A GNN-based Generative Model for Generating Synthetic Cyber-Physical Power System Topology

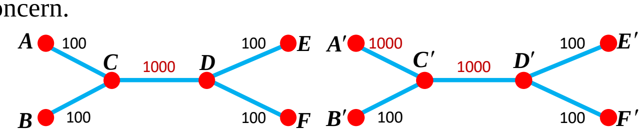
Yigu Liu, Haiwei Xie, Alfan Presekal, Alexandru Stefanov, *Member*, *IEEE*, and Peter Palensky, *Senior Member*, *IEEE*

Abstract**—**Synthetic networks aim at generating realistic projections of real-world networks while concealing the actual system information. This paper proposes a scalable and effective approach based on graph neural networks (GNN) to generate synthetic topologies of Cyber-Physical power Systems (CPS) with realistic network feature distribution. In order to comprehensively capture the characteristics of real CPS networks, we propose a generative model, namely Graph-CPS, based on graph variational autoencoder and graph recurrent neural networks. The method hides the sensitive topological information while maintaining the similar feature distribution of the real networks. We used multiple power and communication networks to prove and assess the effectiveness of the proposed method with experimental results.

Index Terms**— Cyber-Physical Systems, Graph Neural Networks, Synthetic Networks.**

I. 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.



(a) (b)

Fig. 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 [1][2]. 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 [1][2], 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 Fig. 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 Section IV 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.

II. Modeling of Cyber-Physical Power System

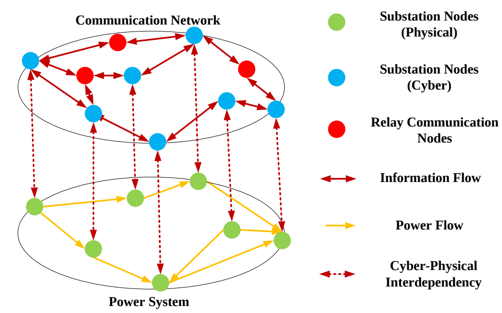


Fig. 2 Illustration of the “partially one-to-one” interdependency of CPS.

As shown in Fig. 2, we model the cyber-physical power system as an interdependent network consisting of two layers, i.e., communication network $\mathbb{G}_C(\boldsymbol{V}_C,\boldsymbol{E}_C)$ and power system $\mathbb{G}_P(\boldsymbol{V}_P,\boldsymbol{E}_P)$, 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 $\mathbb{G}_C$ and $\mathbb{G}_P$.

According to [1], the interdependencies of CPS can be divided into “one-to-one”, “one-to-multiple”, and “multiple-to-multiple” correspondences. In this paper, we follow the typical substation communication structure from [4]. 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.

III. Graph-CPS: Generating Synthetic 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 is the type feature of node. In this paper, we consider three different node types in the power system, i.e., generator, load, and zero injection node, and $x_t=-1,1,0$, respectively. In the communication model, we consider all nodes are substation routers. x_i is the network feature of node *i*. Note that one can perform different types of node/edge features to serve different research goals. In this paper, we use capacity centrality to quantify the feature of the nodes in both the communication network and power system, as shown in (1).

$$x_i=\sum_{j\in \boldsymbol{N}_i} e_j \tag{1}$$

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 $V(x)=P(x_i,x_t\le x)$. 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 (2).

$$KL(V(\hat{x})\|V(x))=-\sum V(\hat{x})\log\frac{V(x)}{V(\hat{x})} \tag{2}$$

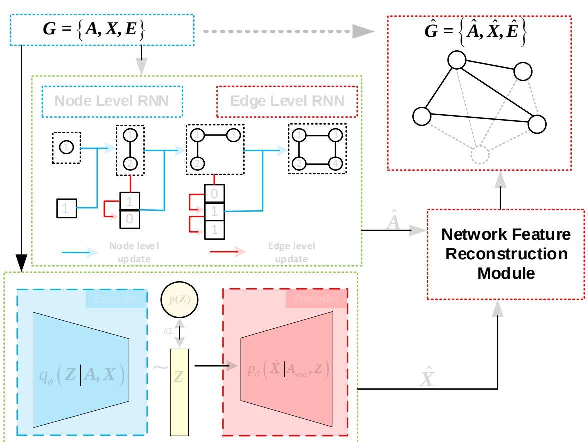


Fig. 3 Framework of Graph-CPS.

As shown in Fig. 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 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)-(4).

$$h_o=f_{trans}(h_{o-1},S_{o-1}^{\pi}) \tag{3}$$

$$\theta_o=f_{out}(h_o) \tag{4}$$

Where h_o encodes the generated graph of current time step, and S_{o-1}^{π} 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 [5]. As in Fig. 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 (5).

$$q_\phi(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{A})=\prod_{r=1}^{N} q_\phi(z_r|\boldsymbol{X},\boldsymbol{A}) \tag{5}$$

For the detailed definition of the two-layer GCN, readers are referred to [6]. 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 $\boldsymbol{A}_{vae}$ as inputs. $\boldsymbol{A}_{vae}$ is the result of the inner-product [6] sampling from $\boldsymbol{Z}$. Then, the generated node attribute $\hat{\boldsymbol{X}}$ is calculated as shown in (6)-(7).

$$p_\theta\left(\hat{\boldsymbol{X}}|\boldsymbol{A}_{vae},\boldsymbol{Z}\right)=\prod_{i=1}^{N} p_\theta(\hat{x}_i,\hat{x}_t|\boldsymbol{A}_{vae},\boldsymbol{Z}) \tag{6}$$

$$p_\theta(\hat{x}_i,\hat{x}_t|\boldsymbol{A}_{vae},\boldsymbol{Z})=\mathcal{N}\left(\hat{x}_i,\hat{x}_t|\boldsymbol{\mu}_i,\mathrm{diag}(\boldsymbol{\sigma}_i^2)\right) \tag{7}$$

Where $\boldsymbol{\mu}=\mathrm{GCN}_\mu(\boldsymbol{A}_{vae},\boldsymbol{Z})$ is the matrix of mean vectors $\boldsymbol{\mu}_i$ and similarly $\log\boldsymbol{\sigma}=\mathrm{GCN}_\sigma(\boldsymbol{A}_{vae},\boldsymbol{Z})$. The GCN in the decoder is defined as $\mathrm{GCN}(\boldsymbol{A}_{vae},\boldsymbol{Z})=\boldsymbol{A}'_{vae}\,\mathrm{ReLU}(\boldsymbol{A}'_{vae}\boldsymbol{X}\boldsymbol{W}_0)\boldsymbol{W}_1$, where $\boldsymbol{W}_0$ and $\boldsymbol{W}_1$ are the trained parameters. $\mathrm{ReLU}(\bullet)=\max(0,\bullet)$ 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 (2) and minimize the variational upper bound $\mathcal{L}$ as shown in equation (8).

$$\mathcal{L}=\mathbb{E}_{q_\phi(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{A})}[-\log_{p_\theta}(\boldsymbol{A}_{vae}|\boldsymbol{Z})]+KL[q_\phi(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{A})\|p(\boldsymbol{Z})]$$
$$+KL[V(\hat{x})\|V(x)] \tag{8}$$

After the RNN and VAE modules, $\hat{\boldsymbol{A}}$ and $\hat{\boldsymbol{X}}$ are obtained. In the NFR module, we use Algorithm 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 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 $\mathrm{Re}(\hat{x}_i)$ is the remaining node attribute of $\hat{x}_i$ that is not assigned to any edge yet. Initially, $\mathrm{Re}(\hat{x}_i)=\hat{x}_i$.

Based on [5][6] and Fig. 3, one can derive that both GraphRNN and GraphVAE use an encoder to learn a distribution $P_{\mathrm{model}}(G)$ based on the input data, which is stored in the latent space. Then, the decoder will interpret $P_{\mathrm{model}}(G)$ by sampling from the latent space and generate the output graphs, where the sampling is random but constrained by $P_{\mathrm{model}}(G)$. 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 1: Network feature reconstruction module

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

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

Step 1 $\hat{E} \leftarrow 0$
Step 2 Sort $\hat{X}$ in descending order
Step 3 Sort V_C / V_p in degree descending order based on $\hat{A}$
Step 4 Assign $\hat{X}$ to V_C / 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 $\hat{e}_j = \mathrm{Re}(\hat{x}_i) / |N_i^0|$
Step 8 Update $\mathrm{Re}(\hat{x}_i)$
Step 9 **End for**
Step 10 **Repeat** Step 5-8 until all $\hat{e}_j > 0$
Step 11 **Return** $\hat{G} = \{\hat{A}, \hat{X}, \hat{E}\}$

IV. Case Study

In this Section, we implement the proposed Graph-CPS on three power systems and three power grid communication networks to demostrate 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 [7]. For cyber layer, we use the communication network for Jiangsu province power grids in China [8] and two validated communication networks for IEEE 39-bus, 118-bus system, respectively [9][10]. 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 I 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 I, 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.

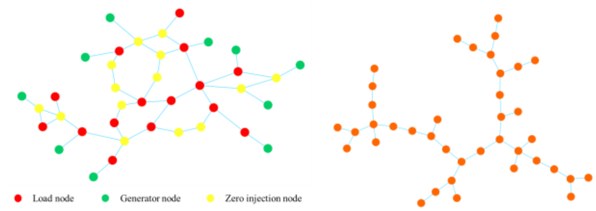


Fig. 4 (a) Generated synthetic power topology for IEEE 39-bus system, (b) Generated synthetic communication network.

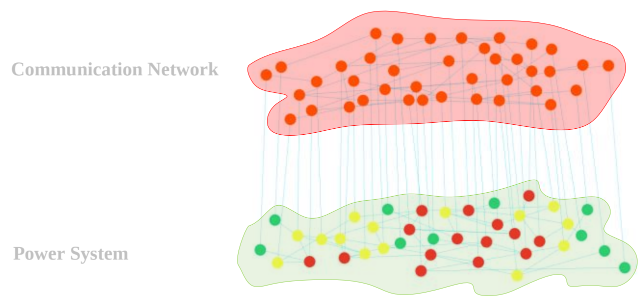


Fig. 5 Generated synthetic CPS for IEEE 39-bus system.

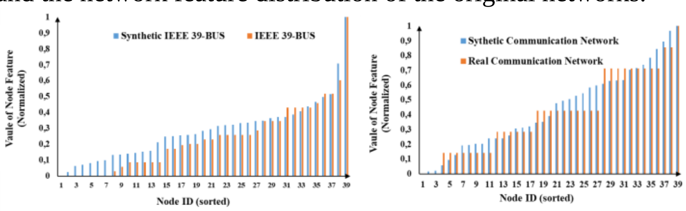
TABLE I Statistical Comparison Between Real and Synthetic CPS

	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)
$\langle bc \rangle$	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)
$Var(\hat{x}_i)$	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, $\langle bc \rangle$: average node betweenness centrality, $\max\langle bc \rangle$: maximum betweenness centrality, $\bar{\hat{x}}_i$: the mean value of the normalized synthetic node features, $Var(\hat{x}_i)$: the variance of the normalized synthetic node features, (*): the number in the brackets represents the difference between synthetic networks and original networks.

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 Fig. 4. Then, we form the interdependency for the synthetic CPS by following the “degree-to-degree” principle in [11] as shown in Fig. 5. In Fig. 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 Fig. 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 Fig. 6. In Fig. 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 Fig. 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.



(a) (b)

Fig. 6 (a) Comparison of node feature for IEEE 39-Bus system, (b) Comparison of node feature for IEEE 39-Bus communication system.

REFERENCES

[1] A. B. Birchfield, T. Xu, K. M. Gegner, K. S. Shetye and T. J. Overbye, “Grid Structural Characteristics as Validation Criteria for Synthetic Networks,” *IEEE Trans. Power Syst*., vol. 32, no. 4, pp. 3258-3265, Jul. 2017.

[2] 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.

[3] X. 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.

[4] “Centralized Protection and Control,” ABB, 2019. [Online]. Available: https://library.e.abb.com/public/6b20916a4d2e412daabb76fbada1268e/Centralized_Protection_and_Control_White_paper_2NGA000256_LRENA.pdf

[5] 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.

[6] T. N. Kipf, M. Welling, “Variational Graph Auto-Encoders,” arXiv: 1611.07308 [stat.ML], Nov. 2016.

[7] “ENTSO-E Transmission System Map,” ENTSOE, 2019. [Online]. Available: https://www.entsoe.eu/data/map/

[8] B. Qi et al., “An Emerging Survivability Technology for Dispatching Service of Electric Power Communication Network,” *IEEE Access*, vol. 6, pp. 21231-21241, May, 2018.

[9] 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,” in *PNNL-28598*, 2019.

[10] Yiwei Zhang, Tianlun Jiang, Qingxin Shi, Wenxia Liu, and Shaofeng Huang, “Modeling and Vulnerability Assessment of Cyber Physical System Considering Coupling Characteristics,” *Int. J. Electr. Power Energy Syst.*, vol. 142, part B, Nov. 2022.

[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.