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

Liu, Yigu; Ștefanov, Alexandru; Palensky, Peter

DOI
10.1109/TII.2024.3438232

Publication date
2024

Document Version
Final published version

Published in
IEEE Transactions on Industrial Informatics

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

Yigu Liu, Alexandru Ştefanov, *Member, IEEE*, and Peter Palensky, *Senior Member, IEEE*

***Abstract*—Synthetic networks aim to generate 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 cyber–physical power systems (CPPS). In this article, we propose a two-stage generative model for generating synthetic communication topologies of large-scale CPPS 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 CPPS 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.**

***Index Terms*—Cyber–physical power system (CPPS), smart grid, synthetic network.**

I. Introduction

With the increasing digitalization of power grids, the cyber–physical power systems (CPPS) are extensively studied [1], [2], [3]. 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 CPPS are missing in the current literature. Under such a background, synthetic networks emerge as a promising method to generate fictitious but realistic projections of power grids and communication networks. A synthetic CPPS avoids revealing sensitive network models and data while providing reliable test networks for research.

Manuscript received 23 January 2024; revised 23 May 2024; accepted 15 July 2024. Date of publication 13 August 2024; date of current version 5 November 2024. This work was supported in part by EU Horizon Europe eFORT Project under Grant 101075665 and in part by China Scholarship Council. Paper no. TII-24-0368. *(Corresponding author: Yigu Liu.)*

The authors are with the Department of Electrical Sustainable Energy, Delft University of Technology, 2628 CD Delft, The Netherlands (e-mail: y.liu-18@tudelft.nl; a.i.stefanov@tudelft.nl; p.palensky@tudelft.nl).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TII.2024.3438232.

Digital Object Identifier 10.1109/TII.2024.3438232

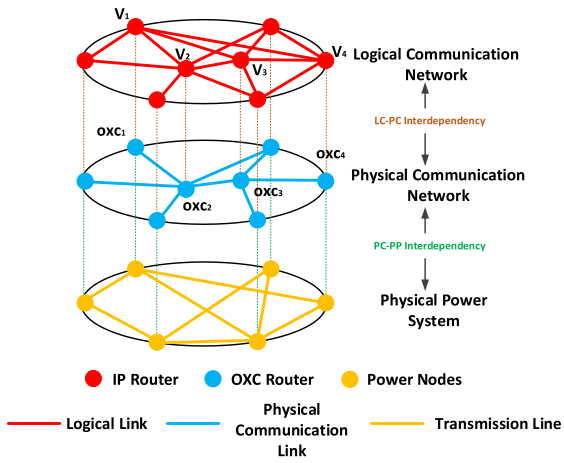

Fig. 1. Three interdependent networks in CPPS.

In recent years, researchers explored methods to mainly generate synthetic power systems. Zhou and Bialek [4] developed a synthetic dc power flow model for the continental power grids in Europe based on the available public data. The authors in [5], [6], [7], [8], and [9] conducted research on large-scale synthetic power systems. Espejo et al. [5] generate and validate the synthetic power systems topology from the perspective of complex network theory. In [6], a learning-based method is proposed to generate synthetic power grids, which are evaluated by considering power flows and vulnerability against failures. In [7], [8], and [9], 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 (PPS). How to generate a large-scale synthetic cyber–physical system is rarely investigated because of two reasons: First, lack of real CPPS data for model validation, i.e., system parameters and structural topologies; and second, increased computational complexity in generating large-scale CPPS models. With the fast power grid digitalization, the power system is now tightly coupled with the cyberinfrastructure 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 CPPS based on the results of synthetic power systems.

In CPPS-related literature, most test cases are restricted to the standard IEEE test systems [10], [11], [12]. Davis et al. [10] propose a framework to model the cyber–physical system dependencies and assess the vulnerabilities of a CPPS with eight remote terminal units. Cai et al. [11] use IEEE 39-bus and China's Guangdong 500-kV system to model the CPPS and analyze cascading failures considering the interactions between cyber and physical layers. Zang et al. [12] analyze the fault propagation mechanism of cyber–physical systems for IEEE 118-bus and 300-bus systems. 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 CPPS 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 [4], [5], [6], [7], [8], [9], and [10], the general process of generating synthetic power systems is given as follows:

1) collect public data, e.g., resident and geographic information;
2) generate synthetic power grids based on the available public data;
3) 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 CPPS. 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, we replicate the existing communication network design process, generating statistically realistic communication topologies. Otherwise, it might lead to significant deviations from realism.

Based on the discussion above, in this article, we propose a two-stage generative model to generate realistic, large-scale synthetic cyber–physical systems based on the existing power grids. For a given power grid, we reproduce the typical communication system design process to generate synthetic communication topologies consisting of physical and logical communication networks (PCN and LCN) for large-scale CPPS 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 CPPS 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 we do not consider the historical evolution of the CPPS. To the best knowledge of the authors, this research is pioneering the generation of large-scale, synthetic CPPS. The main contributions of this article are summarized as follows.

1) We propose a two-stage model for generating realistic, large-scale synthetic communication topologies for CPPS based on the existing power grids. We identify the CPPS as a triple interdependent network consisting of PCN, LCN, and PPS.
2) The first stage is the PCN generator; we sequentially generate the initial topology of the PCN and then add more redundancy 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; we utilize the decentralized communication structure and define a communication hub (CH) index considering both communication traffic volume and node criticality to identify the optimal CHs for the LCN.

The rest of this article is organized as follows. Section II introduces the framework for generating large-scale synthetic CPPS. Sections III and IV present the two stages for generating the PCN and LCN topology, respectively. Section V presents the case study and validation results. Finally, Section VI concludes this article.

II. Framework: Generating a Triple Interdependent Cyber–Physical System

Typically, researchers consider that the cyber–physical system comprises 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 [13], [14], consisting of physical and logical communications. These cyber system interdependencies are essential for the overall operation of CPPS. In this research, the cyber–physical system is considered a triple interdependent network, as represented in Fig. 1. It consists of the PCN, LCN, and PPS. The complex interdependencies among the three layers are defined in the framework for generating a large-scale synthetic CPPS.

A. Three Interdependent Networks in CPPS

PCN: At the PCN layer, each node is an optical cross-connect (OXC) router or synchronous digital hierarchy 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, wireless communication, and satellite communication [15]. 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.

LCN: The logical communication layer represents the interactions between PCN nodes. The LCN topology is predetermined to satisfy the system operation requirements. Each node in the LCN is an IP router, which corresponds to a node in the PCN, e.g., an OXC router. The nodes in the LCN are connected by logical links. It is worth mentioning that the LCN is a virtual

network configured by CPPS designers. In the communication process, a logical link may pass through multiple nodes in the PCN for successful information delivery. For example, nodes V_1 and V_4 in the LCN are adjacent, as represented in Fig. 1. However, in the PCN, the traffic between OXC_1 and OXC_4 will pass through nodes OXC_2 and OXC_3. Note that, in this article, we consider wired networks, in particular wavelength-division multiplexing optical networks, when generating the LCN, because large-scale wireless communication networks are usually not used in typical CPPS. Also, in this article, we only consider static routing. The dynamic routing is beyond the scope of this article.

PPS: In this article, we focus on the 380–400 kV high-voltage transmission network. Therefore, each node in the PPS is a substation, while the transmission lines and transformers between substations represent the edges.

B. Complex Interdependencies Among CPPS Layers

As shown in Fig. 1, there are two types of interdependencies in CPPS, i.e., LCN and PCN interdependency (LC-PC), and PCN and PPS interdependency (PC-PP).

LC-PC interdependency: The interdependency between LCN and PCN 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 PCN impact the operational cost of data transmission, which is decided by the topology of the LCN. Meanwhile, the topology of the LCN also has a significant impact on the operational performance of the PCN. 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 $\boldsymbol{V}$, $\boldsymbol{E}_L$, and $\boldsymbol{E}_P$ are the set of nodes and edges in two networks. The LC-PC interdependency is presented as follows:

$$L_u = \begin{cases} 1, & \text{if } u \in \boldsymbol{E}_L \\ 0, & \text{otherwise} \end{cases} \tag{1.1}$$

$$L_{ur} = \begin{cases} 1, & \text{if } u \text{ passes through } r \\ 0, & \text{otherwise} \end{cases} \tag{1.2}$$

$$L_{ur} \le L_u \quad \forall u \in \boldsymbol{E}_L,\ r \in \boldsymbol{E}_P \tag{1.3}$$

$$\sum_{r\in\Theta_o(n)} L_{ur} - \sum_{r\in\Theta_i(n)} L_{ur} = \begin{cases} 1, & \text{if } O(u) = n \\ -1, & \text{if } D(u) = n \\ 0, & \text{otherwise} \end{cases} \tag{1.4}$$

$$\sum_{u\in\boldsymbol{E}_L} L_{ur} \le W_r \quad \forall r \in \boldsymbol{E}_P \tag{1.5}$$

where u is a logical link in the LCN, and r is a physical link in the PCN. L_u is the logical link variable and L_{ur} is the logical link routing variable. Equation (1.3) indicates the mapping relationship between logical and physical links. Equation (1.4) 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 $n \in \boldsymbol{V}$, and $O(u)$ 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, (1.5) 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.

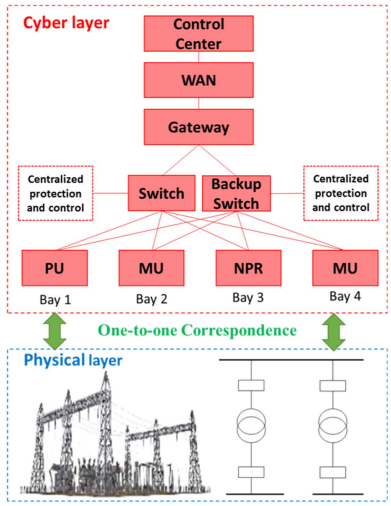

Fig. 2. Substation communication hybrid architecture (NPR: Numerical protection relay, MU: Merging unit, and PU: Process unit).

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 [16]. However, in a real-world scenario, the PC-PP interdependency is more complex. Fig. 2 shows the state-of-the-art substation communication architecture deployed in industry [17]. 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 within the substation. They communicate with the CCs through wide area networks (WANs) via the routing gateways in the substations. In this article, each substation is associated with a physical communication node, i.e., gateway. Considering the relay communication nodes [16] 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 predetermined routing path defined in the LCN and is delivered to the CC. The CC will make the optimal decision based on the collected data and send the commands back to the PPS using the same method. In Fig. 5 of Section IV, we present more detailed illustrations to explain this concept.

C. Two-Stage Generative Model for Large-Scale Synthetic CPPS

Generally, the design of a network topology includes the following steps:

1) initial topology design;
2) increase redundancy to enhance the network resilience;

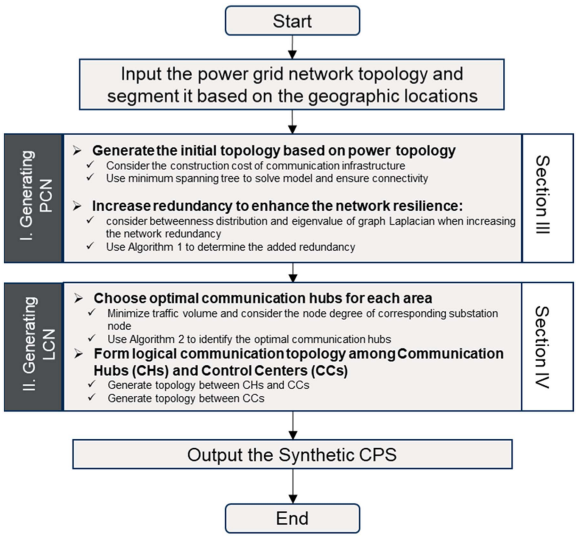

Fig. 3. Diagram of the two-stage generative model.

3) routing configuration [13].

In this article, we follow these sequential steps. The two-stage generative model proposed in this article is presented in Fig. 3. Note that we divide the large-scale network into multiple small-scale subnetworks and denote a subnetwork as a communication area [15]. As indicated in [20], 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. Hu et al. [20] pointed out that the N_C normally scales from 4 to 12. Therefore, by following the segmentation method in [20], stages I and II in Fig. 3 are implemented in each communication area. Besides, the following assumptions are used.

1) The topology of the power system is known. Therefore, the goal of this research is to generate the LCN, PCN, and the complex interdependencies in CPPS.
2) The historical evolution of CPPS is not considered.
3) Only wired communication networks are considered.

III. PCN Generator: Generating the PCN

A. Generating the Initial Topology of PCN

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

1) *Construction cost:* Shahraeini et al. [15] consider two types of costs for generating a communication network, i.e., passive and active costs. The passive costs are attributed 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 Fig. 2, in this article, the active costs are determined by the number of substations. However, the number of substations is fixed because we generate the synthetic CPPS based on the existing power grids. This indicates that we only need to consider the passive cost in our article, i.e., the total length of communication links.
2) *Connectivity:* Generally, optimization models can be used to obtain the initial PCN 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 PCN 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 [18]. 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 [19] is used to compute the initial topology of the PCN. 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 article, the edge weight is set as the length of the distance between any two substations.

B. Redundancy Enhancement Metric (REM): Increasing Network Redundancy to Ensure Resilience

The initial topology of the PCN 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 the Laplacian matrix and betweenness distribution are considered to increase PCN network redundancy.

Eigenvalue of Laplacian matrix: According to Li [13], 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 $\boldsymbol{V_C} = \{V_C | c = 1, 2, 3, \ldots\}$ is the set of PCN nodes and $\boldsymbol{E_F} = \{E_f | f = 1, 2, 3, \ldots\}$ is the set of PCN edges. We denote the Laplacian matrix of $\boldsymbol{G_F}$ as $\boldsymbol{M} = [M_{CC'}]_{n\times n}$, and for the element $M_{CC'}$ in $\boldsymbol{M}$

$$M_{CC'} = \begin{cases} \sum_{C''=1}^{n} A(V_C, V_{C''}), & \text{if C} = \text{C}' \\ -1 & \text{if C} \neq \text{C}' \end{cases} \tag{2}$$

$$\lambda_2 = \min\{\lambda(\boldsymbol{M}) - \min\{\lambda(\boldsymbol{M})\}\} \tag{3}$$

where $\lambda(\boldsymbol{M})$ 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 communication system performance, such as network stability and connectivity. A larger λ_2 indicates a better system performance. Therefore, the objective is to maximize λ_2 when adding communication edges to increase the redundancy.

Betweenness distribution: Hu et al. [20] indicate that the network betweenness has substantial effects on the network congestion. Normally, the data packets in the cyber layer are transmitted through the shortest path between any arbitrary two nodes. This makes the node betweenness an effective index to quantify the data volume that each node processes. Therefore, the betweenness distribution in the PCN is adopted to evaluate

the network congestion. The more uneven the betweenness distribution is, 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, which makes network congestion easier to happen. In this article, we employ the Gini coefficient to quantify the betweenness distribution. The calculation of node betweenness $B(V_C)$ and Gini coefficient G_{ini} is shown in (4) and (5)

$$B\left(V_C\right) = \sum_{V_C, V_{C\prime}, V_{C\prime\prime} \in \mathbf{V_C}, C \neq C\prime \neq C\prime\prime} \frac{N_{C\prime C\prime\prime}\left(V_C\right)}{N_{C\prime C\prime\prime}} \tag{4}$$

$$G_{\text{ini}} = \frac{1}{2n^2 u} \sum_{C=1}^{n} \sum_{C\prime=1}^{n} \left|B\left(V_C\right) - B\left(V_{C\prime}\right)\right| \tag{5}$$

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

A tradeoff 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 the network is constrained by the budget of the network design. Therefore, we assume that the number of added edges is subjected to construction cost and should satisfy the following condition:

$$N_{\text{add}} = \chi N_k,\ 0 < \chi < 1 \tag{6}$$

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 predetermined 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 REM and denote it as E_r to determine how to add N_{add} redundant edges to increase the network performance

$$E_r = \alpha\lambda_2 - \beta G_{\text{ini}} \tag{7}$$

$$\alpha + \beta = 1,\ 0 < \alpha < 1, 0 < \beta < 1 \tag{8}$$

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 Fig. 4. Taking node 2 as an 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 (6)–(8), one can calculate the E_r of all possible combinations. 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 in a communication area, where the number of nodes is limited [21]. Thus, the computational cost is acceptable

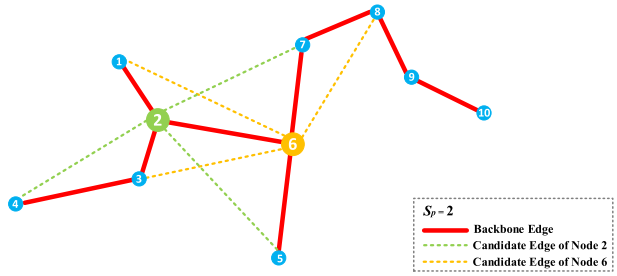

Fig. 4. Generating candidate edges for increasing the redundancy for a communication area.

Algorithm 1: Adding Redundancy for Communication Area.

Input:
Initial topology of the PCN
Parameters: S_p, χ, α, β
Output:
Optimal candidate edge set: $\boldsymbol{C}_{s\_\text{optimal}}$

Step 1 $\boldsymbol{C}_{s\_\text{optimal}} \leftarrow \emptyset$, $\boldsymbol{C_s} \leftarrow \emptyset$
Step 2 **For** $V_c \in \boldsymbol{V}_{CA}(k)$ **do**
Step 3 $\boldsymbol{C_s} \leftarrow$ all $V_C{}'$ satisfy $S_p(V_C, V_C{}') = S_p$
Step 4 **End For**
Step 5 Employ (6) to calculate N_{add}
Step 6 $\boldsymbol{C}_{s\_\text{combination}} \leftarrow$ all combinations ($\boldsymbol{C_s}$,N_{add})
Step 7 **For** combination in $\boldsymbol{C}_{s\_\text{combination}}$ **do**
Step 8 Employ (7) to calculate E_r
Step 9 **End For**
Step 10 $\boldsymbol{C}_{s_\text{optimal}} \leftarrow$combination with highest E_r

and will not be exponentially increased even if we generate large-scale networks. More details about the computational efficiency are discussed in Section V. The algorithm for adding new network redundancy is presented in Algorithm 1, where $S_p(V_C, V_{C'})$ is the shortest path length between V_C and $V_{C'}$, $\boldsymbol{C}_{s\_\text{combination}}$ is the set of all combinations of edges in $\boldsymbol{C_s}$, and $\boldsymbol{V_{CA}}(k) = \{\ldots, V_c, \ldots\}$ is the set of nodes in communication area k.

IV. LCN Generator: Generating the LCN

A. 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 Fig. 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 CC. 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 CH to gather all measurement data from substations. The

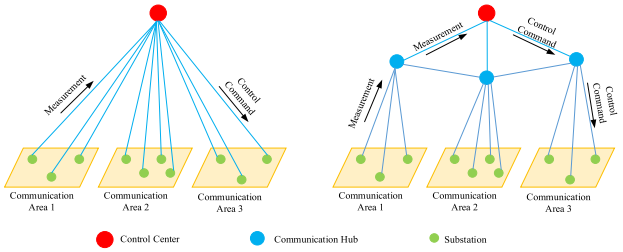

Fig. 5. Centralized communication structure (left) and decentralized communication structure (right) [22].

CHs communicate with the CC. The control commands follow the same routing from the CC 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 [22]. Furthermore, the decentralized communication architecture presents higher reliability with the same construction cost [15]. Therefore, in this article, we adopt the decentralized communication architecture, as illustrated in Fig. 5.

To choose the optimal location for CHs, 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 LCN directly influences the traffic volume, which makes the location of CHs crucial to the performance of the communication network. On the other hand, given that the CHs directly communicate with the CC, it is desirable to connect the CHs 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 CC because the power generation data are crucial to power system operation. Based on the considerations above, we propose the CH index $\boldsymbol{I}_C$ to identify the optimal CHs

$$\boldsymbol{I}_C = \frac{\sum_{C'=1}^{m} A(V_P, V_{P'})}{\sum_{C'=1}^{|V_c|-1} h_{C'} \times p_{CC'}} \tag{9}$$

where m is the number of substations in communication area k, $h_{C'}$ is the number of hops required for the determined communication, and $p_{CC'}$ 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 a higher degree indicates that, once the node is removed, it will pose a serious impact on more nodes in CPPS. 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 [22]. Note that the communication hops between two nodes are decided by the topology of the PCN. By calculating $\boldsymbol{I}_C$ for each substation V_C, the corresponding substation with the largest $\boldsymbol{I}_C$ is identified as the optimal CH. Note that in this article, the CH identification only considers the communication traffic under static routing, which assumes that 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 article, which can be considered as a future study.

B. LCN Topology Between CHs and CCs

After the optimal CHs 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 CH, as shown in Fig. 5. Therefore, the remainder of the LCN topology consists of the topology between CHs and CCs, and the topology between CCs.

The topology between CHs and CCs: In a real-world scenario, backup CCs [20] are extensively deployed to increase system reliability. Therefore, each CH needs to send data packets to both CCs. Note that the difference between them is that, in most of the operational states, main CCs have a high priority to take the actions while the backup CCs work as a redundancy to enhance the resilience of systems in the case of emergency. The communication between CHs and CCs is mostly done through WANs. The WAN topology is beyond the scope of this article. Therefore, we assume that each CH has at least one reliable and cost-efficient path to communicate with the CCs. Generally, the communication topologies between the CHs and CCs have two categories, i.e., double-star and mesh topology. Cai et al. [11] conducted a comparison between these two categories on the IEEE 39-bus system and China's Guangdong 500-kV system, and the experimental results prove that the double-star topology has a 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 a part of the communication nodes fails. Combining all the facts and discussion above, the topology between CHs and CCs is modeled as a 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 follows [23]:

$$p\left(A(V_l, V_{l'})\right) \propto \left[A(V_l, V_{l'})\right]^{-r'} \tag{10}$$

where r' is a constant and satisfies $r' > 1$. V_l and $V_{l'}$ are the nodes in LCN. Equation (10) 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 [20]. On the other hand, the double-star topology is highly vulnerable to cyberattacks 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 [11].

The topology between CCs: The communication among CCs is defined by the intercontrol center communications protocol (ICCP), which is specified worldwide by utilities to provide the services for data exchange, monitoring, and control. The ICCP bilateral tables define the data exchange between two CCs.

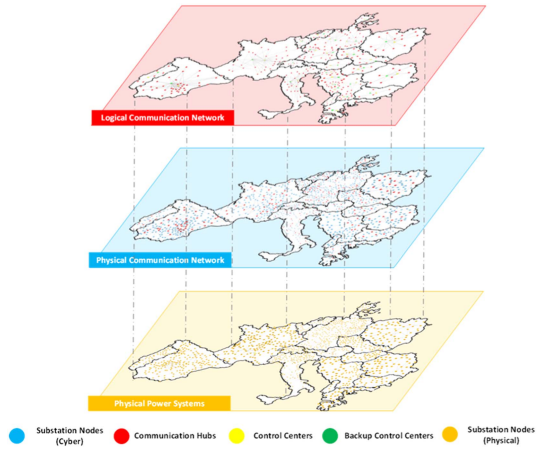

Fig. 6. Generated synthetic CPPS for continental Europe.

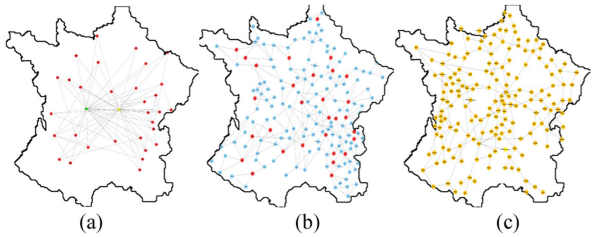

Fig. 7. Generated synthetic CPPS for France. (a) LCN. (b) PCN. (c) PPS.

Normally, all CCs have reliable and efficient communications with their neighboring CCs. Based on the facts above, all CCs are logically reachable by other CCs as long as all power grids are interconnected. Therefore, the LCN topology for CCs is a full connection graph, i.e., each CC is logically connected with other CCs through WANs. It is worth mentioning that the full connection graph in this section represents that, in the LCN, all CCs are logically accessible rather than physically connected.

V. Case Study

In this section, we implement the proposed methods to generate the synthetic CPPS for the interconnected power grids in continental Europe. The parameters for the simulation are $\alpha = 0.5, \beta = 0.5, \chi = 0.3$, and $S_p = 3$. 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.

A. Generated Synthetic CPPS for Continental Europe

The methods proposed in Fig. 3 are used to generate a large-scale, synthetic CPPS based on open-source data from the ENTSO-E website [24]. It provides the 380–400 kV transmission system topologies of the interconnected power grids in continental Europe. The generation results are shown in Figs. 6 and 7 and 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 PCN and LCN is given in Table I. For the clarity of Figs. 6 and 7, although generated in the LCNs, we do not represent the topology between cyber substation nodes and CHs, as well as the direct connection between power plants and CCs. The code and generated models are available online [30].

TABLE I
Statistics of Synthetic Communication Network

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.

At the PCNs layer, we divide the substations into different communication areas. For each area, we randomly allocate N_C substations based on their geographic location and then identify the optimal CH. Based on Hu et al.'s article [20], N_C is set as a random number between 4 and 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 CCs, similar to the CHs in other larger countries.

At the LCNs layer, we decide the number of CCs in each country based on Hu et al.'s article [20]. Typically, a country only has one transmission system operator (TSO), i.e., one main and backup CCs. However, in Germany and Austria, multiple TSOs exist. Therefore, the CHs are divided equally based on the number of TSOs in the country and their geographic location. Besides, all main CCs and backup CCs are logically connected to each other.

B. 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 II, 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,

TABLE II
STATISTICS OF REALISTIC COMMUNICATION NETWORKS IN THE LITERATURE

	Small-Size Syst. [11], [26], [27], [28]	Medium-Size Syst. [20], [26], [29]	Large-Size Syst. [25], [26], [29]	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, and $D = L / N$.

TABLE III
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

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 article.

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 III. One can observe that all complex network theory parameters of the generated communication networks with the proposed method are within the parameter ranges given in Table II. Statistically speaking, in Table II, 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. Comparing with Table III, 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 I, 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 the average shortest path length, the result of the 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.

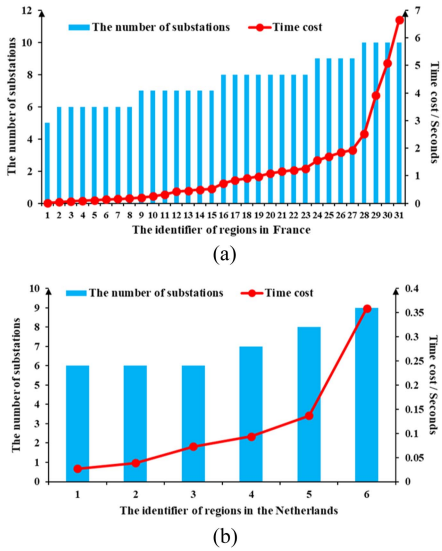

Fig. 8. Generation time of France and The Netherlands. (a) France. (b) The Netherlands.

C. Evaluating Time Efficiency of Proposed Methods

In this part, we evaluate the time efficiency of the proposed methods. Fig. 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 the following equation:

$$T_{\text{total}} = T_{PCN} + T_{LCN} \tag{11}$$

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

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 Hu et al. [20]. 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 Fig. 8, the left axis represents the number of substations in each area, and the subaxis 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 Section V-A, the size of each communication area is limited, which determines that the final time cost of each area will not exceed the maximum time cost, as shown in Fig. 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 article, 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, Fig. 8 proves that the proposed methods

TABLE IV
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

• N/A represents that the corresponding graph is not a connected graph.

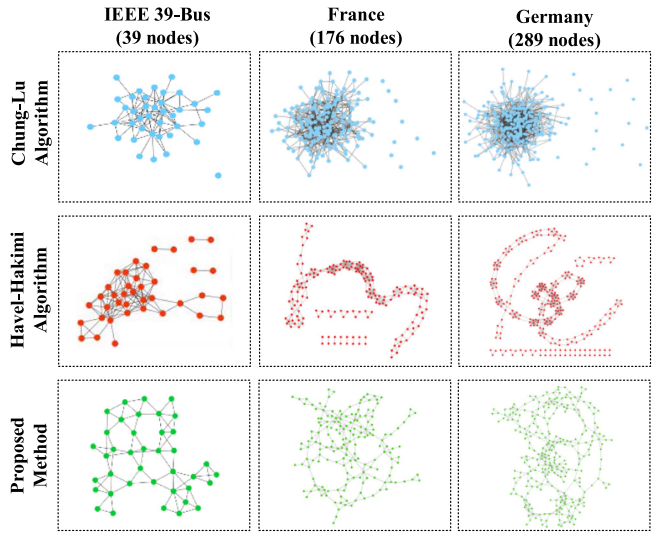

Fig. 9. Comparison with traditional generative algorithms in the literature.

are suitable for generating large-scale, synthetic CPPS in a time-efficient manner.

D. Performance Comparison and Evaluation

In this part, we compare the proposed method with the traditional algorithms in the literature. In Fig. 9, we present the generation results of two traditional generative algorithms with various network sizes, i.e., Chung–Lu and Havel–Hakimi algorithms [31]. Also, the complex network parameters are presented in Table IV. 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 that there is no single isolated node because the Havel–Hakimi algorithm generates the network based on the given degree. For

TABLE V
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

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 parameter distribution, as shown in Table IV. The comparison above proves the good performance of the proposed method.

In the following evaluation, Table V showcases the network parameters of The Netherlands synthetic networks with varying communication area sizes. As noted in [20], 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 V. One can observe that, as the N_C increases, the network average node degree and network density decrease, while the average shortest path length increases. Compared with the data of small-size systems in Table II, only when N_C is in the range of 4–6, all parameters fit to the listed range. Therefore, when N_C is in the range of 4–6, the generated network has the most realistic network parameters.

VI. CONCLUSION

In this article, we focused on generating large-scale synthetic communication topologies for CPPSs. The proposed method circumvented the dilemma of CPPS data availability by reproducing the typical design process of communication networks. It generated synthetic topologies consisting of PCN and LCN for large-scale CPPS. The method was implemented to generate a synthetic CPPS 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 demonstrated its scalability and computational time efficiency. This research pioneered the synthetic CPPS modeling and formed a solid foundation for further investigations to reveal invaluable characteristics, patterns, and mechanisms of CPPSs.

Note that our article focused on generating the communication topologies based on the existing power grids, which is the first and critical step of generating complete synthetic CPPS. In future research, we will add more complexity for synthetic CPPS, e.g., information and cyber–physical interaction models. We will also investigate generating synthetic CPPS for different research purposes based on the results of this article.

REFERENCES

[1] J. Sztipanovits, T. Bapty, X. Koutsoukos, Z. Lattmann, S. Neema, and E. Jackson, "Model and tool integration platforms for cyber–physical system design," *Proc. IEEE*, vol. 106, no. 9, pp. 1501–1526, Sep. 2018.

[2] 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," *IEEE Trans. Ind. Inform.*, vol. 14, no. 9, pp. 4200–4211, Sep. 2018.

[3] Y. Wang, C.-F. Chen, P.-Y. Kong, H. Li, and Q. Wen, "A cyber–physical–social perspective on future smart distribution systems," *Proc. IEEE*, vol. 111, no. 7, pp. 694–724, Jul. 2023.

[4] Q. Zhou and J. W. Bialek, "Approximate model of European interconnected system as a benchmark system to study effects of cross-border trades," *IEEE Trans. Power Syst.*, vol. 20, no. 2, pp. 782–788, May 2005.

[5] R. Espejo, S. Lumbreras, and A. Ramos, "A complex-network approach to the generation of synthetic power transmission networks," *IEEE Syst. J.*, vol. 13, no. 3, pp. 3050–3058, Sep. 2019.

[6] S. Soltan, A. Loh, and G. Zussman, "A learning-based method for generating synthetic power grids," *IEEE Syst. J.*, vol. 13, no. 1, pp. 625–634, Mar. 2019.

[7] K. M. Gegner, A. B. Birchfield, T. Xu, K. S. Shetye, and T. J. Overbye, "A methodology for the creation of geographically realistic synthetic power flow models," in *Proc. IEEE Power Energy Conf. Illinois*, 2016, pp. 1–6.

[8] T. Xu, A. B. Birchfield, K. S. Shetye, and T. J. Overbye, "Creation of synthetic electric grid models for transient stability studies," in *Proc. 10th Bulk Power Syst. Dyn. Control Symp.*, 2017, pp. 1–6.

[9] T. Xu, A. B. Birchfield, and T. J. Overbye, "Modeling, tuning and validating system dynamics in synthetic electric grids," *IEEE Trans. Power Syst.*, vol. 33, no. 6, pp. 6501–6509, Nov. 2018.

[10] K. R. Davis et al., "A cyber-physical modeling and assessment framework for power grid infrastructures," *IEEE Trans. Smart Grid*, vol. 6, no. 5, pp. 2464–2475, Sep. 2015.

[11] Y. Cai, Y. Cao, Y. Li, T. Huang, and B. Zhou, "Cascading failure analysis considering interaction between power grids and communication networks," *IEEE Trans. Smart Grid*, vol. 7, no. 1, pp. 530–538, Jan. 2016.

[12] 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," *Rel. Eng. Syst. Saf.*, vol. 189, pp. 232–241, Sep. 2019.

[13] H. Li, "Network topology design," in *Communications for Control in Cyber Physical Systems*. Amsterdam, The Netherlands: Elsevier, 2016, pp. 149–180.

[14] T. Wang, Q. Long, X. Gu, and W. Chai, "Information flow modeling and performance evaluation of communication networks serving power grids," *IEEE Access*, vol. 8, pp. 13735–13747, 2020.

[15] M. Shahraeini, M. H. Javidi, and M. S. Ghazizadeh, "Comparison between communication infrastructures of centralized and decentralized wide area measurement systems," *IEEE Trans. Smart Grid*, vol. 2, no. 1, pp. 206–211, Mar. 2011.

[16] 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," *Proc. CSEE*, vol. 36, no. 17, pp. 4521–4532, Sep. 2016.

[17] J. Valtari and S. Joshi, "Centralized protection and control," ABB, 2019. [Online]. Available: https://library.e.abb.com/public/6b20916a4d2e412daabb76fbada1268e/Centralized_Protection_and_Control_White_paper_2NGA000256_LRENA.pdf

[18] Y. Yan, Y. Qian, H. Sharif, and D. Tipper, "A survey on smart grid communication infrastructures: Motivations, requirements and challenges," *IEEE Commun. Surv. Tut.*, vol. 15, no. 1, pp. 5–20, Jan./Mar. 2013.

[19] F. M. Preparata and M. I. Shamos, *Computational Geometry: An Introduction*. Berlin, Germany: Springer, 1985.

[20] J. Hu, Z.-H. Li, and X. Z. Duan, "Structural feature analysis of the electric power dispatching data network," *Proc. CSEE*, vol. 29, no. 4, pp. 53–59, Feb. 2009.

[21] G. W. Li, W. Y. Ju, X. Z. Duan, and D. Y. Shi, "Transmission characteristics analysis of the electric power dispatching data network," *Proc. CSEE*, vol. 32, no. 22, pp. 141–148, Aug. 2012.

[22] Y. Wang, P. Yemula, and A. Bose, "Decentralized communication and control systems for power system operation," *IEEE Trans. Smart Grid*, vol. 6, no. 2, pp. 885–893, Mar. 2015.

[23] 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 *Proc. Int. Conf. Commun., Circuits Syst.*, 2006, pp. 2687–2691.

[24] ENTSOE, "ENTSO-E transmission system map," 2019. [Online]. Available: https://www.entsoe.eu/data/map/

[25] X. Fan, D. Wang, S. Aksoy, A. Tbaileh, and J. Ogle, "Coordination of transmission, distribution and communication systems for prompt power system recovery after disasters," Pacific Northwest Nat. Lab., Richland, WA, USA, Tech. Rep. PNNL-28598, 2019.

[26] O. Boyaci, M. R. Narimani, K. Davis, and E. Serpedin, "Generating connected, simple, and realistic cyber graphs for smart grids," in *Proc. IEEE Texas Power Energy Conf.*, 2022, pp. 1–6.

[27] B. Qi et al., "An emerging survivability technology for dispatching service of electric power communication network," *IEEE Access*, vol. 6, pp. 21231–21241, 2018.

[28] R. Atat, M. Ismail, S. S. Refaat, E. Serpedin, and T. Overbye, "Cascading failure vulnerability analysis in interdependent power communication networks," *IEEE Syst. J.*, vol. 16, no. 3, pp. 3500–3511, Sep. 2022.

[29] Y. Zhang, T. Jiang, Q. Shi, W. Liu, and S. Huang, "Modeling and vulnerability assessment of cyber physical system considering coupling characteristics," *Int. J. Elect. Power Energy Syst.*, vol. 142, Nov. 2022, Art. no. 108321.

[30] "Synthetic cyber-physical power systems for continental Europe," 2024. [Online]. Available: https://github.com/Cyber-Resilient-Power-Grids/Synthetic-CPS

[31] F. Chung and L. Lu, "The configuration model for power law graphs," in *Complex Graphs and Networks*. Providence, Rhode Island, AMS Publications, 2006, pp. 223–237.