENGES AND RESEARCH QUESTIONS

In this section, the challenges and research questions are elaborated as follows. From the perspective of synthetic CPS, there are three major challenges:

(1) Model and data availability. The availability of real CPS data is the first challenge, which is not only the basis of synthetic network generation, but also the key to validating the generated synthetic networks.

(2) Model and data confidentiality. Closely related to the first challenge, ensuring the security of real CPS information is also demanding. It is essential to guarantee that adversaries cannot reverse-engineer the synthetic networks to obtain the real CPS data.

(3) Synthetic network validation. It is challenging to validate the generation quality of the synthetic network. On one hand, the network data of a CPS is high-dimensional and the difference quantification of two high-dimensional datasets is difficult. On the other hand, the network validation is not only about validating the generated parameters that can be directly observed, e.g., network topology, system parameters, but also about validating the overall system performance, e.g., communication process, power flow results. Note that minor parameter changes, even if they pass parameter validations, can result in completely different overall system performance.

From the perspective of vulnerability assessment, the challenges are two-fold:

(1) Varied operational states. The current vulnerability assessment methods focus on evaluating the CPS vulnerability at a single time instant, which may lead to a biased assessment result when the operational state is changing. However, evaluating systematic vulnerabilities under various operational states can be computationally expensive.

(2) System interaction mechanisms. The integration of cyber and physical system layers introduces intricate interactions between components, making it difficult to analyze and quantify their correlations. This is particularly challenging as understanding these correlations is critical for identifying the most vulnerable components in CPS.

Based on the discussion and analysis above, the research questions that are answered in this thesis are elaborated as follows:

Q1: How can we accurately generate synthetic CPS topologies by utilizing CPS models and data with varying levels of completeness while ensuring data confidentiality? Furthermore, how can we effectively and thoroughly validate these generated topologies?

Due to national security concerns, detailed CPS information cannot be publicly disclosed, i.e., power grid models, communication network architectures, and data. Usually, researchers need Non-Disclosure Agreements (NDAs) to access the CPS models and data of system operators for specific research purposes. As a result, the quality and completeness of available CPS data can vary significantly. Therefore, it is crucial to develop methodologies that can handle different levels of data completeness. For instance, if communication network data is missing, it should be possible to generate realistic synthetic communication networks based on existing power grid data. Besides, ensuring the data security of real CPS information is also a significant challenge. It is crucial to guarantee that adversaries cannot decipher the synthetic networks and reverse-engineer them to extract the underlying real CPS data. From the perspective of result validation, given that **Q1** mainly focuses on the topological aspects of CPS, the complex network theory parameters are suitable options to thoroughly describe the network characteristics, e.g., node degree distribution, betweenness distribution, average shortest path length, etc. The current topology generation methods are capable of generating networks with specified characteristics, e.g., scale-free property, small-world property, etc. However, they often fail to produce networks that exhibit consistent characteristics across multiple parameters. For example, the Chung-Lu algorithm [16] generates a scale-free network with a given average node degree, while it does not apply any constraints to other parameter distributions such as the betweenness distribution. When evaluating a synthetic network topology, it is essential to ensure that the generated network closely mirrors the real one across a comprehensive set of characteristics, which is also the core problem we need to solve in **Q1**.

Q2: How can we effectively generate digital siblings with fully integrated operational models for both cyber and physical system layers? How can we accurately capture and learn the intricate characteristics of real system patterns, enabling synthetic networks to behave similarly to real systems?

To answer **Q2**, it is essential to understand the differences between digital siblings and digital twins. Digital twins focus on absolute consistency with real networks, while digital siblings aim to capture only the global characteristics of the target system. Metaphorically, digital siblings share the same ”DNA” as digital twins but exhibit different traits. This distinction is essential for safeguarding the authenticity of Critical Infrastructure (CI) data. As shown in Figure 1.2, a qualified synthetic network should serve as an alternative test system and this synthetic system exhibits similar performance and behavior as the real system without disclosing any real system information. Such synthetic networks are capable of bridging the gap of CPS data shortage as well as accelerating the development of CPS. However, a significant challenge in this endeavor is the accurate representation of the interaction dynamics between CPS components, as real-world systems exhibit complex behaviours that are not easily replicated. Besides, the system characteristics of cyber and physical layers are quite different in terms of system modelling and components interactions. Therefore, a multi-faceted approach is required.

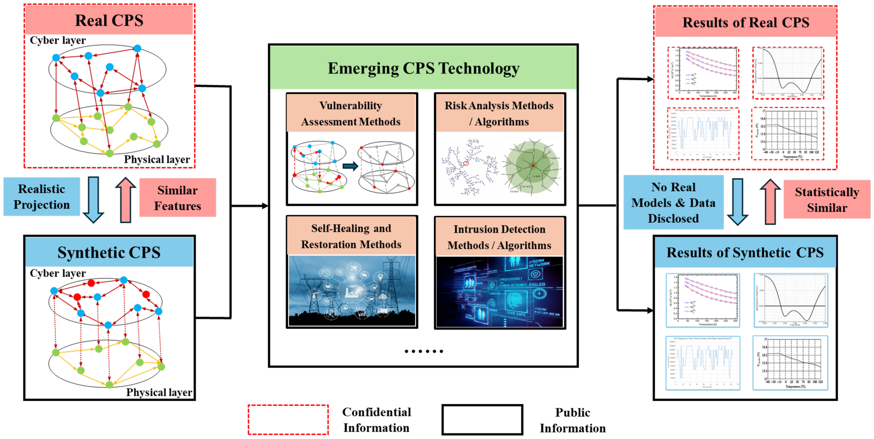


Figure 1.2 Schematic Diagram of Synthetic CPS Generation

In this background, leveraging Graph Neural Network (GNN) algorithms [17], [18] to analyse the real CPS data and generate synthetic models becomes a promising approach. The GNN is capable of learning the underlying patterns and dependencies within graph data. These models are particularly effective in capturing the intricate relationships in graph-based systems, i.e., CPS, enabling the generation of digital siblings that closely resemble real-world networks. The core challenge of implementing GNN lies in ensuring that the digital siblings not only replicate the structural aspects of the real systems but also accurately reflect their operational behaviours under diverse conditions. By focusing on the detailed interaction patterns and validating the models through rigorous testing, we can develop digital siblings that behave in a manner similar to real systems, providing a reliable foundation for research and analysis.

Q3: How to systematically define and analyze the complex interdependencies between cyber and physical layers? How to efficiently evaluate and identify the vulnerability of CPS considering time-varying operational states?

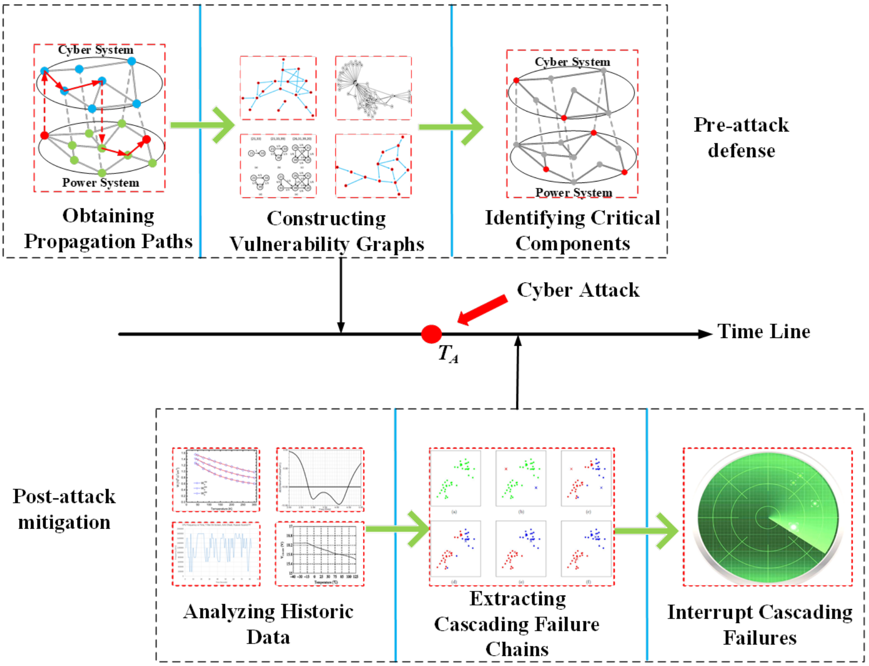

Cyber System
Power System
Obtaining Propagation Paths
Constructing Vulnerability Graphs
Cyber System
Power System
Identifying Critical Components
Pre-attack defense
Cyber Attack
Time Line
T_A
Post-attack mitigation
Analyzing Historic Data
Extracting Cascading Failure Chains
Interrupt Cascading Failures

Figure 1.3 Research Framework of Vulnerability Assessment for CPS

From the perspective of cyber-physical interdependency, two critical aspects should be addressed: structural and operational interdependencies. Structural interdependency focuses on investigating how the cyber and physical layers are physically interconnected and how such coupling methods will influence network performance. On the other hand, the operational interdependency refers to the operational interactions between cyber and physical system layer, which are related to the operational status of CPS. With respect to the well-defined cyber-physical interdependencies, we thoroughly analyze and evaluate the CPS vulnerabilities as presented in Figure 1.3. Figure 1.3 answers the second part of **Q3** from a data-driven perspective. By utilizing the historical data, the fault propagation paths and cascading failure chains can be obtained for the construction of vulnerability graphs. Furthermore, data mining techniques are adopted to identify the frequent cascading failure patterns and critical components. Then, by using limited resources, the identified vulnerabilities are well defended to achieve efficient pre-attack defense and post-attack mitigation.

1.3. Original Contributions

By answering the proposed research questions in Section 1.2, the major contributions of this thesis are summarized as follows:

- A two-stage generative model is proposed to generate large-scale synthetic communication topologies for CPSs. The proposed method circumvents the dilemma of CPS data availability by reproducing the typical design process of communication networks. The method is implemented to generate a synthetic CPS for the interconnected power grids in continental Europe, which is statistically validated by comparing the results with 18 realistic communication networks for power grids. The proposed method pioneers the synthetic CPS modelling and forms a solid foundation for further investigations to reveal invaluable characteristics, patterns, and mechanisms of CPSs.

- A GNN-based generative model, namely Graph-CPS, is proposed to generate a synthetic CPS topology with realistic network feature distribution. Graph-CPS captures not only complex network metrics but also intrinsic attributes of nodes and edges. Experimental results demonstrate that Graph-CPS accurately models various network types and scales, making it highly scalable. This work marks a significant advancement in adding more complexities to synthetic CPS models, contributing to enhanced system security and the protection of sensitive information.

- A hybrid generator, namely SibGen, is proposed to create a digital sibling of real CPS by generating both topological and operational models. The digital sibling produced by SibGen mirrors the fundamental topological characteristics and operational behavior of the real CPS, while exhibiting distinct structural and operational configurations, thereby ensuring the protection of sensitive CPS data. In case study, five different metrics are utilized to validated the synthetic CPS from both topological and operational perspectives. By simultaneously implementing the proposed GraphCCI on the synthetic CPS and the real CPS, it is proved that the generated synthetic CPS can be utilized as alternative test systems and exhibit similar characteristics as the real networks.

- A graph data mining-based critical components identification model named GraphCCI is proposed to evaluates the criticality of CPS components under time-varying operational states. The key innovation of GraphCCI lies in its shift from static vulnerability assessments to a dynamic, data-driven approach that more effectively captures the intricate interdependencies and evolving risks within cyber-physical systems.

To provide a comprehensive analysis of CPS vulnerabilities, GraphCCI introduces the concepts of manifest and latent correlations among components, enabling a deeper exploration of their interrelations. The case study reveals that there are limited numbers of critical components. Once these components are properly defended, the overall system security will be significantly enhanced.

1.4. Outline of the Thesis

The thesis outline is illustrated in Figure 1.4. Chapter 2 and Chapter 3 develop the synthetic CPS generation methodologies based on CPS data with varying levels of completeness and availability. More specifically, Chapter 2 proposes a two-stage generative model that generates large-scale synthetic communication topologies for interconnected power grids, i.e., continental European power systems, based on the existing power systems. In Chapter 3, a scalable generative model is proposed using graph neural networks, namely Graph-CPS, to generate a synthetic CPS topology with realistic network features distribution while ensuring the confidentiality of the input data and models. Then, a hybrid generator, namely SibGen, is proposed to generate the complete synthetic CPS including the topological and operational models of both cyber and physical system layers. Chapter 4 proposes CPS vulnerability assessment methods from the perspectives of Sequential Data Mining (SDM) [19] and Graph Data Mining [20]. First, the SDM algorithm is used to mine the historical cascading failure chain database and analyze the frequent cascading patterns. Then, a weighted cascading graph database is constructed employing graph data mining algorithms to identify the frequent subgraphs containing thorough component correlations. At last, a critical components identification model named GraphCCI is proposed to identify the vulnerabilities of CPS. Chapter 5 concludes the thesis and discusses the future research.

The remainder of this thesis is organized as follows:

Chapter 2 This chapter addresses **Q1** by proposing a two-stage generative model for generating synthetic communication topologies of large-scale CPS based on the existing power grids. It reproduces the existing communication network design process. The method generates statistically realistic networks. The proposed method is implemented to create a realistic, large-scale synthetic CPS for the interconnected power grids in continental Europe. In this chapter, we identify the CPS as a triple interdependent network consisting of Physical Communication Network (PCN), Logical Communication Network (LCN), and Physical Power System (PPS). The first stage is the PCN generator. The initial topology of the PCN is sequentially generated. Redundancy is added by jointly considering network congestion and connectivity. This approach is aimed at increasing the network's resilience, thereby rendering the generated PCN more aligned with realistic scenarios. The second stage is the LCN generator. The decentralized communication structure is utilized. A Communication Hub (CH) index is defined considering both communication traffic volume and node criticality to identify the optimal communication hubs for the LCN.

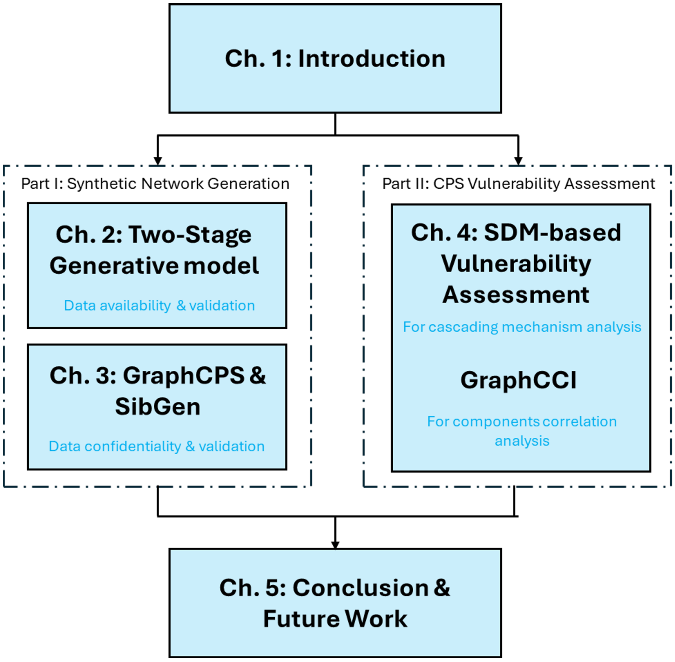

Ch. 1: Introduction
Part I: Synthetic Network Generation
Ch. 2: Two-Stage Generative model
Data availability & validation
Ch. 3: GraphCPS & SibGen
Data confidentiality & validation
Part II: CPS Vulnerability Assessment
Ch. 4: SDM-based Vulnerability Assessment
For cascading mechanism analysis
GraphCCI
For components correlation analysis
Ch. 5: Conclusion & Future Work

Figure 1.4 Thesis Outline

Chapter 3 This chapter aims to answer **Q2** and generate synthetic networks for both cyber and physical system layers. Different from chapter 2, this chapter addresses the scenario where real CPS data is available, e.g., network topologies and operational parameters. Graph neural networks are employed to capture the characteristics of the real networks and generate corresponding synthetic networks. The contributions of this chapter are two-fold: firstly, a scalable generative model, i.e., Graph-CPS, is proposed to generate a synthetic CPS topology with realistic network feature distribution. This model is capable of learning different complex network parameters as well as capturing the distribution of different network features of the input networks. Secondly, a hybrid generator, i.e., SibGen, is proposed to generate the digital sibling of the real CPS. SibGen generates both the topological and operational models of the input network. In SibGen, two effective training strategies are proposed, i.e., dual graph training and prior knowledge-constrained training. In the case study, seven different metrics are evaluated to compare the real CPS and generated synthetic network. The comparison results prove that SibGen is capable of learning the global characteristics of the input network from both topological and operational perspectives.

Chapter 4 This chapter solves the problem defined in **Q3**. In this chapter, a novel cascading failure model is proposed considering the interaction between the cyber and physical system layers for every single time instant. Based on quasi-dynamic simulations, a database of cascading failure chains is generated. This contains various operating conditions. Sequential mining algorithms are used to identify the frequent sequential cascading patterns. Vulnerability indices are constructed based on complex network theory to evaluate the importance of components in the cascading failure process and identify the critical components in CPS. Furthermore, two correlations are defined, i.e., manifest and latent correlations, to better reveal the CPS cascading mechanism and comprehensively investigate the apparent and potential correlations between CPS components. Then, A set of definitions are proposed to map the historical cascading failure datasets into weighted cascading graphs, and construct the weighted cascading graph database for graph data mining to thoroughly capture the cascading features of CPS. By jointly considering the manifest and latent correlations and the graph data mining results, a model is proposed for the critical components identification, i.e., GraphCCI.

Chapter 5 This chapter presents the major conclusions of this thesis and provide promising topics for future research.

2

Generating Large-Scale Synthetic Communication Topologies for Cyber-Physical Power Systems

Synthetic networks aim at generating realistic projections of real-world networks while concealing the actual system information. Researchers have mainly explored methods to create synthetic power systems. However, with the rapid power grid digitalization, new methods are needed for synthetic communication networks of CPS. In this chapter, a two-stage generative model is generated for generating synthetic communication topologies of large-scale CPS based on the existing power grids. It reproduces the existing communication network design process and is capable of generating statistically realistic networks. The proposed method is implemented to create a realistic, large-scale synthetic CPS for the interconnected power grids in continental Europe. The method is validated by comparing the generated communication network with 18 realistic communication network topologies with different system sizes. The experimental results validate the scalability and effectiveness of the generative model.

The content of this chapter has been published in: **Yigu Liu**, Alexandru Ştefanov, Peter Palensky. "Generating Large-Scale Synthetic Communication Topologies for Cyber-Physical Power Systems". *IEEE Transactions on Industrial Informatics*. Early Access, DOI: 10.1109/TII.2024.3438232

2.1. Introduction

With the increasing digitalization of power grids, the Cyber-Physical power Systems (CPS) are extensively studied [21], [22], [23]. However, given the national security concerns, detailed information about CPS cannot be publicly disclosed, i.e., power grid models, communication network architectures, and operational data. Also, standard test systems for CPS are missing in the current literature. Under such background, synthetic networks emerge as a promising method to generate fictitious but realistic projections of power grids and communication networks. A synthetic CPS avoids revealing sensitive network models and data while providing reliable test networks for research.

In recent years, researchers explored methods to mainly generate synthetic power systems. Reference [10] developed a synthetic DC power flow model for the continental power grids in Europe based on available public data. References [11], [12], [13], [14], [15] conducted research on large-scale synthetic power systems. In [11], the authors generate and validate the synthetic power systems topology from the perspective of complex network theory. In [12], a learning-based method is proposed to generate synthetic power grids, which are evaluated by considering power flows and vulnerability against failures. In [13], [14], [15], the synthetic network cases are extended with generator cost data and dynamic models for economic and transient stability studies. One can observe that the current literature on developing synthetic networks is mainly focused on the physical power system. How to generate a large-scale synthetic cyber-physical system is rarely investigated because of two reasons: (i) lack of real CPS data for model validation, i.e., system parameters and structural topologies, and (ii) increased computational complexity in generating large-scale CPS models. With the fast power grid digitalization, the power system is now tightly coupled with the cyber infrastructure in an unprecedented way. This makes the industrial communication networks, i.e., operational technologies, indispensable for power system operation. Therefore, we are motivated to investigate how to generate a synthetic CPS based on the results of synthetic power systems.

In CPS-related literature, most test cases are restricted to the standard IEEE test systems [24], [25], [26]. Reference [24] proposes a framework to model the cyber-physical system dependencies and assess the vulnerabilities of a CPS with eight remote terminal units. Reference [25] uses IEEE 39-bus and China's Guangdong 500-kV system to model the CPS and analyze cascading failures considering the interactions between cyber and physical layers. Reference [26] analyzes the fault propagation mechanism of cyber-physical systems for IEEE 118-bus and 300-bus systems. Currently, there is no standard CPS test system. All CPS-related test systems are generated based on the subjective assumptions of researchers, which may lead to biased experimental results. Besides, the dimensions of such test systems are far from the actual size of a real cyber-physical system, which leads to the following question. Are the experimental results obtained by using small-scale systems applicable to real, large-scale systems? The answer is debatable. Therefore, to serve as a better study case, a large-scale synthetic CPS model is surely desirable.

Ideally, the generated synthetic networks should have consistent characteristics with the original systems, i.e., size and structural features. Based on [10], [11], [12], [13], [14], [15], [24], the general process of generating synthetic power systems is: (i) collect public data, e.g., resident and geographic information; (ii) generate synthetic power grids based on available public data; and (iii) compare the synthetic networks with the actual power systems or standard test systems in terms of power flow results or complex network features. Given the fact that the actual communication network architectures and data are highly confidential, it is difficult to compare the characteristics of generated synthetic networks and real CPS. Also, it is worth mentioning that the major difference between the synthetic network generation and communication network design is that the synthetic network focuses on mirroring the realism of existing communication networks closely rather than pursuing optimal operational performance of the network. Therefore, to generate the synthetic communication network, the existing communication network design process is replicated, generating statistically realistic communication topologies. Otherwise, it might lead to significant deviations from realism.

Based on the discussion above, in this chapter, a two-stage generative model is proposed to generate realistic, large-scale synthetic cyber-physical systems based on the existing power grids. For a given power grid, the typical communication system design process is reproduced to generate synthetic communication topologies consisting of physical and logical communication networks for large-scale CPS assuming that the actual cyber system is designed to be functional in terms of network performance. The proposed method is implemented to generate the synthetic CPS model of the interconnected power grids in continental Europe, which is statistically validated by comparing the results with 18 realistic communication networks for power grids. It is worth mentioning that the historical evolution of the CPS is not considered in this chapter. To the best knowledge of the authors, this research is pioneering the generation of large-scale, synthetic CPS. The main contributions of this paper are summarized as follows:

1) A two-stage model is proposed for generating realistic, large-scale synthetic communication topologies for CPS based on existing power grids. The CPS is identified as a triple interdependent network consisting of Physical Communication Network (PCN), Logical Communication Network (LCN), and Physical Power System (PPS).

2) The first stage is the PCN generator, the initial topology of the PCN is sequentially generated and then more redundancies are added by jointly considering network congestion and connectivity. This approach is aimed at increasing the network's resilience, thereby rendering the generated PCN more aligned with realistic scenarios.

3) The second stage is the LCN generator, decentralized communication structure is utilized and a Communication Hub (CH) index is defined considering both communication traffic volume and node criticality to identify the optimal communication hubs for the LCN.

2.2. Framework: Generating a Triple Interdependent Cyber-Physical System

Typically, researchers consider that the cyber-physical system comprises of two interdependent layers, i.e., physical power grid and communication network infrastructure. However, the current literature overlooks the fact that the cyber system is also an interdependent network [27], [28], consisting of physical and logical communications. These cyber system interdependencies are essential for the overall operation of CPS. In this research, the cyber-physical system is considered a triple interdependent network as represented in Figure 2.1. It consists of the physical communication network, logical communication network, and physical power system. The complex interdependencies among the three layers are defined in the framework for generating a large-scale synthetic CPS.

2.2.1. The Three Interdependent Networks in CPS

Physical communication network: At the PCN layer, each node is an Optical Cross-Connect (OXC) router or Synchronous Digital Hierarchy (SDH) device installed in a substation. The edges in PCN are the physical communication media such as Digital Power Line Carrier (DPLC), Optical Power Ground Wire (OPGW), Broadband Power Line (BPL), wireless communication, and satellite communication [29]. Note that the DPLC and OPGW are frequently used in power systems due to the low operational costs. Furthermore, they do not require additional authorization from third parties. Normally, when a data packet is transmitted to a PCN node it either passes through the node without stopping or outputs from the optical domain to the local clients.

Logical communication network: The logical communication layer represents the interactions between PCN nodes. The LCN topology is pre-determined to satisfy the system operation requirements. Each node in the logical communication network is an IP router, which corresponds to a node in the physical communication network, e.g., OXC router. The nodes in the logical communication network are connected by logical links. It is worth mentioning that the logical communication network is a virtual network configured by CPS designers. In the communication process, a logical link may pass through multiple nodes in the physical communication network for a successful information delivery. For example, nodes V1 and V4 in the logical communication network are adjacent as represented in Figure 2.1. However, in the physical communication network, the traffic between OXC1 and OXC4 will pass through node OXC2 and OXC3. Note that in this chapter, wired networks are considered, in particular Wavelength-Division Multiplexing (WDM) optical networks, when generating the LCN, because large scale wireless communication networks are usually not used in typical CPS. Also, in this chapter, only static routing is considered. The dynamic routing is beyond the scope of this chapter.

Physical power system: This chapter focuses on the 380-400 kV high voltage transmission network. Therefore, each node in the physical power system is a substation while the transmission lines and transformers between substations represent the edges.

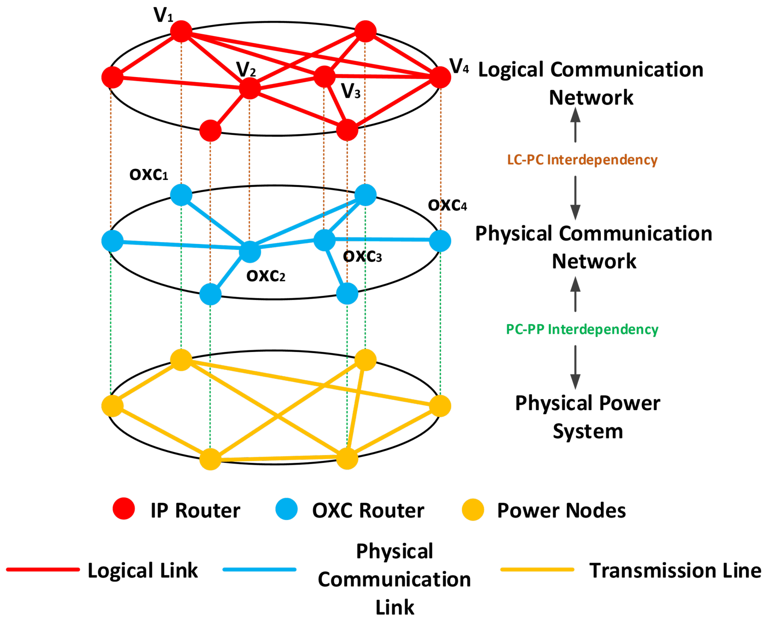


Figure 2.1 Three Interdependent Networks in CPS.

2.2.2. Complex Interdependencies among CPS Layers

As shown in Figure 2.1, there are two types of interdependencies in CPS, i.e., LCN and PCN interdependency (LC-PC), and PCN and PPS interdependency (PC-PP).

LC-PC Interdependency: The interdependency between logical and physical communication networks is essential for efficiently delivering control commands to actuators in the power grid and reporting operational data to Control Centers (CCs). The congested or invalid edges and nodes in the physical communication network impact the operational cost of data transmission, which is decided by the topology of the logical communication network. Meanwhile, the topology of the logical communication network also has a significant impact on the operational performance of the physical communication network. To thoroughly describe the LC-PC interdependency, we denote the LCN and PCN as $G_L(\boldsymbol{V}, \boldsymbol{E}_L)$ and $G_P(\boldsymbol{V}, \boldsymbol{E}_P)$, respectively, where $\boldsymbol{V}, \boldsymbol{E}_L, \boldsymbol{E}_P$ are

$$L_u = \begin{cases} 1, & \textit{if } u \in \boldsymbol{E}_L \\ 0, & \textit{otherwise} \end{cases} \tag{2.1}$$

$$L_{ur} = \begin{cases} 1, & \textit{if } u \textit{ passes through } r \\ 0, & \textit{otherwise} \end{cases} \tag{2.2}$$

$$L_{ur} \leq L_u, \quad \forall u \in \boldsymbol{E}_L, \; r \in \boldsymbol{E}_P \tag{2.3}$$

$$\sum_{r\in\Theta_o(n)} L_{ur} - \sum_{r\in\Theta_i(n)} L_{ur} = \begin{cases} 1, & \textit{if } O(u)=n \\ -1, & \textit{if } D(u)=n \\ 0, & \textit{otherwise} \end{cases} \tag{2.4}$$

$$\sum_{u\in \boldsymbol{E}_L} L_{ur} \le W_r, \quad \forall r \in \boldsymbol{E}_P \tag{2.5}$$

where u is a logical link in the LCN, r is a physical link in the PCN. L_u is the logical link variable and L_{ur} is the logical link routing variable. (2.3) indicates the mapping relationship between logical and physical links. (2.4) reveals the continuity of the logical links over the physical links, where $\Theta_o(n)$ and $\Theta_i(n)$ are the set of physical links outgoing and entering node $n \in V$, $O(u)$ and $D(u)$ are the origin and destination nodes of logical link u. Given that the number of wavelengths of each physical link is limited, (2.5) indicates that the sum of the logical link routing variables over r is constrained by W_r, the upper bound for the number of wavelengths on the corresponding optical fiber.

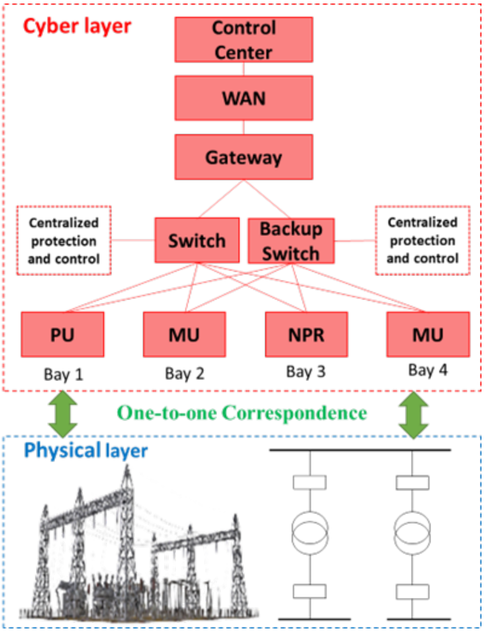


Figure 2.2 The Substation Communication Hybrid Architecture (NPR: Numerical Protection Relay, MU: Merging Unit, PU: Process Unit).

PC-PP Interdependency: Based on the interconnection of the cyber and physical nodes, the PC-PP interdependency is divided into “one-to-one”, “one-to-multiple”, and “multiple-to-multiple” correspondences [30]. However, in a real-world scenario, the PC-PP interdependency is more complex. Figure 2.2 shows the state-of-the-art substation communication architecture deployed in industry [31]. In the hybrid architecture of the substation communication, the Numerical Protection Relays (NPRs), Merging Units (MUs), and Process Units (PUs) send or receive data on a Local Area Network (LAN) within the substation. They communicate with the control centers through Wide Area Networks (WANs) via the routing gateways in the substations. In this chapter, each substation is associated with a physical communication node, i.e., gateway. Considering the relay communication nodes [30] in WAN, we define the PC-PP interdependency as "partially one-to-one" correspondence. Each substation node is exclusively associated with a communication node, i.e., routing gateway, while not all cyber nodes are connected with the substation nodes.

It is worth mentioning that the interdependency between the LCN and PPS is achieved through the PCN. That is, the measurement data of PPS are uploaded to the PCN. Then, the data packet follows the pre-determined routing path defined in the LCN and is delivered to the control center. The control center will make the optimal decision based on the collected data and send the commands back to the PPS using the same method. In the Figure 2.5 of section 2.4.1, we present more detailed illustrations to explain this concept.

2.2.3. Two-Stage Generative Model for Large-Scale Synthetic CPS

Generally, the design of a network topology includes the following steps: (i) initial topology design, (ii) increase redundancy to enhance the network resilience, and (iii) routing configuration [27]. In this chapter, we follow these sequential steps. The two-stage generative model proposed in this chapter is presented in Figure 2.3. Note we divide the large-scale network into multiple small-scale subnetworks and denote a subnetwork as a communication area [29]. As indicated in [32], the communication area is segmented based on the geographic locations, that is, for each communication area, one should allocate N_C substations that are geographically next to each other. Reference [32] pointed out that the N_C normally scales from 4 to 12. Therefore, by following the segmentation method in [32], the stage I and II in Figure 2.3 are implemented on each communication area. Besides, the following assumptions are used: (i) The topology of the power system is known. Therefore, the goal of this research is to generate the logical communication network, physical communication network, and the complex interdependencies in CPS. (ii) The historical evolution of CPS is not considered. (iii) Only wired communication networks are considered.

2.3. PCN Generator: Generating the Physical Communication Network

2.3.1. Generating the Initial Topology of PCN

The initial PCN topology is generated by considering the construction costs and network connectivity.

1) Construction cost. In [29], the authors consider two types of costs for generating a communication network, i.e., passive and active costs. The passive costs are attributed

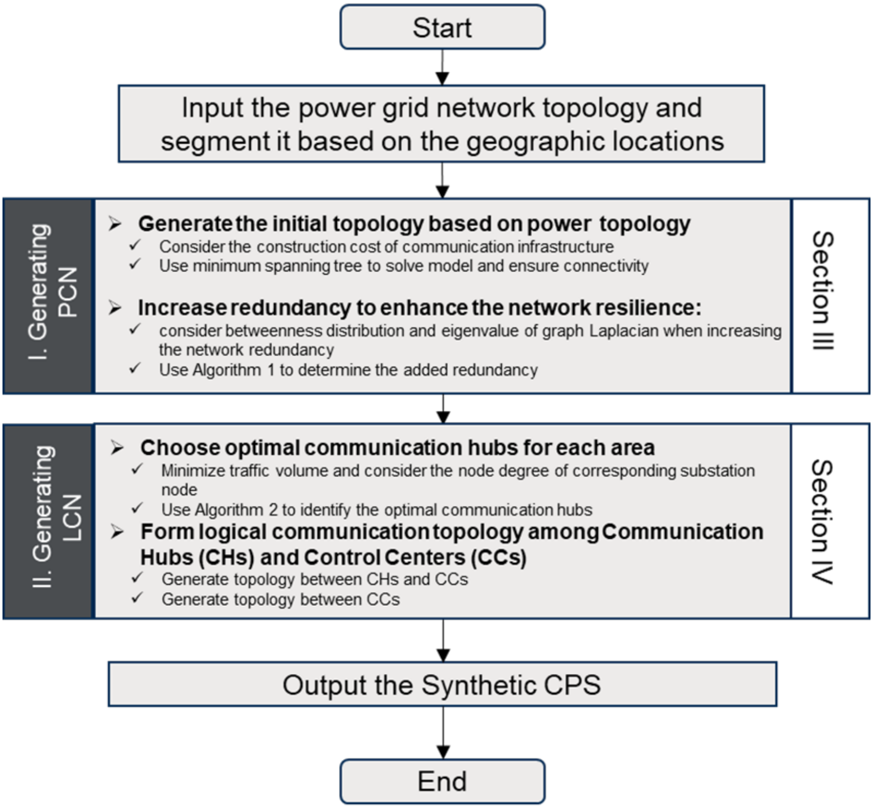

Start
Input the power grid network topology and segment it based on the geographic locations
I. Generating PCN
Generate the initial topology based on power topology
✓ Consider the construction cost of communication infrastructure
✓ Use minimum spanning tree to solve model and ensure connectivity
Increase redundancy to enhance the network resilience:
✓ consider betweenness distribution and eigenvalue of graph Laplacian when increasing the network redundancy
✓ Use Algorithm 1 to determine the added redundancy
Section III
II. Generating LCN
Choose optimal communication hubs for each area
✓ Minimize traffic volume and consider the node degree of corresponding substation node
✓ Use Algorithm 2 to identify the optimal communication hubs
Form logical communication topology among Communication Hubs (CHs) and Control Centers (CCs)
✓ Generate topology between CHs and CCs
✓ Generate topology between CCs
Section IV
Output the Synthetic CPS
End

Figure 2.3 The Diagram of Two-Stage Generative Model.

to passive components in the fiber optical network, which mainly depend on the length of the communication medium. Active costs are determined by the number of network switches and routers installed in the system. According to Figure 2.2, the active costs are determined by the number of substations. However, the number of substations is fixed because we generate the synthetic CPS based on the existing power grids. This indicates that we only need to consider the passive cost in our chapter, i.e., total length of communication links.

2) Connectivity. Generally, optimization models can be used to obtain the initial physical communication network topology with minimum construction cost. However, we need to ensure that the PCN remains a connected graph. Furthermore, we also need to make sure that the physical communication network topology remains connected in each communication area.

It is noted that the PCN is tightly coupled with power grids, and their topologies are highly similar [33]. Therefore, we take the substation nodes in the power grids and collect the distance data between different substations for the generation of the initial

PCN topology. Based on the discussion above, to satisfy the construction cost and network connectivity requirements, the minimum spanning tree [34] is used to compute the initial topology of the physical communication network. It generates a subgraph of a connected, edge-weighted undirected graph that connects all the vertices together, without any cycles and with the minimum total edge weight. In this chapter, the edge weight is set as the length of the distance between any two substations.

2.3.2. REM: Increasing Network Redundancy to Ensure Resilience

The initial topology of the physical communication network only satisfies the basic requirements for network design. In a real industrial scenario, communication network redundancy is needed to deal with contingencies. It provides backup communication paths for data transfer. Changing the design of the PCN topology improves the communication network performance such as network stability, connectivity, and congestion issues. Therefore, the eigenvalue of Laplacian matrix and betweenness distribution are considered to increase PCN network redundancy.

Eigenvalue of Laplacian matrix: According to [27], the network connectivity is related to the second smallest eigenvalue of the corresponding Laplacian matrix. For a PCN $\boldsymbol{G_F} = (\boldsymbol{V_C}, \boldsymbol{E_F})$, where $\boldsymbol{V_C} = \{V_C | C = 1, 2, 3...\}$ is the set of PCN nodes and $\boldsymbol{E_F} = \{E_f | f = 1, 2, 3...\}$ is the set of PCN edges. We denote the Laplacian matrix of $\boldsymbol{G_F}$ as $\boldsymbol{M} = [M_{CC'}]_{n\times n}$, where C and C' are the identifiers of nodes, and for the element $M_{CC'}$ in $\boldsymbol{M}$,

$$M_{CC'} = \begin{cases} \sum_{C''=1}^{n} A(V_C, V_{C''}), & if\ C = C' \\ -1 & if\ C \neq C',\ V_C\ and\ V_{C'}\ are\ connected \\ 0 & if\ C \neq C',\ V_C\ and\ V_{C'}\ are\ not\ connected \end{cases} \tag{2.6}$$

$$\lambda_2 = \min\left\{\lambda\left(\boldsymbol{M}\right) - \min\left\{\lambda\left(\boldsymbol{M}\right)\right\}\right\} \tag{2.7}$$

where C'' is the identifier of the neighbor node of V_C, $\sum_{C''=1}^{n} A(V_C, V_{C''})$ represents the node degree of V_C. $\lambda(\boldsymbol{M})$ is the set of eigenvalues of $\boldsymbol{M}$. The second smallest eigenvalue of $\boldsymbol{M}$ is denoted as λ_2, which is highly related to the performance of the communication system, such as network stability and connectivity. A larger λ_2 indicates better system performance. Therefore, the objective is to maximize λ_2 when adding communication edges to increase redundancy.

Betweenness distribution: reference [32] indicates that the network betweenness has substantial effects on the network congestion. Normally, data packets in the cyber layer are transmitted through the shortest path between any two arbitrary nodes. This makes the node betweenness an effective index to quantify the data volume that each node processes. Therefore, the betweenness distribution in the physical communication network is adopted to evaluate network congestion. The more uneven the betweenness distribution, the easier the system can be congested. If the betweenness is unevenly distributed, it means a small number of communication nodes will frequently be on the communication paths. Meanwhile, the communication capacity of a node is limited, making network congestion easier to happen. In this chapter, we employ the Gini coefficient to quantify the betweenness distribution. The calculation of the node betweenness $B(V_C)$ and the Gini coefficient G_{ini} are shown in equations (2.8)-(2.9):

$$B\left(V_C\right) = \sum_{V_C, V_{C'}, V_{C''} \in \boldsymbol{V_C}, C \neq C' \neq C''} \frac{N_{C'C''}\left(V_C\right)}{N_{C'C''}} \tag{2.8}$$

$$G_{ini} = \frac{1}{2n^2 u} \sum_{C=1}^{n} \sum_{C'=1}^{n} \left|B\left(V_C\right) - B\left(V_{C'}\right)\right| \tag{2.9}$$

where $N_{C'C''}(V_C)$ is the number of all shortest paths between node V_C' and V_C'' that go through V_C. $N_{C'C''}$ is the number of all shortest paths between node V_C'. C', C'' are the identifiers of node V_C' and V_C''. u is the average betweenness of all nodes in the physical communication network. A large G_{ini} represents the uneven distribution of betweenness, therefore, our goal is to minimize the G_{ini} of PCN.

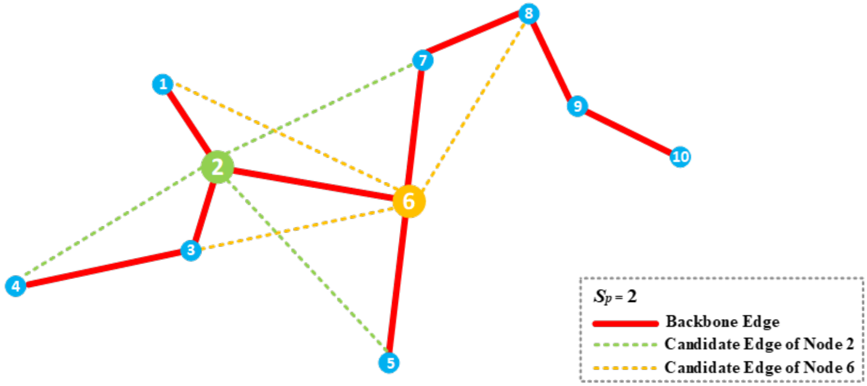


Figure 2.4 Generating Candidate Edges for Increasing the Redundancy for a Communication Area.

A trade-off between λ_2 and G_{ini} occurs when adding new communication edges to increase the communication network connectivity and redundancy. Ideally, a complete graph is desirable from the perspective of system performance. However, the construction cost of network is constrained by the budget of the network design. Therefore, we assume the number of added edges is subjected to construction cost and should satisfy the following condition:

$$N_{add} = \chi N_k, \; 0 < \chi < 1 \tag{2.10}$$

where N_{add} is the number of added redundancy edges. N_k is the number of initial edges in communication area k. χ is the redundancy coefficient of N_k subjected to the pre-determined budget. For each communication area, the number of added edges should not exceed a certain portion of the number of initial edges. Subsequently, we propose the Redundancy Enhancement Metric (REM) and denote it as E_r to determine how to add N_{add} redundant edges to increase the network performance.

$$E_r = \alpha\lambda_2 - \beta G_{ini} \tag{2.11}$$

$$\alpha + \beta = 1, \quad 0 < \alpha < 1, \quad 0 < \beta < 1 \tag{2.12}$$

where α and β are the weighted factors for λ_2 and G_{ini}, respectively. Based on all the constraints and parameters proposed above, we generate the candidate edges for each node in each communication area as shown in Figure 2.4. Taking node 2 as example, we generate the candidate edges, i.e., green dotted lines, by connecting the target node and its neighbor nodes whose shortest path length to the target node is S_P. By traversing all the nodes in the communication area, the candidate set $\boldsymbol{C_S}$ is obtained. Then based on (2.10)-(2.12), one can calculate the E_r of all possible combination. The combination with the highest E_r value contains the edges that are suitable for increasing network redundancy. Note that the process mentioned above will only be implemented on a communication area, where the number of nodes is limited [35]. Thus, the computational cost is acceptable and will not be exponentially increased even if we generate large-scale networks. More details about the computational efficiency are discussed in Section 2.5. The algorithm of adding new network redundancy is presented in Algorithm 2.1, where $S_P(V_C, V_{C'})$ is the shortest path length between V'_C and V''_C, $\boldsymbol{C_{S\_combination}}$ is the set of all combinations of edges in $\boldsymbol{C_S}$, and $\boldsymbol{V_{CA}}(k) = \{..., V_C, ...\}$ is the set of nodes in communication area k.

Algorithm 2.1: Adding Redundancy for Communication Area

Input:

Initial topology of the physical communication network
Parameters: S_P, χ, α, β

Output:

Optimal candidate edge set: $\boldsymbol{C_{s\_optimal}}$

```
Step 1   $\boldsymbol{C_{s\_optimal}} \leftarrow \varnothing$, $\boldsymbol{C_s} \leftarrow \varnothing$
Step 2   For $V_C \in \boldsymbol{V_{CA}}(k)$ do
Step 3       $\boldsymbol{C_s} \leftarrow$ all $V_{C'}$ that satisfy $S_P(V_C, V_{C'}) = S_P$
Step 4   End For
Step 5   Employ equation (2.10) to calculate $N_{add}$
Step 6   $\boldsymbol{C_{s\_combination}} \leftarrow$ all combinations $(\boldsymbol{C_s}, N_{add})$
Step 7   For combination in $\boldsymbol{C_{s\_combination}}$ do
Step 8       Employ equation (2.11) to calculate $E_r$
Step 9   End For
Step 10  $\boldsymbol{C_{s\_optimal}} \leftarrow$ combination with highest $E_r$
```

2.4. LCN Generator: Generating the Logical Communication Network

2.4.1. Choose Optimal CHs for Each Communication Area

Generally, there are two types of communication architectures for power systems, i.e., centralized and decentralized architectures as shown in Figure 2.5. In a centralized architecture, power system measurements are encapsulated into data packets in substations using various standards, e.g., C37.118, IEC 104, and DNP 3, and are communicated directly to a control center. After data processing, appropriate control commands are communicated to the controlled power elements in substations. In a decentralized architecture, data packets also follow the same standards, but the communication structure is different. First, the communication system is divided into multiple communication areas. Each area has a communication hub to gather all measurement data from substations. The communication hubs communicate with the control center. The control commands follow the same routing from the control center to the controlled power elements in substations. Compared with centralized communication, the decentralized architecture has a better performance in terms of time delays even with lower network bandwidths [36]. Furthermore, the decentralized communication architecture presents higher reliability with the same construction cost [29]. Therefore, in this chapter, we adopt the decentralized communication architecture as illustrated in Figure 2.5.

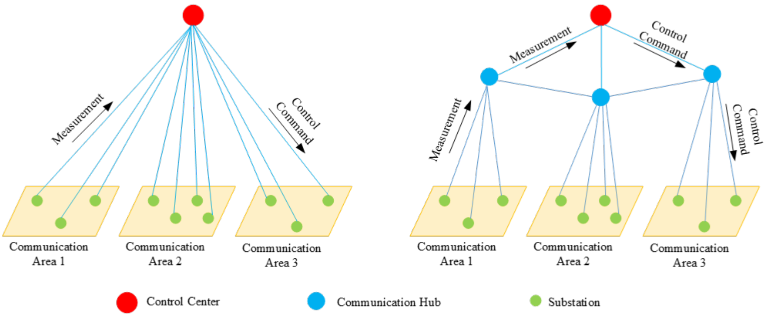


Figure 2.5 Centralized Communication Structure (Left) and Decentralized Communication Structure (Right) [36].

To choose the optimal location for communication hubs, we consider both the overall traffic volume and the importance of the corresponding power nodes of substations. On the one hand, the logical topology of logical communication network directly influences the traffic volume, which makes the location of communication hubs crucial to the performance of the communication network. On the other hand, given that the communication hubs directly communicate with the control center, it is desirable to connect the communication hubs with the critical substations. This ensures that anomalous behaviors and contingencies are directly and effectively monitored to maximize communication system reliability. Note that the power plant communication nodes directly communicate with the control center because the power generation data is crucial to power system operation. Based on the considerations above, we propose the CH index $\boldsymbol{I_C}$ to identify the optimal CHs.

$$\boldsymbol{I_C} = \frac{\sum_{C'=1}^{m} A(V_P, V_{P'})}{\sum_{C'=1}^{|V_C|-1} h_{C'} \times p_{CC'}} \tag{2.13}$$

where m is the number of substations in communication area k, $h_{C'}$ is the number of hops required for the determined communication, $p_{CC'}$ is the size of transmitted data packet. We use the node degree $A(V_P, V_{P'})$ to quantify the importance level of a substation in PPS. A node with higher degree indicates that once the node is removed, it will pose serious impact on more nodes in CPS. Therefore, the response time delay can be reduced if direct monitoring and control are implemented to those nodes and thus systematic security can be increased. For the consideration of the traffic volume, $\boldsymbol{I_C}$ depends on the number of required communication hops and the data packet size [36]. Note that the communication hops between two nodes are decided by the topology of the physical communication network. By calculating $\boldsymbol{I_C}$ for each substation V_C, the corresponding substation with the largest $\boldsymbol{I_C}$ is identified as the optimal communication hub. Note that in this chapter, the communication hub identification only considers the communication traffic under static routing, which assumes the system is under normal operation. In case of contingencies, optimal dynamic routing strategies can be considered to increase the overall system resilience. However, the dynamic routing is beyond the scope of this chapter, which can be considered as a future study.

2.4.2. LCN Topology among CHs and CCs

After the optimal communication hubs are identified in each communication area, the logical topology of all substation nodes is determined, i.e., each substation in the area has a direct logic link to the communication hub as shown in Figure 2.5. Therefore, the remainder of the logical communication network topology consists of the (i) topology between communication hubs and control centers, and (ii) topology between control centers.

The topology between CHs and CCs: in a real-world scenario, backup control centers [32] are extensively deployed to increase system reliability. Therefore, each communication hub needs to send data packets to both control centers. Note that the difference between them is that in the most of the operational states, main control centers have high priority to take the actions while the backup control centers work as a redundancy to enhance the resilience of systems in case of emergency. The communication between CHs and control centers is mostly done through WANs. The WAN topology is beyond the scope of this chapter. Therefore, we assume that each communication hub has at least one reliable and cost-efficient path to communicate with the control centers. Generally, the communication topologies between the CHs and CCs have two categories, i.e., double-star and mesh topology. Reference [25] conducted a comparison between these two categories on IEEE 39-bus system and China's Guangdong 500-kV system, and the experimental results prove that the double-star topology has lower probability of catastrophic failures than with the mesh topology. This is because the double-star topology is capable of maintaining its functionality even when part of the communication nodes fails. Combining all the facts and discussion above, the topology between communication hubs and control centers are modelled as double-star topology. The double star topology is normally a scale-free network, whose degree distribution has the power-law distribution characteristics and can be written as in (2.14) [37].

$$p\left(A(V_l, V_{l'})\right) \propto \left[A(V_l, V_{l'})\right]^{-r'} \tag{2.14}$$

where r' is a constant and satisfies $r' > 1$. $V_l, V_{l'}$ are the nodes in LCN. (2.14) indicates that there is a small number of critical nodes in the network, and the systematic connectivity is dramatically decreased once those nodes are removed. The current literature suggests that the double-star topology has a better communication performance in terms of transmission ability and network congestion compared with the mesh topology [32]. On the other hand, the double-star topology is highly vulnerable to cyber-attacks if adversaries have enough system information, e.g., system topology and operational data. However, given that system information is highly confidential, the double-star topology is more suitable than the mesh topology. The preferential attachment algorithm is adopted to generate the double-star topology between CHs and CCs [25].

The topology between CCs: the communication among control centers is defined by the Inter-Control Center Communications Protocol (ICCP), which is specified world-wide by utilities to provide the services for data exchange, monitoring, and control. The ICCP bilateral tables define the data exchange between two control centers. Normally, all control centers have reliable and efficient communications with their neighboring control centers. Based on the facts above, all control centers are logically reachable by other control centers as long as all power grids are interconnected. Therefore, the LCN topology for control centers is a full connection graph, i.e., each control center is logically connected with other control centers through WANs. It is worth mentioning that the full connection graph in this section represents that in the logical communication network, all control centers are logically accessible rather than physically connected.

2.5. Case Study

In this section, we implement the proposed methods to generate the synthetic CPS for the interconnected power grids in continental Europe. The parameters for the simulation are $\alpha = 0.5$, $\beta = 0.5$, $\chi = 0.3$, $S_P = 3$. The methods are coded in Python and simulations are run on a computer equipped with an Intel i7-8750H CPU at 2.2 GHz and 16 GB RAM.

2.5.1. Generated Synthetic CPS for Continental Europe

The methods proposed in Figure 2.3 are used to generate a large-scale, synthetic CPS based on open-source data from the ENTSO-E website [38]. It provides the 380-400 kV transmission system topologies of the interconnected power grids in continental Europe. The generation results are shown in Figure 2.6 and Figure 2.7 give a clearer demonstration of the LCN and PCN of the French power system. Detailed information about the number of nodes and edges in both physical and logical communication networks is given in Table 2.1. For the clarity of Figure 2.6 and 2.7, although generated in the LCNs, we do not represent the topology between cyber substation nodes and communication hubs, as well as the direct connection between power plants and control centers. The code and generated models are available online [39].

At the physical communication networks layer, we divide the substations into different communication areas. For each area, we randomly allocate substations based on their geographic location and then identify the optimal communication hub. Based on [32], N_C is set as a random number between 4 to 12. However, in several small countries, e.g., Albania, Croatia, Slovenia, and Macedonia, the number of substations is not enough for initiating multiple communication areas. Therefore, we consider that the substations in these countries directly communicate with the control centers, similar to the communication hubs in other larger countries.

At the logical communication networks layer, we decide the number of control centers in each country based on [32]. Typically, a country only has one Transmission System Operator (TSO), i.e., one main and backup control centers. However, in Germany and Austria, multiple TSOs exist. Therefore, the communication hubs are divided equally based on the number of TSOs in the country and their geographic location. Besides, all main control centers and backup control centers are logically connected to each other.

2.5.2. Statistical Analysis and Validation of Generation Results

In this part, we use realistic communication network data of power grids to verify the generation results of our proposed method. In Table 2.2, we collect 18 communication networks for power grids with different system sizes from the current literature. They are categorized into small, medium, and large size systems comprising of 7, 6, and 5 communication networks, respectively. Based on these networks, we compute complex network parameters in terms of the number of nodes and edges, average node degree, average shortest path length, network diameter, and network density. For each parameter we calculate the range based on the given realistic communication network. Note that these complex network parameters depict the global features of the target networks. Therefore, the local network features are not discussed in this chapter.

To verify the effectiveness and scalability of the proposed method, we implement it to power systems with different system sizes scaling from 14 to 289 node systems. The generation results are shown in Table 2.3. One can observe that all complex network

Table 2.1: The History of Cyber-Physical Events

Countries	**No. of nodes in PCN**	**No. of edges in PCN**	**No. of nodes in LCN**	**No. of edges in LCN**
Albania	6	7	8	17
Austria	31	37	35	39
Belgium	38	50	40	47
BiH	11	13	136	28
Bulgaria	23	32	25	29
Croatia	9	10	11	24
Czechia	44	69	46	56
Denmark	31	42	33	40
France	176	268	178	223
FYROM	9	10	11	24
Germany	289	355	297	373
Greece	31	45	33	38
Hungary	26	38	28	33
Italy	150	230	152	191
Montenegro	8	9	10	22
Poland	59	93	61	78
Portugal	41	60	43	50
Romania	51	77	53	63
Serbia	28	37	30	37
Slovakia	27	38	29	34
Slovenia	9	11	11	24
Spain	179	299	181	234
Switzerland	41	62	43	52
Netherlands	35	48	37	44

*The communication relay nodes are not included in the number of PCN nodes and LCN nodes.

Table 2.2: Statistics of Realistic Communication Networks in the Literature

	Small Size Syst. [25], [40], [41], [42]	**Medium Size Syst.** [32], [40], [43]	**Large Size Syst.** [44], [40], [43]	**Overall**
N	(18, 49)	(103, 182)	(236, 404)	(18, 404)
L	(29, 98)	(124, 232)	(357, 608)	(29, 608)
$\langle k \rangle$	(2.833, 4)	(2.551, 3.546)	(2.119, 3.01)	(2.119, 4)
$\langle l \rangle$	(2.433, 3.681)	(3.169, 6.697)	(6.721, 11.67)	(2.433, 11.67)
d	(5, 8)	(12, 15)	(22, 28)	(5, 28)
D	(0.933, 2)			

N: number of nodes, L: number of edges, $\langle k \rangle$: average node degree, $\langle l \rangle$: average shortest path length, d: network diameter, D: network density, $D = L/N$.

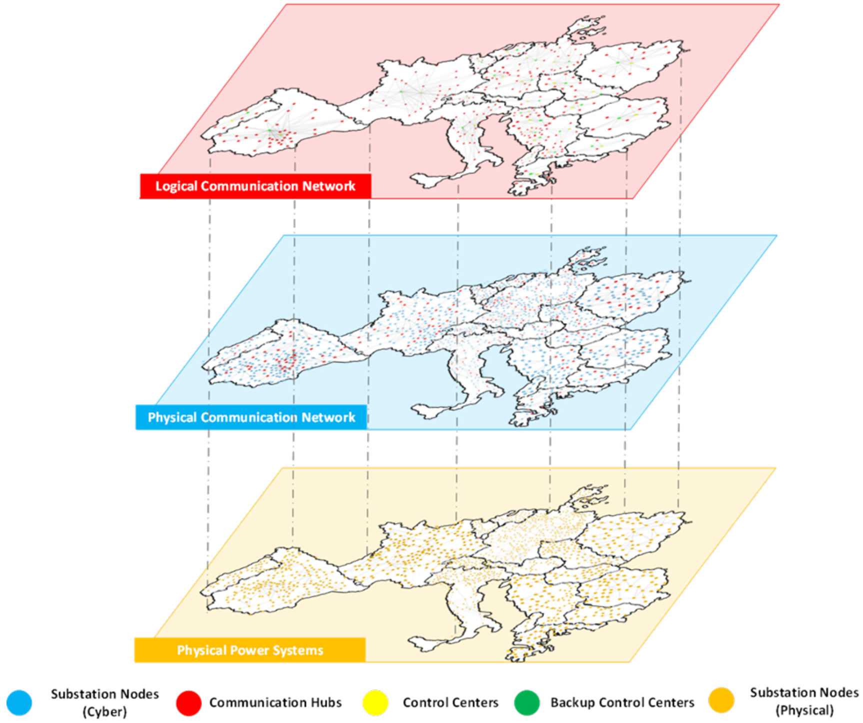
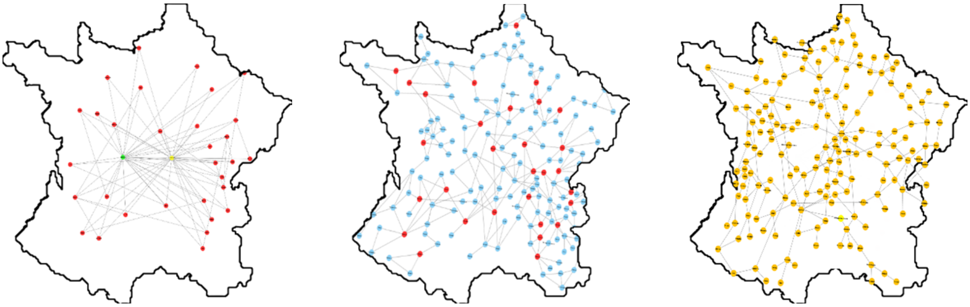


Figure 2.6 Generated Synthetic CPS for Continental Europe.



(a) LCN

(b) PCN

(c) PPS

Figure 2.7 Generated Synthetic CPS for France.

theory parameters of the generated communication networks with the proposed method are within the parameter ranges given in Table 2.2. Statistically speaking, in Table 2.2, the average node degree decreases when the system size increases, while the average shortest path length and network diameter increase when the system size increases.

Table 2.3: Statistics of Generated Communication Networks

	Small Size Syst.		**Medium Size Syst.**		**Large Size Syst.**
	IEEE 14-bus	**IEEE 39-bus**	**IEEE 118-bus**	**France**	**Germany**
N	14	39	118	176	289
L	26	67	204	268	355
$\langle k \rangle$	3.714	3.648	3.458	3.045	2.456
$\langle l \rangle$	2.078	3.836	6.159	7.706	11.547
d	4	7	14	18	28
D	1.857	1.718	1.729	1.522	1.228

Comparing with Table 2.3, similar patterns can be observed. By comparing each parameter, one can observe that in the case of IEEE 14-bus, the average shortest path length is slightly out of the given range. This is because in Table 2.1, the given system size is from 18 to 49, while 14-bus system is smaller than 18-node system. Based on our former discussion of average shortest path length, the result of IEEE 14-bus system still follows the same pattern. Based on the discussion above, the effectiveness of the proposed method is verified. The case study on systems with different sizes also shows that our method has excellent performance on scalability.

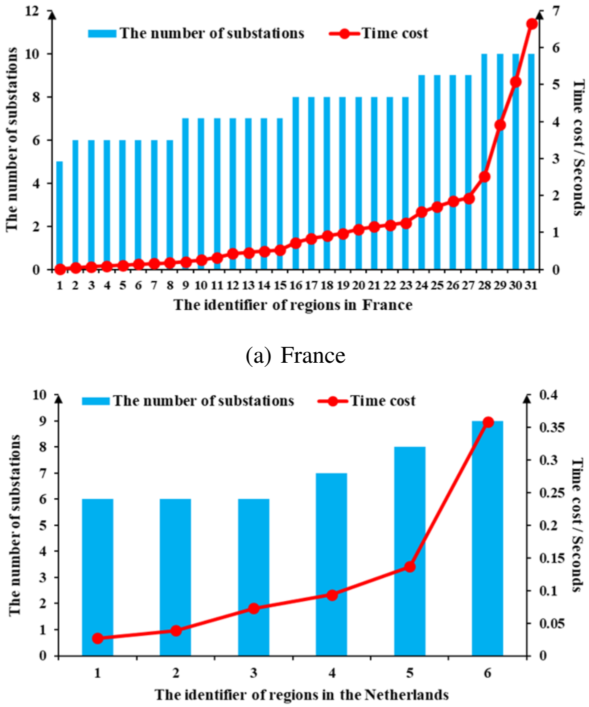
2.5.3. Evaluating Time Efficiency of Proposed Methods

In this part, we evaluate the time efficiency of the proposed methods. Figure 2.8 presents the time cost of generating large-scale, synthetic cyber-physical systems. France and The Netherlands are selected to evaluate the performance of the proposed methods on different power system sizes. The time cost consists of two parts as shown in equation (2.15).

$$T_{\text{total}} = T_{PCN} + T_{LCN} \tag{2.15}$$

where T_{total} is the total time cost of proposed method, T_{PCN} is the run time of the stage I proposed in Section 2.3, and T_{LCN} is the run time of stage II as shown in Section 2.4.

The algorithm complexity of T_{PCN} is $O(n!)$ and the algorithm complexity of T_{LCN} is $O(n)$. Although the complexity of T_{PCN} is high, the input size, i.e., number of PCN nodes in each communication area, is limited according to [32]. Therefore, the time cost of the proposed method will not be exponentially increased even when the input size increases. Furthermore, we present the time cost of generating synthetic communication topology for France and the Netherlands to further prove the scalability of the proposed methods. In Figure 2.8, the left axis represents the number of substations in each area, and the sub axis on the right represents the cumulative time cost. We can observe that as the area size increases, the time cost also increases, but the increment is at an acceptable level. As discussed in part A of this section, the size of each communication area is limited, which determines that the final time cost of each area will not exceed



(a) France

(b) The Netherlands

Figure 2.8 Generation Time of France and The Netherlands.

the maximum time cost as shown in Figure 2.8. Typically, when generative models are applied to large-scale networks, the computational cost grows exponentially with the increase of system size. However, in this chapter, the cost problem is addressed by applying the decentralized communication structure. The proposed generative model adopts the idea of divide and conquer rather than generating the entire network in one batch. Therefore, Figure 2.8 proves that the proposed methods are suitable for generating large-scale, synthetic CPS in a time-efficient manner.

2.5.4. Performance Comparison and Evaluation

In this part, we compare the proposed method with the traditional algorithms in the literature. In Figure 2.9, we present the generation results of two traditional generative algorithms with various network sizes, i.e., Chung-Lu and Havel-Hakimi algorithms [16]. Also, the complex network parameters are presented in Table 2.4. By observing the generated networks, one can notice that the network connectivity is the major issue of the traditional methods. As the network size increases, more isolated networks show up. For Chung-Lu algorithm, all the isolated parts are discrete nodes. This phenomenon is caused because the Chung-Lu algorithm generates a network based on the given distribution of node degrees. The larger the network size, the more difficult it is to guarantee the network connectivity while keeping the given distribution. Therefore, the scalability of this method is limited. The Havel-Hakimi algorithm also has the same issue as in Chung-Lu algorithm. The difference is there is no single isolated node because the Havel-Hakimi algorithm generates the network based on the given degree. For communication networks, the overall network connectivity is the first priority, because it provides an alternative communication path when the system is suffering from contingencies. Compared with the proposed methods, the Chung-Lu and Havel-Hakimi algorithms also fail to generate networks with realistic parameters distribution as shown in Table 2.4. The comparison above proves the good performance of the proposed method.

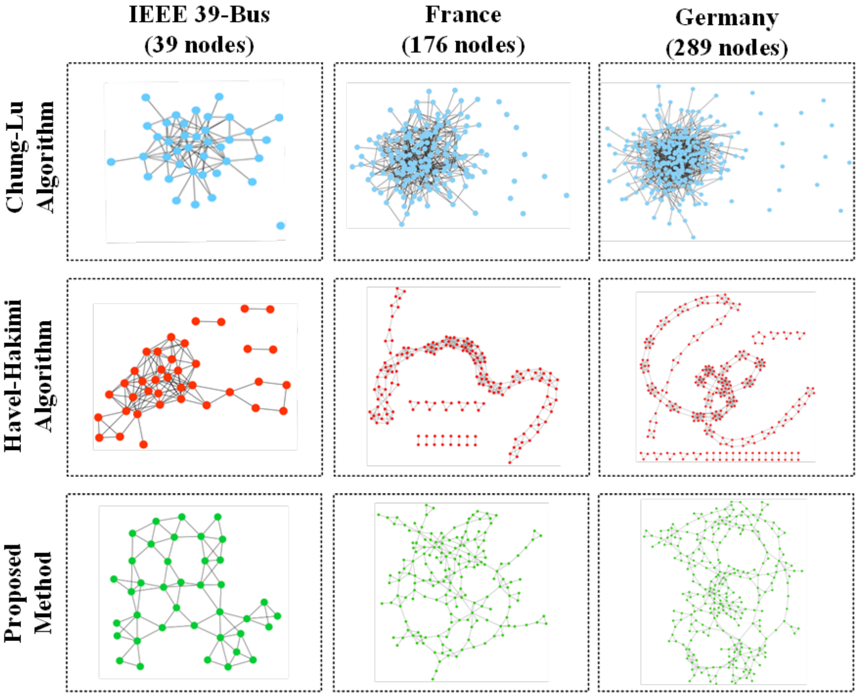


Figure 2.9 Comparison with traditional generative algorithms in the literature.

In the following evaluation, Table 2.5 showcases the network parameters of the Netherlands synthetic networks with varying communication area sizes. As noted in [32], N_C, the number of substations within each communication area, ranges from 4 to 12. To examine the impact of communication area segmentation size on the accuracy of generation results, we have segmented this range into four distinct categories, as detailed in Table 2.5. One can observe that as the N_C increases, the network average node degree and network density decreases while the average shortest path length increases. Compared with the data of small size systems in Table 2.2, only when N_C is at the range of 4 to 6, all parameters fit to the listed range. Therefore, when N_C is at the range of 4 to 6, the generated network has the most realistic network parameters.

Table 2.4: Statistics of Generated Communication Networks

	IEEE 39-Bus System (39 nodes)			France (176 nodes)			Germany (289 nodes)		
	Chung-Lu	**Havel-Hakimi**	**Proposed method**	**Chung-Lu**	**Havel-Hakimi**	**Proposed method**	**Chung-Lu**	**Havel-Hakimi**	**Proposed method**
N	39	39	39	176	176	176	289	289	289
L	86	113	67	486	493	268	820	796	355
$\langle k \rangle$	4.3	5.795	3.648	5.254	5.602	3.045	5.39	5.509	2.456
$\langle l \rangle$	N/A	N/A	3.836	N/A	N/A	7.706	N/A	N/A	11.547
d	N/A	N/A	7	N/A	N/A	18	N/A	N/A	28
D	2.205	2.897	1.718	2.761	2.801	1.522	2.837	2.754	1.228

Table 2.5: Statistics of Generated Communication Networks for the Netherlands

N_C	**4-6**	**6-8**	**8-10**	**10-12**
N	35	35	35	35
L	58	55	52	52
$\langle k \rangle$	3.314	3.142	2.971	2.972
$\langle l \rangle$	3.615	4.159	4.661	4.642
d	7	10	7	10
D	1.657	1.571	1.486	1.486

2.6. Conclusion

This chapter focuses on generating large-scale synthetic communication topologies for CPSs. The proposed method circumvents the dilemma of CPS data availability by reproducing the typical design process of communication networks. It generates synthetic topologies consisting of physical and logical communication networks for large-scale CPS. The method is implemented to generate a synthetic CPS for the interconnected power grids in continental Europe, which is statistically validated by comparing the results with 18 realistic communication networks for power grids. Furthermore, the experimental results demonstrate its scalability and computational time efficiency. This research pioneers the synthetic CPS modelling and forms a solid foundation for further investigations to reveal invaluable characteristics, patterns, and mechanisms of CPSs. Note that this chapter focuses on generating the communication topologies based on existing power grids, which is the first and critical step of generating complete synthetic CPS. In the next chapter, more complexity will be added to synthetic CPS, e.g., information and power flow models.

3

GNN-BASED GENERATIVE MODELS FOR GENERATING DIGITAL SIBLINGS FOR CYBER-PHYSICAL POWER SYSTEMS

This chapter aims at generating digital siblings for both cyber and physical layers. Different from chapter 2, this chapter addresses the scenario where the real CPS model and data, such as network topologies and operational parameters, are available. Graph Neural Networks are employed to capture the characteristics of the real networks and generate corresponding synthetic networks. A scalable generative model, called Graph-CPS, is introduced to create synthetic CPS topologies that maintain realistic network feature distributions. Additionally, a hybrid generator, named SibGen, is proposed to generate a digital sibling of the real CPS. SibGen is capable of producing both topological and operational models based on the input network. Both methods have been carefully implemented and validated through extensive experimental simulations. The results demonstrate that Graph-CPS is not only scalable but also accurate in preserving the essential characteristics of input networks, regardless of the network type or size. Moreover, SibGen effectively captures the global characteristics of the input network from both topological and operational perspectives.

Parts of this chapter have been published in: **Yigu Liu**, Haiwei Xie, Alfan Presekal, Alexandru Ştefanov, Peter Palensky. "A GNN-Based Generative Model for Generating Synthetic Cyber-Physical Power System". *IEEE Transactions on Smart Grid*, vol. 14, no. 6, pp. 4968-4971, 2023.

3.1. Introduction

With the increasing digitalization of modern power grids, the operation characteristics of the Cyber-Physical power System (CPS) have significantly changed. To accurately analyze the new system behavior, reliable models are needed for CPS research. The models should have consistent network characteristics with the real CPS to ensure the accuracy of simulation results. Meanwhile, the models should avoid revealing any sensitive system information that may be exploited by the adversaries, e.g., system topology, network features. To this end, synthetic networks, which can comprehensively mimic the characteristics of actual networks, became the answer to this concern. Based on the current literature, all the possible network generation methods is divided into three categories and analyze the pros and cons of each category.

Complex network-based methods: Represented by Barabasi-Albert [45] and Erdos-Renyi [46] models, the generative algorithms in this category are capable of generating networks that satisfy given requirements, e.g., scale free property, small world property. Besides, the complex network parameters [47] of the generated synthetic network, e.g., degree [37], [48], closeness [49], betweenness [50], [51], etc., can be customized. However, these algorithms have two major limitations: (a) Network connectivity. These algorithms have unsatisfying performance on generating connected network while only focusing on the global complex network parameter distribution [16], which is not acceptable in generating synthetic CPS. (b) Incapability of generating operational models. The complex network-based methods only focus on the topological aspects while it fails to generate the operational models, e.g., power flow models, information flow models.

Open source data-based methods: This category utilizes the open source data, e.g., population distribution, historical electricity consumption, geographical information, etc., to plan the locations of synthetic nodes. Then, the network edges are arranged based on various assumptions, e.g., construction cost [52], power flow convergence [53], etc. This category is capable of generating both topological and operational models of synthetic networks, as well as guaranteeing the reasonable parameter distribution to a certain extent. However, it completely ignores the real-world networks and the generated synthetic networks may have different system behaviors compared to the real networks. Therefore, the conducted experiments on these synthetic networks may lead to biased results.

Machine learning-based methods: Enabled by the development of graph neural networks [54], [55], [56], [57], [58], this category trains the machine learning models and learns the characteristics distribution of the input real-world networks. Then, based on the learned knowledge, the generative models output the realistic projection of the input network, which maintains the similar system characteristics while concealing all the sensitive information of the real networks. There are two major challenges in this category: (1) Access to the real network data. Due to national security concerns, the real data of CPS is highly confidential and thus inaccessible. (2) Data security. Although Machine learning-based methods can generate synthetic networks with altered topology and system parameters, we need to make sure that no sensitive data can be obtained by reverse engineering.

In this chapter, we aim at generating synthetic networks for both cyber and physical layers. Different from chapter 2, this chapter addresses the scenario where real CPS data, such as network topologies and operational parameters, is accessible. Graph Neural Networks are employed to capture the characteristics of the real networks and generate corresponding synthetic networks. The contributions of this chapter are listed as follows:

(1) A scalable generative model, namely Graph-CPS, is proposed to generate a synthetic CPS topology with realistic network feature distribution. This model is capable of learning different complex network parameters as well as capturing the distribution of different network features of the input networks. In case study, the Graph-CPS is implemented on various sizes of network scaling from 18 nodes to 1225 nodes. The experimental results prove that the Graph-CPS is scalable and accurate to preserve the characteristics of input networks with not only different network types but also different network sizes.

(2) A hybrid generator, namely SibGen, is proposed to generate the digital sibling of the real CPS. SibGen can generate both topological and operational models of the input network. In SibGen, two effective training strategies are proposed, i.e., dural graph training and prior knowledge-constrained training. In case study, five different metrics are evaluated to compare the real CPS and the generated synthetic network. The comparison results prove that the SibGen is capable of learning the global characteristics of the input network from both topological and operational perspectives.

3.2. A GNN-Based Generative Model for Generating Synthetic Cyber-Physical Power System Topology

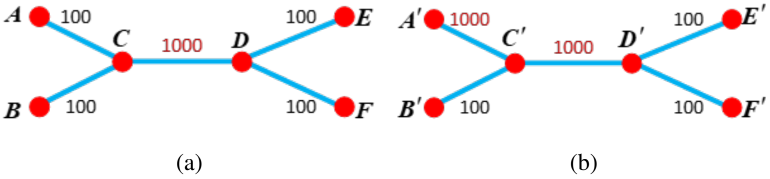


Figure 3.1 Illustration of Networks with Same Topological Parameter Distribution but with Different Network Features.

The current research of synthetic networks mainly focuses on the power grids, the corresponding research on cyber aspects are insufficient [53], [11]. Besides, the common philosophy in the literature is to generate a statistically realistic network in terms of complex network parameters, e.g., degree distribution, average path length [53], [11], etc. Such consideration, although it captures the system characteristics to a certain extent, neglects the inherent system attributes of the nodes and edges such as the bandwidth of communication links and capacity of transmission lines in the CPS. Taking Figure 3.1 as an example, although the networks in (a) and (b) have the same topology, the edge attributes are different. Consequently, one can obtain different results if they run power or information flow models on two networks.

Based on the discussion above, we propose a scalable generative model, namely Graph-CPS, to generate a synthetic CPS topology with realistic network feature distribution. This model is capable of learning different complex network parameters as well as capturing the distribution of different network features of the input networks. The experimental results in thoroughly prove the effectiveness and scalability of Graph-CPS. It can accurately capture the characteristics of input networks with not only different network types, but also different network sizes. To the best knowledge of the authors, our paper is a pioneer work of its kind in generating synthetic topologies for CPS.

3.2.1. Topological Modeling of Cyber-Physical Power System

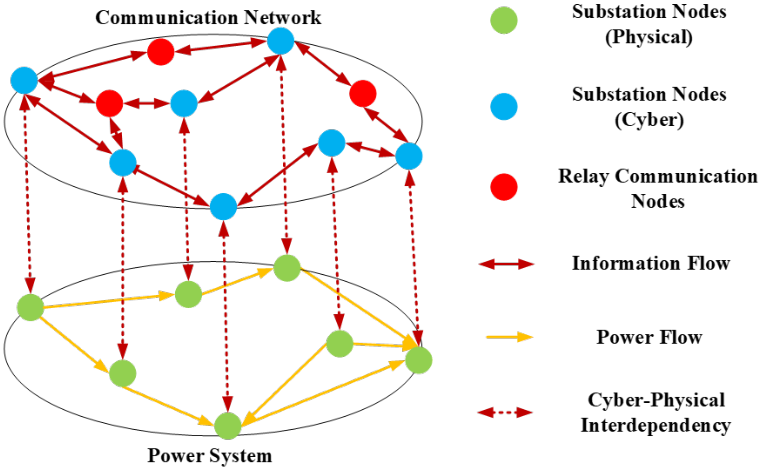


Figure 3.2 Illustration of the “Partially One-to-one” Interdependency of CPS.

As shown in Figure 3.2, we model the cyber-physical power system as an interdependent network consisting of two layers, i.e., communication network $\mathbb{G}_{\mathbb{C}}(\boldsymbol{V_C}, \boldsymbol{E_C})$ and power system $\mathbb{G}_{\mathbb{P}}(\boldsymbol{V_P}, \boldsymbol{E_P})$, where $\boldsymbol{V_C} = \{..., V_C, ...\}$, $|\boldsymbol{V_C}| = m$, $\boldsymbol{V_P} = \{..., V_P, ...\}$, $|\boldsymbol{V_P}| = n$ are the cyber/physical substation node sets of the two layers and $\boldsymbol{E_C} = \{..., E_C, ...\}$, $|\boldsymbol{E_C}| = h$, $\boldsymbol{E_P} = \{..., E_P, ...\}$, $|\boldsymbol{E_P}| = k$ are the communication/transmission edge sets of $\mathbb{G}_{\mathbb{C}}$ and $\mathbb{G}_{\mathbb{P}}$.

According to [53], the interdependencies of CPS can be divided into “one-to-one”, “one-to-multiple”, and “multiple-to-multiple” correspondences. In this chapter, we follow the typical substation communication structure from [31]. That is, the Numerical Protection Relays (NPRs), Merging Units (MUs), and Process Units (PUs) communicate through a Local Area Network (LAN) within the substation. They access the control centers through Wide Area Networks (WANs) via the routing gateways in the substations and relay communication nodes. Therefore, the CPS interdependency is defined as "partially one-to-one" interdependency, i.e., each physical substation node is associated with a cyber substation node, i.e., routing gateway, while not all cyber nodes are connected with the physical substation nodes.

3.2.2. Graph-CPS: Generating Synthetic Topology of Cyber-Physical Power Systems

For an input network $\boldsymbol{G} = \{\boldsymbol{A}, \boldsymbol{X}, \boldsymbol{E}\}$, $\boldsymbol{A}$ is the adjacent matrix of the network, $\boldsymbol{X} = \{(x_t, x_i)|t, i = 1, 2, 3, ...\}$ is the node attribute set of all nodes, and $\boldsymbol{E} = \{e_j|j = 1, 2, 3, ...\}$ is the edge attribute set of all edges. x_t represents the node types. In this chapter, we consider three different node types in the power system, i.e., generator load, and zero injection node, and the corresponding $x_t = -1, 1, 0$, respectively. In the communication model, we consider all nodes are substation routers. x_i is the node feature of node i while e_j is the edge feature of edge j. Note that one can perform different types of node/edge features to serve different research goals. In this chapter, we use capacity centrality to quantify the feature of the nodes in both the communication network and power system, as shown in (3.1).

$$x_i = \sum_{j \in \boldsymbol{N}_i} e_j \tag{3.1}$$

where $\boldsymbol{N}_i$ is the neighbor edge set of node i, and e_j is defined as the capacity of the edge, e.g., transmission line capacity in power system and bandwidth of communication links in cyber layer. To comprehensively capture the global network features, we covert the node attribute vector X into a probability distribution $V(x)$. When comparing the network feature distribution of the two different networks, we use the Kullback-Leibler divergence to quantify the difference between the two different probability distributions as shown in (3.2).

$$KL(V(\hat{x}) \| V(x)) = -\sum V(\hat{x}) \log \frac{V(x)}{V(\hat{x})} \tag{3.2}$$

As shown in Figure 3.3, the Graph-CPS consists of three modules, i.e., Recurrent Neural Network (RNN), Variational Autoencoder (VAE), and Network Feature Reconstruction (NFR). The RNN and VAE modules generate the synthetic CPS topology and network features, separately. Then, the NFR module integrates the generated data and forms the new synthetic network. The RNN module, leveraging the strong sequential modeling capabilities of GraphRNN, recurrently and cooperatively generates the synthetic network topology at both the node and edge levels. GraphRNN can generate

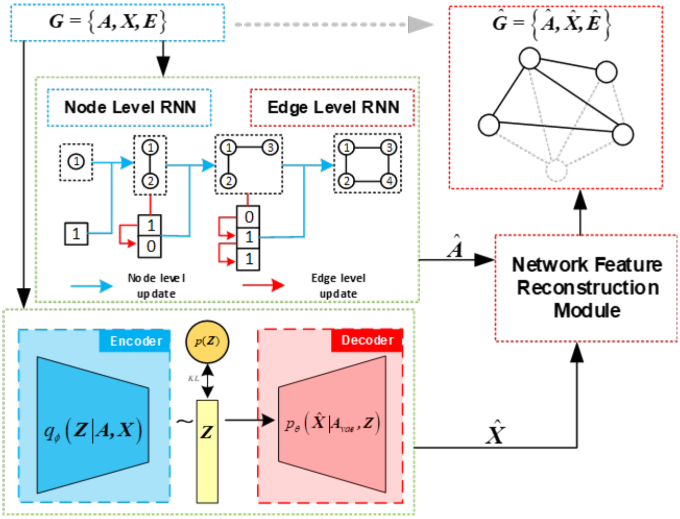


Figure 3.3 Framework of Graph-CPS.

graphs with various topological properties such as different sizes, degree distributions, clustering coefficients, etc. Besides, GraphRNN is capable of accurately capturing both global and local graph properties [17]. Meanwhile, the VAE module employs VAE's powerful probabilistic modeling and latent space representation to generate diverse and realistic network features [18]. The RNN module sequentially and recurrently generates the synthetic network topology by cooperatively using two RNNs, i.e., node level RNN and edge level RNN. Both of the RNNs consist of state-transition function and an output function as in (3.3)-(3.4).

$$h_o = f_{trans}(h_{o-1}, S^{\pi}_{o-1}) \tag{3.3}$$

$$\theta_o = f_{out}(h_o) \tag{3.4}$$

where h_o encodes the generated graph of current time step, and S^{π}_{o-1} is the adjacency vector for the $o-1$ nodes of last time step. θ_o indicates the distribution of binary adjacency vector for node o. f_{trans} and f_{out} can be arbitrary neural networks. For more details of RNN modeling, readers are referred to [17]. As in Figure 3.3, the output of RNN module is the synthetic topology $\hat{\boldsymbol{A}}$.

The encoder of VAE module takes $\boldsymbol{A}$ and $\boldsymbol{X}$ as inputs, and it uses a two-layer Graph Convolutional Network (GCN) to project the inputs into the latent space $\boldsymbol{Z}$, which is expressed in (3.5).

$$q_\phi(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{A}) = \prod\nolimits_{r=1}^{N} q_\phi(Z_r|\boldsymbol{X},\boldsymbol{A}) \tag{3.5}$$

For the detailed definition of the two-layer GCN, readers are referred to [18]. The latent space $\boldsymbol{Z}$ is regularized by a simplistic isotropic Gaussian prior $P(\boldsymbol{Z}) = N(0, I)$. The decoder is also a two-layer GCN which takes $\boldsymbol{Z}$ and $\boldsymbol{A}_{voe}$ as inputs. $\boldsymbol{A}_{voe}$ is the result of the inner-product [18] sampling from $\boldsymbol{Z}$. Then, the generated node attribute $\hat{\boldsymbol{X}}$ is calculated as shown in (3.6)-(3.7).

$$P_\theta\left(\hat{\boldsymbol{X}}|\boldsymbol{A}_{vae},\boldsymbol{Z}\right) = \prod\nolimits_{i=1}^{N} P_\theta\left(\hat{x}_i,\hat{x}_t|\boldsymbol{A}_{vae},\boldsymbol{Z}\right) \tag{3.6}$$

$$P_\theta\left(\hat{x}_i,\hat{x}_t|\boldsymbol{A}_{vae},\boldsymbol{Z}\right) = \mathcal{N}\left(\hat{x}_i,\hat{x}_t|\boldsymbol{\mu}_i,\operatorname{diag}(\boldsymbol{\sigma}_i^2)\right) \tag{3.7}$$

where $\boldsymbol{\mu} = GCN_\mu(\boldsymbol{A}_{vae},\boldsymbol{Z})$ is the matrix of mean vectors $\boldsymbol{\mu}_i$ and similarly $log\boldsymbol{\sigma} = GCN_\sigma(\boldsymbol{A}_{vae},\boldsymbol{Z})$. The GCN in the decoder is defined as $GCN(\boldsymbol{A}_{vae},\boldsymbol{Z}) = \boldsymbol{A}^{'}_{vae}ReLU(\boldsymbol{A}^{'}_{vae}\boldsymbol{X}\boldsymbol{W}_0)\boldsymbol{W}_1$, where $\boldsymbol{W}_0$ and $\boldsymbol{W}_1$ are the trained parameters. $ReLU(\boldsymbol{\cdot}) = max(0,\cdot)$ and $\boldsymbol{A}^{'}_{vae} = \boldsymbol{D}^{-1/2}\boldsymbol{A}_{vae}\boldsymbol{D}^{-1/2}$ is the symmetrically normalized adjacency matrix. $\boldsymbol{D}$ is the degree matrix of $\boldsymbol{A}_{vae}$.

The goal of the proposed method is to generate synthetic networks with consistent network feature distribution to the input graph. Therefore, during the training process, we consider the equation (3.2) and minimize the variational upper bound $\mathcal{L}$ as shown in equation (3.8).

$$\mathcal{L} = \mathbb{E}_{q_\phi(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{A})}\left[-\log_{P_\theta}(\boldsymbol{A}_{vae}|\boldsymbol{Z})\right] + KL\left[q_\phi(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{A})\|P(\boldsymbol{Z})\right] + KL\left[V\left(\hat{x}\right)\|V\left(x\right)\right] \tag{3.8}$$

After the RNN and VAE modules, $\hat{\boldsymbol{A}}$ and $\hat{\boldsymbol{X}}$ are obtained. In the NFR module, we use Algorithm 3.1 to map node attribute obtained. In the NFR module, we use Algorithm 3.1 to map node attribute $\hat{\boldsymbol{X}}$ to$\hat{\boldsymbol{A}}$ and reconstruct the edge attribute $\hat{\boldsymbol{E}}$. Note that when mapping $\hat{\boldsymbol{X}}$ to $\hat{\boldsymbol{A}}$, we assume that the nodes with higher degree have higher node attribute. In Algorithm 3.1, $\boldsymbol{V}_C/\boldsymbol{V}_P$ is the node set for cyber layer and physical layer. N_i^0 is the neighbor edge set of node $\hat{x}_i$ whose $\hat{e_j} = 0$ and $Re(\hat{x}_i)$ is the remaining node attribute of $\hat{x}_i$ that is not assigned to any edge yet. Initially, $Re(\hat{x}_i) = \hat{x}_i$.

Based on [17], [18] and Figure 3.3, one can derive that both GraphRNN and GraphVAE use an encoder to learn a distribution $P_{model}(G)$ based on the input data, which is stored in the latent space. Then, the decoder will interpret $P_{model}(G)$ by sampling from the latent space and generate the output graphs, where the sampling is random but constrained by $P_{model}(G)$. Therefore, if one wants to back solve from the output and obtain the exact real input data, at least the following information is needed: (1) exact sampling probabilities used by our method to generate the synthetic network,

(2) exact learned parameters of the encoder, and (3) learned distribution. Note that for condition (1), each generation is an independent event with different random probabilities and thus is inaccessible. Also, conditions (2) and (3) are unfeasible without condition (1). Although the adversaries may use brute force to back solve from the output data, it is still unfeasible to back solve the model because: (i) in CPS minor differences in network topology and node/edge attributes leads to different power flow results, and (ii) the adversaries do not know the real CPS. It means they have no reference and cannot control the difference between their back solving results and real CPS, which leads back to issue (i). Therefore, to the best knowledge of the authors, it is unlikely to back solve the generation process with only knowing the generated synthetic network.

Algorithm 3.1: Network feature reconstruction module

Input: Generated adjacent matrix $\hat{\boldsymbol{A}}$ and node attributes $\hat{\boldsymbol{X}}$

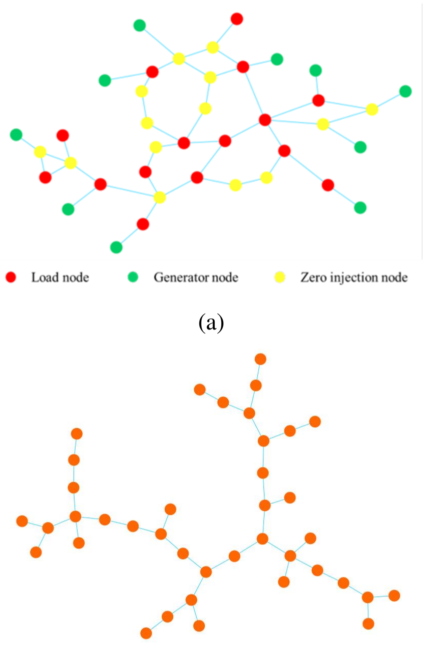
Output: $\hat{\boldsymbol{G}} = \{\hat{\boldsymbol{A}}, \hat{\boldsymbol{X}}, \hat{\boldsymbol{E}}\}$

```
Step 1   $\hat{\boldsymbol{E}} \leftarrow 0$
Step 2   Sort $\hat{\boldsymbol{X}}$ in descending order
Step 3   Sort $\boldsymbol{V}_C/\boldsymbol{V}_P$ in degree descending order based on $\hat{\boldsymbol{A}}$
Step 4   Assign $\hat{\boldsymbol{X}}$ to $\boldsymbol{V}_C/\boldsymbol{V}_P$
Step 5   Locate the node $\hat{x}_i$ with the smallest degree
Step 6   **For** $j \in N_i^0$ **do:**
Step 7       $\hat{e_j} = Re(\hat{x}_i)/|N_i^0|$
Step 8       Update $Re(\hat{x}_i)$
Step 9   **End For**
Step 10  **Repeat** Step 5-8 until all $\hat{e}_j > 0$
Step 11  **Return** $\hat{\boldsymbol{G}} = \{\hat{\boldsymbol{A}}, \hat{\boldsymbol{X}}, \hat{\boldsymbol{E}}\}$
```

3.2.3. Implementation and Evaluation of Graph-CPS

In this Section, we implement the proposed Graph-CPS on three power systems and three power grid communication networks to demonstrate and assess the model effectiveness and scalability. For physical layer, we used the IEEE 39-bus standard test system, Italian and German transmission systems (380kV- 400kV), the European continental power grids [38]. For cyber layer, we use the communication network for Jiangsu province power grids in China [41] and two validated communication networks for IEEE 39-bus, 118-bus system, respectively [44], [43]. The size of the networks mentioned above were scaled from 18 nodes to 1225 nodes and the networks contain both IEEE standard test systems and the real systems.

Table 3.1 provides the statistical comparison between real and synthetic CPS. From the topological perspective, we evaluate the quality of the generated synthetic network based on multiple complex network parameters, i.e., average node degree, average shortest path length, network diameter, network density, average and maximum node betweenness. These parameters reflect the global structural characteristics of a network. From the perspective of network features, we evaluate the generation quality by comparing the mean value and the variance of the normalized generated features. Based on Table 3.1, one can observe that all generated parameters have small differences compared with the original networks. Therefore, it is proved that the Graph-CPS is scalable and accurate to preserve the characteristics of input networks with not only different network types, i.e., power and communication networks, but also different network sizes.



(a)

(b)

Figure 3.4 (a) Generated Synthetic Power Topology for IEEE 39-bus System, (b) Generated Synthetic Communication Network.

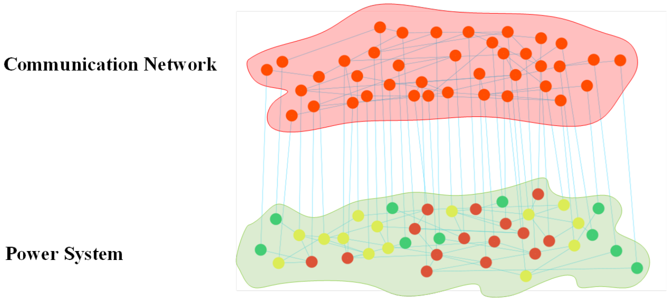
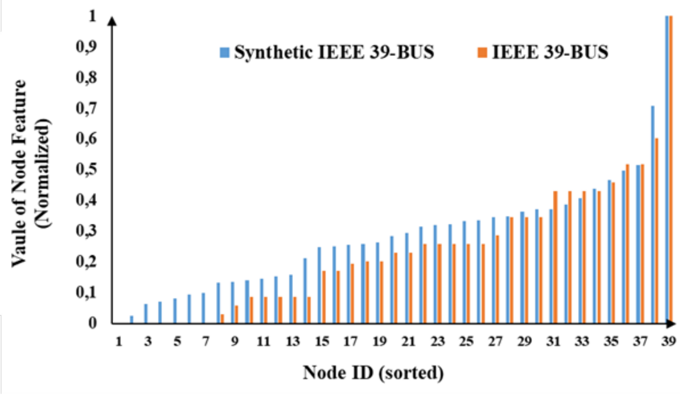


Figure 3.5 Generated Synthetic CPS for IEEE 39-bus System.

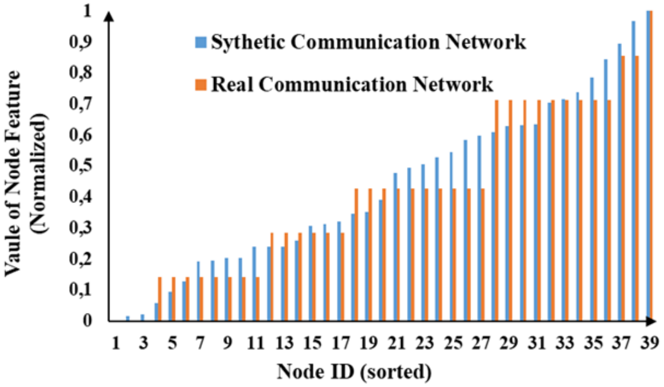
Table 3.1: Statistical Comparison between Real and Synthetic CPS

	Synth. IEEE 39-bus power syst.	**Synth. Italian power syst.**	**Synth. German power syst.**	**Synth. European power syst.**	**Synth. IEEE 39-bus comm. syst.**	**Synth. Jiangsu comm. syst.**	**Synth. 128 nodes comm. syst.**
N	39 (0)	151 (0)	289 (0)	1226 (-1)	39 (0)	18 (0)	128 (0)
L	46 (-1)	192 (+4)	345 (+10)	1598 (-11)	38 (0)	29 (0)	160 (+4)
$\langle k \rangle$	2.308 (+0.051)	2.543 (+0.053)	2.388 (+0.069)	2.602 (-0.016)	1.949 (-0.001)	3.222 (+0.001)	2.481 (+0.082)
$\langle l \rangle$	4.761 (-0.012)	9.731 (-0.205)	11.756 (-0.209)	23.396 (-0.186)	6.874 (0.125)	2.523 (-0.020)	6.654 (+0.097)
d	11 (-1)	28 (-1)	30 (-2)	63 (-4)	16 (0)	5 (0)	16 (-2)
D	0.062 (-0.02)	0.016 (+0.0004)	0.008 (+0.0003)	0.002 (0)	0.051 (0)	0.189 (0)	0.019 (+0.0004)
$\langle bv \rangle$	0.102 (-0.001)	0.058 (-0.001)	0.038 (-0.001)	0.018 (0)	0.159 (0.003)	0.095 (-0.002)	0.045 (+0.001)
$\max \langle bc \rangle$	0.494 (-0.018)	0.318 (+0.073)	0.393 (-0.099)	0.236 (+0.061)	0.596 (0.043)	0.334 (-0.002)	0.541 (+0.063)
$\bar{\hat{x}}_i$	0.287 (-0.048)	0.348 (-0.109)	0.387 (-0.119)	0.314 (-0.033)	0.436 (-0.018)	0.490 (-0.028)	0.516 (-0.094)
$Var(\hat{\boldsymbol{X}}_i)$	0.036 (+0.008)	0.036 (+0.008)	0.038 (+0.007)	0.048 (-0.003)	0.074 (-0.003)	0.118 (-0.002)	0.061 (+0.009)

N: number of nodes, L: number of edges, $\langle k \rangle$: average node degree, $\langle l \rangle$: average shortest path length, d: network diameter, D: network density, $D = L/N$, $\langle bv \rangle$: average node betweenness centrality, $\max \langle bc \rangle$: maximum betweenness centrality, $Var(\hat{\boldsymbol{X}}_i)$: the mean value of the normalized synthetic node features, $Var(\hat{\boldsymbol{X}}_i)$: the variance of the normalized synthetic node features, $(*)$: the number in the brackets represents the difference between synthetic networks and original networks



(a)



(b)

Figure 3.6 (a) Comparison of Node Feature for IEEE 39-Bus System, (b) Comparison of Node Feature for IEEE 39-Bus Communication System.

To better present the generation results, we give a more detailed study case for IEEE 39-bus system and its communication model. The generation results are given as shown in Figure 3.4. Then, we form the interdependency for the synthetic CPS by following the "degree-to-degree" principle in [25] as shown in Figure 3.5. In Figure 3.4(a), the numbers of load, generator, and zero injection nodes are 15, 10, 14, respectively. In IEEE 39-bus system, the numbers are 17, 10, 12, which have the close distribution of node type. Besides, in Figure 3.4(b), the synthetic communication network has a clear tree structure as the input communication networks does, and it proves that our method can effectively learn the global structure characteristics of the input network. Moreover, we compare and visualize the generated node features as shown in Figure 3.6. In Figure 3.6(a), the mean value (normalized) of the node features in IEEE 39-bus system is 0.239, while in synthetic result the value is 0.287. In Figure 3.6(b), the mean value (normalized) of the node features in real communication model is 0.418, while in synthetic result the value is 0.436. Meanwhile, the difference of the variances for two networks are 0.008 and 0.003, respectively. Therefore, it proves that the Graph-CPS can generate realistic synthetic network features. Therefore, the experimental results prove that Graph-CPS is capable of capturing both the different topological statistics and the network feature distribution of the original networks.

3.3. SibGen: A Hybrid Generator for Generating the Digital Sibling of Cyber-Physical Power System

The rising demand for advanced research on CPS compels the creation of realistic and reliable test systems. However, the national security concerns prevent the public sharing of real Critical Infrastructure (CI) data and models, including the CPS data. Therefore, it is desirable to have a realistic projection of the real system, which conceals the sensitive system model and data, e.g., system topologies and parameter configurations, while maintaining the similar global characteristics as the real systems. To this end, the concept of Digital Siblings (DSs) is proposed as a potential and promising solution.

Similar to Digital Twins (DTs) [59], DSs serve as virtual representations of physical systems that reflect their real-world counterparts in a digital environment. However, the primary distinction between DSs and DTs lies in their objectives: while DTs strive for absolute consistency with the real network, DSs focus on capturing the overall characteristics of the target system. Metaphorically, DSs share the same ”genetic materials” as DTs but display different ”traits”, which are critical for safeguarding the confidentiality of CPS models and data. Given this background, developing an effective and reliable generative model to create realistic DSs without exposing real CPS models and data is of paramount importance.

As discussed in 3.1, existing generative models commonly used to create synthetic networks for real-world systems can be classified into three categories, i.e., complex network-based methods, open access data-based methods, and machine learning-based methods. However, both complex network-based and open access data-based methods fail to comprehensively reflect the operational and behavioral characteristics of real CPS. While complex network-based methods primarily focus on topology, open-source data-based methods rely on external assumptions, leading to potential deviations from the actual system behavior. To overcome these limitations, machine learning-based methods have emerged as a promising solution. By learning the characteristic distributions of real-world systems, these methods can generate DSs that accurately represent real-world CPS characteristics while preserving data confidentiality. However, how to comprehensively and accurately learn the CPS characteristics and output realistic DS are challenging: (1) Heterogeneous graph learning [60]. In graph theory, the CPS can be interpreted as a heterogeneous graph that contains diverse types of nodes and edges. This diversity makes it challenging to define a unified approach for message passing or feature aggregation across the graph during the learning process. (2) Inductive learning [61]. To preserve the data confidentiality of CPS, we need to transform the real system topology and parameters into a new shape based on the learned network information. This is an inductive learning process, which poses great challenges to learn the complex and high-dimensional representations of the input CPS.

Based on the discussion above, Graph Recurrent Neural Networks (GraphRNN) [17] and Graph Attention Networks (GATs) [62] become promising approaches to address the proposed question. From the perspective of topology generation, GraphRNN can generate graphs with various topological properties such as different sizes, degree distributions, clustering coefficients, etc. Besides, GraphRNN is capable of accurately capturing both global and local graph property [17]. From the perspective of operational parameters generation, GATs successfully address the two challenges mentioned above. By leveraging attention mechanisms [63], GATs can handle situations where different nodes or edges in the graph have varying levels of influence, making it well-suited for heterogeneous graphs. Besides, the multihead attention strategy [62] in GATs enhances its representation ability and thus increases the model performance on understanding the complex input data. Furthermore, unlike traditional graph embedding methods that rely on predefined global structures (e.g., adjacency matrices), GATs operate with local node features and neighborhood information, making it more suitable for inductive learning tasks. Therefore, in this paper, A hybrid generator, namely SibGen, is proposed to create a digital sibling of real CPS by generating both topological and operational models. The main contributions of this section are summarized as follows:

1. This section proposes SibGen, a hybrid digital sibling generator built upon GraphRNN and GAT. SibGen generates digital siblings that replicate the topological characteristics and operational behaviors of real CPS, while exhibiting distinct structural and operational configurations, thereby ensuring the protection of sensitive CPS data.

2. To address the edge feature generation limitation in existing models, the concept of dual graph is introduced to enable SibGen to generate synthetic attributes for both nodes and edges, thereby capturing CPS characteristics in a more comprehensive manner.

3. To effectively incorporate physical constraints of CPS into the generation process, the prior knowledge-constrained training strategy is proposed to enhance the realism and reliability of the digital sibling's topological and operational representations.

4. To thoroughly evaluate the similarities between the input real CPS and the generated digital sibling, 7 metrics are implemented to evaluate both cyber and physical systems from the both the topological and operational perspectives.

3.3.1. The Interdependent Graph Model of Cyber-Physical Power System

In Section 3.2.3, the CPS modelling only considers the topological aspects of the system. However, a complete synthetic CPS requires not only the topology information but also the operational models of each subnetwork. Therefore, in this chapter, the cyber-physical power system is mapped into an interdependent graph consisting of two heterogeneous subnetworks, i.e., communication network and power system, each of which is denoted as $\boldsymbol{G}_u = (\boldsymbol{V}_u, \boldsymbol{E}_u, \boldsymbol{v}_{ua}, \boldsymbol{e}_{ua})$, where $\boldsymbol{V}_u$ represents the node set of $\boldsymbol{G}_u$, $\boldsymbol{E}_u$ represents the edge set of $\boldsymbol{G}_u$, $\boldsymbol{v}_{ua}$ represents the node attributes of $\boldsymbol{V}_u$. The sub-index $u \in \{p, c\}$ represents the power system or communication network, respectively. The detailed system models of the two subnetworks are given as follows.

Communication Network: $\boldsymbol{V}_c = \{\boldsymbol{V}_{ctrl}, \boldsymbol{V}_{cn}\}, |\boldsymbol{V}_c| = N_{vc}$, where $\boldsymbol{V}_{ctrl}$, $\boldsymbol{V}_{cn}$ represent the control centers and communication nodes, respectively. N_{vc} is the number of communication nodes. $\boldsymbol{E}_c = \{..., E_c, ...\}, |\boldsymbol{E}_c| = N_{ec}$, where E_c is the communication link and N_{ec} is the number of communication links. Besides, $\boldsymbol{v}_{ca} = \{\boldsymbol{P}_c\}$ includes the node processing capacity $\boldsymbol{P}_c$ while $\boldsymbol{e}_{ca} = \{\boldsymbol{b}, \boldsymbol{L}\}$ contains the bandwidth $\boldsymbol{b}$ and the communication media latency $\boldsymbol{L}$.

Power System: $\boldsymbol{V}_p = \{\boldsymbol{V}_{pg}, \boldsymbol{V}_{pz}, \boldsymbol{V}_{pl}\}, |\boldsymbol{V}_p| = N_{vp}$, where $\boldsymbol{V}_{pg}$, $\boldsymbol{V}_{pz}$, $\boldsymbol{V}_{pl}$ represent the sets of generator, zero injection, and load nodes, respectively. N_{vp} is the number of nodes in the power system. $\boldsymbol{E}_p = \{..., E_p, ...\}, |\boldsymbol{E}_p| = N_{ep}$, where E_p is the branch and N_{ep} is the number of branches. N_{ep} is the number of edges in power system. Besides, $\boldsymbol{v}_{pa} = \{\boldsymbol{P}_d, \boldsymbol{P}_g\}$ includes the active power demand $\boldsymbol{P}_d$, and active power output $\boldsymbol{P}_g$ while $\boldsymbol{e}_{pa} = \{\boldsymbol{X}, \boldsymbol{C}_p\}$ contains the reactance $\boldsymbol{X}$ and branch capacitance $\boldsymbol{C}_{\boldsymbol{p}}$.

Based on the discussion above, the cyber-physical power system is denoted as $\mathfrak{S} = (\boldsymbol{G}_c, \boldsymbol{G}_p, \boldsymbol{\psi})$, where $\boldsymbol{\psi}$ represents the mapping relationship between the communication network and power system. In this chapter, we employ the "partially one-to-one" mapping relationship as described in references [52].

3.3.2. Operational Models of Cyber and Physical System Layers

In this section, the operational models are introduced, which are implemented in both the cyber and physical system layers. For the communication network, the OT communication network model is used based on our previous work in [64] as represented in Figure 3.7. For the power system, the DC power flow model is implemented.

Communication Network: the OT communication network in a cyber-physical power system is used to monitor and control the physical processes of the power grid in real-time. The OT communication network is critical for ensuring the reliability of power system operation. As shown in Figure 3.7, the access layer between the OT communication network and power system are the Intelligent Electronic Devices (IEDs), comprising of Numerical Protection Relays (NPRs), Merging Units (MUs), and Process Unit (PUs). These devices are used for the monitoring, protection, and control of the power grid. The bidirectional communication between IEDs and control center is achieved through the substation switches, gateway, and Wide Area Networks (WANs), where the Centralized Control and Protection systems (CPC) are implemented

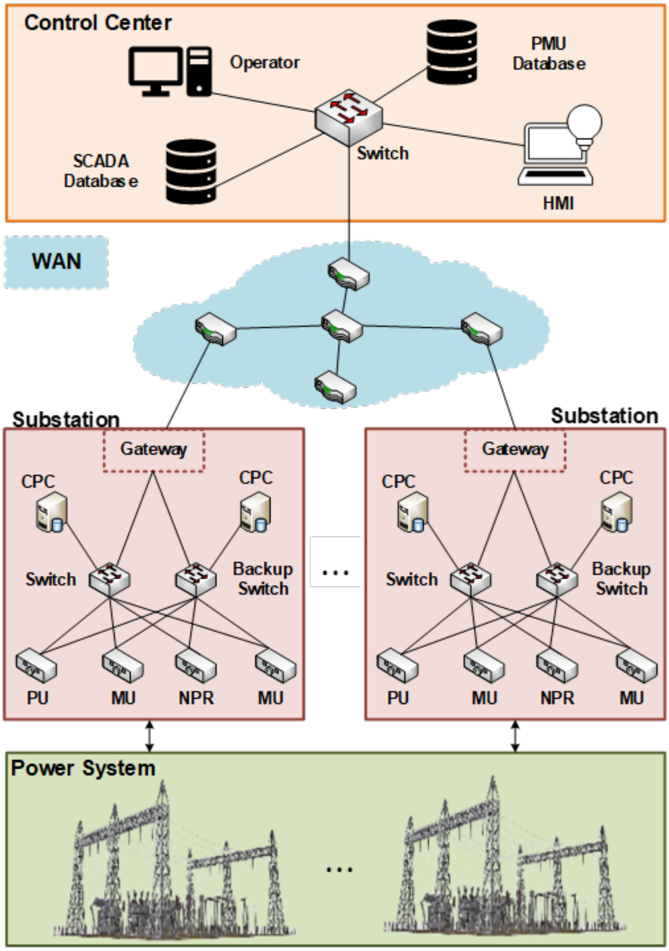


Figure 3.7 The Architecture of OT Communication Network Model.
(PMU: phasor measurement unit, HMI: human machine interface, CPC: centralized protection and control, NPR: numerical protection relay, MU: merging unit, PU: process unit)

to facilitate the system operation. Eventually, all the collected measurements and data are processed and analyzed in the control center. System operators send from the control center command and control messages to the substations. In the OT communication network, the wide area communications are simulated between substations and control centers. The communication links possess inherent attributes that reflect the real communication infrastructure, including factors such as latency and bandwidth capacity. The communication link design is determined upon the most optimal media available, utilizing fiber optics with a bandwidth range of 1-5 Gbps and a latency of 0.5ms per 100 km [65]. The proposed model incorporates Transmission Control Protocol (TCP) and User Datagram Protocol (UDP) for transmitting information over a wide area network. TCP and UDP serve as transport protocols for transmitting measurement and control data in power grid OT communication networks [66]. Then, the network performance is evaluated by using the Round-Trip Time (RTT) R_{tt} and packet loss rate η, which are calculated by the equations (3.9)-(3.10).

$$R_{tt} = \sum_{c \in \boldsymbol{p}} L_c + \sum_{c \in \boldsymbol{p}} \frac{D_s}{b_c} + t_p + t_q \tag{3.9}$$

$$\eta = \frac{\sum\limits_{t=t_0}^{T_e} \sum\limits_{i=1}^{N_i} L_i\left(t\right) + \sum\limits_{t=t_0}^{T_e} \sum\limits_{j=1}^{N_j} L_j\left(t\right)}{\sum\limits_{t=t_0}^{T_e} \sum\limits_{i=1}^{N_i} \varphi_i^t + \sum\limits_{t=t_0}^{T_e} \sum\limits_{j=1}^{N_j} \varphi_j^t} \tag{3.10}$$

where R_{tt} contains the propagation latency $\sum_{c \in \boldsymbol{p}} L_c$, transmission delay $\sum_{c \in \boldsymbol{p}} D_s/b_c$, processing delay t_p and queuing delay t_q. L_c is the latency of a communication link E_c, $\boldsymbol{p}$ is the set of communication links on a communication path. D_s is the size of the data packet, b_c is the bandwidth of a communication link E_c. $L_i(t)$ is the number of lost data packets of node i at t, φ_i^t is the data packet generated by node i at t, T_e is the total communication time.

Power System: In this chapter, we implement the DC power flow model [67] in the power system, which calculates the bus voltage angles $\boldsymbol{\theta}$, by using the known active power injection $\boldsymbol{P}_{inj}$ and susceptance $\boldsymbol{B}$. It can be calculated as in (3.11).

$$\boldsymbol{\theta} = -\boldsymbol{B}^{-1} \boldsymbol{P}_{inj} \tag{3.11}$$

Then, using the calculated $\boldsymbol{\theta}$, one can traverse all the branches in the power system and use equation (3.12) to calculate the active power flow for each branch.

$$P_{mn} = \frac{\theta_m - \theta_n}{X_{mn}} \tag{3.12}$$

In this section, the generated synthetic power system model is validated by comparing it with the input network. To thoroughly analyze the similarity of two networks, the topological and operational features are jointly considered from five perspectives, i.e., distribution of normalized node power injections $\boldsymbol{P}_{inj}^{'}(V_p)$ [68], distribution of node vulnerability features $I(V_p)$ [69], node distance degrees $D(V_p)$ [70], node capacity degrees $B(V_p)$, and N-1 contingency analysis. The calculations of six metrics are given as follows.

$$P_{inj}^{'}\left(V_p\right) = \frac{P_{inj}\left(V_p\right)}{S_B} \tag{3.13}$$

$$I\left(V_p\right)=1-\frac{\left(N_{vp}-k_p\right)W_{avg}\left(\boldsymbol{G}_u^{'}\left(V_p\right)\right)}{N_{vp}W_{avg}\left(\boldsymbol{G}_u\right)} \tag{3.14}$$

$$W_{avg}(\boldsymbol{G}_u)=\frac{1}{N_{vp}\left(N_{vp}-1\right)}\sum\nolimits_{i\neq j\in\boldsymbol{V}_p}X_{ij} \tag{3.15}$$

$$D\left(V_p\right)=\sum\nolimits_{E_p\in\boldsymbol{n}}X_{ij} \tag{3.16}$$

$$B\left(V_p\right)=\max\left(\sum\nolimits_{E_p\in\boldsymbol{n}^{+}}C_p,\sum\nolimits_{E_p\in\boldsymbol{n}^{-}}C_p\right) \tag{3.17}$$

where S_B is the system base power, $\boldsymbol{G}_u^{'}(V_p)$ is the cohesion graph [69] of node V_p, $W_{avg}(\boldsymbol{G}_u)$ is the weighted average path length, k_p is the number of neighbour nodes of V_p (including V_p). $X_{ij}\in\boldsymbol{X}$, $C_p\in\boldsymbol{C}_p$. $\boldsymbol{n}=\{\boldsymbol{n}^{+},\boldsymbol{n}^{-}\}$ is the set of neighbour edges of V_p, where $\boldsymbol{n}^{+}$ is the set of neighbour edges flowing into V_p and $\boldsymbol{n}^{-}$ is the set of neighbour edges flowing out of V_p. The $\boldsymbol{P}_{inj}^{'}(V_p)$ reflects the importance of node V_p in electricity transmission. $W_{avg}(\boldsymbol{G}_u)$ reflects the transmission performance and efficiency of a power grid. Reference [69] suggests that when the $W_{avg}(\boldsymbol{G}_u)$ decreases, the system becomes more vulnerable to cascading failures. Based on $W_{avg}(\boldsymbol{G}_u)$, $I(V_p)$ indicates the importance of a node when the system experiences cascading failures. $D(V_p)$ and $B(V_p)$ reflect the power flow transmission capabilities of nodes from the perspectives of branch reactance, branch capacitance and power flow directions, respectively. Furthermore, the synthetic model is also validated from the perspective of N-1 contingency analysis. Each power system branch is removed and the power flow variance is observed over each branch. Then, the average power flow variance A_{pfv} of the system is used to indicate the importance of the branch, as shown in (3.18):

$$A_{pfv}\left(E_p\right)=\frac{\sum_{\boldsymbol{E}_p-\{E_p\}}\triangle P_f}{\left|\boldsymbol{E}_p-\{E_p\}\right|} \tag{3.18}$$

where $\triangle P_f$ is the power flow variance cause by the removal of branch E_p.

3.3.3. The Architecture of SibGen

The proposed framework of SibGen is illustrated in Figure 3.8. SibGen consists of two modules, i.e., Graph Recurrent Neural Network (GraphRNN) module and Graph Attention Networks (GAT) module. The objective of SibGen is to learn the characteristics of the input graph data $\boldsymbol{G}_u=(\boldsymbol{V}_u,\boldsymbol{E}_u,\boldsymbol{v}_{ua},\boldsymbol{e}_{ua})$, and output the realistic synthetic network $\widehat{\boldsymbol{G}}_u=(\widehat{\boldsymbol{V}}_u,\widehat{\boldsymbol{E}}_u,\widehat{\boldsymbol{v}}_{ua},\widehat{\boldsymbol{e}}_{ua})$. In GraphRNN module, the GraphRNN algorithm is employed [17] to generate the synthetic topology of CPS, where the new nodes and edges are alternately generated based on the learned characteristics of input graph. In GAT module, a novel bi-level GAT model embedded with prior constraints is proposed to generate the network attributes for both nodes and edges, where the concept of dual graph is defined to facilitate the training process, and the multihead masked attention mechanism is implemented.

GraphRNN module: The GraphRNN module is defined as a deep autoregressive model and follows the essential philosophy of the Recurrent Neural Network (RNN). It sequentially generates the new network topology by alternately using node-level RNN and edge-level RNN as shown in Figure 3.8. The GraphRNN module is defined as in equations (3.19)-(3.23).

$$\boldsymbol{S}^{\pi}=f_s(\boldsymbol{G}_u,\pi)=\left\{S_1^{\pi},\ldots,S_o^{\pi},\ldots,S_{|\boldsymbol{V}_u|}^{\pi}\right\} \tag{3.19}$$

$$\boldsymbol{G}_u=f_G(\boldsymbol{S}^{\pi}) \tag{3.20}$$

$$p\left(\boldsymbol{G}_u\right)=\sum\nolimits_{\boldsymbol{S}^{\pi}} p(\boldsymbol{S}^{\pi})\prod\left[f_G(\boldsymbol{S}^{\pi})=\boldsymbol{G}_u\right] \tag{3.21}$$

$$p\left(\boldsymbol{S}^{\pi}\right)=\prod\nolimits_{o=1}^{|\boldsymbol{V}_u|+1} p\left(S_o^{\pi}\mid S_1^{\pi},\ldots,S_{o-1}^{\pi}\right) \tag{3.22}$$

$$p\left(S_o^{\pi}\mid S_{<o}^{\pi}\right)=\prod\nolimits_{l=1}^{o-1} p\left(S_{o,l}^{\pi}\mid S_{o,<l}^{\pi},S_{<o}^{\pi}\right) \tag{3.23}$$

where $\boldsymbol{S^{\pi}}$ represents the sequence of graph $\boldsymbol{G}_u$, $f_s(*)$ and $f_G(*)$ represents the mapping relationships between the given graph $\boldsymbol{G}_u$ and its corresponding sequence $\boldsymbol{S}^{\pi}$, and π is the given node ordering. We denote Π as the set of all $|\boldsymbol{V}_u|!$ possible node permutations. $S_{<o}^{\pi}$ is the set of sequences that the node identifier is less than o, and $S_{o,<l}^{\pi}$ is the set of sequences that the edge identifier is less than l. For a given undirected input graph $\boldsymbol{G}_u=(\boldsymbol{V}_u,\boldsymbol{E}_u,\boldsymbol{v}_{ua},\boldsymbol{e}_{ua})$, one can use equation (3.19) and (3.20) to map it into a specific sequence $\boldsymbol{S}^{\pi}$ with a given node ordering π. Then, according to equation (3.21), the likelihood of a graph $\boldsymbol{G}_u$ is represented by the likelihood of the sequence. Equation (3.22) and (3.23) represent the node-level and edge level RNN, respectively. Combining the process above, one can learn the characteristics of the input graph $\boldsymbol{G}_u$ and sequentially generate the synthetic topology $\widehat{\boldsymbol{G}}_u=(\widehat{\boldsymbol{V}}_u,\widehat{\boldsymbol{E}}_u)$. Compared with traditional graph topology generation algorithms, GraphRNN is proved to have better performance on the graph statistics in terms of degrees, clustering coefficients and orbit counts [71]. Besides, thorough Maximum Mean Discrepancy (MMD) evaluations are conducted on various datasets, and it proves that GraphRNN achieves 80% decrease of MMD on average compared with traditional baselines and 90% decrease of MMD on average compared with deep learning baselines [17].

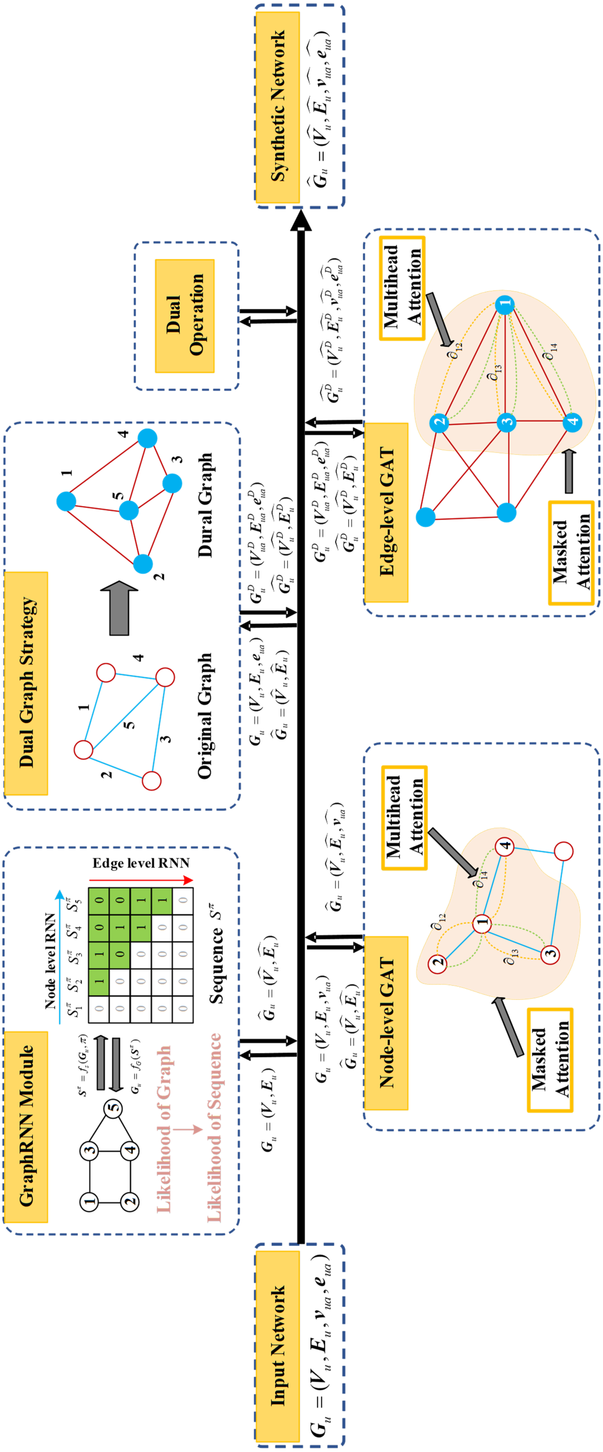

Input Network
$G_u = (V_u, E_u, v_{ua}, e_{ua})$
GraphRNN Module
Likelihood of Graph
Likelihood of Sequence
$S^\pi = f_S(G_u, \pi)$
$G_u = f_G(S^\pi)$
Node level RNN
Edge level RNN
Sequence S^π
$G_u = (V_u, E_u)$
$\hat{G}_u = (\hat{V}_u, \hat{E}_u)$
Node-level GAT
Multihead Attention
Masked Attention
∂_{12}
∂_{13}
∂_{14}
$G_u = (V_u, E_u, v_{ua})$
$\hat{G}_u = (\hat{V}_u, \hat{E}_u, v_{ua})$
Dual Graph Strategy
Original Graph
Dural Graph
$G_u = (V_u, E_u, e_{ua})$
$\hat{G}_u = (\hat{V}_u, \hat{E}_u)$
$G^D_u = (V^D_u, E^D_{ua}, e^D_{ua})$
$\hat{G}^D_u = (\hat{V}^D_u, \hat{E}^D_u)$
Edge-level GAT
$\hat{G}^D_u = (\hat{V}^D_u, \hat{E}^D_u, v^D_{ua}, \hat{e}^D_{ua})$
Dual Operation
Synthetic Network
$\hat{G}_u = (\hat{V}_u, \hat{E}_u, \hat{v}_{ua}, \hat{e}_{ua})$

Figure 3.8 SibGen Framework.

Graph attention network module: the graph attention network (GAT) module first trains on the input data $\boldsymbol{G}_u$ and then generates the synthetic node and edge attributes based on the generated synthetic topology $\widehat{\boldsymbol{G}}_u$ from the GraphRNN module. In the single graph attention layer, the input is a node attribute vector set $\boldsymbol{h} = \left\{\overrightarrow{h_1}, ..., \overrightarrow{h_k}, ..., \overrightarrow{h_N}\right\}$, and the output is the synthetic node attribute vector set $\boldsymbol{h}' = \left\{\overrightarrow{h'_1}, ..., \overrightarrow{h'_k}, ..., \overrightarrow{h'_N}\right\}$. The detailed process is presented as in equations (3.24)-(3.27).

$$\ell_{kl} = a\left(\boldsymbol{W}\overrightarrow{h_k}, \boldsymbol{W}\overrightarrow{h_l}\right) \tag{3.24}$$

$$\partial_{kl} = \mathrm{softmax}_l\left(\ell_{kl}\right) = \frac{\exp\left(\ell_{kl}\right)}{\sum_{m\in\mathbb{N}_k}\exp\left(\ell_{km}\right)} \tag{3.25}$$

$$\partial_{kl} = \frac{\exp\left(\mathrm{LeakyReLU}\left(\overrightarrow{\boldsymbol{a}}^T\left[\boldsymbol{W}\overrightarrow{h_k} \parallel \boldsymbol{W}\overrightarrow{h_l}\right]\right)\right)}{\sum_{m\in\mathbb{N}_k}\exp\left(\mathrm{LeakyReLU}\left(\overrightarrow{\boldsymbol{a}}^T\left[\boldsymbol{W}\overrightarrow{h_k} \parallel \boldsymbol{W}\overrightarrow{h_m}\right]\right)\right)} \tag{3.26}$$

$$\overrightarrow{h'_k} = \sigma\left(\sum\nolimits_{l\in\mathbb{N}_k}\partial_{kl}\boldsymbol{W}\overrightarrow{h_l}\right) \tag{3.27}$$

where $\overrightarrow{h_k} \in \mathbb{R}^F$, $\overrightarrow{h'_k} \in \mathbb{R}^{F'}$, and F represents the dimension of the feature vector. a is a single-layer feedforward neural network and is defined as $\mathbb{R}^F \times \mathbb{R}^{F'} \to \mathbb{R}$, $\boldsymbol{W} \in \mathbb{R}^{F' \times F}$ is a weight matrix shared by all $\overrightarrow{h_k}$. ℓ_{kl} represents the importance of node l to k, i.e., attention coefficients. ∂_{kl} represents the normalized ℓ_{kl}. $\overrightarrow{\boldsymbol{a}}^T \in \mathbb{R}^{2F'}$ is a weight vector to parametrize a. $\mathbb{N}_k$ is the neighbour set of nodes k. T is the transposition and $||$ is the concatenation operation. σ is a nonlinearity. Equation (3.24) describes the process of self-attention mechanism [63], which transforms the input features into higher-level features and computes the attention coefficient. In (3.25), the $softmax$ function [72] is used to normalize the ℓ_{kl} so that the attention coefficients of different nodes are comparable. Meanwhile, the LeakyReLU nonlinearity is employed as the activation function. As shown in Figure 3.8, the GAT module distributes the attention of each node only to its neighbour nodes rather than all the nodes in the graph, which is denoted as masked attention mechanism [62]. Then, by aggregating all the calculated information, the synthetic attributes are generated. The implementation of the masked attention mechanism enables the model to focus on the neighbour nodes rather than the nodes with a long distance, which helps the model to learn better about the topological information of graph.

3.3.4. Operational Model Generation based on Dual Graph and Prior Knowledge-Constrained Training

Dual Graph: Despite that the GAT algorithm can effectively generate the synthetic network attributes, it can only process the node attributes while fails to also generate the synthetic attributes for edges. To address this issue, the dual graph strategy is proposed to transform the input network topology as well as the network attributes into a dual graph for training while maintaining the same network information. The concept of dual graph is given as follows.

Definition 1 (Dual Graph): For a given graph $\boldsymbol{G}_u$, the relationship between $\boldsymbol{G}_u$ and its dual graph $\boldsymbol{G}_u^D$ is denoted as $\boldsymbol{G}_u \triangleleft \boldsymbol{G}_u^D$, where each element of $\boldsymbol{G}_u^D = (\boldsymbol{V}_u^D, \boldsymbol{E}_u^D, \boldsymbol{v}_{ua}^D, \boldsymbol{e}_{ua}^D)$ satisfies:

$$\boldsymbol{V}_u^D = \{..., E_u, ...\}\,,\, E_u \in \boldsymbol{E}_u \tag{3.28}$$

$$\boldsymbol{E}_u^D = \{..., V_u, ...\}\,,\, V_u \in \boldsymbol{V}_u \tag{3.29}$$

$$v_{ua}^D\left(E_u\right) \in \boldsymbol{v}_{ua}^D,\, v_{ua}^D\left(E_u\right) = e_{ua}\left(E_u\right) \tag{3.30}$$

$$e_{ua}^D(V_u) \in \boldsymbol{e}_{ua}^D,\, e_{ua}^D(V_u) = v_{ua}\left(V_u\right) \tag{3.31}$$

where $e_{ua}(E_u) \in \boldsymbol{E}_u$ and $v_{ua}(V_u) \in \boldsymbol{V}_u$. It is noted that for two arbitrary nodes in $\boldsymbol{G}_u^D$, if their corresponding edges in $\boldsymbol{G}_u$ are connected to the same node, we add an edge between these two nodes in $\boldsymbol{G}_u^D$. In Figure 3.8, a detailed example is given to describe the process.

Prior knowledge-constrained training: When generating synthetic attributes for the dual graphs, the generation results must adhere to specific requirements to ensure compliance with the physical rules of cyber-physical power systems. For example, if a node $V_p \in \boldsymbol{V}_{pz}$ is a zero injection node, it suggests that the corresponding node attributes P_d and P_g are 0. Similarly, other node types also have different constraints. Therefore, the GAT structure is modified. A training strategy is proposed embedded with prior knowledge constraints. Assume $\widetilde{\boldsymbol{h}}$ is the output of the GAT layer, and $\boldsymbol{h}'$ is the adjusted output after the constraints are implemented. The relationship between $\widetilde{\boldsymbol{h}}$ and $\boldsymbol{h}'$ is given as in (3.32).

$$\boldsymbol{h}' = f_{cons}(\widetilde{\boldsymbol{h}}) \tag{3.32}$$

where $f_{cons}(*)$ represents the different constraint operation. As discussed above, the P_d and P_g are set to 0 for zero injection nodes. For generator nodes, only P_d is set to 0. For load nodes, only the P_g is set to 0. During training, the loss function L is typically defined as the error between the predicted output and true attributes. In this paper, the Mean Squared Error (MSE) is used as the loss function:

$$L = \frac{1}{N}\sum\nolimits_{k=1}^{N}\left\|\overrightarrow{h'_k} - \overrightarrow{h_k}\right\|^2 = \frac{1}{N}\sum\nolimits_{k=1}^{N}\left\|f_{cons}\left(\overrightarrow{\widetilde{h_k}}\right) - \overrightarrow{h_k}\right\|^2 \tag{3.33}$$

Given that the model is trained based on the gradient descent algorithm, the introduction of changes the gradient calculation as shown in (3.34).

$$\frac{\partial L}{\partial \boldsymbol{W}} = \frac{\partial L}{\partial \overrightarrow{h'_k}} \cdot \frac{\partial \overrightarrow{h'_k}}{\partial \overrightarrow{\widetilde{h_k}}} \cdot \frac{\partial \overrightarrow{\widetilde{h_k}}}{\partial \boldsymbol{W}} = \sum\nolimits_{x=d,g}\left(\frac{\partial L}{\partial \overrightarrow{h'_k}(P_x)} \cdot \frac{\partial \overrightarrow{h'_k}(P_x)}{\partial \overrightarrow{\widetilde{h_k}}(P_x)} \cdot \frac{\partial \overrightarrow{\widetilde{h_k}}(P_x)}{\partial \boldsymbol{W}}\right) \tag{3.34}$$

where $\partial\overrightarrow{h'_k}/\partial\overrightarrow{\widetilde{h_k}}$ captures the influence of the constraint operations on the gradient. When any value of P_d and P_g are set to 0, the corresponding $\partial\overrightarrow{h'_k}(P_x)/\partial\overrightarrow{\widetilde{h_k}}(P_x)$ is 0. It indicates that the model stops learning and updating the corresponding parameters while it does not interfere the learning and updating the items that are not set to 0. Besides, to improve the fitting ability of the model, we use the multi-head attention strategy [62], as shown in Figure 3.8 and equation (3.35).

$$\overrightarrow{h'_k} = \|_{m=1}^{M}\, \sigma\left(\sum\nolimits_{l\in\mathbb{N}_k} \partial_{kl}^{m} \boldsymbol{W}^m \overrightarrow{h_l}\right) \tag{3.35}$$

where $||$ represents the concatenation operation, ∂_{kl}^{m} represents the calculation results of $\boldsymbol{W}^m$ with the m-th head. According to [62], the multihead attention strategy enhances the expression capability of the models. It enables GAT to simultaneously focus on the different aspects of input data so that the data characteristics can be better and thoroughly captured.

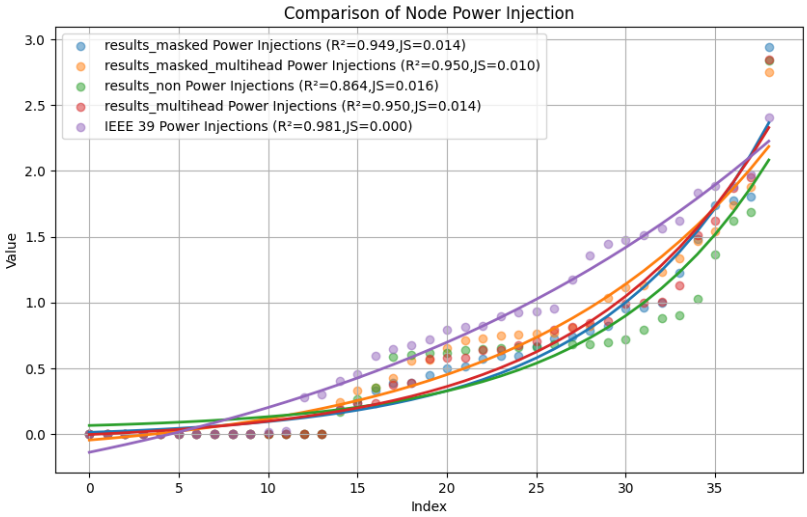
3.3.5. Implementation and Evaluation of SibGen

In this section, the proposed SibGen is implemented on the IEEE 39-bus system and its validated OT communication network in [44]. The interdependency between cyber and physical system layers follows the principle of "degree-to-degree" [25]. The methods are coded in Python and power system simulations are run on a computer equipped with an Intel i7-8750H CPU at 2.2 GHz and 16 GB RAM. The OT communication network simulations are performed using Mininet, which runs on Ubuntu 22.04, RAM 64 GB, Intel(R) Xeon(R) W-2123 CPU 3.60GHz.

In this part, the generated synthetic power system is validated by comparing it with the input network. To thoroughly analyze the similarity of two networks, the topological and operational features are jointly considered based on the five metrics mentioned in section 3.3.2. Furthermore, the model performance is also evaluated with different attention strategy combinations, i.e., Non-Strategy (NS), Only Masked Attention (OMA), only Multihead Attention (MA), and Masked Attention combined with Multihead Attention (MAMA). The comparison of relative errors (using IEEE 39-bus system as baseline) on the five metrics are given in Table 3.2. The comparison results of the metrics distribution are given in Figure 3.9. To better describe the comparison results, the Jensen-Shannon (JS) divergence is used to quantify the difference between the two different probability distributions as shown in equation (3.2). From an overall perspective, the MAMA strategy outperforms other combinations on both relative error and JS divergence. In Table 3.2, the average relative error of MAMA on different metrics is on average 33.9% less than the other three strategies. In Figure 3.9, the MAMA strategy generates the synthetic network with similar metrics distribution as the input network while maintaining acceptable differences, which satisfies the requirements of a digital sibling.

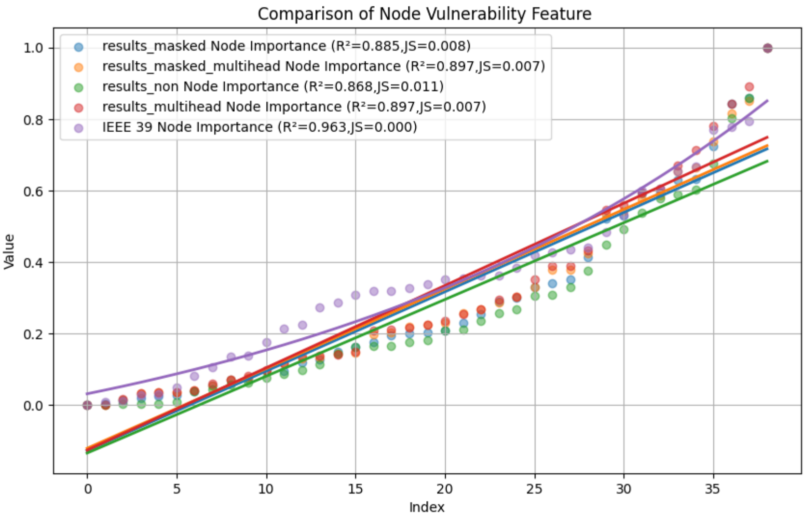
Table 3.2: Statistics Comparison between Generated Results and IEEE 39 Bus System

Relative Error	Node Power Injection	Node Vulnerability Feature	N-1 Contingency	Node Distance Degree	Node Capacity Degree	Average Relative Error
MA	0,264103	0,209016	0,615955	0,682464	0,172535	0,388815
MAMA	0,196154	0,241803	0,164391	0,696682	0,084507	**0, 276708**
NS	0,244872	0,122951	0,503682	0,71564	0,18662	0,354753
OMA	0,242308	0,278689	1,370014	0,725118	0,200704	0,563367

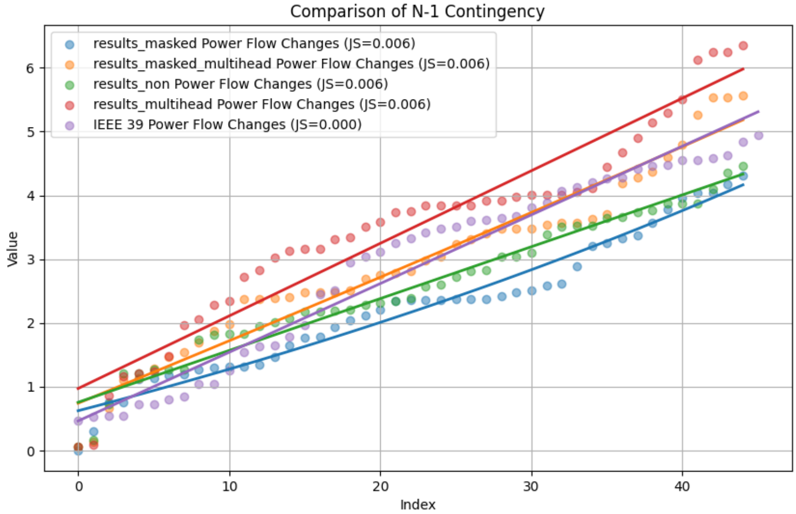


(a)

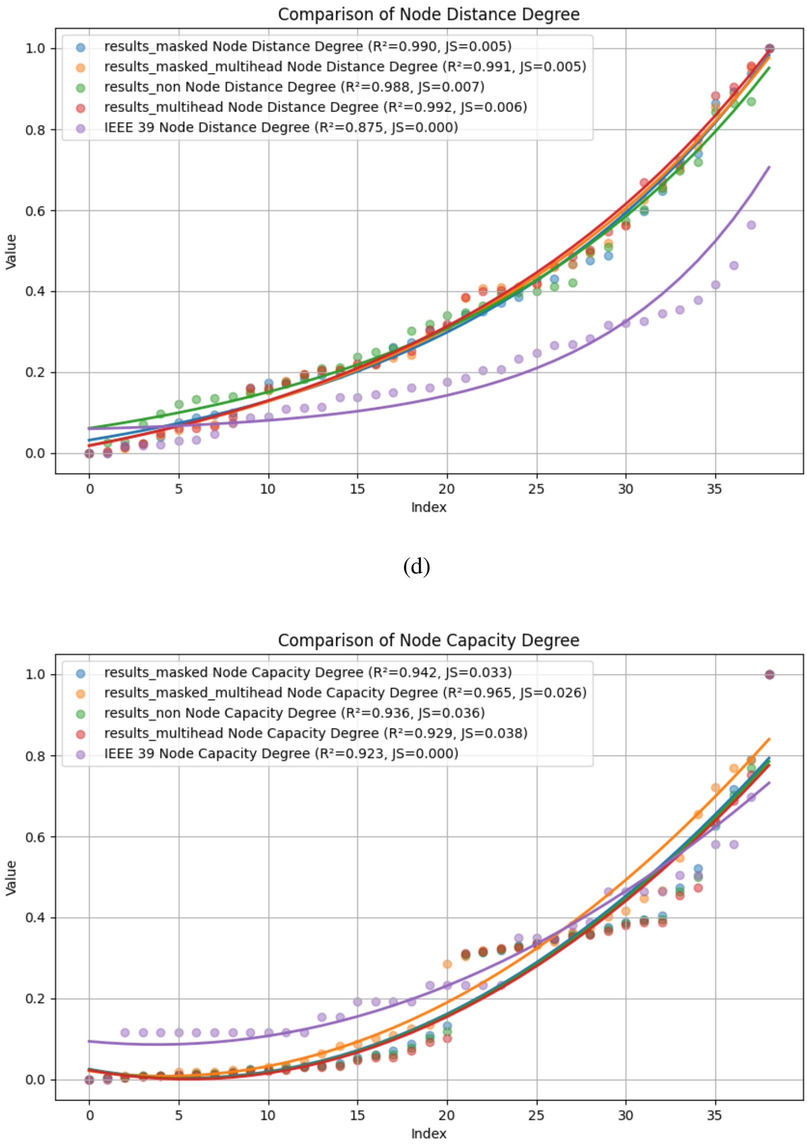
The OT simulations are performed using Mininet, which implements operating-system-level virtualization. Mininet is a network emulator primarily utilized for the purpose of testing computer networks [73]. It facilitates the creation of virtual networks on a computer and enables test network structures without having physical devices. Mininet runs the simulation on Ubuntu 22.04, RAM 64 GB, Intel(R) Xeon(R) W-2123 CPU 3.60GHz. Mininet provides features to set the bandwidth link between nodes. This mechanism allows the simulation to implement link bandwidth variations for the real and


Comparison of Node Vulnerability Feature
results_masked Node Importance (R²=0.885,JS=0.008)
results_masked_multihead Node Importance (R²=0.897,JS=0.007)
results_non Node Importance (R²=0.868,JS=0.011)
results_multihead Node Importance (R²=0.897,JS=0.007)
IEEE 39 Node Importance (R²=0.963,JS=0.000)
Value
1.0
0.8
0.6
0.4
0.2
0.0
0
5
10
15
20
25
30
35
Index

(b)


Comparison of N-1 Contingency
results_masked Power Flow Changes (JS=0.006)
results_masked_multihead Power Flow Changes (JS=0.006)
results_non Power Flow Changes (JS=0.006)
results_multihead Power Flow Changes (JS=0.006)
IEEE 39 Power Flow Changes (JS=0.000)
Value
6
5
4
3
2
1
0
0
10
20
30
40
Index

(c)



(d)

(e)

Figure 3.9 Comparison Between IEEE 39-bus System and Synthetic Network. (a) Node Power Injection, (b) Node Vulnerability Feature, (c) N-1 Contingency, (d) Node Distance Degree, (e) Node Capacity Degree

synthetic network scenarios. The Mininet topologies are derived from real and synthetic networks, including substations, wide area networks, and control centers.

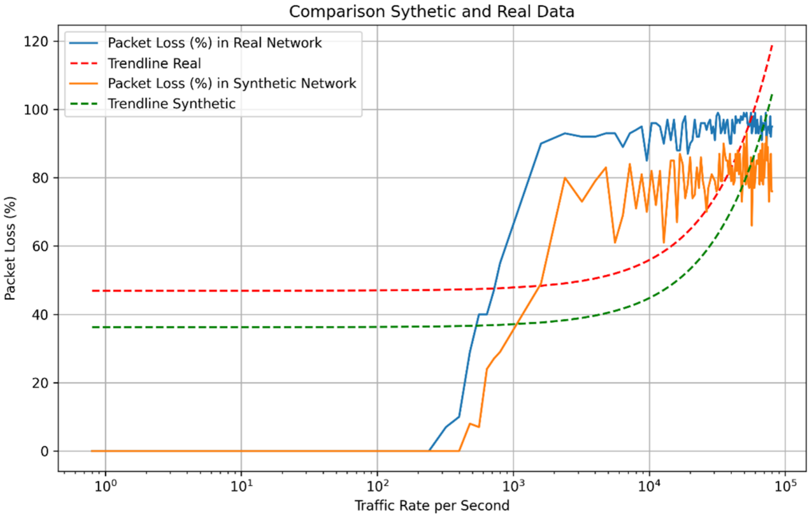


Figure 3.10 Packet Losses Comparison between Real and Synthetic Networks with Variation in Traffic Intensities.

The OT traffic is simulated using the OT communication network topologies implemented in Mininet. The OT network traffic originates from the nodes within the Mininet network, which have been customized to simulate the functionalities of IEDs, MUs, RTUs, and the control center. In this experiment, the simulated traffic is implemented using UDP, which carries the payloads of power system measurements and control commands between substations and the control center. In order to evaluate the performance of the OT communication network under various traffic loads, a range of traffic rates are simulated, which are transmitted from substations to the control center, ranging from one to one million packets per second. The network performance is examined by measuring the average RTT and packet losses under varying traffic intensities. Figures 3.10 and 3.11 show the comparison of packet losses and average RTT between the real and synthetic networks. Similar to the synthetic power system, both figures demonstrate that the original communication network and synthetic communication network possess similar characteristics while maintaining acceptable discrepancy. In Figure 3.10, both the real and synthetic networks maintain a low packet loss rate at lower traffic levels, which indicates that both network configurations are effective at handling traffic with a minimal packet loss. Besides, both networks exhibit an overall trend of gradually increasing packet loss rates as the traffic increases. As traffic reaches a certain threshold, i.e., around 10ˆ4 packets per second, the packet loss rates for both networks start to rise significantly, demonstrating a common challenge in managing large data flows. Figure 3.11 also reflects similar patterns for both real and synthetic

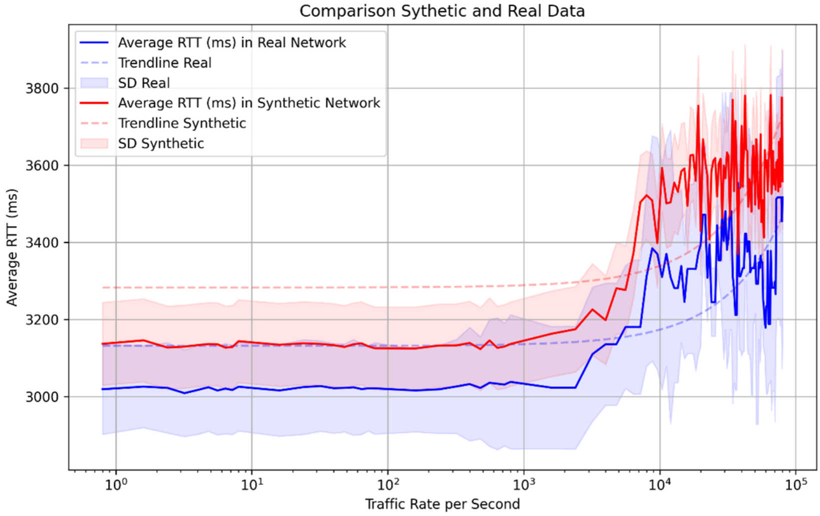


Figure 3.11 Average RTT Comparison between Real and Synthetic Networks with Variation in Traffic Intensities.

communication networks. The shaded areas representing the Standard Deviation (SD) for both real and synthetic networks illustrate the variability in RTT measurements. Both networks exhibit increasing variability as traffic rates increase, suggesting a common increase in unpredictability of network performance under higher traffic loads.

3.4. CONCLUSION

In this chapter, we assume that both the cyber and physical system data are available. In section 3.2, we propose a scalable generative model, namely Graph-CPS, to generate a synthetic CPS topology with realistic network feature distribution. This model is capable of learning different complex network parameters as well as capturing the distribution of different network features of the input networks. In experimental results, we implement the proposed Graph-CPS on three power systems and three power grid communication networks to demonstrate and assess the model effectiveness and scalability. The size of the networks mentioned above were scaled from 18 nodes to 1225 nodes and the networks contain both IEEE standard test systems and the real systems. The validation results thoroughly prove that Graph-CPS can accurately capture the characteristics of input networks with not only different network types, but also different network sizes.

In section 3.3, we clarify the differences between the definitions of digital twins and digital siblings. To capture the intricate system behaviors from the high-dimensional

CPS data, we propose a hybrid generator, namely SibGen, to generate the digital sibling of the real CPS. SibGen can generate both topological and operational models of the input network. In SibGen, two effective training strategies are proposed, i.e., dural graph training and prior knowledge-constrained training. The dural graph training strategy solves the problem that the current generation model can only process the node attributes while fails to also generate the synthetic attributes for edges. The prior knowledge-constrained training injects the physical constraints of CPS to make the generation results more realistic. In case study, five different metrics are evaluated to compare the real CPS and the generated synthetic network. The comparison results prove that the SibGen is capable of learning the global characteristics of the input network from both topological and operational perspectives.

4

Vulnerability Assessment for Cyber-Physical Power Systems Considering Time-Varying Operational States

Cyber security risks are emerging in CPS due to the increasing integration of cyber and physical infrastructures. Critical component identification is a crucial task for the mitigation and prevention of catastrophic blackouts. While efforts have been made to study the vulnerability features of power systems under the occurrence of a single, discrete disturbance or failure at a specific time instant, this chapter focuses on identifying the critical components of the cyber-physical system considering time-varying operational states. To this end, this chapter investigate the CPS vulnerability features from the perspectives of manifest and latent component correlations, providing an in-depth analysis to reveal the cascading mechanisms in CPS.

Parts of this chapter have been published in:

1. **Yigu Liu**, Alexandru Ştefanov, Ioannis Semertzis, Peter Palensky. "GraphCCI: Critical Components Identification for Enhancing Security of Cyber-Physical Power Systems". *IEEE Transactions on Industrial Cyber-Physical Systems.* , vol. 2, pp. 340-349, 2024.

2. **Yigu Liu**, Ioannis Semertzis, Alexandru Ştefanov, Peter Palensky. "Critical components identification for cyber-physical power systems considering time-varying operational states". *9th Workshop on Modeling and Simulation of Cyber-Physical Energy Systems, MSCPES 2021, Held as part of the Cyber-Physical Systems and Internet-of-Things Week, Proceedings*, Nashville, TN, USA, 2021.

4.1. Introduction

Due to the rapid integration of cyber and physical infrastructures, modern power systems are becoming more efficient while also exhibiting increased vulnerabilities. This emerging risk was starkly demonstrated by the three major cyber attacks on the Ukrainian power grid in 2015, 2016, and 2022 [6], [7], [74], underscoring the critical need for enhanced security measures in this landscape. The evolving communication infrastructures have significantly altered the propagation mechanisms of cascading failures in the Cyber-Physical power Systems (CPS) [75]. These changes present novel challenges in ensuring safe system operation. Consequently, it is imperative to thoroughly investigate the new cascading mechanisms and pinpoint the critical components of CPS, which will enable the implementation of timely mitigation strategies, thereby enhancing the overall security and resilience of CPS.

As an interdependent network, the functionality of CPS can be interrupted by various means of adversaries targeting any of the coupled network. In order to investigate the detailed mechanism of how CPS is corrupted, it is beneficial to divide the invalidation scenarios of CPS into different categories. In this chapter, we classify the cause of cyber-physical contingencies into two major groups based on the original source, namely cyber-originated contingencies and physical-originated contingencies, which are shown as in Figure 4.1.

Cyber-Originated Contingencies: Cyber-originated contingencies are solely caused by the failures or errors from cyber layer [76]. The main cause of cyber-originated contingencies is cyber attacks, which are better concealed and less noticeable compared with traditional physical causes of contingencies. Cyber attacks are conducted at the cyber system layer, which impact the functionalities of ICT devices and OT systems. As shown in 4.1, we consider cyber-based and network-based contingencies in this category. The cyber-based contingency refers to the attacks launched from only cyber layer. Code and command manipulations are considered the main attack in this category, which sabotage the functionalities of software and firmware of the system based on adversary's purpose. Malware injection injects worm or virus to the systems. The two blackouts of Ukrainian power grids in 2015 and 2016 are the typical real-life examples of malware injection. Other cyber-based contingencies include password cracking, supply chain attacks and database manipulation, etc. Another kind of cyber-originated contingency is network-based contingency, which are constructed through the virtual network access without affecting the software or the firmware of the system nor the physical communication links [76]. False data injection attacks utilize the network information and manipulate the uploaded measurement data of power grids and misleading the operator to conduct wrong actions. Man-in-the-middle attacks insert a controllable computer into the communication network to access and control traffic or change the transmitted data. Denial-of-service attacks temporarily or indefinitely disrupt OT services. Other network-based contingencies include packet sniffing, rogue node, etc.

Physical-Originated Contingencies: In physical-originated contingencies, main causes can be classified into three aspects: random failure, natural hazards, and

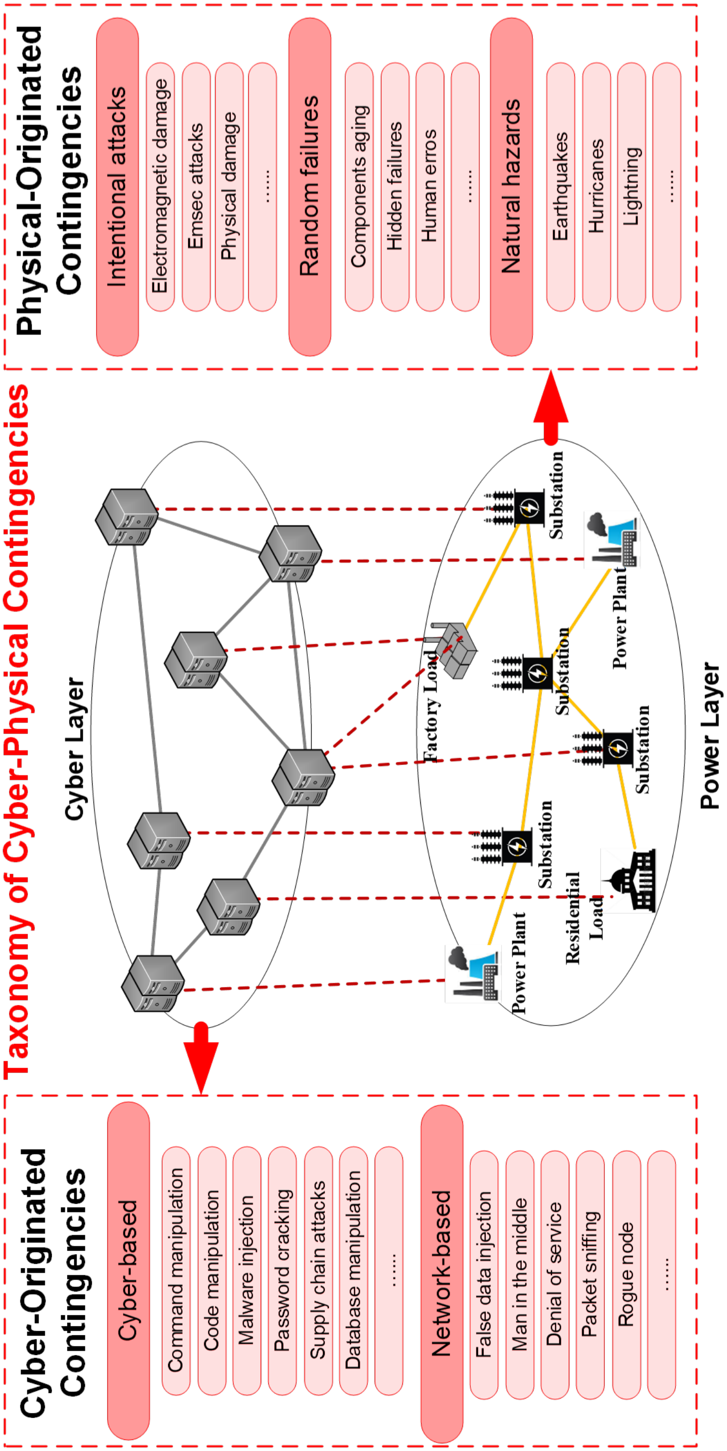


Figure 4.1 The Taxonomy of Cyber-Physical Contingencies

intentional attacks. For random failure, it affects all similar components with the same probability distribution, that is, other factors (e.g., location of nodes) have no impact on the probability of failure. On the other hand, intentional attacks usually attack the CPS with a clear target, from the perspective of attackers, they will investigate the system information (e.g., system topology, parameter configuration) and attack the weakest points to maximize the overall impact on CPS. Besides, natural hazards, e.g., hurricanes, snowstorm, and earthquakes, will also damage the operational status of CPS, especially considering the contingency of perturbing the critical components in system.

In recent years, growing attention has been paid to the system resilience of CPS, with an emphasis on extreme events, e.g., severe weather [77], [78], and cyber attacks [79],[80]. Vulnerability assessment plays a crucial role in enhancing the overall system resilience. It offers crucial insights to system operators during decision-making, particularly when defensive resources for the system are limited. Vulnerability assessment usually includes the different steps [81]: (1) System representation, which defines the structural, logical, and functional interdependencies among system components. (2) System mathematical modeling, which quantifies the performance indicators based on various assumptions (e.g., different attack mechanisms, different systematic interdependencies, different research objectives). (3) Model solving, which explores the system behavior under different operational and accidental conditions. Furthermore, in [82], authors propose that the problem of vulnerability assessment can be considered from three levels: system, scenarios, and access points.

System: the system level vulnerability assessment refers to determine the critical components based on the inherent system property (e.g., equipment location, device function, carried load). For example, in [83], authors propose a new centrality index based on the maximum flow from generator to load to identify the critical components in system.

Scenarios: the scenario level vulnerability assessment captures the inherent mechanism (e.g., fault propagation mechanism, system levels at each time instant) of system under disturbance (e.g., malicious attacks, natural hazard). It reflects the system vulnerability during the process of different system operation scenarios. More specifically, efforts have been made to rank the priority of components based on the consequence of its loss (e.g., load loss [26], unsupplied energy [84]). Researchers [85], [86] also pay attention to the existing defensive strategies embedded in system and evaluate the effectiveness of defense. Besides, in [87], authors propose a novel framework CPIndex to calculate the security level of system at each time instant.

Access points: access points refer to the system vulnerability that can be exploited by adversaries, such as common vulnerabilities and exposures (CVEs) [88] and accessible external equipment.

Researchers have developed various methods to evaluate the vulnerability of CPS, and these methods have different research background regarding different objectives. In this chapter, we divide all the methods into two categories: topology-based methods and operation-based methods. In the following content, we will present detailed discussion about the two categories, respectively.

Topology-based Methods: The topology-based methods focus on investigating the system vulnerability purely based on the topological structure of CPS. During this process, complex network theory (CNT) [81] is frequently adopted to analyze the intricate relationships among large number of nodes. CNT abstracts different components in complex systems into nodes or vertices, which are interconnected by links or edges. In cyber-physical power system, nodes or vertices can be generators, transformers, routers, operation centers, etc. Meanwhile, links or edges can be communication media, power transmission lines, or the interdependency between cyber and physical layers. In general, CNT uses metrics and indices (e.g., degree [37], [48], closeness [49], betweenness [50], [51]) to capture the importance of a node in system.

Operation-based Methods: Topology-based methods naturally neglect the heterogeneity of nodes in both cyber and physical layer and focus only on the structure of interdependent network. Consequently, the inherent physical mechanisms (e.g., load redistribution, routing protocols) of both layers will be ignored, which may result in unrealistic conclusions. To this end, researchers start to jointly consider the operational mechanisms of cyber and physical layer during the process of vulnerability assessment. In power system, the extensively concerned operational mechanism is power flow, which calculate the steady-state solutions of power grid [81]. Besides, when considering the vulnerability assessment on the cyber side, researchers have various assumptions. From the perspective of information transmission, simple communication system is always considered [89], [90], [91], [92], in which information packets are transmitted through an end-to-end network. Once failures occur on the path, the transmission process is immediately considered interrupted. Furthermore, with respect to detailed operational conditions in cyber system, transmission errors and delays are considered in information transmission model [93]. The detailed approaches adopted to solve the problems in this category are two fold, i.e., model-based approaches and machine learning-based approaches. On the one hand, the model-based methods consider the system operation models, e.g., power and information flow models, and evaluate the criticality of each component based on the system operational data, e.g., stability analysis [94], [95], [96], historical cascading failure data [97], [98], [99]. On the other hand, machine learning-based methods tend to train and learn the system features from the historical data, e.g., cascading failure data [100], [101], [102], where graph neural networks [101], reinforcement learning [103], [104], and data mining algorithms [79] are used to extract the system features and identify the vulnerable CPS components.

Table 4.1: The History of Cyber-Physical Events

ID	Coupling method	Interdependency	Heterogeneity	Physical System Size	Cyber System Size	Name of tested systems	Methodology
[5]	One-to-one	Structural	N	25000	25000	Italian power grids	Percolation theory
[25]	One-to-one	Operational	N	39, /	39, /	IEEE 39-bus system, China's Guangdong 500-kv power system	Dynamic power flow, complex network theory
[105]	Multiple-to-multiple	Structural	Y	1000	-	Self-generated synthetic network	Deep learning
[106]	One-to-multiple	Operational	N	349	-	Italian power grid network: HVIET	DC power flow model
[107]	One-to-multiple	Structural	N	10000	-	Self-generated network based on Barabasi-Albert model	Percolation theory
[108]	One-to-multiple	Operational	Y	26, 60, 101, 228, 564	25, 59, 98, 159, 288	IEEE 14, 30, 57, 118, and 300-Bus systems	AC power flow model
[109]	Multiple-to-multiple	Structural	N	1000	10000	Generalized Barabási-Albert model	Percolation theory
[110]	Multiple-to-multiple	Structural	N	106	106	Self-generated network	Percolation theory
[111]	One-to-one	Structural	N	6.4×10^5	6.4×10^5	Self-generated network	Percolation theory
[112]	One-to-one	Operational	Y	4	10	Microgrid	Linear programming model
[26]	One-to-one	Operational	N	-	-	IEEE 118 and 300-bus systems	AC power flow, Complex network theory
[113]	One-to-one	Structural	N	104	104	Self-generated network	Percolation theory
[114]	One-to-one	Structural	N	106	106	Self-generated network	Percolation theory
[115]	One-to-one	Structural	N	8×10^5	8×10^5	Self-generated network	Percolation theory
[116]	One-to-one	Structural	N	105	105	Self-generated network	Complex network theory
[117]	One-to-multiple	Operational	Y	310	39	HVIET, GARR	DC power flow

As shown in table 4.1, the representative papers in the domain of CPS vulnerability assessment are listed. Among these methods, 52.9% of all methods adopt the one-to-one coupling method to model CPS, and the percentages of one-to-multiple and multiple-to-multiple coupling method are 23.5% and 23.6%, respectively. From the perspective of interdependency, 41.2% of the methods focus on analyzing the operational interdependency and the structural methods are about 58.8%. Besides, this chapter also concerns about whether researchers take the heterogeneity of nodes in CPS into consideration and 23.5% investigate the impact on CPS of node heterogeneity. Considering the system size of case study, current literature conducts the methods on both small- and large-scale systems. Regarding the detailed methods in vulnerability assessment, the most frequently used methods in topology-based methods is percolation theory, while for the operation-based methods, AC and DC power flow models are frequently considered. The current literature has yielded fruitful results in identifying critical CPS components, yet each methodological category has notable limitations, which are two-fold. On the one hand, The existing work only evaluates the CPS at a single time instant. However, we argue that this may not always be the case. Instead of considering CPS disturbances or failures as single-occurrence events, in this chapter we treat them as a set of sequential discrete events. Disturbances and failures can occur at any time instant during CPS operation over a certain time period. Meanwhile, the operational states, e.g., loads and power flows, are constantly varying in time. Under such assumption, the vulnerability features generated by the existing static methods, which aim at a particular time instant may not be applicable to time-varying CPS operational states. To this end, a fundamentally new approach is needed to systematically capture the vulnerability characteristics and identify the most critical CPS components of whole operation time period to develop effective and economic mitigation strategies. On the other hand, topology-based methods partially unravel network structural features but overlook the complexity models [53] and heterogeneity [30] inherent in CPS as industrial systems, potentially skewing identification results. The operation-based methods consider CPS's operational facets, analyzing historical data to discern inter-component correlations. However, these methods typically extract correlations solely from the known data. Although some works consider different operational states, no historical data can cover all possible system conditions and capture all possible correlations between components. In these two categories, commonly employed statistical methods, like machine learning algorithms and graph theory indices, are limited to quantifying correlations presented in the historical data. This process overlooks latent correlations under unrepresented operational states, introducing significant bias in identifying critical components. Therefore, this chapter also aims to introduce a methodology that not only analyzes apparent component correlations but also quantifies latent ones, ensuring more accurate and realistic identification outcomes.

To address the issues above, in this chapter we made the following contributions:

(1) we propose a novel cascading failure model considering the interaction between cyber and physical layers for every single time instant. Based on quasi-dynamic simulations, we generate a database of cascading failure chains. This contains various operating conditions. We adopt the sequential mining algorithms to identify the frequent

sequential cascading patterns. Vulnerability indices are constructed based on complex network theory to evaluate the importance of components in the cascading failure process and identify the critical components in CPS.

(2) We define two correlations, i.e., manifest and latent correlations, to better reveal the cascading mechanism of CPS and comprehensively investigate the apparent and potential correlations between CPS components.

(3) We propose a set of definitions to map the historical cascading failures datasets into weighted cascading graphs, and then construct the weighted cascading graph database for graph data mining to thoroughly capture the cascading features of CPS. By jointly considering the manifest and latent correlations and the graph data mining results, we propose a critical components identification model named GraphCCI.

4.2. Critical Components Identification for Cyber-Physical Power Systems Considering Time-Varying Operational States

4.2.1. System Vulnerability Considering Time-varying Operational States

In previous discussion, we argue that the current vulnerability assessment methods may not be applicable or even feasible when considering the change of CPS operational status. As shown in Figure 4.2, in a real-world scenario, the operational states of CPS are constantly changing, which means the system will react to failures or disturbances differently at various time instants. More concretely, the cyber-physical system may show different cascading failure patterns under time-varying operational states, which will directly change the vulnerability features. In this context, we first model a failure, e.g., line tripping, in CPS to trigger the cascading failures at a specific time instant, e.g., t_2, t_4, t_6 or as represented in Figure 4.2. To thoroughly investigate the vulnerability characteristics of CPS at a specific time instant, we consider that any component in the cyber-physical system may fail, and we generate possible cascading failure chains for all components. These cascading failure chains contain the detailed vulnerability features of CPS at the time instant. By combining cascading failure chains of all-time instants, a cascading failure chain database is generated, which captures the intricate relationships among components and reveals the fault propagation mechanism of CPS under different operating conditions. For instance, for a certain time interval $[t_1, t_u]$, suppose the cascading failure chain set includes $\boldsymbol{X}_{CF}(t_1)$ at t_2, $\boldsymbol{X}_{CF}(t_2)$ at t_2,..., $\boldsymbol{X}_{CF}(t_u)$ at t_u, then the cascading failure chain database can be presented as:

$$\boldsymbol{X}_D = \{\boldsymbol{X}_{CF}(t_u) \,|1 \leq u \leq U\} \tag{4.1}$$

The definition of $\boldsymbol{X}_{CF}(t_u)$ can be found in (4.9). At last, we intend to employ sequential data mining algorithms to mine the cascading failure database and identify the critical components of CPS. Generally, the sequential data mining algorithms return the patterns that are frequently shown in the database. For cyber-physical systems, if a cascading failure pattern frequently appears in $\boldsymbol{X}_D$, it means that the corresponding components play a critical role in the cascading process. If such critical components are reinforced and cyber secure, the system resilience will be greatly improved.

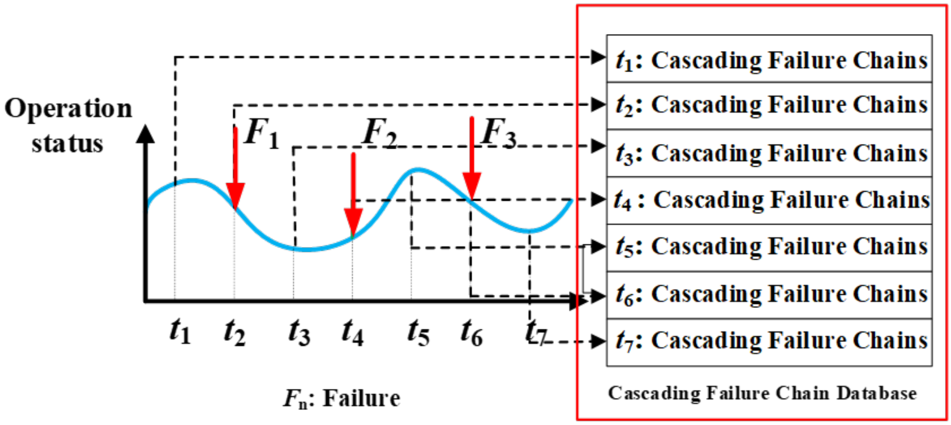

Operation status
F_1
F_2
F_3
t_1
t_2
t_3
t_4
t_5
t_6
t_7
F_n: Failure
t_1: Cascading Failure Chains
t_2: Cascading Failure Chains
t_3: Cascading Failure Chains
t_4 : Cascading Failure Chains
t_5: Cascading Failure Chains
t_6: Cascading Failure Chains
t_7: Cascading Failure Chains
Cascading Failure Chain Database

Figure 4.2 The Time-varying Operational States of CPS.

4.2.2. Modeling of CPS and Cascading Failures

In this Section, we investigate CPS modeling from both topological and operational perspective. We model the cascading failures at each time instant to show how CPS will react to disturbances under different operating conditions. Then, by collecting the cascading failure chains at each time instant, a database is generated to further reveal the systematic vulnerability features of the cyber-physical system.

In this section, we abstract the CPS into an interdependent network, in which nodes and edges are used to represent the cyber-physical system components and interconnections among them, respectively.

Physical Layer: the generators, substations and loads are considered as physical nodes, while the transmission lines and transformers are considered as physical edges. Consequently, we can directly map a power grid into an undirected and unweighted graph based on its own topology.

Cyber Layer: the SCADA system in the control center and station control systems in substations are abstracted into cyber nodes, while their communication links are considered as cyber edges. It is worth mentioning that for the cyber layer we only consider the influence of the cyber layer topology on the physical layer operation. In this research, we do not consider the detailed communication mechanisms, e.g., routing protocols. Typically, the communication networks for power grids are implemented as double-star or mesh networks [35], [32] From the perspective of complex network theory, double-star networks are scale-free networks [45]. The control centers are considered hub nodes with higher degrees in the system. If one of these nodes fail, the cyber-physical system will suffer severe consequence. The double star networks are sensitive to intentional cyber-physical attacks, but resilient to random failures. On the other hand, mesh networks, as opposite to double-star networks, show the feature of small-world [118], which indicates that mesh networks have a broader degree distribution and are more vulnerable to random failures. Generally, a broader degree distribution increases the robustness of complex networks. However, when cyber and physical layers are coupled to form an interdependent network, a broader degree distribution increases the vulnerability of the interdependent networks to random failures [119]. Meanwhile, the research of Ye et al. [25] also shows that power grids coupled with double-star communication network have a lower probability of catastrophic failures than with mesh networks. Therefore, in this section, we adopt the double-star network to model the topology of the cyber system.

Structural Interdependency: in this section, we consider the interdependence between cyber and physical layers as a "one-to-one" correspondence [119]. The number of nodes in the cyber layer is the same as in the physical layer, and a cyber node is exclusively interconnected with a physical node. Parshani et al. [116] defines the interdependency of networks as intersimilarity from a topology perspective and investigates the robustness of interdependent networks under different intersimilarities. The results show that for scale-free networks, the interdependency should be "degree-to-degree", which means that the node with the highest degree in the cyber layer should be interconnected with the node with the highest degree in the physical layer.

Failures such as protection maloperation or loss of communications may trigger cascading effects in the cyber-physical system. Furthermore, when power grids are tightly coupled with communication infrastructures, the extent of fault propagation in CPS may be significantly increased considering the complex interdependencies between the cyber and physical layers. For example, one disturbance in one network may simultaneously have an influence within the network and on its interdependent networks. In this subsection, we present the simulation process of generating the cascading failure chains for every time instant used to generate the cascading failure chain database.

When the power system is congested, system operators redispatch generation or even shed load to ensure that the power grid is securely and economically operated. Therefore, an optimal DC power flow model represented by equations (4.2) – (4.7) is used to minimize the load shedding when disturbances occur in the cyber-physical system.

$$\min \sum_{y \in D} W_y \left| p_y - P_{dy} \right| \tag{4.2}$$

$$\boldsymbol{F} = \boldsymbol{A}\boldsymbol{P} \tag{4.3}$$

$$\sum_{x=1}^{n} p_x = 0 \tag{4.4}$$

$$P_{dy} \leq p_y \leq 0, y \in \boldsymbol{D} \tag{4.5}$$

$$P_{gx}^{\min} \leq p_x \leq P_{gx}^{\max}, x \in \boldsymbol{G} \tag{4.6}$$

$$-F_l^{\max} \leq F_l \leq F_l^{\max}, L_l \in \boldsymbol{L} \tag{4.7}$$

where $\boldsymbol{G}$ and $\boldsymbol{D}$ are the set of generators and loads, respectively, W_y is the cost of load shedding, $\boldsymbol{L} = \{L_l \,|l = 1, 2, ..., N_l\}$ is the set of branches in the power grid and $\boldsymbol{P} = [p_1, p_2, ..., p_k, ...]^T$ is the vector of power node injections. Equation (4.3) represents the DC power flow equation. $\boldsymbol{A}$ is the nodal admittance matrix and $\boldsymbol{F} = [F_1, F_2, ..., F_l, ...]$ is the vector of branch power flows. p_y represents the load of node y. P_{dy} represents the rated load at node y. p_x represents the output power of generator x. $P_{gx}^{\max}$ and $P_{gx}^{\min}$ are the upper and lower limits of the output power of generator x, respectively. $F_l^{\max}$ is the transmission capacity of the *l-th* branch.

Ye et al. [25] propose an interaction model and analyses the system performance under both intentional attacks and random failures. Dong et al. [120] propose a probabilistic failure model to simulate the cascading process between cyber and physical layers. Based on these works, an interactive model is used to capture the main features of both cyber and physical layers and give a rough approximation to describe the interdependency between the two layers, which is presented as follows.

Cascading failures in the same layer: we consider that cascading failures in power grids are mainly caused by load redistribution when branches are disconnected and by hidden failures. Due to a hidden failure [121], the outage of branch L_l may cause the failure of its neighbors with a low probability P_1. When a branch is overloaded due to system load redistribution, we assume that the branch will be disconnected with a probability P_2. We do not consider the mutual influence among cyber nodes, i.e., the failure of a cyber node will only influence the data communication and will not cause a failure of other cyber nodes.

The impact of disturbances in the cyber layer to the physical layer: we consider that the cyber nodes are directly coupled with the physical nodes of power grids. When a cyber node is out of service, the control center loses the remote monitoring and control capabilities of the physical node and all corresponding branches in the substation. Consequently, when these branches are overloaded, they will operate in an insecure state and will be eventually disconnected by system protection after a period of time. On the other hand, a failed cyber node may be on the communication path between the control center and another cyber node. Under such circumstances, we consider that the control center also loses the monitoring and control capabilities of the associated physical nodes.

In this section, we investigate systematic cyber-physical system vulnerabilities. Therefore, we include various cascading failure scenarios by assuming that each component is possible to fail at every time instant. More specifically, we trip all the branches one by one to collect all possible cascading failure chains at every time instant. Then, by repeating the same process, the cascading failure chains are combined to generate the cascading failure chain database as shown in Figure 4.2. The detailed simulation process of one single time instant is presented in Figure 4.3. A disconnected branch is removed from the power grid topology. The updated topology is represented by N_{real}. Furthermore, we consider N_{control} to be a subset of N_{real} for which the system operator still has monitoring and control capabilities. The branches connected to the physical nodes affected by the failure of their corresponding cyber nodes are removed from N_{control}. We consider that the cyber nodes are vulnerable to cyber attacks and some will fail due to malicious attacks or other contingencies in each iteration. The cyber nodes will be removed with a small probability P_3.

The cascading failure process at time instant t_U starts by disconnecting branch L_l and scanning for cyber and hidden failures. The N_{real} and N_{control} CPS topologies are updated. The DC power flow is first calculated based on the updated N_{real}. If there are overloaded branches, we calculate the optimal DC power flow based on the updated N_{control}. The results of the optimal DC power flow give the power injections for the physical nodes in N_{control}. The redispatch of generation with minimum load shedding costs is implemented using N_{real}. We calculate load redistribution based on the new power injections and previously available measurements for the physical nodes affected by the failure of their cyber nodes. The overloaded branches are disconnected with their corresponding probabilities. It is worth mentioning that a branch may be disconnected based on local measurements by protection relays and control commands from the control center. When a branch is overloaded, system operators will adjust the generation or initiate load shedding. If the overload is not mitigated, the branch will be tripped by overload protection. Therefore, in our section, we assume that when a branch is overloaded, it is tripped by local protection with a probability P_2. The process is repeated until there are no further overloaded branches. The cascading failure chain is exported to the database.

It is worth mentioning that the simulation process illustrated in Figure 4.3 is used to generate the cascading failure chain $\boldsymbol{X}_{CF}^{L_l}(t_u)$ initiated by the disconnection of branch L_l at t_u. To thoroughly capture the vulnerability features of CPS and generate the cascading failure chain $\boldsymbol{X}_{CF}(t_u)$ at t_u, this simulation should be conducted for every branch in $\boldsymbol{L}$. This can be represented by equations (4.8) and (4.9).

$$\boldsymbol{X}_{CF}^{L_l}(t_u) = \rho\left(C_1, C_2, \ldots, C_n\right), C_k \in \boldsymbol{C} = \boldsymbol{V}_C \cup \boldsymbol{L} \tag{4.8}$$

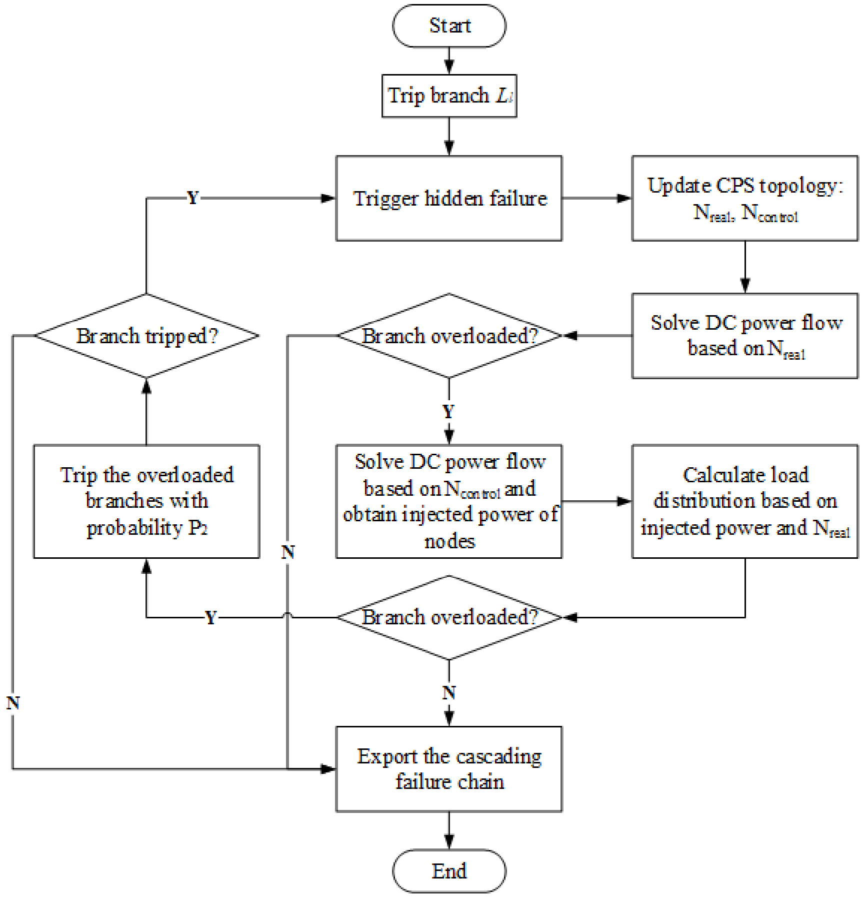

Start
Trip branch L_l
Trigger hidden failure
Update CPS topology: N_{real}, $N_{control}$
Y
Branch tripped?
Branch overloaded?
Solve DC power flow based on N_{real}
Y
Trip the overloaded branches with probability P_2
Solve DC power flow based on $N_{control}$ and obtain injected power of nodes
Calculate load distribution based on injected power and N_{real}
N
Y
Branch overloaded?
N
N
Export the cascading failure chain
End

Figure 4.3 Simulation Process of Cascading Failures.

$$\boldsymbol{X}_{CF}\left(t_u\right)=\left\{\boldsymbol{X}_{CF}^{L_l}\left(t_u\right)\left|L_l\in\boldsymbol{L}\right.\right\}\tag{4.9}$$

where $\rho\left(C_1,C_2,\ldots,C_n\right)=C_1\rightarrow C_2\rightarrow,\ldots,\rightarrow C_n$. $\boldsymbol{V}_C=\left\{v_g\left|g=0,1,2,...,N_g\right.\right\}$ represents the set of cyber nodes at the cyber layer. The cascading failure chain database $\boldsymbol{X}_D$ can be generated based on equations (4.1) and (4.9).

4.2.3. Critical Components Identification from a Data Mining Perspective

In this section, we take advantage of the fact that $\boldsymbol{X}_{CF}^{L_l}\left(t_u\right)$ can be viewed as a sequence for data mining and employ the *PrefixSpan* sequential data mining algorithm to capture the most frequent cascading failure sequence, i.e., CPS vulnerable sequence. Based on the identified patterns, we propose a vulnerability metric to further quantify the vulnerability of each component in the cyber-physical system. For a cyber-physical system, the cascading failure chain database can be very large, in which some cascading failure patterns may show up repeatedly. We use the frequency of these patterns to quantify the vulnerability of each CPS component. The cascading failure patterns are defined as candidate sequences waiting to be evaluated whether they are vulnerable sequences or not.

Definition 4.1 (Candidate Sequence): Based on the definition of $\boldsymbol{X}_{CF}^{L_l}(t_u)$, if there exists $\{C_{j1}, C_{j2}, \ldots, C_{jz}\} \subseteq \{C_1, C_2, \ldots, C_n\}$, a sequence $\alpha{=}\rho\,(C_{j1}, C_{j2}, \ldots, C_{jz})$ is called a *subsequence* of a cascading failure chain $\boldsymbol{X}_{CF}^{L_l}(t_u)$, which can be denoted as $\alpha \triangleright \boldsymbol{X}_{CF}^{L_l}(t_u)$.

Definition 4.2 (Vulnerability Degree): For a candidate sequence $\alpha{=}\rho\,(C_{j1}, C_{j2}, \ldots, C_{jz})$, the vulnerability degree is defined as:

$$\boldsymbol{V}_D\,(\alpha) = |\{\rho \,|\, (\rho \in \boldsymbol{X}_D) \wedge (\alpha \triangleright \rho)\}\;| \tag{4.10}$$

Based on the definitions above, *PrefixSpan* can be adopted to identify the vulnerable sequence with higher vulnerability degrees. The details of *PrefixSpan* are reported in [19]. Based on the vulnerable sequences identified above, in this part, we propose a vulnerability metric to further quantify the vulnerability of each CPS component. As discussed in Section 4.2.2, for each cascading failure chain $\boldsymbol{X}_{CF}^{L_l}(t_u)$, the components highly positioned in the chain result in high vulnerabilities. Therefore, we propose a metric named *total sequential vulnerability* to identify the critical components in the cyber-physical system.

Definition 4.3 (Total Sequential Vulnerability): For a vulnerable sequence $\beta_m{=}\rho\,(..., C_i, ...)$, the sequential vulnerability $S_{\beta_m}(C_i)$ of component C_i in β_m is defined as:

$$S_{\beta_m}(C_i) = N_{\beta_m} - \delta_{\beta_m}(C_i) + 1 \tag{4.11}$$

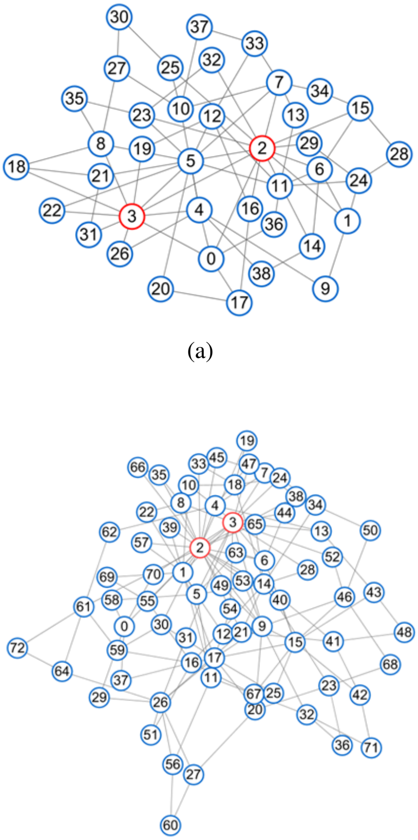
where N_{β_m} is the number of components in β_m and $\delta_{\beta_m}(C_i)$ is the order of C_i in β_m. Based on equation (4.11), by combining the sequential vulnerability of component C_i in all M vulnerable sequences containing C_i, the total sequential vulnerability of C_i can be represented as:

$$S\,(C_i) = \sum_{m=1}^{M} S_{\beta_m}(C_i) \tag{4.12}$$

4.2.4. Case Study

In this section, we conduct experiments on IEEE 39-bus and IEEE RTS-96 models to evaluate the effectiveness of the proposed method. Their cyber-physical systems and the proposed method are implemented in Python. The probabilities for the simulation of cascading failure chains are set as follows: $P_1 = 0.05$, $P_2 = 0.95$, $P_3 = 0.01$.

As discussed in Section 4.2.2, we use a scale-free network to simulate the cyber layer. Based on the Barabási–Albert (BA) model [45], Figure 4.4 shows the generated cyber topologies of IEEE 39-bus and IEEE RTS-96 system, respectively.

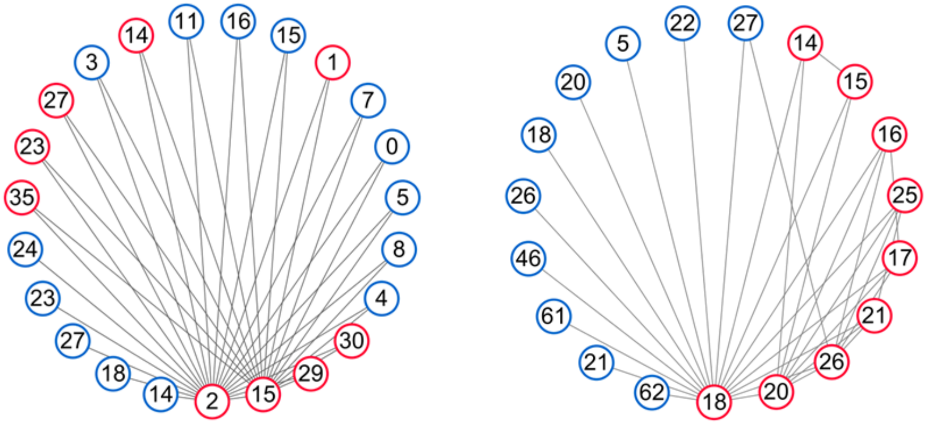


(a)

(b)

Figure 4.4 Cyber Layer Topology: (a) IEEE 39-Bus System, (b) IEEE RTS-96 Bus System.

The method proposed in Section 4.2.2 is used to generate the vulnerable sequences of IEEE 39-bus and IEEE RTS-96 system. For IEEE RTS-96 system, we use the peak



(a)

(b)

Figure 4.5 Vulnerable Sequence Identification: (a) IEEE 39-Bus System, (b) IEEE RTS-96 Bus System. (The cyber nodes are represented with blue, while the power system branches are represented with red.)

loads of each week for a 52-week load profile to simulate the time-varying operational states of CPS. For IEEE 39-bus system, we change the load proportionally in each simulation over 52 weeks. In the final database, there are 1901 cascading failure chains for IEEE 39-bus system and 6479 cascading failure chains for IEEE RTS-96 system. Figure 4.5 shows all the vulnerable sequences identified for the two test systems. Furthermore, based on equations (4.11)-(4.12), the total sequential vulnerabilities are calculated to quantify the vulnerabilities of CPS components in the test systems. Table 4.2 and 4.3 show the top 5 components in both cyber and physical layers with the highest total sequential vulnerabilities.

Table 4.2: Vulnerable Components of IEEE 39-Bus System Sorted by Total Sequential Vulnerability

Branches in Physical Layer			**Nodes in Cyber Layer**		
Ranking	**ID of Branches**	$S(C_i)$	**Ranking**	**ID of Nodes**	$S(C_i)$
1	2	50	1	3	5
2	15	24	2	16	4
3	1	5	3	11	3
4	35	4	4	15	3
5	23	4	5	8	3

From the perspective of degree distribution, in Figure 4.5(a), the components with the highest degree are branches 2, 15 and 29. This ranking is different from the ranking of total sequential vulnerability. This is because the total sequential vulnerability also considers the position of components in a vulnerable sequence. When a component

Table 4.3: Vulnerable Components of IEEE RTS-96 System Sorted by Total Sequential Vulnerability

Branches in Physical Layer			**Nodes in Cyber Layer**		
Ranking	**ID of Branches**	$S(C_i)$	**Ranking**	**ID of Nodes**	$S(C_i)$
1	18	93	1	27	3
2	20	64	2	5	2
3	16	25	3	18	1
4	26	23	4	21	1
5	17	17	5	20	1

frequently appears at the start position of a sequence, it means this component has a more significant impact on other components in the system. If the cyber-physical security of such components can be strengthened, then the scale of cascading failures will be reduced and thus the system will be more resilient. It is worth mentioning that although the degree distribution and total sequential vulnerability of power nodes are much higher than the ones of the cyber nodes, they are equally important for cyber-physical systems.

On the other hand, as shown in Table 4.2 and 4.3, we can observe that the span of $S(C_i)$ is quite large, which means, taking IEEE 39-bus system as an example, branch 2 is more vulnerable than branch 23, and by extension, other branches ranked behind branch 23 in the system. Such results indicate that for cyber-physical systems, there is a limited number of critical components, which must be reinforced and cyber secure. In our case, Table 4.2 and 4.3 give the top 5 critical components in both cyber and physical layers of the IEEE 39-bus and IEEE RTS-96 systems.

4.3. GraphCCI: Critical Components Identification for Enhancing Security of Cyber-Physical Power Systems

Based on the cascading models in Section 4.2, we expand the CPS cascading modeling details to further investigate the interactions between CPS components. This section proposes a graph data mining-based critical components identification model named GraphCCI, which evaluates the criticality of CPS components from the perspectives of manifest and latent correlations. First, we abstract the cascading failure data under different operational states into a weighted cascading graph database. Then, the TKG algorithm is adopted to identify the frequent subgraphs in the constructed graph database. Meanwhile, the definition of CC-Graph is proposed to model the overall cascading features based on the graph mining results. Finally, the NC-Index is proposed to evaluate the criticality of each CPS component. Our case study reveals that the cyber-physical system shows different cascading features under different system conditions. Verifications on the IEEE 39-bus test system demonstrate the effectiveness of our method. The identification results can provide an important reference to enhance

CPS security and prevent cascading failures and even a blackout.

4.3.1. Cyber-Physical Cascading Model Considering Time-Varying Operational States

In power systems, cascading failures can be described as a rapid, uncontrolled sequence of power equipment disconnections from the power grid, which may result in a blackout. In general, the fundamental idea of generating cascading failure datasets in power systems is based on simulations with existing cascading models [25], [101]. In cyber-physical power systems, the cascading process described above is further influenced by the cyber-physical interactions. The cyber-physical interplay can amplify the cascading effects. For instance, cyber attacks, e.g., false data injection and distributed denial of service, can misguide the decision-making in the control center and pose a significant threat to power system operation. Furthermore, a power system outage can disrupt communication networks affecting the power grid monitoring and control capabilities, which can further destabilize the CPS. In this section, we adopt the cyber-physical cascade model developed in our previous work [101]. Note that to use the methodology proposed in this section for analyzing the cyber-physical cascading mechanism, the cascade data can also be generated based on other cyber-physical system models in the literature. The cyber-physical cascading failures chain C_{CF} can be represented as in (4.13).

$$C_{CF} = \langle \boldsymbol{C}_1 \rangle \rightarrow \langle \boldsymbol{C}_2 \rangle \rightarrow \langle \boldsymbol{C}_3 \rangle \rightarrow \ldots \langle \boldsymbol{C}_i \rangle \rightarrow \ldots \langle \boldsymbol{C}_n \rangle \tag{4.13}$$

where $\boldsymbol{C}_i = \{C_{i1}, C_{i2}, ..., C_{ik}, ...C_{im}\}$ represents a set of components in CPS and the element C_{ik} can be either a cyber or physical component. The transmission lines represent the physical components, while the Supervisory Control and Data Acquisition (SCADA) system in the control center, communication network components, and the protection, automation and control systems in substations are abstracted into cyber components. $\{C_{i1}, C_{i2}, ..., C_{ik}, ...C_{im}\}$ indicates that after the removal of prefixed components $\langle \boldsymbol{C}_{i-1} \rangle$, multiple components can be disabled simultaneously. In general, a cascading failure chain as (4.13) contains information about components correlation and transitivity. (i) Components correlation: in (4.13), the relationship between $\langle \boldsymbol{C}_1 \rangle$ and $\langle \boldsymbol{C}_2 \rangle$ can be considered as the causality correlation, which indicates that the failure of the components in $\langle \boldsymbol{C}_2 \rangle$ is caused by the removal of all components in $\langle \boldsymbol{C}_1 \rangle$. (ii) Transitivity: in [80], the transitivity of a cascading failure chain is defined as: if there exist $\langle \{C_{11}\} \rangle \rightarrow \langle \{C_{21}, C_{22}\} \rangle \rightarrow \langle \{C_{31}\} \rangle$, the components C_{11} and C_{31} are correlated even if the failure of C_{31} is not directly cause by C_{11}. Note that if the correlations $\langle \{C_{11}\} \rangle \rightarrow \langle \{C_{21}, C_{22}\} \rangle$ and $\langle \{C_{21}, C_{22}\} \rangle \rightarrow \langle \{C_{31}\} \rangle$ originate from two different cascading failure chains, the transitivity property cannot be used directly. We will further discuss this issue in Definition 4.6.

In this section, we further investigate the correlations among CPS components. Based on the cascading failure data, in the following content, we construct the cascading graph database and mine the frequent subgraph to further reveal the cascading mechanism of CPS. By utilizing the cascading model in [101], we generate N cascading chains at a given operational state as in (4.13) and construct a weighted cascading graph. The definitions and detailed generation process are as follows.

Definition 4.4 (Manifest Correlation): For two given CPS components $C_{ik} \in \boldsymbol{C}_i$ and $C_{jk} \in \boldsymbol{C}_j$, if $\boldsymbol{C}_i$ and $\boldsymbol{C}_j$ are in the same cascading failure chain, then we define the correlation between C_{ik} and C_{jk} as manifest correlation, and it is denoted as $C_{ik} \rightarrow C_{jk}$.

Definition 4.5 (Latent Correlation): For three given CPS components $C_{ik} \in \boldsymbol{C}_i$, $C_{jk} \in \boldsymbol{C}_j$ and $C_{lk} \in \boldsymbol{C}_l$, if it satisfies $C_{ik} \rightarrow C_{jk}$, $C_{jk} \rightarrow C_{lk}$ and $\boldsymbol{C}_i$ is not in the same cascading failure chain with $\boldsymbol{C}_l$, then we define the correlation between C_{ik} and C_{lk} as latent correlation, and it is denoted as $C_{ik} \Rightarrow C_{lk}$.

Example 4.1. Let two cyber-physical cascading failure chains both with the length of 3 be $C_{CF}^{(1)} = \langle\{C_{11}\}\rangle \rightarrow \langle\{C_{21}, C_{22}\}\rangle \rightarrow \langle\{C_{31}\}\rangle$ and $C_{CF}^{(2)} = \langle\{C_{21}, C_{22}\}\rangle \rightarrow \langle\{C_{31}\}\rangle \rightarrow \langle\{C_{41}\}\rangle$, where $C_{CF}^{(1)}$ and $C_{CF}^{(2)}$ are generated under the same system condition. In this example, C_{11} and C_{31} have the manifest correlation. C_{11} and C_{41} have the latent correlation.

Definition 4.6 (Transitivity of Cascading Correlation): We define the symbol $\triangleright$ to indicate the cascading correlation between any two components C_{ik} and C_{jk}, and it is denoted as:

$$R\left(C_{ik}, C_{jk}\right) = C_{ik} \triangleright C_{jk} \tag{4.14}$$

Note that $C_{ik} \triangleright C_{jk}$ indicates that C_{ik} and C_{jk} either satisfy $C_{ik} \rightarrow C_{jk}$ or $C_{ik} \Rightarrow C_{lk}$. Then, the transitivity of cascading correlation is defined as if $\exists$ $C_{1k}, C_{2k}, C_{3k}, ..., C_{ik}, ...C_{nk}$ satisfy $C_{1k} \triangleright C_{2k}, C_{2k} \triangleright C_{3k}, ..., C_{ik} \triangleright C_{(i+1)k}, ...,$ $C_{(n-1)k} \triangleright C_{nk}$, then

$$R\left(C_{1k}, C_{2k}, C_{3k}, ..., C_{ik}, ..., C_{nk}\right) = C_{1k} \triangleright C_{2k} \triangleright C_{3k} ... \triangleright C_{ik} ... \triangleright C_{nk} \tag{4.15}$$

Note that once (4.15) is satisfied, there is a transitivity property between any two components in (4.15).

Definition 4.7 (Mapping a Cascading Chain into a Graph): We define a mapping operator $F : R\left(C_{CF}^{(i)}\right) \mapsto \boldsymbol{G}_{CF}^{(i)}$, and $\boldsymbol{G}_{CF}^{(i)} = F\left(R\left(C_{CF}^{(i)}\right)\right) = \left\langle \boldsymbol{V}_{CF}^{(i)}, \boldsymbol{E}_{CF}^{(i)}, \boldsymbol{w}^{(i)}, \boldsymbol{\phi}_w^{(i)} \right\rangle$ is a directed graph, where $\boldsymbol{V}_{CF}^{(i)}$ is the set of vertices in $\boldsymbol{G}_{CF}^{(i)}$ and is mapped from all the components in $C_{CF}^{(i)}$, $\boldsymbol{w}^{(i)}$ is the weight set of all edges mapped by the mapping relationship $\boldsymbol{\phi}_w^{(i)}$.

Based on definitions 4.4-4.7, one can map a cascading failure chain $C_{CF}^{(i)}$ into a directed and weighted graph. Note that in definition 4.4, the weights of all edges are set to 1 by default because for one cascading failure chain, each component can only be removed once, and the weights of edges represent the frequency of the corresponding correlation in the cascading data. To thoroughly evaluate the importance of each component in the system, one can construct N cyber-physical cascading failure chains, i.e., $C_{CF}^{(1)}, C_{CF}^{(2)}, ..., C_{CF}^{(N)}$. Then, based on definitions 4.4-4.7, we can construct N directed graphs, i.e., $\boldsymbol{G}_{CF}^{(1)}, \boldsymbol{G}_{CF}^{(2)}, ..., \boldsymbol{G}_{CF}^{(N)}$. Furthermore, these graphs can be combined to generate a weighted cascading graph $\boldsymbol{G}_{CF}(t_x)$ for a single operational state t_x as follows:

$$\boldsymbol{G}_{CF}\left(t_x\right) = \left\langle \boldsymbol{V}_{CF}^{(t_x)}, \boldsymbol{E}_{CF}^{(t_x)}, \boldsymbol{w}^{(t_x)}, \boldsymbol{\phi}_w^{(t_x)} \right\rangle \tag{4.16}$$

$$\boldsymbol{V}_{CF}^{(t_x)} = \bigcup_{i=1}^{N} \boldsymbol{V}_{CF}^{(i)} \tag{4.17}$$

$$\boldsymbol{E}_{CF}^{(t_x)} = \bigcup_{i=1}^{N} \boldsymbol{E}_{CF}^{(i)} \tag{4.18}$$

$$\boldsymbol{w}^{(t_x)} = \left\{ w_{E_{CF}^{(t_x)}} \middle| w_{E_{CF}^{(t_x)}} = f(E_{CF}^{(t_x)}) \right\} \tag{4.19}$$

where $f(E_{CF}^{(t_x)})$ is the frequency of edge $E_{CF}^{(t_x)}$ among $\boldsymbol{G}_{CF}^{(1)}, \boldsymbol{G}_{CF}^{(2)}, ..., \boldsymbol{G}_{CF}^{(N)}$. By following definitions 4.4-4.7 and equations (4.14)-(4.19), the cascading correlations are captured and emerged into the weighted cascading graph. The transitivity of cascading correlations is also converted into the connectivity of components. If there exists a path between two vertices in the weighted cascading graph, it indicates that there is a manifest or latent correlation between the two components. In Algorithm 4.1, we present the detailed generation process of the weighted cascading graph.

The cascading characteristics captured in $\boldsymbol{G}_{CF}(t_x)$ contain only the system information under one specific operational state, which fail to capture the overall cascading features of CPS under different operational states [79], [101]. For example, the critical components identified under a specific operational state may not apply to other operational states. Therefore, to capture the overall cascading characteristics, we define a weighted cascading graph database that contains the cascading characteristics under different time-varying operational states. As in Figure 4.6, for a certain time interval $[t_0, t_u]$, the weighted cascading graph database $\boldsymbol{G}_D$ can be represented as:

$$\boldsymbol{G}_D = \{\boldsymbol{G}_{CF}\left(t_x\right) \mid t_0 \leq t_x \leq t_u\} \tag{4.20}$$

Algorithm 4.1: Generation of weighted cascading graph

Input:

$C_{CF}^{(1)}, C_{CF}^{(2)}, ..., C_{CF}^{(N)}$ at t_x

Output:

Optimal candidate edge set: $\boldsymbol{G}_{CF}(t_x)$

```
Step 1  $\boldsymbol{V}_{CF}^{(t_x)} \leftarrow \varnothing$, $\boldsymbol{E}_{CF}^{(t_x)} \leftarrow \varnothing$
Step 2  **For** each $C_{CF}^{(i)}$ **do**
Step 3        Convert $C_{CF}^{(i)}$ into $\boldsymbol{G}_{CF}^{(i)}$ based on definition 4.4-4.7
Step 4  **End For**
Step 5  **For** each $\boldsymbol{G}_{CF}^{(i)}$ **do**
Step 6        $\boldsymbol{V}_{CF}^{(t_x)} \leftarrow \boldsymbol{V}_{CF}^{(i)} \cup \boldsymbol{V}_{CF}^{(t_x)}$
Step 7        $\boldsymbol{E}_{CF}^{(t_x)} \leftarrow \boldsymbol{E}_{CF}^{(i)} \cup \boldsymbol{E}_{CF}^{(t_x)}$
Step 8  **End For**
Step 9  Employ equation (4.19) to calculate $\boldsymbol{w}^{(t_x)}$
Step 10 **Return** $\boldsymbol{G}_{CF}(t_x)$
```

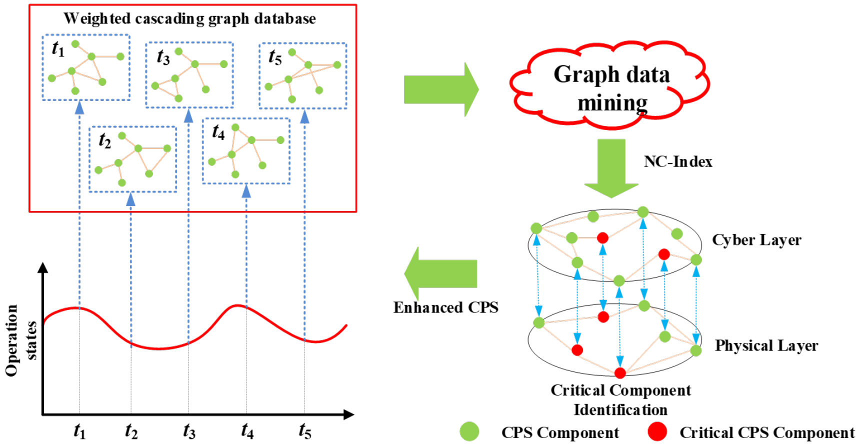


Figure 4.6 The Framework of GraphCCI.

In this section, we propose a model for critical components identification, i.e., GraphCCI. As represented in Figure 4.6, we first collect the cascading failure data under different operational state. Then, by adopting the methods proposed in Section 4.3.1, we map the cascading information into a weighted cascading graph database. Note that to increase the accuracy of the critical component evaluation results, one should simulate different operational states as much as possible so that $\boldsymbol{G}_D$ can comprehensively cover the cascading failures information. The next step is to utilize graph data mining algorithms to identify the critical subgraphs. In this section, we focus on the frequency aspect of subgraphs and adopt the TKG algorithm [122] to identify the top-K frequent subgraphs from $\boldsymbol{G}_D$. Then, by using the proposed NC-Index, we identify the critical CPS components and enhance the security level of CPS.

4.3.2. Graph Mining-Based Critical Component Evaluation

To better reveal the cascading characteristics of CPS, we adopt graph data mining algorithms to mine the frequent subgraphs in the weighted cascading graph database constructed in Section 4.3.1. The definitions of graph data mining are given as follows:

Definition 4.8 (Cascading Subgraphs): For a given cascading graph $\boldsymbol{G}_{CF}(t_x) = \left\langle \boldsymbol{V}_{CF}^{(t_x)}, \boldsymbol{E}_{CF}^{(t_x)}, \boldsymbol{w}^{(t_x)}, \boldsymbol{\phi}_w^{(t_x)} \right\rangle$, if there exists a graph $\boldsymbol{g}_{CF}^{(i)} = \left\langle \boldsymbol{v}_{CF}^{(i)}, \boldsymbol{e}_{CF}^{(i)}, \boldsymbol{w}_g{}^{(i)}, \boldsymbol{\phi}_{gw}^{(i)} \right\rangle$ that satisfies $\boldsymbol{v}_{CF}^{(i)} \subseteq \boldsymbol{V}_{CF}^{(t_x)}$, $\boldsymbol{e}_{CF}^{(i)} \subseteq \boldsymbol{E}_{CF}^{(t_x)}$, $\boldsymbol{w}_g{}^{(i)} \subseteq \boldsymbol{w}^{(t_x)}$, $\boldsymbol{\phi}_{gw}^{(i)} \subseteq \boldsymbol{\phi}_w^{(t_x)}$, then $\boldsymbol{g}_{CF}^{(i)}$ is a subgraph of $\boldsymbol{G}_{CF}(t_x)$, which is denoted as $\boldsymbol{g}_{CF}^{(i)} \subseteq \boldsymbol{G}_{CF}(t_x)$.

Definition 4.9 (Frequent Cascading Subgraphs): For a given weighted cascading graph database $\boldsymbol{G}_D$ and a subgraph $\boldsymbol{g}_{CF}^{(i)} \subseteq \boldsymbol{G}_{CF}(t_x)$, the support (occurrence frequency) of $\boldsymbol{g}_{CF}^{(i)}$ is calculated by (4.21):

$$sup(\boldsymbol{g}_{CF}^{(i)}) = \left| \left\{ \boldsymbol{G}_{CF}\left(t_x\right) \,\middle|\, \boldsymbol{G}_{CF}\left(t_x\right) \in \boldsymbol{G}_D \cap \boldsymbol{g}_{CF}^{(i)} \subseteq \boldsymbol{G}_{CF}\left(t_x\right) \right\} \right| \tag{4.21}$$

If $sup(\boldsymbol{g}_{CF}^{(i)})$ is greater than a user-defined minimum threshold $minsup$, then $\boldsymbol{g}_{CF}^{(i)}$ is considered a frequent cascading subgraphs, and is denoted as $\boldsymbol{g}_f^{(i)}$.

In general, graph data mining algorithms require a user-defined $minsup$ to determine whether a subgraph is frequent or not. However, how to set an appropriate $minsup$ is challenging. If the is too high, few or even no subgraphs can be discovered. If the $minsup$ is too low, plenty of useless subgraphs will be included in the results and thus decrease the accuracy of identifying critical components for CPS. Therefore, to address the mentioned issue, we adopted a Top-K structure [123]. For a user-defined $K \geq 1$ and a graph database $\boldsymbol{G}_D$, the Top-K graph mining problem is to find a set $\boldsymbol{F}_g = \left\{ \boldsymbol{g}_f^{(i)} \mid \boldsymbol{g}_f^{(i)} = \left\langle \boldsymbol{v}_f^{(i)}, \boldsymbol{e}_f^{(i)}, \boldsymbol{w}_f{}^{(i)}, \boldsymbol{\phi}_{fw}^{(i)} \right\rangle \right\}$ consists of K subgraphs that their support is greater or equal to that of any other subgraphs not in $\boldsymbol{F}_g$. There is a fundamental distinction between the $minsup$ and Top-K approaches. Compared with the Top-K method, the $minsup$ approach does not prioritize the results according to the frequency of subgraphs. As a result, modifying the $minsup$ parameter might result in the omission of important information. However, in Top-K method, adjusting the K value ensures the consistent retrieval of the top K most frequent subgraphs, irrespective of the adjustments. That is, the most critical components are always prioritized. Note that K is a parameter defined by the user, which should be set with consideration to the defensive capabilities of the CPS operator. This means that the CPS operator must select K by considering the number of critical components that can be simultaneously defended or enhanced. In Section 4.3.3, a thorough analysis of how to determine an appropriate value for K are presented.

In this section, we adopt the TKG algorithm [122] to mine the Top-K frequent cascading subgraphs from the constructed database $\boldsymbol{G}_D$. The critical questions that need to be answered during the graph data mining process are how to effectively traverse all the possible subgraphs and how to efficiently calculate the support of each subgraph. To do so, we utilize the rightmost path extension strategy [20] to traverse the target graphs without missing any nodes and edges. Then, the canonical Depth-First Search (DFS) code [122] is used to represent the graphs in a unified format so that it can significantly facilitate the mining process. The reason we employ DFS rather than Breadth-First Search (BFS) is that BFS is less efficient than DFS when traversing the graph data and generating subgraph candidates [122]. In [20], the authors thoroughly compared the DFS and BFS strategies, focusing on two classic algorithms: FSG (which uses a BFS strategy) [124] and *gSpan* (which uses a DFS strategy) [20]. The test dataset comprises 340 different graphs, each containing an average of 27 nodes and 28 edges, with the largest graph containing 214 nodes and 214 edges. The experimental results indicate that *gSpan* using DFS consumes significantly less computational memory and achieves a better performance, i.e., 15 to 100 times, than FSG using BFS. Therefore, we choose DFS over BFS in our method. Also, this is the reason why we choose the rightmost path extension strategy because it can avoid using BFS and it allows to explore the search space while avoiding generating extra candidates.

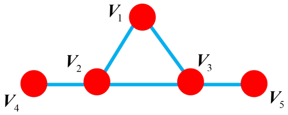


Figure 4.7 The Rightmost Path Extension Strategy.

Rightmost path extension strategy: This strategy follows the principle of depth-first search, and it is implemented over a graph using a recursive stack. In this stack, nodes are used as the basis for an extension, and the currently processed node is called the rightmost node. In general, there are two types of extensions: forward extensions and backward extensions, where forward extensions are used to form an edge to visit new nodes and backward extensions are the opposite. Note that this strategy always implements backward extensions before forward extensions to avoid missing edges. Figure 4.7 gives an example of how the rightmost path extension traverses a graph. Assuming that we start from node V_1, one can randomly choose from its neighbors V_2 and V_3 for the next extension. Taking V_2 as an example, V_4 is next to be visited. Then, because V_4 does not have other neighbors, the strategy will go back to V_2 and visit V_3. At this moment, V_3 has two available neighbors, i.e., V_1 and V_5. Given that the extension between V_3 and V_1 is the backward extension, the rightmost path strategy will visit V_1 first and then V_5. The eventual visiting order of the edges is E_{12}, E_{24}, E_{23}, E_{31} and E_{35}.

Canonical DFS: the depth-first search of a graph is defined as a sequence of the extended edges, sorted in the depth-first search order. Continuing the previous example of Figure 4.7, the sequence of E_{12}, E_{24}, E_{23}, E_{31} and E_{35} is the DFS of the graph. To make sure that each graph and subgraphs in the database can be represented by only a specific DFS during the mining process, the *total order of extended edges* is used to unify the expression of each graph. For the definition of *total order of extended edges*, readers are referred to [122] for details. For a graph, the DFS with the smallest *total order of extended edges* is the canonical DFS. Algorithm 4.2 presents how TKG mines the Top-K frequent cascading subgraphs from the constructed database $\boldsymbol{G}_D$, where *RightMostPathExtension*(*) and *isCanonical*(*) represent the strategy and method implementation for the corresponding targets as discussed. Note that all the components included in Q_K are considered as critical CPS components, and they are denoted as $\boldsymbol{C}_c = \boldsymbol{v}_f^{(1)} \cup \boldsymbol{v}_f^{(2)} \ldots \cup \boldsymbol{v}^{(i)}{}_f \ldots \cup \boldsymbol{v}_f^{(K)}$.

Algorithm 4.2: Mining top-k frequent subgraphs

Input:

$\boldsymbol{G}_D$
K
Q_K: For storing the current top-k frequent subgraphs
Q_C: For storing candidate subgraphs for next extension

Output:

The set of frequent subgraphs: $\boldsymbol{G}_f = \left\{ \boldsymbol{g}_f^{(i)} \mid i = 1, 2, \ldots \right\}$

```
Step 1   $minsup = 1$
Step 2   While $Q_C$ is not empty do
Step 3       $\boldsymbol{g} \leftarrow$ the subgraph with the highest support in $Q_C$
Step 4       $\epsilon \leftarrow RightMostPathExtension(\boldsymbol{g})$
Step 5       For extension $\in \epsilon$ do
Step 6           $\boldsymbol{g}' \leftarrow \boldsymbol{g} \cup extension$
Step 7           If $sup(\boldsymbol{g}') \geq minsup$ and $isCanonical(\boldsymbol{g}')$
Step 8               $Q_K \leftarrow \boldsymbol{g}'$
Step 9               If $|Q_K| > K$
Step 10                  $minsup = min(\boldsymbol{sup}(\boldsymbol{g}_{CF}^{(i)}))$
Step 11              End
Step 12              $Q_C \leftarrow \boldsymbol{g}'$
Step 13          End
Step 14      End
Step 15  End
Step 16  Return $Q_K$
```

In this part, we quantify the correlations between the identified critical components to further evaluate the criticality of each component from the perspectives of manifest and latent correlations as defined in Section 4.3.1. For the convenience of calculation, we merge all the identified frequent subgraphs in $\boldsymbol{F}_g$ into a Cascading Characteristics Graph (CC-Graph).

Definition 4.10 (CC-Graph): For a given frequent subgraph set $\boldsymbol{F}_g$ the corresponding CC-Graph is defined as in (4.22)-(4.25):

$$\boldsymbol{G}_{CC} = \langle \boldsymbol{V}_{CC}, \boldsymbol{E}_{CC}, \boldsymbol{w}_{CC}, \boldsymbol{\phi}_{CC} \rangle \tag{4.22}$$

$$\boldsymbol{V}_{CC} = \bigcup_{i=1}^{I} \boldsymbol{v}_f^{(i)} \tag{4.23}$$

$$\boldsymbol{E}_{CC} = \bigcup_{i=1}^{I} \boldsymbol{e}_f^{(i)} \tag{4.24}$$

$$\boldsymbol{w}^{(t_x)} = \{ w_{\boldsymbol{E}_{CC}} \mid w_{\boldsymbol{E}_{CC}} = f(\boldsymbol{E}_{CC}) \} \tag{4.25}$$

The definition of CC-Graph is similar to the definition of $\boldsymbol{G}_{CF}(t_x)$. Note that $\boldsymbol{G}_{CC}$ is not necessarily a connected graph. In $\boldsymbol{G}_{CC}$, all the edges represent the manifest correlations among the identified critical components. To quantify the manifest correlations, we define the manifest correlation coefficient as in Definition 4.11.

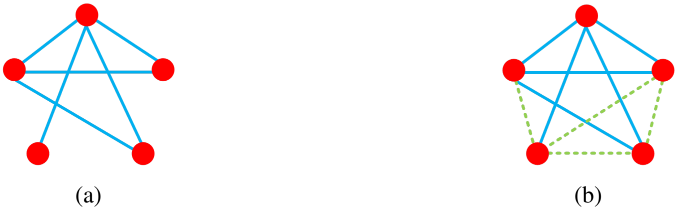
Definition 4.11 (Manifest Correlation Coefficient): For an edge $e_p = (v_q, v_r) \in \boldsymbol{E}_{CC}$, the manifest correlation coefficient is defined as in (4.26)-(4.28):

$$\boldsymbol{C}_{CF}(t_x) = \left\{ R\left(C_{CF}^{(1)}(t_x), C_{CF}^{(2)}(t_x), ..., C_{CF}^{(N)}(t_x) \right) \right\} \tag{4.26}$$

$$\boldsymbol{C}_D = \{ \boldsymbol{C}_{CF}(t_x) \mid t_0 \leq t_x \leq t_u \} \tag{4.27}$$

$$M_{Ce_p}(\boldsymbol{C}_D) = \frac{|\{ \boldsymbol{C}_{CF}(t_x) \mid \exists\ v_q \to v_r \}|}{|\{ \boldsymbol{C}_{CF}(t_x) \mid v_q \in \boldsymbol{C}_{CF}(t_x) \}|} \tag{4.28}$$

By calculating the manifest correlation for all edges in $\boldsymbol{G}_{CC}$, the CC-Graph is updated to $\boldsymbol{G}_{CC} = \left\langle \boldsymbol{V}_{CC}, \boldsymbol{E}_{CC}, \boldsymbol{w}_{CC} \boldsymbol{M}_{\boldsymbol{C}}{}^{T}, \boldsymbol{\phi}_{CC} \right\rangle$, where $\boldsymbol{M}_{\boldsymbol{C}}{}^{T}$ are the sets of M_{Cep} for all edges and they share the same mapping relationship $\boldsymbol{\phi}_{CC}$ as for $\boldsymbol{w}_{CC}$. Then, to thoroughly investigate the cascading characteristics of CPS, we evaluate the latent correlations among the identified critical components. Figure 4.8(a) is an example of a CC-graph, where the blue edges represent the manifest correlations. For the nodes that are not directly connected, they may or may not have latent correlations, as demonstrated in the green edges in Figure 4.8(b). To examine the latent correlation features, we extend $\boldsymbol{G}_{CC}$ to a full connection graph $\boldsymbol{G}'_{CC} = \left\langle \boldsymbol{V}_{CC}, \boldsymbol{E}'_{CC}, \boldsymbol{w}_{CC}\boldsymbol{M_C}^{T} \oplus \boldsymbol{L_C}^{T}, \boldsymbol{\phi}'_{CC} \right\rangle$, where $\boldsymbol{L_C}^{T} = \left\{ L_{Ceq} \middle| \, e_q \in \boldsymbol{E}'_{CC} \right\}$, and the latent correlation coefficient of the extended edges are calculated by (4.29).


(a)

(b)

Figure 4.8 (a) Example of CC-Graph. (b) Latent Correlation Calculation.

Definition 4.12 (Latent Correlation Coefficient): For an edge $e_q = (v_q, v_r) \in \boldsymbol{E}'_{CC}$, the latent correlation coefficient L_{Ceq} is defined as in (4.29):

$$L_{C_{e_q}} = \frac{|\{\boldsymbol{G}_{CF}\left(t_x\right) \mid \exists \, v_q \Rightarrow v_r\}|}{|\boldsymbol{G}_D|} \tag{4.29}$$

Based on the extended CC-Graph, we propose the node criticality index to quantify the importance of each identified critical component. The definition of NC-index is given as follows.

Definition 4.13 (NC-Index): For a critical component $v_q \in \boldsymbol{V}_{CC}$, the NC-index of v_q is denoted as $\boldsymbol{N_C}_q$, and is calculated by (4.30):

$$\boldsymbol{N_C}_q = \sum_{\boldsymbol{E}_y} M_{C_{e_y}} + \sum_{\boldsymbol{E}_y} L_{C_{e_y}} \tag{4.30}$$

where $\boldsymbol{E}_y = \{e_1, e_2, ...e_y, ..., e_Y\}$ consists of all the edges that are connected to v_q including the extended edges. For each critical component, $\boldsymbol{N_C}_q$ evaluates its criticality considering both its manifest and latent correlations. The higher the $\boldsymbol{N_C}_q$ value, the more important the component is for enhancing CPS security.

4.3.3. Case Study

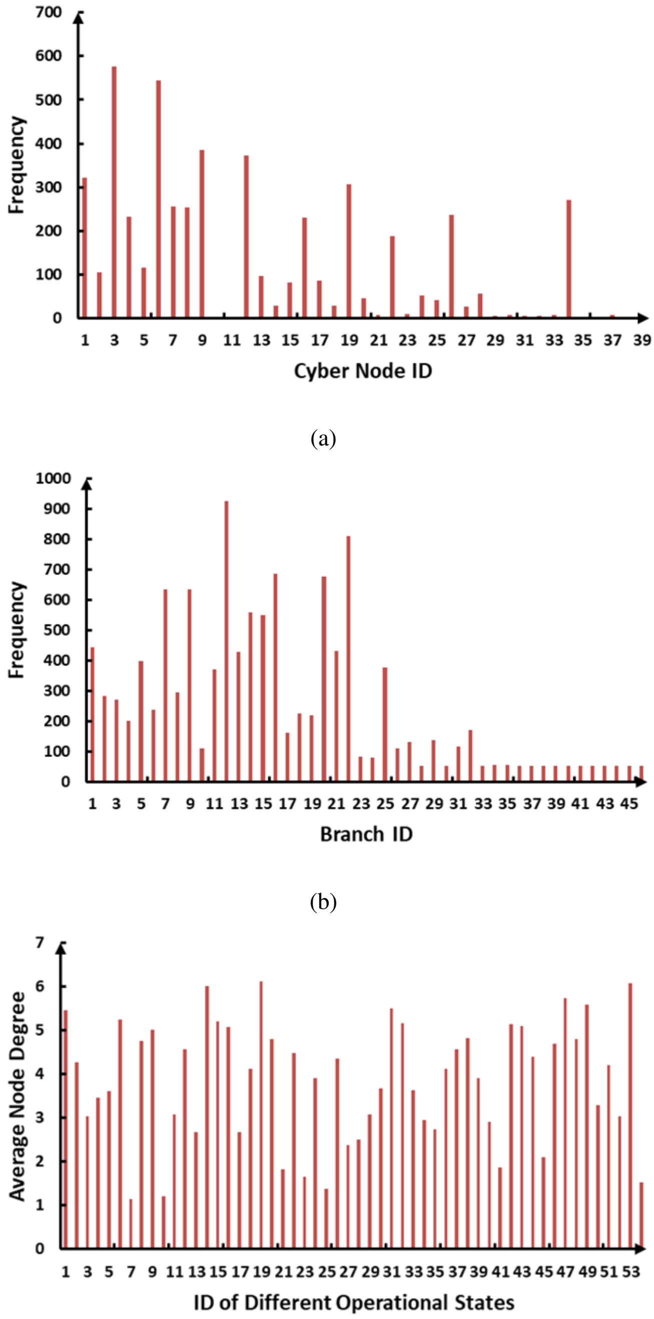
In this section, we implement the proposed methodology to the IEEE 39-bus test system. The modeling details and cyber-physical cascading model can be found in our previous work [101], which contains 78 nodes in total. In this section, we simulate the cascading model for 54 weeks and collect 2,483 cascading chains. For each week, we construct a weighted cascading graph to form the graph database. The simulations are conducted in Python running on a laptop, which is equipped with an Intel i7-8750H CPU @ 2.2 GHz and 16 GB RAM.

From the graph database perspective, Figures 4.9(a) and 4.9(b) present the frequencies of cyber-physical components in the database. The frequencies reflect the extent to which the components contribute to the cascading process. At the cyber system layer, the 5 most critical cyber components are nodes 3, 6, 12, 1, and 19, while at the physical system layer the 5 most critical components are branches 12, 22, 16, 20, and 7. To analyse the constructed weighted graph database, we adopt the average node degree and average node betweenness to describe the graph features of the graph database. The average node degree defines the average amount of nodes connected to a selected node, and it reflects the connectivity of the graph. A high average node degree means that the information or resource can be exchanged in a more efficient manner. On the other hand, the average node betweenness in a graph reflects the extent to which nodes act as bridges in the transmission of information or resources. This metric measures the importance of each node as an intermediary on the shortest paths connecting other pairs of nodes within the network, on average. The results of the graph feature are given in Figures 4.9(c) and 4.9(d). By analyse the results, one can observe that the weighted cascading graphs under different operational states exhibit distinctly different characteristics. In Figure 4.9(c), the average node degree scales from 1.143 (operational state 7) to 6.119 (operational state 19), while in Figure 4.9(d), the highest value (0.045309) is 278 times bigger than the smallest value (0.000255). Such significant variation further proves our argument in the Introduction that the experimental results under one single operational state may not be applicable under different system statuses.

Then, we present the construction results of CC-Graph using the methodology proposed in Section 4.2.2. During the implementation process of the TKG algorithm, we investigate the impact of different K values on the number of identified critical components. In Figure 4.10, as the K value increases, the number of critical components increases along with it. However, the increasing rate has a visible decrease at K=40. On the other hand, in Figure 4.11, we present the relationship between K value and the structural entropy [125] of CC-Graph. The structural entropy $E_{entropy}(\boldsymbol{G}_{CC})$ is used to quantify the information amount contained in each CC-Graph that is constructed based on a given K value, and it can be calculated by following equation (4.31):

$$E_{entropy}(\boldsymbol{G}_{CC}) = -\sum_{i=1}^{I} \Big(P_d(\boldsymbol{v}_f^{(i)}) \times \log_2 P_d(\boldsymbol{v}_f^{(i)}) \Big) \tag{4.31}$$

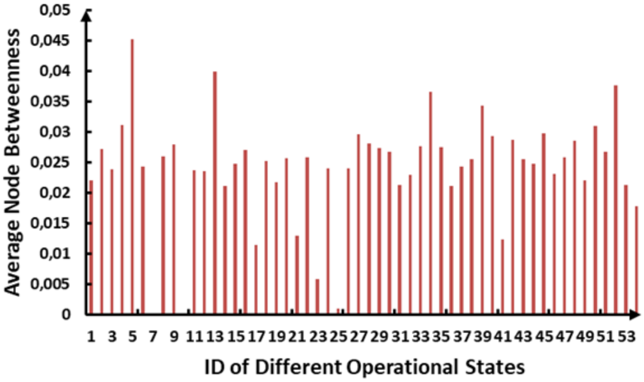
where $P_d(\boldsymbol{v}_f^{(i)})$ is the probability distribution of the degree of node $\boldsymbol{v}_f^{(i)}$. When the structural entropy of a graph is higher, it indicates that the graph is more complex and contains more information. In our case, it is desirable to analyze the CC-Graph with the highest structural entropy, because it means that the corresponding CC-Graph contains the most thorough information of components correlation. In Figure 4.11, one can observe that the $E_{entropy}(\boldsymbol{G}_{CC})$ quickly increases when K is small and is eventually stabilized. This process indicates that as the K increases, the CC-Graph contains more


Frequency
700
600
500
400
300
200
100
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39
Cyber Node ID
Frequency
1000
900
800
700
600
500
400
300
200
100
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45
Branch ID
Average Node Degree
7
6
5
4
3
2
1
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49 51 53
ID of Different Operational States

(a)

(b)

(c)



(d)

Figure 4.9 (a) Frequency of Cyber Nodes in Graph Database (b) Frequency of Physical Branches in Graph Database. (c) Average Node Degree of Weighted Cascading Graphs (d) Average Node Betweenness of Weighted Cascading Graphs

information of component correlation. Also, when the K increases beyond a certain point, the increase of K will not add new information to the CC-Graph and only causes small changes to the $E_{entropy}(\boldsymbol{G}_{CC})$. Therefore, when K value is too low, some critical components correlation information may be missed in the CC-Graph. On the other hand, when K value is too high, it does not add new and useful information to the CC-Graph while it also increases the cost of defending critical components. Based on the discussion above, the optimal K value is determined when the corresponding $E_{entropy}(\boldsymbol{G}_{CC})$ reaches the maximum. In Figure 4.11, the optimal K is 40.

Figure 4.12 presents the generated CC-Graph when K=40. In this graph, there are in total 21 critical cyber nodes and 38 critical physical branches. For each pair of nodes that are directly connected, apparent manifest correlations exist. For each pair of nodes that are indirectly connected but have an accessible path, latent correlations exist. Note that the latent correlations in Figure 4.12 only consider the mined frequent cascading subgraphs. They frequently appear in the graph database, and it does not prove that there are no latent correlations between those node pairs having no accessible path. For example, node 1 and node 24 on the top of the CC-Graph are not directly or indirectly connected, but there is still a possible latent correlation between them. From a global perspective, the CC-Graph in Figure 4.11 is not a connected graph, and the node degree of each node is not high (the maximum value is 3). It indicates that the range of the frequent cascading patterns is not extensive. However, by observing the marked area, this is a comparatively large connected graph, which indicates that if any node in this area fails, it may cause a severe impact on the system operation. In the next part, we will further quantify the criticality of each node in Figure 4.12 by using the proposed

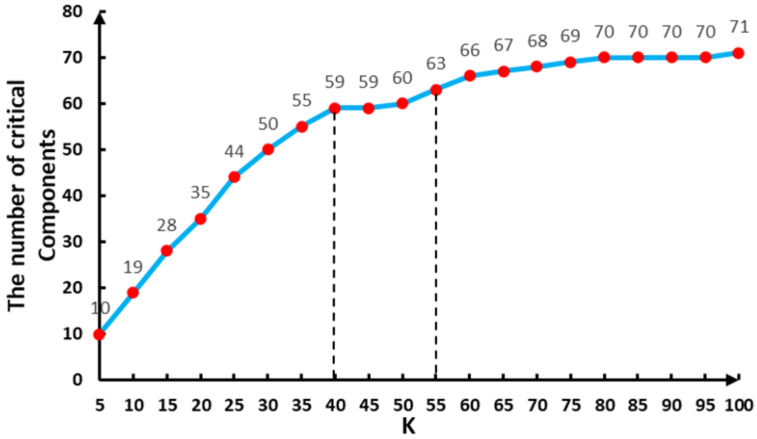


Figure 4.10 The Number of Identified Critical Components under Different K.

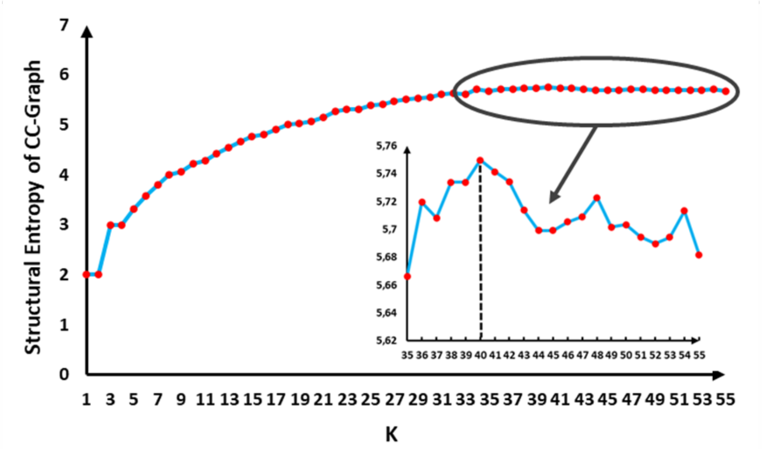


Figure 4.11 The Relationship between K Value and the Structural Entropy of CC-Graph.

NC-Index.

In Figure 4.13 and Table 4.4, we present the calculation results of all the indices we proposed in Section 4.2.2. In Figure 4.13(a), we only consider the manifest correlation. The ranking results of the manifest correlation coefficients are decided by two factors, i.e., the evident support in the historical data and the node degree of the nodes in the CC-Graph. Figure 4.13(a) also proves this point and the nodes with a higher degree are comparatively more critical than the other low-degree nodes. Also, the results indicate that the most high-ranked components are in the largest subgraph. This indicates that these nodes have tighter connections with the other nodes and a wider range to propagate

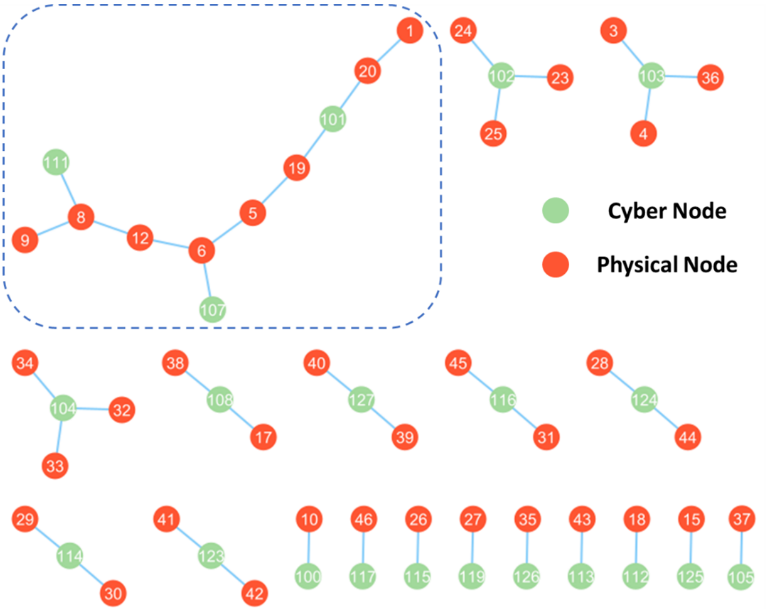
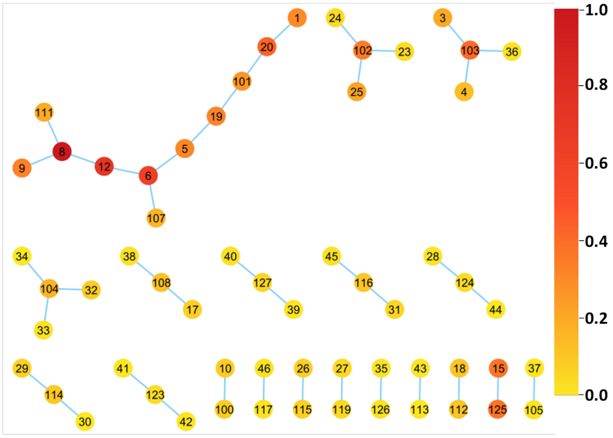
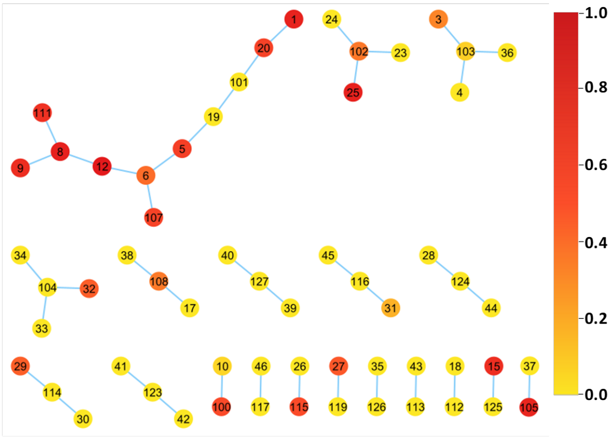


Figure 4.12 Constructed CC-Graph when K=40.



(a)

the failures. The detailed ranking information is given in Table 4.4. In Figure 4.13(b), we jointly consider the manifest correlations and latent correlations. Compared with Figure 4.13(a), the most critical components are still mainly distributed in the largest subgraph. However, part of the critical components from the largest subgraph rank lower, while some components from smaller subgraphs rank higher. This is because the latent correlation considers the global relationships among components, and it quantifies



(b)

Figure 4.13 (a) CC-Graph with Only Manifest Correlation Coefficients (b) CC-Graph with NC-Index

the risk of indirectly triggering a cascading failure.

Table 4.4: Ranking of Critical Components Considering Different Indices

Ranking	Considering only manifest correlation coefficient	Considering NC-Index
1	8 (physical)	12 (physical)
2	12 (physical)	8 (physical)
3	6 (physical)	25 (physical)
4	20 (physical)	1 (physical)
5	103 (cyber)	105 (cyber)
6	15 (physical)	9 (physical)
7	125 (cyber)	15 (physical)
8	102 (cyber)	111 (cyber)
9	9 (physical)	5 (physical)
10	19 (physical)	20 (physical)

In this part, we compare the proposed method with the existing literature to prove its effectiveness. We compare the performance of methods from two aspects: load loss and network efficiency. For the load loss, we implement each method to identify the top-5 critical components for the CPS of IEEE 39-bus system as explained in [101]. Then, we traverse the possible combination of those components and disconnect them to observe the load loss after the cascading failures. For each method, we record the highest load loss. Similarly, we use the same approach to calculate the network efficiency of each remaining network after the cascading failures. It is worth mentioning that unlike load loss, the network efficiency only indicates the topological features of the network, and it quantifies the network connectivity.

We compare the proposed method with reference [101] and [126], where [101] considered the cascading failure data and identified the critical components based on the proposed index while [126] evaluated the nodes importance for power system from the perspective of centrality measures. In Figure 4.14, we present the comparison results. From the perspective of load loss, one can observe that the removal of the critical components indentified by the proposed method can cause a much higher load loss, while there is no load loss in the results of [126]. The reason behind the results is that reference [126] neglects the node heterogeneity of CPS and only consider the topological aspects of networks. In real industrial scenarios, we place greater emphasis on factors that can directly lead to security issues and financial losses, such as load loss. Besides, by analyzing the network efficiency results, one can observe that there is a clear decrease in all three methods compared with the initial network. However, by combining the results of load loss and network efficiency, the critical components identified by our method can cause more catastrophic cascading failures by inflicting a comparable degree of damage on the network. The comparison results effectively confirm the precision of our method in identifying critical components compared with the existing literature.

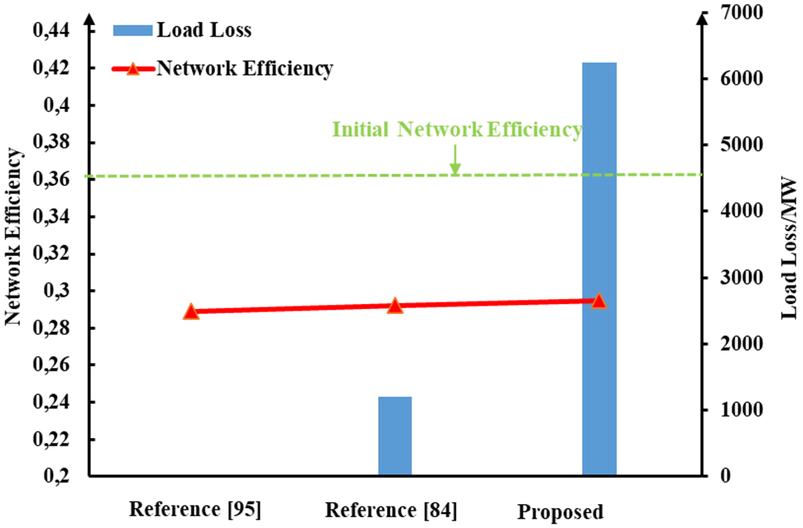


Figure 4.14 The Comparison Results.

4.3.4. System Behavior Comparisons between Generated Digital Sibling and the Real System

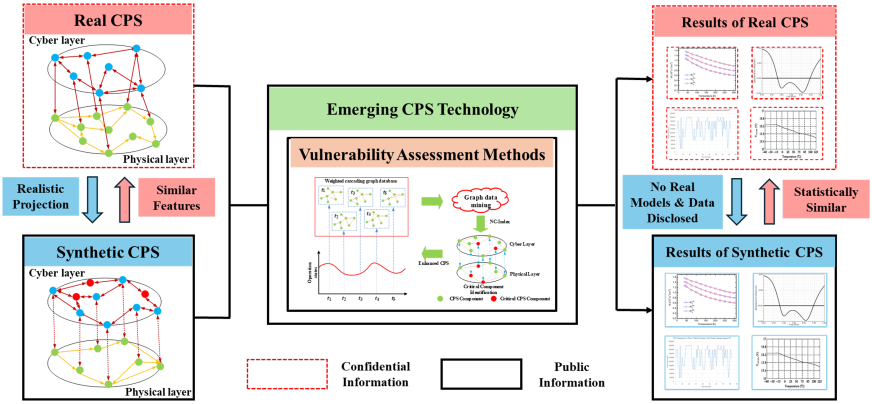


Figure 4.15 System Behavior Comparisons between Digital Siblings and Real Systems

The ultimate goal of a digital sibling is to enable researchers to develop practical algorithms and methodologies for CPS while ensuring the confidentiality of real models and data. In this section, we apply the proposed GraphCCI on both the generated digital sibling, as described in Section 3.3, and its corresponding real system, i.e., the IEEE 39-bus system. As shown in Figure 4.15, we compare the vulnerability assessment results of these two systems to prove that the results are statistically similar. This similarity allows researchers to conduct practical, realistic research while the confidentiality of the real systems are well preserved.

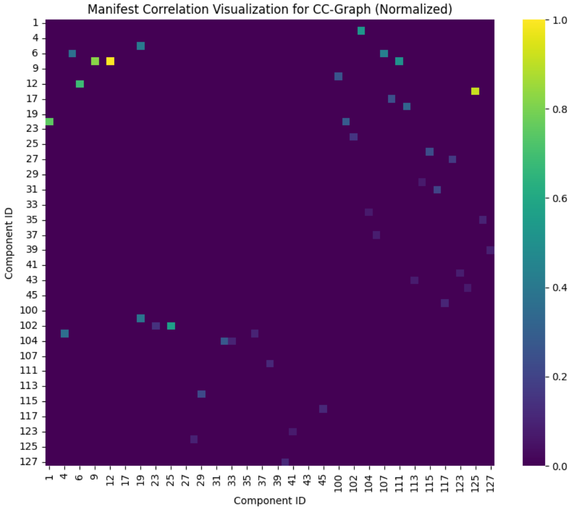


Figure 4.16 Manifest Correlation of Cyber-Physical Components in CC-Graph (IEEE 39-Bus System)

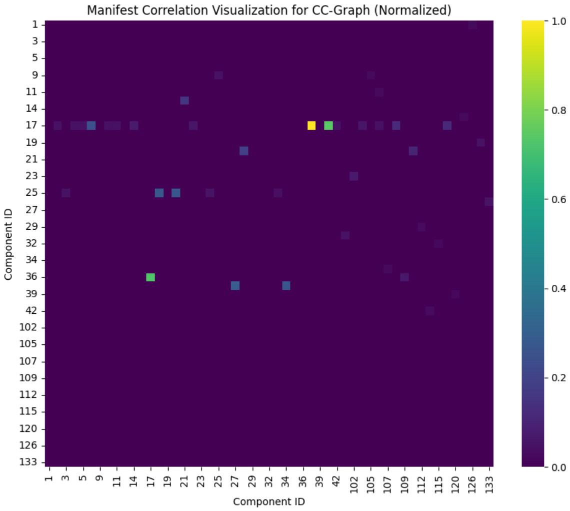


Figure 4.17 Manifest Correlation of Cyber-Physical Components in CC-Graph (Digital Sibling)

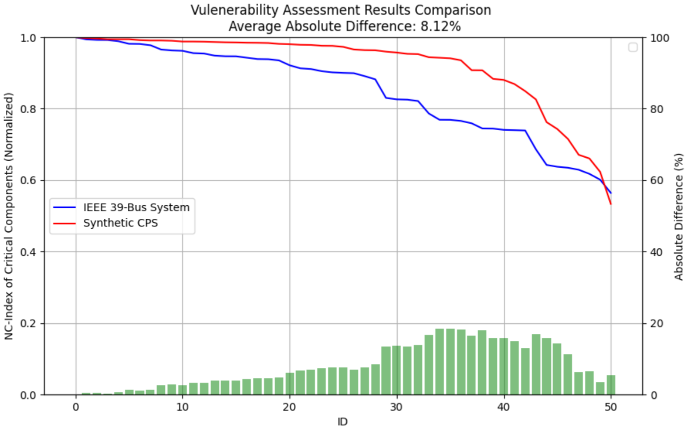


Figure 4.18 Vulnerability Characteristics Comparison

In Figure 4.16 and Figure 4.17, the manifest correlations [127] of CPS components are visualized. From the results of the two diagrams, the distribution of the manifest correlation in both exhibits similar characteristics, indicating that only a small number of key manifest correlation in the system have extremely high probability weights. That is, there is a limited number of critical components in the system that have high probability to propagate the cascading failures. Moreover, in Figure 4.18, the vulnerability parameters, i.e., NC-Index [127], of the top 50 critical CPS components identified by GraphCCI are compared. On average, the difference in vulnerability assessment results between the synthetic CPS (digital sibling) and IEEE 39-bus system is 8.12%. More specifically, in the highest ranking critical components, the difference of NC-Index is even lower than the average, indicating the similar vulnerability characteristics of the synthetic CPS and real CPS. Overall, it is demonstrated that to test the proposed methods and investigate the systematic characteristics of CPS, one can implement the proposed methods on the synthetic networks generated by SibGen, thereby fully preserving the confidentiality of real system models and data.

4.4. Conclusion

This chapter investigates the vulnerability assessment methods for CPS. It identifies the research gap that the existing work only evaluates the CPS at a single time instant while fails to capture the CPS vulnerability characteristics under time-varying operational states. Furthermore, to thoroughly investigate the vulnerability features of CPS, we propose the definition of manifest and latent correlations of CPS components.

From the perspective of manifest correlations, we model the cascading failures considering the interaction of cyber and physical layers. By combining cascading failure chains of all-time instants, a cascading failure chain database is generated. This captures the intricate manifest relationships among components and reveals the fault propagation mechanism of CPS under different operating conditions. The sequential data mining algorithms are adopted to identify the vulnerable sequences. The total sequential vulnerability metric is proposed to quantify the vulnerabilities of CPS components. The simulation results show that there is only a limited number of critical CPS components. The resilience of the cyber-physical system can be greatly improved if these critical components are reinforced and cyber secured.

From the perspective of latent correlations, we propose a graph data mining-based critical components identification model named GraphCCI, which jointly evaluates the criticality of CPS components from the perspectives of manifest and latent correlations. First, we abstract the cascading failure data under different operational states into a weighted cascading graph database which captures both the latent and manifest correlations of CPS components. Then, the TKG algorithm is adopted to identify the frequent subgraphs in the constructed graph database. Meanwhile, the definition of CC-Graph is proposed to model the overall cascading features based on the graph mining results. Finally, the NC-Index is proposed to evaluate the criticality of each CPS component. Our case study reveals that the cyber-physical system shows different cascading features under different system conditions. Verifications on the IEEE 39-bus test system demonstrate the effectiveness of our method. The identification results can provide an important reference to enhance CPS security and prevent cascading failures and even a blackout.

5

Conclusions and Future Research

5.1. Conclusions

This dissertation focuses on the synthetic network generation and vulnerability assessment of CPS. With the rapid power grid digitization, substantial research is essential to address its emerging CPS challenges. This thesis responds to the growing need for reliable CPS test systems, as real CPS models and data are highly confidential due to national security concerns. The synthetic networks generated in this thesis allow researchers to test newly developed methods and obtain feedback comparable to that from real CPS, facilitating innovation without compromising security. First, this thesis highlights methods for generating and validating synthetic networks under varying levels of data completeness and availability while ensuring the confidentiality of the real CPS data and models. Then, it also evaluates the vulnerability characteristics of CPS under time-varying operational states and conducts in-depth analysis on the correlations between CPS components. Furthermore, the proposed vulnerability assessment method is implemented on both the generated synthetic CPS and the real system. The comparison results prove that the synthetic CPS can be utilized as alternative test system and exhibits similar characteristics as the real network. The main contributions of this dissertation are the development of three generative models tailored to different implementation scenarios and two methods for identifying critical components using data-driven techniques. The proposed algorithms and models are rigorously validated through extensive simulation experiments. Overall, this dissertation offers realistic test systems for CPS researchers, thoroughly analyses and quantifies the correlations between cyber-physical components, and reveals CPS vulnerability characteristics under varying operational states. The detailed research outputs are listed below:

Implementation of a large-scale synthetic topology generation method for continental Europe (Q1):

Chapter 2 addresses the challenge of incomplete CPS data, where only power system data is available. To overcome this, a two-stage generative model is proposed, enabling the creation of realistic, large-scale synthetic CPS based on existing power grids. The generated synthetic networks are thoroughly validated using complex network parameters, which is a powerful tool to quantitatively compare the difference between the synthetic networks and the real networks. It can comprehensively capture the characteristics of a cyber-physical power system from both topological and operational perspectives. Furthermore, the proposed method addresses the connectivity issues commonly found in existing literature. This work marks a pioneering step in synthetic CPS modelling, establishing a robust foundation for future research aimed at uncovering critical characteristics, patterns, and mechanisms inherent to cyber-physical systems.

Scalable generative model for generating synthetic CPS with realistic network features (Q2):

The first half of Chapter 3 assumes the availability of both cyber and physical system data and introduces Graph-CPS, a scalable generative model designed to generate synthetic CPS topologies. This method accurately reflects realistic network feature distributions while ensuring the confidentiality of the real models and data. Graph-CPS is capable of learning and reproducing various complex network parameters, capturing the distributions of different network features from input networks. Validation results demonstrate that Graph-CPS can accurately capture the characteristics of input networks, not only across different network types but also across varying network sizes. The guiding philosophy of this approach lies in its ability to generalize across diverse network configurations, ensuring that synthetic topologies retain the inherent complexity and scalability of real CPS networks. This method paves the way for a deeper understanding of CPS behaviour, offering valuable insights into CPS structures and network parameter configurations.

Digital sibling generator for CPS (Q2)

Chapter 3 focuses on generating both topological and operational models for CPS, assuming that complete CPS data is available. To capture complex system behaviors from high-dimensional CPS data, a hybrid generator, SibGen, is introduced to create a digital sibling of the real CPS. A key distinction is made between the concepts of digital twins and digital siblings, highlighting their fundamental differences. The core idea behind digital siblings lies in balancing the fidelity of synthetic models with the need for data confidentiality. SibGen not only learns the topological features of CPS but also effectively captures the operational characteristics. Besides, SibGen is capable of delivering alternative synthetic test systems. The digital siblings replicate the characteristics of the real systems without disclosing any real information. This dual capability allows research conducted on the generated digital sibling to closely mirror real-world scenarios, making the research findings more practical and convincing. By enabling studies that are more aligned with actual conditions, SibGen empowers further research in CPS, laying a solid foundation for the rapid development of CPS technologies.

CPS vulnerability assessment under time-varying operational states (Q3):

Chapter 4 investigates the vulnerability assessment methods for CPS. It identifies the research gap that the existing work only evaluates the CPS at a single time instant while fails to capture the CPS vulnerability characteristics under time-varying operational states. The goal of chapter 4 is to shift from static vulnerability assessments to dynamic, data-driven approaches that better capture the intricate interdependencies and evolving risks within CPS. To provide a more comprehensive analysis to CPS vulnerability, the concepts of manifest and latent correlations among CPS components are proposed:

Manifest correlations represent the observable relationships between CPS components, such as those evident in cascading events. By combining cascading failure chains over multiple time instants, a cascading failure chain database is constructed, capturing the complex interdependencies among components. Sequential data mining algorithms are then applied to identify vulnerable sequences within these chains, and a total sequential vulnerability metric is introduced to quantify the vulnerabilities of individual CPS components. Simulation results reveal that only a small subset of components is critical, and reinforcing these key components can significantly improve system resilience.

Latent correlations describe the statistically inferred relationships between CPS components, which may not be directly observable but are essential for understanding system behavior. Building on the concepts of both manifest and latent correlations, a graph data mining-based model, GraphCCI, is introduced to identify critical components by evaluating both types of correlations. The case study demonstrates that CPS exhibits distinct cascading behaviors under different operational conditions. Validations conducted on the IEEE 39-bus test system confirm the effectiveness of this method. The identification results offer valuable insights for enhancing CPS security, preventing cascading failures, and mitigating the risk of blackouts.

5.2. Future Work

Based on the research output of the thesis and the discussion in Section 6.1, the potential research topics of synthetic CPS and CPS vulnerability assessment are listed as follows:

Data confidentiality: In Chapter 3, we discussed the data confidentiality issues when generating synthetic networks based on the real CPS data. Essentially, this is a trade-off problem between the data confidentiality and the similarity of synthetic and real systems. Therefore, how to control the difference between the synthetic and real systems is a promising research topic. Ideally, it is desirable to minimize the differences between synthetic networks and real systems, especially the overall system behaviors. However, higher similarity means higher risk to expose the real CPS data. Although the topology and the parameter configuration of synthetic networks are different from the real ones, it is still possible for the adversaries to initiate catastrophic attacks based on the overall system characteristics. For example, if the adversaries learn that the real system is a mesh network, then they only need to initiate multiple random attacks to achieve high attack impact. Because the mesh network is resilient to attacks with specific target while it is vulnerable to multiple random attacks due to its damage to the network connectivity. Therefore, we need to find a balance point where the real CPS data is well secured and the synthetic network can still perform as an effective test system.

More complexity models in synthetic networks: In Chapter 2 and 3, we have successfully generated the power flow and information flow models for CPS. However, to achieve more accurate simulation results, more complexity models are required. For example, the power dynamic models, the communication protocols, etc. On the other hand, more complexity models means higher dimensions of the input data, which significantly increases the difficulties of generating the realistic synthetic systems. Currently, the GNN-based generative models are capable of generating similar parameter distributions. However, the power systems and the communication networks are sophisticated manufactured systems, and small deviations in the parameter settings can cause different simulation results. For example, small changes in the load parameters can result into a different set of power flow solutions. Under such considerations, if more complexity models are included, it is challenging to learn the accurate characteristics of the high-dimensional real data while maintaining the similar operational behaviors.

Evolution mechanism of CPS vulnerability assessment: In Chapter 4, we clarify that the CPS vulnerability characteristics changes under different system operating states, based on which we identify the critical components for CPS. However, the vulnerability evolution mechanism of the system is still unclear and needs to be further explored. How do the vulnerability characteristics of a system change over time? Additionally, in case of contingencies, how will these vulnerability characteristics evolve? In chapter 4, we managed to construct the weighted vulnerability graph for each time instant. Therefore, it is desirable to analyze the connections between these time-series graph data and further explore the vulnerability features of CPS. Graph data mining algorithms are promising to identify the time-related patterns, e.g., periodical patterns, sequential patterns, etc. Based on the identified time-related patterns, one can analyze the evolution process of the vulnerability features and further reveal the systematic mechanisms of CPS.

Bibliography

[1] O. P. Veloza and R. H. Cespedes. “Vulnerability of the Colombian electric system to blackouts and possible remedial actions”. In: *2006 IEEE Power Engineering Society General Meeting*. 2006.

[2] E. Zio and T. Aven. “Uncertainties in smart grids behavior and modeling: What are the risks and vulnerabilities? How to analyze them?” In: *Energy Policy* 39.10 (Oct. 2011), pp. 6308–6320.

[3] S. Kamali and T. Amraee. “Blackout prediction in interconnected electric energy systems considering generation re-dispatch and energy curtailment”. In: *Applied Energy* 187 (Feb. 2017), pp. 50–61.

[4] M. Rong, C. Han, and L. Liu. “Critical infrastructure failure interdependencies in the 2008 Chinese winter storms”. In: *2010 International Conference on Management and Service Science*. Wuhan, China, 2010, pp. 1–4.

[5] O. P. Veloza and F. Santamaria. “Analysis of major blackouts from 2003 to 2015: Classification of incidents and review of main causes”. In: *The Electricity Journal* 29.7 (Sept. 2016), pp. 42–49.

[6] R. M. Lee, M. J. Assante, and T. Conway. *Analysis of the Cyber Attack on the Ukrainian Power Grid*. Available: https://ics.sans.org/ics-library. Mar. 2016.

[7] M. J. Assante, R. M. Lee, and T. Conway. *Modular ICS Malware*. Available: https://ics.sans.org/ics-library. Aug. 2017.

[8] A. B. D. Costa and G. Suroyo. *Power Restored to Some Areas in Indonesia Capital, Parts of Java After 9 Hours*. [Online]. Aug. 2019. URL: https://www.reuters.com/article/us-indonesia-power-idUSKCN1UU060.

[9] J. Pearson. *Russian spies behind cyber attack on Ukraine power grid in 2022 - Researchers*. [Online]. Nov. 2023. URL: https://www.reuters.com/technology/cybersecurity/russian-spies-behind-cyberattack-ukrainian-power-grid-2022-researchers-2023-11-09/.

[10] Q. Zhou and J. W. Bialek. “Approximate model of European interconnected system as a benchmark system to study effects of cross-border trades”. In: *IEEE Transactions on Power Systems* 20.2 (May 2005), pp. 782–788. DOI: 10.1109/TPWRS.2005.846622.

[11] R. Espejo, S. Lumbreras, and A. Ramos. “A Complex-Network Approach to the Generation of Synthetic Power Transmission Networks”. In: *IEEE Systems Journal* 13.3 (Sept. 2019), pp. 3050–3058. DOI: 10.1109/JSYST.2018.2879701.

[12] S. Soltan, A. Loh, and G. Zussman. “A learning-based method for generating synthetic power grids”. In: *IEEE Systems Journal* 13.1 (Mar. 2019), pp. 625–634. DOI: 10.1109/JSYST.2018.2883902.

[13] K. M. Gegner, A. B. Birchfield, T. Xu, K. S. Shetye, and T. J. Overbye. “A methodology for the creation of geographically realistic synthetic powerflow models”. In: *2016 IEEE Power and Energy Conference at Illinois (PECI)*. Urbana, IL, Apr. 2016, pp. 1–6. DOI: 10.1109/PECI.2016.7596202.

[14] T. Xu, A. B. Birchfield, K. S. Shetye, and T. J. Overbye. “Creation of Synthetic Electric Grid Models for Transient Stability Studies”. In: *2017 IREP Symposium Bulk Power System Dynamics and Control*. Espinho, Portugal, 2017. DOI: 10.1109/IREP.2017.7989848.

[15] T. Xu, A. B. Birchfield, and T. J. Overbye. “Modeling, Tuning and Validating System Dynamics in Synthetic Electric Grids”. In: *IEEE Transactions on Power Systems* 33.6 (Nov. 2018), pp. 6501–6509. DOI: 10.1109/TPWRS.2018.2871842.

[16] F. Chung and L. Lu. “The Configuration Model for Power Law Graphs”. In: *Complex Graphs and Networks*. CBMS Lecture Series. Providence, RI: American Mathematical Society, 2006, pp. 223–237. DOI: 10.1090/conm/302/05619.

[17] J. You, R. Ying, X. Ren, W. L. Hamilton, and J. Leskovec. “GraphRNN: Generating Realistic Graphs with Deep Auto-regressive Model”. In: *International Conference on Machine Learning*. Stockholm, Sweden, 2018. DOI: 10.1145/3209978.3210010.

[18] T. N. Kipf and M. Welling. “Variational Graph Auto-Encoders”. In: *arXiv preprint arXiv:1611.07308* (Nov. 2016).

[19] J. Pei, J. Han, and H. Tong. “Mining Sequential Patterns by Pattern-Growth: The PrefixSpan Approach”. In: *IEEE Transactions on Knowledge and Data Engineering* 16.11 (Nov. 2004), pp. 1424–1440. DOI: 10.1109/TKDE.2004.77.

[20] X. Yan and J. Han. “gSpan: Graph-based Substructure Pattern Mining”. In: *Proceedings of the 2002 IEEE International Conference on Data Mining*. Japan, 2002, pp. 721–724. DOI: 10.1109/ICDM.2002.1047831.

[21] J. Sztipanovits, T. Bapty, X. Koutsoukos, Z. Lattmann, S. Neema, and E. Jackson. “Model and Tool Integration Platforms for Cyber–Physical System Design”. In: *Proceedings of the IEEE* 106.9 (Sept. 2018), pp. 1501–1526. DOI: 10.1109/JPROC.2018.2854278.

[22] Y. Feng, B. Hu, H. Hao, Y. Gao, Z. Li, and J. Tan. “Design of Distributed Cyber–Physical Systems for Connected and Automated Vehicles with Implementing Methodologies”. In: *IEEE Transactions on Industrial Informatics* 14.9 (Sept. 2018), pp. 4200–4211. DOI: 10.1109/TII.2018.2874321.

[23] Y. Wang, C. F. Chen, P. Y. Kong, H. Li, and Q. Wen. “A Cyber–Physical–Social Perspective on Future Smart Distribution Systems”. In: *Proceedings of the IEEE* 111.7 (July 2023), pp. 694–724. DOI: 10.1109/JPROC.2023.1234567.

[24] K. R. Davis, J. Smith, L. Johnson, and M. Brown. “A Cyber-Physical Modeling and Assessment Framework for Power Grid Infrastructures”. In: *IEEE Transactions on Smart Grid* 6.5 (Sept. 2015), pp. 2464–2475. DOI: 10.1109/TSG.2015.2401234.

[25] Y. Cai, Y. Cao, Y. Li, T. Huang, and B. Zhou. “Cascading Failure Analysis Considering Interaction Between Power Grids and Communication Networks”. In: *IEEE Transactions on Smart Grid* 7.1 (Jan. 2016), pp. 530–538. DOI: 10.1109/TSG.2015.2491234.

[26] T. Zang, S. Gao, B. Liu, T. Huang, T. Wang, and X. Wei. “Integrated Fault Propagation Model-Based Vulnerability Assessment of the Electrical Cyber-Physical System Under Cyber-Attacks”. In: *Reliability Engineering System Safety* 189 (Sept. 2019), pp. 232–241. DOI: 10.1016/j.ress.2019.03.012.

[27] H. Li. “Network Topology Design”. In: *Communications for Control in Cyber Physical Systems*. Amsterdam, the Netherlands: Elsevier, Dec. 2016, pp. 149–180.

[28] T. Wang, Q. Long, X. Gu, and W. Chai. “Information Flow Modeling and Performance Evaluation of Communication Networks Serving Power Grids”. In: *IEEE Access* 8 (Jan. 2020), pp. 13735–13747.

[29] M. Shahraeini, M. H. Javidi, and M. S. Ghazizadeh. “Comparison Between Communication Infrastructures of Centralized and Decentralized Wide Area Measurement Systems”. In: *IEEE Transactions on Smart Grid* 2.1 (Mar. 2011), pp. 206–211.

[30] X. P. Ji, B. Wang, D. Liu, and T. Zhao. “Review on Interdependent Networks Theory and Its Applications in the Structural Vulnerability Analysis of Electrical Cyber-physical System”. In: *Proceedings of the China Society for Electrical Engineering (CSEE)*. Vol. 36. 17. Sept. 2016, pp. 4521–4532.

[31] J. Valtari and S. Joshi. *Centralized Protection and Control*. Tech. rep. ABB, 2019.

[32] J. Hu, Z.-H. Li, and X. Z. Duan. “Structural Feature Analysis of the Electric Power Dispatching Data Network”. In: *Proceedings of the China Society for Electrical Engineering (CSEE)*. Vol. 29. 4. Feb. 2009, pp. 53–59.

[33] Y. Yan, Y. Qian, H. Sharif, and D. Tipper. “A Survey on Smart Grid Communication Infrastructures: Motivations, Requirements and Challenges”. In: *IEEE Communications Surveys Tutorials* 15.1 (Jan. 2013), pp. 5–20.

[34] F. M. Preparata and M. I. Shamos. *Computational Geometry: An Introduction*. New York, NY, USA: Springer-Verlag, 1985.

[35] G. W. Li, W. Y. Ju, X. Z. Duan, and D. Y. Shi. “Transmission Characteristics Analysis of the Electric Power Dispatching Data Network”. In: *Proceedings of the China Society for Electrical Engineering (CSEE)*. Vol. 32. 22. Aug. 2012, pp. 141–148.

[36] Y. Wang, P. Yemula, and A. Bose. “Decentralized Communication and Control Systems for Power System Operation”. In: *IEEE Transactions on Smart Grid* 6.2 (Mar. 2015), pp. 885–893.

[37] W. J. Bai, T. Zhou, Z. Q. Fu, Y. H. Chen, X. Wu, and B. H. Wang. “Electric Power Grids and Blackouts in Perspective of Complex Networks”. In: *2006 International Conference on Communications, Circuits and Systems*. Guilin, China, 2006, pp. 2687–2691.

[38] ENTSO-E. *ENTSO-E Transmission System Map*. Available: https://www.entsoe.eu/data/map/. 2019.

[39] *Synthetic Cyber-Physical Power Systems for Continental Europe*. Available: https://github.com/Cyber-Resilient-Power-Grids/Synthetic-CPS.

[40] O. Boyaci, M. R. Narimani, K. Davis, and E. Serpedin. “Generating Connected, Simple, and Realistic Cyber Graphs for Smart Grids”. In: *2022 IEEE Texas Power and Energy Conference (TPEC)*. College Station, TX, USA, 2022, pp. 1–6.

[41] B. Q. et al. “An Emerging Survivability Technology for Dispatching Service of Electric Power Communication Network”. In: *IEEE Access* 6 (2018), pp. 21231–21241.

[42] R. Atat, M. Ismail, S. S. Refaat, E. Serpedin, and T. Overbye. “Cascading Failure Vulnerability Analysis in Interdependent Power Communication Networks”. In: *IEEE Systems Journal* 16.3 (Sept. 2022), pp. 3500–3511.

[43] Y. Zhang, T. Jiang, Q. Shi, W. Liu, and S. Huang. “Modeling and Vulnerability Assessment of Cyber Physical System Considering Coupling Characteristics”. In: *International Journal of Electrical Power Energy Systems* 142.Part B (Nov. 2022).

[44] X. Fan, D. Wang, S. Aksoy, A. Tbaileh, Q. H., T. Fu, and J. Ogle. *Coordination of Transmission, Distribution and Communication Systems for Prompt Power System Recovery After Disasters*. Tech. rep. PNNL-28598, 2019.

[45] A.-L. Barabási and R. Albert. “Emergence of Scaling in Random Networks”. In: *Science* 286.5439 (Oct. 1999), pp. 509–512. DOI: 10.1126/science.286.5439.509.

[46] P. Erdős and A. Rényi. “On Random Graphs I”. In: *Publicationes Mathematicae Debrecen* 6 (1959), pp. 290–297.

[47] S. Boccaletti and et al. “Complex Networks: Structure and Dynamics”. In: *Physics Reports* 424.4 (2006), pp. 175–308.

[48] M. A. Klopotek, S. T. Wierzchon, and K. Trojanowski. “Intelligent Information Processing and Web Mining”. In: *Proceedings of the International IIS: IIPWM 06 Conference*. Ustron, 2006.

[49] F. Gutierrez and et al. “Vulnerability Analysis of Power Grids Using Modified Centrality Measures”. In: *Discrete Dynamics in Nature and Society* 2013 (2013).

[50] E. Bompard, D. Wu, and F. Xue. “Structural Vulnerability of Power Systems: A Topological Approach”. In: *Electric Power Systems Research* 81.7 (July 2011), pp. 1334–1340.

[51] C. Li and et al. “Method for Evaluating the Importance of Power Grid Nodes Based on PageRank Algorithm”. In: *IET Generation, Transmission & Distribution* 8.11 (Nov. 2014), pp. 1843–1847.

[52] Y. Liu, A. Stefanov, and P. Palensky. “Generating Large-Scale Synthetic Communication Topologies for Cyber-Physical Power Systems”. In: *IEEE Transactions on Industrial Informatics* (2024). Early Access. DOI: 10.1109/TII.2024.3438232.

[53] A. B. Birchfield, T. Xu, K. M. Gegner, K. S. Shetye, and T. J. Overbye. “Grid Structural Characteristics as Validation Criteria for Synthetic Networks”. In: *IEEE Transactions on Power Systems* 32.4 (July 2017), pp. 3258–3265.

[54] M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst. “Geometric Deep Learning: Going beyond Euclidean data”. In: *IEEE Signal Processing Magazine* 34.4 (July 2017), pp. 18–42. DOI: 10.1109/MSP.2017.2693418.

[55] M. Gori, G. Monfardini, and F. Scarselli. “A New Model for Learning in Graph Domains”. In: *Proceedings of the 2005 IEEE International Joint Conference on Neural Networks*. Vol. 2. Montreal, QC, Canada, 2005, pp. 729–734. DOI: 10.1109/IJCNN.2005.1555942.

[56] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini. “The Graph Neural Network Model”. In: *IEEE Transactions on Neural Networks* 20.1 (Jan. 2009), pp. 61–80. DOI: 10.1109/TNN.2008.2005605.

[57] K. Ishaque, Z. Salam, M. Amjad, and S. Mekhilef. “An Improved Particle Swarm Optimization (PSO)–Based MPPT for PV With Reduced Steady-State Oscillation”. In: *IEEE Transactions on Power Electronics* 27.8 (Aug. 2012), pp. 3627–3638. DOI: 10.1109/TPEL.2012.2185713.

[58] R. Levie, F. Monti, X. Bresson, and M. M. Bronstein. “CayleyNets: Graph Convolutional Neural Networks With Complex Rational Spectral Filters”. In: *IEEE Transactions on Signal Processing* 67.1 (Jan. 2019), pp. 97–109. DOI: 10.1109/TSP.2018.2879624.

[59] S. Mihai, M. Yaqoob, D. V. Hung, *et al.* “Digital Twins: A Survey on Enabling Technologies, Challenges, Trends and Future Prospects”. In: *IEEE Communications Surveys Tutorials* 24.4 (Sept. 2022), pp. 2255–2291.

[60] J. Melton and S. Krishnan. “muxGNN: Multiplex Graph Neural Network for Heterogeneous Graphs”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.9 (Sept. 2023), pp. 11067–11078.

[61] Z. Zhao, H. Zhou, L. Qi, L. Chang, and M. Zhou. “Inductive Representation Learning via CNN for Partially-Unseen Attributed Networks”. In: *IEEE Transactions on Network Science and Engineering* 8.1 (Jan. 2021), pp. 695–706.

[62] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio. “Graph Attention Networks”. In: *International Conference on Learning Representations* (2018). URL: https://openreview.net/forum?id=rJXMpikCZ.

[63] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. “Attention is All you Need”. In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Vol. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[64] A. Presekal, A. Ştefanov, V. S. Rajkumar, and P. Palensky. “Attack Graph Model for Cyber-Physical Power Systems Using Hybrid Deep Learning”. In: *IEEE Transactions on Smart Grid* 14.5 (Sept. 2023), pp. 4007–4020.

[65] M. Perez-Molina, D. Larruskain, P. E. Lopez, G. Buigues, and V. Valverde. “Review of Protection Systems for Multi-terminal High Voltage Direct Current Grids”. In: *Renewable and Sustainable Energy Reviews* 144 (2021), p. 111037. DOI: 10.1016/j.rser.2020.111037.

[66] N. Ortiz, A. A. Cardenas, and A. Wool. “Scada World: An Exploration of the Diversity in Power Grid Networks”. In: *Proceedings of the ACM on Measurement and Analysis of Computing Systems*. Vol. 8. 1. 2024, pp. 1–32.

[67] A. J. Wood, B. F. Wollenberg, and G. B. Sheble. “The ’DC’ or Linear Power Flow”. In: *Power Generation, Operation, and Control*. 3rd. Hoboken, NJ, USA: Wiley, 2014. Chap. 6, p. 277.

[68] G. Zhang, J. Zhang, J. Yang, C. Wang, Y. Zhang, and M. Duan. “Vulnerability Assessment of Bulk Power Grid Based on Weighted Directional Graph and Complex Network Theory”. In: *Electric Power Automation Equipment* 29.4 (Apr. 2009), pp. 21–29.

[69] Q. Y. Xie, C. Deng, H. Zhao, and Y. X. Weng. “Evaluation Method for Node Importance of Power Grid Based on the Weighted Network Model”. In: *Automation of Electric Power Systems* 33.4 (Feb. 2009).

[70] X. Jingyou, C. Chong, L. Chunjian, C. Xi, X. Wei, and L. Xiangning. “Identification of Power Grid Key Parts Based on Improved Complex Network Model”. In: *Automation of Electric Power Systems* 40.10 (May 2016). DOI: 10.7500/AEPS20150824006.

[71] T. Hocevar and J. Demšar. “A Combinatorial Approach to Graphlet Counting”. In: *Bioinformatics* 30.4 (2014), pp. 559–565. DOI: 10.1093/bioinformatics/btt651.

[72] R. S., A. S. Bharadwaj, D. S. K., M. S. Khadabadi, and A. Jayaprakash. “Digital Implementation of the Softmax Activation Function and the Inverse Softmax Function”. In: *2022 4th International Conference on Circuits, Control, Communication and Computing (I4C)*. Bangalore, India, 2022, pp. 64–67. DOI: 10.1109/I4C57141.2022.10057747.

[73] Mininet. *Mininet, An Instant Virtual Network on your Laptop (or other PC)*. n.d.

[74] K. Proska, J. Wolfram, J. Wilson, D. Black, K. Lunden, D. K. Zafra, N. Brubaker, T. Mclellan, and C. Sistrunk. *Sandworm Disrupts Power in Ukraine Using a Novel Attack Against Operational Technology*. Online. Nov. 2023.

[75] V. S. Rajkumar, A. Ştefanov, A. Presekal, P. Palensky, and J. L. R. Torres. “Cyber Attacks on Power Grids: Causes and Propagation of Cascading Failures”. In: *IEEE Access* 11 (Sept. 2023), pp. 103154–103176.

[76] A. S. Musleh, G. Chen, and Z. Y. Dong. “A Survey on the Detection Algorithms for False Data Injection Attacks in Smart Grids”. In: *IEEE Transactions on Smart Grid* 11.3 (May 2020), pp. 2218–2234.

[77] J. Xu, R. Yao, and F. Qiu. “Mitigating Cascading Outages in Severe Weather Using Simulation-Based Optimization”. In: *IEEE Transactions on Power Systems* 36.1 (Jan. 2021), pp. 204–213.

[78] Y. K. Wu, Y. C. Chen, H. L. Chang, and J. S. Hong. “The Effect of Decision Analysis on Power System Resilience and Economic Value During a Severe Weather Event”. In: *IEEE Transactions on Industrial Applications* 58.2 (Mar. 2022), pp. 1685–1695.

[79] Y. Liu, S. Gao, J. Shi, X. Wei, and Z. Han. “Sequential-Mining-Based Vulnerable Branches Identification for the Transmission Network Under Continuous Load Redistribution Attacks”. In: *IEEE Transactions on Smart Grid* 11.6 (Nov. 2020), pp. 5151–5160.

[80] Y. Liu, S. Gao, J. Shi, X. Wei, Z. Han, and T. Huang. “Pre-Overload-Graph-Based Vulnerable Correlation Identification Under Load Redistribution Attacks”. In: *IEEE Transactions on Smart Grid* 11.6 (Nov. 2020), pp. 5216–5226.

[81] A. Abedi, L. Gaudard, and F. Romerio. “Review of Major Approaches to Analyze Vulnerability in Power System”. In: *Reliability Engineering System Safety* 183 (2019), pp. 153–172.

[82] C. Ten, C. Liu, and G. Manimaran. “Vulnerability Assessment of Cybersecurity for SCADA Systems”. In: *IEEE Transactions on Power Systems* 23.4 (Nov. 2008), pp. 1836–1846.

[83] A. Dwivedi and X. Yu. “A Maximum-Flow-Based Complex Network Approach for Power System Vulnerability Analysis”. In: *IEEE Transactions on Industrial Informatics* 9.1 (Feb. 2013), pp. 81–88.

[84] J. Johansson, H. Hassel, and E. Zio. “Reliability and Vulnerability Analyses of Critical Infrastructures: Comparing Two Approaches in the Context of Power Systems”. In: *Reliability Engineering System Safety* 120 (2013), pp. 27–38.

[85] J. X. Gao, B. Barzel, and A. L. Barabási. “Universal Resilience Patterns in Complex Networks”. In: *Nature* 530 (2016), pp. 307–312.

[86] S. Hosseini, K. Barker, and J. E. Ramirez-Marquez. “A Review of Definitions and Measures of System Resilience”. In: *Reliability Engineering System Safety* 145 (2016), pp. 47–61.

[87] C. Vellaithurai, A. Srivastava, S. Zonouz, and R. Berthier. “CPIndex: Cyber-Physical Vulnerability Assessment for Power-Grid Infrastructures”. In: *IEEE Transactions on Smart Grid* 6.2 (Mar. 2015), pp. 566–575.

[88] *Common Vulnerabilities and Exposures*. Accessed: 2024-04-27, https://cve.mitre.org/.

[89] T. Zhao, D. Wang, D. Lu, Y. Zeng, and Y. Liu. “A Risk Assessment Method for Cascading Failure Caused by Electric Cyber-Physical System (ECPS)”. In: *Proceedings of the 5th International Conference on Electric Utility Deregulation and Restructured Power Technology*. Changsha, China, 2015, pp. 787–791.

[90] K. Marashi, S. S. Sarvestani, and A. R. Hurson. “Consideration of Cyber-Physical Interdependencies in Reliability Modeling of Smart Grids”. In: *IEEE Transactions on Sustainable Computing* 3.2 (Apr. 2017), pp. 73–83.

[91] G. Celli, E. Ghiani, F. Pilo, and G. G. Soma. “Impact of ICT on the Reliability of Active Distribution Networks”. In: *Proceedings of the CIRED Workshop on Integrating Renewables into Distribution Grid*. Lisbon, Portugal, 2012, pp. 1–4.

[92] D. Schacht, D. Lehmann, H. Vennegeerts, S. Krahl, and A. Moser. “Modelling of Interactions between Power System and Communication Systems for the Evaluation of Reliability”. In: *Proceedings of the Power Systems Computation Conference*. Genoa, Italy, 2016, pp. 1–7.

[93] C. Wang, T. Zhang, F. Luo, F. Li, and Y. Liu. “Impacts of Cyber System on Microgrid Operational Reliability”. In: *IEEE Transactions on Smart Grid* 10.1 (Jan. 2019), pp. 105–115. DOI: 10.1109/TSG.2017.2732484.

[94] Y. Wang, C. Liu, M. Shahidehpour, and C. Guo. “Critical Components for Maintenance Outage Scheduling Considering Weather Conditions and Common Mode Outages in Reconfigurable Distribution Systems”. In: *IEEE Transactions on Smart Grid* 7.6 (Nov. 2016), pp. 2807–2816.

[95] J. V. Milanovic and W. Zhu. “Modeling of Interconnected Critical Infrastructure Systems Using Complex Network Theory”. In: *IEEE Transactions on Smart Grid* 9.5 (Sept. 2018), pp. 4637–4648.

[96] S. Fattaheian-Dehkordi, M. Fotuhi-Firuzabad, and R. Ghorani. “Transmission System Critical Component Identification Considering Full Substations Configuration and Protection Systems”. In: *IEEE Transactions on Power Systems* 33.5 (Sept. 2018), pp. 5365–5373.

[97] Y. Zhao, S. Liu, Z. Lin, L. Yang, Q. Gao, and Y. Chen. “Identification of Critical Lines for Enhancing Disaster Resilience of Power Systems with Renewables based on Complex Network Theory”. In: *IET Generation, Transmission & Distribution* 14.20 (2020), pp. 4459–4467.

[98] L. Li, H. Wu, Y. Song, and Y. Liu. “A State-Failure-Network Method to Identify Critical Components in Power Systems”. In: *Electric Power Systems Research* 181 (Jan. 2020), pp. 1–10.

[99] Q. Gao, Y. Wang, X. Cheng, J. Yu, X. Chen, and T. Jing. “Identification of Vulnerable Lines in Smart Grid Systems based on Affinity Propagation Clustering”. In: *IEEE Internet of Things Journal* 6.3 (June 2019), pp. 5163–5171.

[100] Y. Jia, Z. Xu, L. Lai, and et al. “Risk-based Power System Security Analysis considering Cascading Outages”. In: *IEEE Transactions on Industrial Informatics* 12.2 (Mar. 2016), pp. 872–882.

[101] Y. Liu, N. Zhang, D. Wu, and et al. “Searching for Critical Power System Cascading Failures with Graph Convolutional Network”. In: *IEEE Transactions on Control and Network Systems* 8.3 (Sept. 2021), pp. 1304–1313.

[102] X. Wu, D. Wu, and E. Modiano. “Predicting Failure Cascades in Large Scale Power Systems via the Influence Model Framework”. In: *IEEE Transactions on Power Systems* 36.5 (Sept. 2021), pp. 4778–4790.

[103] Z. Zhang, S. Huang, S. Mei, and et al. “Key Branches Identification for Cascading Failure based on Q-learning Algorithm”. In: *Proceedings of the IEEE International Conference on Power Systems Technology*. Wollongong, NSW, Australia, 2016, pp. 1–6.

[104] Z. Zhang, R. Yao, and S. Huang. “An Online Search Method for Representative Risky Fault Chains based on Reinforcement Learning and Knowledge Transfer”. In: *IEEE Transactions on Power Systems* 35.3 (May 2020), pp. 1856–1867.

[105] A. Sturaro, S. Silvestri, M. Conti, and S. K. Das. “A Realistic Model for Failure Propagation in Interdependent Cyber-Physical Systems”. In: *IEEE Transactions on Network Science and Engineering* 7.2 (Apr. 2020), pp. 817–831.

[106] D. Z. Tootaghaj, N. Bartolini, H. Khamfroush, T. He, N. R. Chaudhuri, and T. L. Porta. “Mitigation and Recovery from Cascading Failures in Interdependent Networks Under Uncertainty”. In: *IEEE Transactions on Control of Network Systems* 6.2 (June 2019), pp. 501–514.

[107] Z. Huang, C. Wang, M. Stojmenovic, and A. Nayak. “Characterization of Cascading Failures in Interdependent Cyber-Physical Systems”. In: *IEEE Transactions on Computers* 64.8 (Aug. 2015), pp. 2158–2168.

[108] B. Moussa, P. Akaber, M. Debbabi, and C. Assi. “Critical Links Identification for Selective Outages in Interdependent Power-Communication Networks”. In: *IEEE Transactions on Industrial Informatics* 14.2 (Feb. 2018), pp. 472–483.

[109] Z. Huang, C. Wang, M. Stojmenovic, and A. Nayak. “Balancing System Survivability and Cost of Smart Grid Via Modeling Cascading Failures”. In: *IEEE Transactions on Emerging Topics in Computing* 1.1 (June 2013), pp. 45–56.

[110] J. Shao, S. V. Buldyrev, S. Havlin, and H. E. Stanley. “Cascade of Failures in Coupled Network Systems with Multiple Support-Dependence Relations”. In: *Physical Review E* 83.3 (Mar. 2011), p. 036116. DOI: 10.1103/PhysRevE.83.036116.

[111] D. Zhou, J. Gao, H. E. Stanley, and S. Havlin. “Percolation of Partially Interdependent Scale-Free Networks”. In: *Physical Review E* 87.5 (May 2013), p. 052812. DOI: 10.1103/PhysRevE.87.052812.

[112] B. Falahati, Y. Fu, and L. Wu. “Reliability Assessment of Smart Grid Considering Direct Cyber-Power Interdependencies”. In: *IEEE Transactions on Smart Grid* 3.3 (Sept. 2012), pp. 1515–1524.

[113] D. Zhou, H. E. Stanley, G. D’Agostino, and et al. “Assortativity Decreases the Robustness of Interdependent Networks”. In: *Physical Review E* 86.6 (2012), p. 06610362. DOI: 10.1103/PhysRevE.86.06610362.

[114] G. Dong, L. Tian, R. Du, and et al. “Analysis of Percolation Behaviors of Clustered Networks with Partial Support–Dependence Relation”. In: *Physica A: Statistical Mechanics and its Applications* 394 (2014), pp. 370–378.

[115] R. Parshani, S. V. Buldyrev, and S. Havlin. “Interdependent Networks: Reducing the Coupling Strength Leads to a Change from a First to Second Order Percolation Transition”. In: *Phys. Rev. Lett.* 105 (4 July 2010), p. 048701. DOI: 10.1103/PhysRevLett.105.048701.

[116] R. Parshani, C. Rozenblat, D. Ietri, C. Ducruet, and S. Havlin. “Intersimilarity between coupled networks”. In: *Europhysics Letters* 92.6 (2010). DOI: 10.1209/0295-5075/92/68002.

[117] V. Rosato, L. Issacharoff, F. Tiricco, and et al. “Modelling Interdependent Infrastructures Using Interacting Dynamical Models”. In: *International Journal of Critical Infrastructures* 4.1/2 (2008), pp. 63–79.

[118] D. Watts and S. Strogatz. “Collective dynamics of ‘small-world’ networks”. In: *Nature* 393 (June 1998), pp. 440–442. DOI: 10.1038/30918.

[119] S. V. Buldyrev, R. Parshani, G. Paul, H. E. Stanley, and S. Havlin. “Catastrophic Cascade of Failures in Interdependent Networks”. In: *Nature* 464 (Apr. 2010), pp. 1025–1028. DOI: 10.1038/4641025a.

[120] D. Zhengcheng, F. Yanjun, and T. Meng. “Influences of Various Coupled Patterns and Coupling Strength on Power-communication Coupled Networks”. In: *High Voltage Engineering* 41.10 (Oct. 2015), pp. 3464–3469. DOI: 10.13336/j.1003-6520.hve.2015.10.038.

[121] F. Yang, A. P. S. Meliopoulos, G. J. Cokkinides, and Q. B. Dam. “Effects of Protection System Hidden Failures on Bulk Power System Reliability”. In: *2006 38th North American Power Symposium*. Carbondale, IL, USA, 2006, pp. 517–523. DOI: 10.1109/NAPS.2006.359621.

[122] P. Fournier-Viger, C. Cheng, J. C.-W. Lin, U. Yun, and U. Iran. “TKG: Efficient Mining of Top-K Frequent Subgraphs”. In: *Proceedings of the 7th International Conference on Big Data Analytics (BDA 2019)*. Cham: Springer, 2019, pp. 209–226. DOI: 10.1007/978-3-030-12345-6_XX.

[123] V. Tseng, C. Wu, P. Fournier-Viger, and P. S. Yu. “Efficient Algorithms for Mining Top-K High Utility Itemsets”. In: *IEEE Transactions on Knowledge and Data Engineering* 28.1 (Jan. 2016), pp. 54–67.

[124] M. Kuramochi and G. Karypis. “Frequent Subgraph Discovery”. In: *Proceedings of the 2001 IEEE International Conference on Data Mining*. USA, 2001, pp. 313–320. DOI: 10.1109/ICDM.2001.1008417.

[125] B. Wang, H. Tang, C. Guo, and Z. Xiu. “Entropy Optimization of Scale-Free Networks’ Robustness to Random Failures”. In: *Physica A: Statistical Mechanics and its Applications* 363.2 (2006), pp. 591–596. DOI: 10.1016/j.physa.2006.03.031.

[126] F. Cadini, E. Zio, and C. Petrescu. “Using Centrality Measures to Rank the Importance of the Components of a Complex Network Infrastructure”. In: *Proceedings of the Critical Information Infrastructure Security: Third International Workshop (CRITIS 2008)*. Italy, 2009, pp. 155–167.

[127] Y. Liu, A. Ştefanov, I. Semertzis, and P. Palensky. “GraphCCI: Critical Components Identification for Enhancing Security of Cyber-Physical Power Systems”. In: *IEEE Transactions on Industrial Cyber-Physical Systems* 2 (2018), pp. 340–349.

Acknowledgements

Completing this PhD thesis has been a challenging and rewarding journey, and I am deeply grateful to all those who have supported me along the way. This accomplishment would not have been possible without the collective efforts, guidance, and encouragement of many individuals.

First and foremost, I would like to express my sincere gratitude to my supervisors, Prof. Peter Palensky and Dr. Alexandru Ştefanov, for their unwavering support, insightful guidance, and invaluable feedback throughout my research. Their expertise and encouragement have been instrumental in shaping this work. Prof. Palensky's profound knowledge and his ability to inspire innovative thinking have greatly influenced my approach to research. Dr. Alexandru Ştefanov's meticulous attention to detail and his dedication to fostering a collaborative research environment have provided me with the tools necessary to overcome numerous challenges. Their mentorship has not only enhanced my academic growth but also my personal development, instilling in me a passion for lifelong learning and inquiry.

I am also thankful to the members of my dissertation committee, Prof. Chen-Ching Liu, Prof. Pieter van Gelder, Dr. Subhash Lakshminarayana, Dr. Zekeriya Erkin, Dr. Junho Hong and Prof. Arno H.M. Smets, for their constructive criticism and thoughtful suggestions, which have greatly enhanced the quality of this thesis. Their diverse perspectives and expertise have broadened my understanding of CPS, and their feedback has been invaluable in refining my research methodology and analysis. I appreciate the time and effort each committee member has dedicated to reviewing my work and providing guidance that has been crucial in shaping the final outcome of this thesis.

A special thank you goes to my master supervisor Prof. Shibin Gao and Dr. Xiaoguang Wei. Your support and stimulating discussions have made this journey enjoyable and intellectually enriching. Our shared experiences, both academic and personal, have created lasting memories and a network of support that I will always cherish.

On a personal note, I am profoundly grateful to my family and friends for their unconditional love and support. Especially for my parents and sister, thank you for believing in me and encouraging me to pursue my dreams. Your patience, understanding, and constant encouragement have been my source of strength. Your ability to provide emotional support has allowed me to focus on my research without undue worry.

Lastly, I would like to thank everyone who has contributed to my growth and success in ways big and small. Your support has been invaluable, and I am truly appreciative. Whether through academic guidance, personal encouragement, or practical assistance, each of you has played a role in this achievement.

CURRICULUM VITÆ

Yigu Liu was born on September 3rd, 1995, in Hengyang, Hunan Province, China. He received the M.Sc degree and the B.Sc degree in electrical engineering from Southwest Jiaotong University, Chengdu, China, in 2020. He is currently working toward the Ph.D. degree in electrical engineering from the Department of Electrical Sustainable Energy, Delft University of Technology, Delft, Netherlands. His research interests include synthetic network generation and vulnerability assessment of CPS.