Programa de Doctorado en Energía Eléctrica

A COMPLEX-NETWORK APPROACH TO SUPPORT TRANSMISSION EXPANSION PLANNING

Autor:
Rafael Espejo González

Directores:
Dr. Andrés Ramos Galán
Dra. Sara Lumbreras Sancho

Instituto de Investigación Tecnológica
Escuela Técnica Superior de Ingeniería ICAI
Universidad Pontificia Comillas

Madrid, 2019

A mis padres

Agradecimientos

A mis padres, por apoyarme en todas las decisiones que he tomado, incluida la de hacer el doctorado. Siempre habéis estado detrás ayudándome y animándome para conseguir todos los objetivos que me proponía. Gracias por enseñarme la cultura del esfuerzo y del trabajo.

A Sara y Andrés, aunque me advertisteis de que con esta tesis podría llegar a odiar a mis supervisores en algún momento, no se puede odiar a quien con cariño te guía y te ilumina en la oscuridad. Sara, tu inquietud investigadora siempre con un reto y una pregunta ha sido parte fundamental para no perder la motivación. Andrés, tu tranquilidad y experiencia ha sido la brújula perfecta para alcanzar los objetivos de esta tesis. Gracias a los dos. Os echaré de menos en esta nueva etapa en la que me embarco.

A mis compañeros IeFs, inteligencia en la que encontrar inspiración para la tesis y compañía para desconectar de ella. Fernando, Guille, gracias por sacar un hueco y contribuir a enriquecer este trabajo. A Marta, gracias por contagiarnos siempre con su optimismo. Espero que estés pronto de vuelta. Marta y Fernando, gracias por acompañarme desde el primer día en esta aventura. Otaola, un amigo con el que confrontar opiniones, gracias por esas largas horas de conversaciones y divagaciones.

A mis amigos Alejandro, Erika y Quique, gracias por hacer que me evadiera en los momentos de frustración. Enrique, espero que te animes y algún día poder leer los agradecimientos de tu Tesis. Erika, gracias por ser un buen ejemplo de amistad en una ciudad desconocida. Alejandro, aun en la distancia siempre has dado impulso a esta tesis, gracias amigo.

Abstract

An increase in power-grid complexity leads to more sophisticated and computationally intensive models for grid analysis, operation, and control. Despite computer advances, traditional power-system methodologies such as power-flow analyses may result in computationally hard problems. This thesis proposes the introduction of complex-network techniques for the research into power systems, leading to new models that provide good approximations with lower computational requirements. The thesis focuses on the generation of synthetic power grids and power-network vulnerability analyses.

Synthetic power grids are non-real power grids that are statistically similar to real power networks from a topological and electrical point of view. This work introduces a new algorithm to generate synthetic transmission power grids. The algorithm considers economic and technical factors in order to mimic the topology of real power networks. Results are tested on selected European transmission power networks.

The thesis also introduces a new metric for the analysis of power-network vulnerability. This is of particular interest in cases such as deliberate attacks. Betweenness centrality, a network-topological metric, is endowed with electrical parameters. It results in a hybrid metric, the Electrical Line Centrality, that measures the impact of line failure on the network. This metric improves prior results while reducing computational times. This is crucial in order to include the protection against deliberate attacks in the network design problem.

Finally, the analysis of power network topology is a necessary prior step in the generation of synthetic power grids and the assessment of power-network vulnerability. In this work, the power-network structure is characterized by global metrics traditionally used in complex-networks. Furthermore, a new framework is introduced to characterize network structure, enhancing network description, and classification. This framework will also allow for the topological validation of synthetic power grids.

Resumen

Un aumento en la complejidad de las redes eléctricas conduce a la necesidad de modelos más sofisticados para el análisis, operación y control de estas. Este aumento en la sofisticación de los modelos implica un incremento del coste computacional de los mismos. A pesar de los avances informáticos, las metodologías que tradicionalmente se han aplicado a los sistemas de energía eléctrica, como los análisis de flujo de carga, pueden tener tiempos de ejecución muy elevados. Esto podría llegar a comprometer la utilidad de estos. Esta tesis propone la introducción de técnicas de redes complejas en los problemas relacionados con los sistemas de energía eléctrica. La introducción de conceptos propios de la teoría de redes complejas permitiría desarrollar modelos que dieran soluciones aproximadas con un coste computacional reducido. Esta tesis se centra en la aplicación de técnicas redes complejas para la generación de redes eléctricas sintéticas y para analizar la vulnerabilidad de la red eléctrica.

Las redes eléctricas sintéticas son redes eléctricas no reales que tienen propiedades eléctricas y topológicas similares a las redes reales. Es decir, son redes ficticias pero cuya operación y control es similar al de las redes reales. Esta tesis propone un nuevo modelo para la generación de redes eléctricas sintéticas. Este nuevo modelo utiliza consideraciones económicas y eléctricas para generar redes sintéticas con una topología similar al de las redes reales. El modelo es validado con la red de transporte eléctrico de España, Portugal y Francia.

Además, la tesis introduce una nueva medida para el análisis de la vulnerabilidad de la red eléctrica. Esto es de especial interés en casos como los ataques deliberados. Una medida propia de redes complejas, la centralidad de intermediación es completada con información eléctrica. Esto da como resultado una métrica híbrida, la Centralidad de la Línea Eléctrica, que mide el impacto del fallo de una línea en la red. Esta nueva medida permite mejorar los resultados obtenidos con las medidas propuestas anteriormente en la literatura, al tiempo que reduce el coste computacional. Esto es crucial de cara a la inclusión de la protección de la red eléctrica contra ataques deliberados en el problema de diseño de la red.

Finalmente, el análisis de la topología de la red eléctrica es un paso previo a la generación de redes eléctricas sintéticas y a la evaluación de la vulnerabilidad de la red eléctrica. En esta tesis se hace una descripción pormenorizada sobre la estructura de la red eléctrica con las medidas tradicionalmente utilizadas en redes complejas. Además, se introduce un nuevo modelo que permite caracterizar la estructura de la red de forma sistemática y detallada, mejorando la descripción y comparación de redes complejas. Este nuevo modelo se utilizará para la validación topológica de redes eléctricas sintéticas.

Contents

List of Figures

List of Tables

Nomenclature

Indices, sets, parameters, and abbreviations used in Chapter 6.

A. Indices and sets:

n	Electrical nodes
l	Transmission lines
$r(n)$	Subset of nodes without slack bus
$G(n)$	Subset of generation nodes
$D(n)$	Subset of demand nodes
$I(n)$	Subset of interconnection nodes

B. Parameters

F	Vector of power flows
X	Vector of line reactance
θ	Vector of voltage angle differences
A	Line incidence matrix
B	Matrix of network susceptances
P	Vector of power injections
S	Matrix of power transfer distribution factors
ΔP	Vector of changes in power injections
$BC(u)$	Betweenness centrality of a node or a line u
$\sigma_{i,j}(u)$	Number of shortest paths from node i to node j that go through u
$\sigma_{i,j}$	Number of shortest paths from node i to node j
$ELC(l)$	Electrical line centrality vector

$ENC(n)$	Electrical node centrality vector
w_g	Vector of node generation capacity
w_d	Vector of node demand
C	Total generation capacity installed in the network

C. Abbreviations

PNS	Power not supplied
DCPF	DC Power Flow
DCOPF	DC Optimal Power Flow
MAE	Mean absolute error
PTDF	Power transfer distribution factor

1

INTRODUCING COMPLEX NETWORKS TO POWER SYSTEMS

1.1. Increasing power-system complexity

The structure of power networks has experienced substantial changes since their appearance. From a set of local low-voltage networks, power grids evolved to become large-scale high-voltage networks that extend over several countries. Moreover, power grids interact with other networks, such as gas or communication networks [1], [2]. Furthermore, relatively recent developments in power systems such as the spread of flexible alternating current transmission systems (FACTS), high voltage DC lines or distributed generation, are shifting the traditional vision of power systems while increasing system complexity [3].

An increase in system complexity calls for more sophisticated and computationally intensive models for grid analysis, operation, and control. Despite computer advances, traditional power-system methodologies, such as power-flow analyses, may result in computationally hard problems, requiring large computational resources and computing times. Moreover, the interconnection among networks increases power-grid vulnerability: a failure in one system can propagate to other systems leading to blackouts with severe economic consequences. This means that systems cannot be studied in isolation, which increases the size of the networks under study. New challenges such as the integration of renewable energy sources or demand response further increase the complexity of the problem.

1.2. The lack of power-grid data hinders innovation

The increase in network size, as well as the introduction of innovative solutions, require the development of new algorithms for network operation, control, and design. Despite a large number of contributions published every year, power-system research is often hindered by the lack of public data. Real data, such as network models, are crucial to test and validate theoretical developments.

Validation processes may require the comparison of new algorithms with existing ones. However, when public data are scarce, re-using a minimal set of test cases is problematic, since the performance of some algorithms, such as heuristics or metaheuristics, might be case dependent. Another solution is to use private data for the construction of test cases. This hampers transparency and replicability. Furthermore, the lack of public data might not encourage interdisciplinary research, since the barriers to entry for experts from other disciplines, who might be tempted to research into power systems are high. Moreover, research might be conditioned by the interest of data owners, which might not be aligned with other stakeholders' interests. These factors combine to one conclusion: the lack of power-grid data is a barrier for research and innovation.

Although some TSOs have started to publish data in Europe within the INSPIRE directive [4], network models are scarce, and information is only partial. Consequently, there is a long haul before the publication of detailed network models. In the U.S., access to real network models is almost null because of security concerns. There are cautions against the disclosure of the real location and the topology of power networks because of terror concerns.

Recent advances in power systems as well as the higher degree of connection with other networks have increased the complexity of power networks. New models are needed to operate and control power networks. Public network models are therefore necessary to enhance research into power systems. However, public data are scarce, and the disclosure of real information might run into security issues.

1.2.1. Existing test cases

The lack of power-network models conditions the testing and validation of theoretical algorithms. We usually see algorithm proposals that are applied to small and old-fashioned test cases such as the IEEE-118-bus standard. This test case stands for a portion of the North American power network in the year 1961. This network cannot represent the complexity of current power grids. Forty-eight years later, transmission networks include higher voltage levels and new types of power generators. After the installation of the first wind farm in 1980, power generation is shifting from large power plants connected to transmission networks to small renewable power plants that are distributed along the transmission and distribution networks. Moreover, these standard cases do not consider recent technologies such as storage systems that can alter the operation of power networks. Finally, the small size of those systems cannot replicate the real behavior of real large networks that expand beyond the borders of each country.

The IEEE-118-buses test case, as well as other IEEE standards, are publicly available in the Power System Test Archive of Washington University [5]. There, repository owners specified the drawbacks of those systems concerning the power-network conditions in 1993. A list of available test cases, network location, date and limitations is shown in Table 1-1. This repository

also includes three dynamic test cases with similar deficiencies. The University of Edinburgh also has an open-access repository that includes four power-flow test cases (information is detailed in Table 1-2) [6]. The unsuitability of test cases is also an issue in distribution networks [7]. Accordingly, test cases should be updated to reflect the complexity of the current power network.

Table 1-1. Test cases available in the Power System Test Archive of Washington University
Source: Washington University [4]

IEEE test case	Location	Date	Comments
14 buses	The midwestern US	February 1962	No line limits. Low base voltage. An overabundance of voltage control capability.
30 buses	The midwestern US	December 1961	No line limits. Line impedance may be wrong.
57 buses	The midwestern US	Early 1960's	No line limits. Line impedance may be wrong.
118 buses	The midwestern US	December 1962	kV levels defined as a bad guess. MVA limits were not part of the first data.
300 buses	No Info	1993	No comments

Table 1-2. Test cases available in the Power System Test Archive of the University of Edinburgh
Source: University of Edinburgh [5]

Test case	Location	Date	Comments
39 bus test case	New England	No Info	Same cost for all generators
Iceland network	Iceland	Published 2011	No cost information Voltage bounds were assumed
Reduced GB network	Great Britain	Published 2010	No comments
GB network	Great Britain	No Info	Obtained from official publicly available data

All prior cases are also available in Matpower [8]. Matpower is a Matlab-based power-system simulation package that includes novel realistic systems such as the PEGASE networks (five instances whose size ranges from 89 to 13,659 nodes) or a few demand-generation scenarios for the Polish grid. The NESTA archive also adds network operation constraints to a set of existing test cases to evaluate and validate power-system optimization algorithms [9]. However, those cases continue to be insufficient for research purposes because they are only used for analyzing power flows in an existing network; additional information should be provided to complete them. Crucially, they do not provide the location of nodes. This information is critical in applications such as transmission expansion planning, where the location of nodes is necessary to calculate the installation cost of new lines. Similarly, those cases usually give just one demand scenario. An extensive set of demand scenarios may be required in specific studies that research into demand response.

Recent manuscripts have published new test cases that are the result of combining several datasets. This is the case of the European transmission network or the Australian power grid [10], [11]. Although some TSOs are starting to publish some limited information, several datasets are usually needed to build a complete test case. For instance, we may combine the electrical parameters of the European transmission network provided by ENTSO-e with the geographical coordinates of the French network provided by RTE to build a case for the France power network [12], [13]. However, this is not a distinct task. Although both networks stand for the same power grid, the number of nodes is not the same in both of them due to different model assumptions. Besides, no details about the generation are provided. The development of those test cases is therefore conditioned by third-party data, and it might not be possible to update them easily. New test cases need to be functional to have the chance to include further power-system developments as well as new detailed information.

Traditionally used test cases do not replicate the complexity of existing real power networks. Although novel network models have been proposed, they continue to be insufficient. New efforts and approaches are required to generate functional and accurate network models that include the complexity of real power grids.

1.2.2. New initiatives

There are two main groups of initiatives that aim to develop functional network models. On the one hand, OpenStreetMap-based algorithms try to create real network models based on crowdsourced data. On the other hand, the GRID DATA project of ARPA-e encourages the development of algorithms to generate non-real, albeit realistic network models.

A. Open-Access-Map initiatives

The OpenStreetMap, OSM, is an initiative to create an open-access map of the world [14]. This map is built with crowdsourced data; everybody can contribute to add information about the real location of roads or transmission lines. Several projects have tried to build network models extracting the information related to power networks, such as substations, generators or transmission lines location, from OSM. All those elements need to be connected (the user must set up the connection of transmission lines within substations). In addition, no electrical parameters are given, so the data are not directly implementable into a model.

Another initiative is SciGRID. The SciGRID project was started by the Next Energy research group and funded by the German Federal Ministry of Education [15]. It aims to build a European transmission-network model. They also have a similar initiative with the gas network. They filter the power information obtained from OSM and abstract the topological information to add electrical parameters based on typical cable values. Similarly, the osmTGmod project uses OSM to build the German transmission network; they complete the information that is missing using heuristics [16].

Rivera et al. propose the automation of network model generation from OSM data [17]. The automation avoids the manual introduction of power relations among elements such as substations, transmission lines, or generators required by prior works. This model is offered at the OSM platform [18]. They have tested the accuracy of this algorithm with 14 real networks, and it ranges from 31% to 94%. The accuracy is the ratio between the line length inferred by the proposed models and the length officially reported by TSOs. This validation does not include any electrical parameter testing. They have also developed a mobile app to enhance users to update the location of power-system components with their smartphone.

The main drawbacks of these initiatives are the lack of electrical information and the errors and missing data in OSM [19]. Accordingly, the development of realistic network models will depend on the time users need to complete the information of all power-system-components location. The accuracy of the models is therefore conditioned by third-party information and the assumptions made to endow network models with electrical parameters.

Finally, these initiatives imply the disclosure of the real location and the topology of power networks that might run into additional issues.

B. Generation of synthetic power grids

The Generating Realistic Information for the Development of Distribution and Transmission Algorithms (GRID DATA) initiative aims to develop novel network models to be used as test cases [20]. Those new test cases should reproduce the characteristics of North American power networks. This initiative is funded by the Advanced Research Projects Agency within the U.S. Department of Energy with $11.3 M. The motivation of this initiative is the development of new Optimal Power-Flow algorithms that will contribute to the increase in network efficiency and reliability. They will also support operation-cost reduction and integration of renewable resources. Furthermore, they highlighted that access to public data would stimulate optimization competitions, encouraging novel contributions.

To generate new network models, ARPA-E proposes two alternatives. First, the anonymization of real data provided by utilities. Second, the development of new algorithms to build realistic network models. In both cases, network models will be published into an open-access repository. Those projects are already underway, and their contributions are expected in the next few years.

Although the generation of algorithms to build test cases is a relatively new topic for transmission networks, they have been used in the analysis of geomagnetic disturbances or to test communication and control networks for smart grids [21], [22]. Some works are focused on the generation of synthetic distribution networks (e.g., Reference Network Model developed at the Institute for Research in Technology) [23].

1.2.3. Public does not mean real data

The lack of public data slows down research and hinders innovation. However, publishing real data raises security issues. For instance, real power-system data can be used to develop new control algorithms, but also to plan an attack to maximize the probabilities of a blackout. Accordingly, the publication of real information, as proposed in OSM initiatives might be controversial.

However, research does not need real information. Researchers need truthful, realistic information: non-real datasets are just as good as real ones, as long as they have the same properties. By the use of realistic test cases, theoretical models and algorithms can be tested in networks that replicate the conditions of real power grids.

The two procedures proposed to provide public data by ARPA-E (network anonymization and network model generation) find a balance between data availability and security issues. This is a strength that is crucial concerning OSM initiatives. Beyond the limits of OSM approaches such as the lack of electrical parameters or the missing information, their success is conditioned by third-party data. This problem is also present in the anonymization of utilities' information. We, therefore, think that the generation of synthetic power grids is the best alternative to the lack of network models. Furthermore, synthetic power grids allow for the introduction of novel developments or data. ARPA-E is the seed of a new research line based on the generation of synthetic power networks. Although ARPA-E defines the research question with clarity, the choice of methods to generate the networks is a crucial further step. The use of traditional power-methods (e.g., optimization problems) may not be an accurate tool to generate those synthetic networks because of network size. New approaches are required to generate those synthetic power grids.

The generation of synthetic power grids, non-real albeit realist network models, is a suitable solution for the lack of publicly available network models. Those systems are not real and do not disclose information about real power networks. However, they replicate the complexity of real networks and have their same properties. Operation and control are similar to the ones of real networks. They can be used in research projects.

1.3. New threats to power-grid robustness

Not only is the lack of network models a problem that cannot be addressed with traditional power-system techniques, but the assessment of network vulnerability is also a challenge due to the large size of power networks and the increasing interconnection among systems. Vulnerability assessment is a crucial step in the design of robust networks. A failure in power networks may lead to a blackout with severe consequences. For instance, in 2012, 620 million people were affected by a power outage in India [24].

Power systems are designed to have redundant lines and extra generation capacity in order to be able to meet demand in case of failures. Therefore, TSOs try to analyze the adequacy of the network to design those redundancies. The adequacy of a power system is defined as "the ability to supply the aggregate electrical demand and energy requirements of the end-use customers at all times, taking into account scheduled and reasonably expected unscheduled outages of system elements" by the National Electric Reliability Council (NERC) [25].

Traditionally, N-1 analysis has been used in electrical engineering to study network adequacy [3]. This analysis provides detailed results of how power flows through the network in case of line or generator failure. It also quantifies the energy that would not be supplied in the system in case of failure. Consequently, TSOs have accurate information to ensure power supply after component failures. However, the analysis of extensive power grids may consume large computational resources. Network design usually considers the failure of one or two components in power networks. Thus, power networks are supposed to be robust against component failure.

However, if a power network does not respond quickly in case of component failure, it may suffer a cascade of failures that lead to massive blackouts [26]. Furthermore, the blackout might also be aggravated by other network failures, such as telecommunication networks.

Furthermore, power networks might be the target of deliberate attacks. These are targeted attacks that aim to collapse power networks, as the cyber-attack that caused the Ukrainian blackout in 2015, affecting 225,000 customers [27]. This collapse was caused by a cyber-attack in which substation breakers were remotely opened. This caused the failure of several components simultaneously affecting 30 different substations.

Accordingly, the N-1 criterion used to design networks is not an accurate tool in case of cascade failures or deliberate attacks. Although it is not possible to completely mitigate the risk of a blackout, network design might contribute to reducing the size and cost of those blackouts. Novel approaches have tried to evolve from the traditional power-network vulnerability assessment. Based on N-1 analysis, high-risk N-k analysis proposes the creation of a list with the most vulnerable elements in the network [28]. Adequacy analysis of power systems can also be addressed as an optimization problem by formulating the problem of optimal interdiction of a power grid in order to identify critical elements [29]. This is a max-min programming problem in which a terrorist tries to attack the system maximizing loss of load. This problem can also be formulated as a Mixed-Integer Non-Linear (MINLP), Bi-level problem, or as Mixed-Integer linear problem (MIP) [30], [31]. However, computational requirements continue limiting the analysis of the power-network vulnerability.

The N-1 analyses are used to assess power-network adequacy. Those studies support the design of robust networks against component failures. However, both the large size of real networks and the high degree of interconnection with other networks limit results obtained from this approach. Furthermore, it is not a suitable approach for deliberate attacks. New methodologies are required to protect power networks of new threats without compromising computational requirements.

1.4. The role of complex-network techniques

Complex networks are systems composed of a large number of connected units that interact among them [32]. Complex-network techniques arose from graph theory to study these interactions -graph edges- among system units -graph nodes-. From biology to social science, complex networks have been applied to different goals, such as the study of protein-protein interaction or the prediction of the currency market in online gaming [33], [34]. The analysis of complex-network topology allows us to understand the principles that guide network evolution and condition its behavior [32].

Power grids can be modeled as complex networks with a set of substations that are connected through transmission lines. The interaction among those substations is the power that flows through power lines when power is injected or withdrawn in each node.

Unlike power-system methodologies, complex-network techniques are relatively light in computational requirements. However, they only consider the topological structure of the system and, in principle, they disregard their nature (e.g., in power networks they do not consider power flows or additional electrical information). In the case of power networks, mere network analysis does not consider the Kirchhoff's circuit laws that govern power flows. Recent studies have started to adapt complex-network methodologies to power systems, for instance by the inclusion of power-flow analyses [35]. Accordingly, by merging complex-network techniques and tools from the power-system domain, new models may provide good approximations with lower computational requirements.

The combination of complex networks and power systems is therefore accurate for the generation of synthetic grids and power-network vulnerability analysis, two problems that cannot be addressed with traditional power-systems methodologies.

1.4.1. Generating synthetic power grids

The need for new power-network models leads us to take up the problem of generating synthetic power grids. Synthetic power grids are non-real power grid cases that are similar to real power networks from a topological and electrical point of view. They are fictitious networks, so information about the real network is not disclosed, but they are similar in terms of operation and control.

In the complex-network field, several algorithms were proposed to generate synthetic networks with low-computationally intensive models [36]. An example of those algorithms is the *preferential attachment model* [37]. Based on the idea "the rich get richer", the algorithm proposed by Barabási and Albert generates networks in which degree distribution follows a power law. The only information considered when generating the synthetic network is the number of connections per node. Thus, those algorithms only focus on network topology disregarding their nature. While those models may be accurate in social networks, they cannot be used in power networks. We may generate a network in which the transmission capacity of transmission lines connected to a power plant is lower than the generation capacity of that generator.

Furthermore, some of those algorithms assumed that networks have specific topological properties. Although several studies have analyzed the structure of power networks, results have led to controversial conclusions [38]. It is not clear whether the topology of power networks fits with the topological features of those synthetic networks or not, and this question has not been assessed satisfactorily to the best of our knowledge. Consequently, as an earlier step to the generation of synthetic power grids, a sound topological analysis of power networks is required.

The analysis of network topology will determine how synthetic networks are generated and, at the same time, will be a measure of network accuracy (one of the conditions to state that a synthetic network is realistic is that its topology is similar to a real power network structure). Topological conclusions may be incorporated in two ways. First, the algorithm may choose which lines would be installed incorporating only topological considerations. That is, the algorithm installs those lines that minimize the error concerning the target topological properties. Second, when lines are added based on an electrical criterion, candidate lines will be filtered considering the target network properties. Finally, synthetic networks should be considered valid only if they share their topological properties with real power networks.

Consequently, complex networks may support the generation of synthetic power grids since they do not require computationally intensive models. New algorithms should combine complex-network techniques with electrical criteria to build realistic test cases.

1.4.2. Assessing power-network vulnerability

Unlike the generation of synthetic networks, which is a relatively new and promising line of research, we find several works that approach the assessment of power-network vulnerability from a complex-network perspective.

The first works try to determine the most critical elements in networks by using exclusively topological metrics (see Chapter 6). However, the results were not accurate enough. Some authors question the ability of complex-network metrics to assess power-network vulnerability since they disregard the electrical nature of power systems. Accordingly, they will never provide accurate results when analyzing network vulnerability [39].

Later on, complex-network metrics were adapted to power systems by including electrical considerations such as transmission lines parameters. Cuadra et al. provide an extensive review of the analysis of robustness in power networks by applying complex-network concepts [40]. Although those hybrid metrics obtained improved results, new enhancements should be introduced to be the right approach and substitute for traditional power-system analyses. Those metrics are promising because of the low computational requirements.

Despite not considering network operation explicitly, vulnerability indices may be effectively incorporated into network design. Transmission expansion planning, designing the transmission network with optimization methods [41], may benefit from these indices in two ways: by introducing them as a partial objective in the optimization function (it penalizes high values of vulnerability indices) or by including them as constraints (it establishes maximum values for the indices). These topological metrics can also be used to select those lines that are potentially promising candidates to be installed in the network [42].

New improvements should be introduced to improve the results provided by complex-network metrics and to be included in the network design problem. Furthermore, as in the case of synthetic power networks, a topological analysis is the previous step to analyze the properties of complex networks. It is also necessary to clearly define how power networks should be modeled as complex networks.

Complex network techniques can support the analysis of power systems with computationally light models. Complex-network studies analyze system topologies. This is a good approach to the generation of synthetic power grids and the assessment of power-network vulnerability. In both cases, complex network models should be completed with electrical information that captures the electrical nature of power grids.

1.5. How can this thesis contribute to research into power systems?

The application of complex-network techniques to the power grid is a line of research that may support the generation of synthetic grids as well as power network vulnerability studies. Those techniques contribute to face two of the problems existing in power systems: the lack of available network models, and the vulnerability assessment in case of cascade failures or deliberate attacks.

Although a few models have been proposed to generate synthetic power grids, the topology of the resulting networks is not consistent with real grids (as it will be further discussed in Chapter 4). This thesis proposes a new algorithm for the generation of synthetic power grids that combines complex-network techniques with electrical considerations. The resulting networks are tested against the real transmission grids of Spain, Portugal, and France.

Furthermore, it proposes a novel framework to validate synthetic power grids. This framework also allows for the understanding and description of the complex-network structure, independently of the nature of those networks. This is a great achievement in the field of complex networks that supports network classification and comparison.

Besides, we analyze complex-network metrics used to asses network vulnerability. As explained, existing methods cannot capture the electrical nature of power systems and results are worse than strictly electrical models. We propose a new hybrid metric that reduces computational requirements while improving results.

To sum up, the thesis objectives are the following:

- The development of a comprehensive analysis of transmission-power-network topology. This will support the understanding of the power-network structure in order to guide the generation of synthetic power grids and the assessment of power-network vulnerability.
- The proposal for a new algorithm to generate synthetic power grids. The novel algorithm will focus on transmission power networks. Furthermore, the resulting synthetic networks should be validated from a topological point of view.
- The proposal for a new hybrid metric to assess power-network vulnerability. This new metric should combine complex-network metrics with electrical parameters.

A detailed summary of each chapter is presented in the section below.

Research Questions & Objectives:

This thesis proposes a **novel algorithm to generate synthetic power networks** by combining complex network techniques with electrical considerations. This merge is also the origin of a **new complex network metric to assess power network vulnerability**. To provide a foundation for this, the thesis develops an extended analysis of power-network topology. It also introduces a **novel framework to describe complex-network structure**.

1.6. A quick guide to the rest of this document

The content of the chapters is described as follows:

- Chapter 2 introduces the topological analysis of power networks. It applies a set of global statistics: network size, degree distribution, characteristic-path length, network diameter, betweenness centrality and network average clustering coefficient to fifteen European transmission networks (400 kV and 220 kV). This analysis tries to find topological patterns and differences among networks by analyzing metric scalability. The analysis focuses on voltage level and network location. Finally, this chapter discusses the characterization of the power network as a scale-free network and a small-world network. This topological analysis has been published as:

  - R. Espejo, S. Lumbreras, and A. Ramos, "Analysis of transmission-power-grid topology and scalability, the European case study," Physica A: Statistical Mechanics and its Applications, vol. 509, pp. 383–395, Nov. 2018.

- Chapter 3 presents an innovative approach to describe complex-network topology from graphlet decomposition, which improves existing approaches for network characterization. This new framework exploits the local information provided by graphlets to give a global explanation of network topology. We propose a twelve-dimensional metric that analyzes how 2- and 3-node graphlets describe the structure of networks. The twelve dimensions are independent of network size, so they allow for direct comparisons of different networks regardless of size. It also reduces the complexity of graphlet counting, since it does not use 4- and 5- node graphlets. The application of the novel framework to five real networks demonstrates its potential to explain both global and local network topological properties. We apply the proposed metrics to a broad set of networks to show that it can classify networks of different nature based on their topological properties. In order to further simplify the interpretation of our graphlet analysis, we reduce the twelve dimensions to their main principal components. This paves the way for a connection between complex-network analyses and diverse real-world applications. This novel framework and the application to real networks have been included in a working paper as:
  - R. Espejo, G. Mestre, F. Postigo, S. Lumbreras, A. Ramos, T. Huang, and E. Bompard, "Exploiting graphlet-decomposition to explain the structure of complex networks."
- Chapter 4 applies the novel approach proposed in Chapter 3 to the European transmission power network. The twelve-dimensional metric supports a better understanding of power-network topology. It explains the similarities and differences we find among networks considering network location and voltage level. Furthermore, it is proven to be an adequate tool to assess the topological consistency of synthetic power networks. The use of this framework clearly shows if the topology of a synthetic network is consistent with real power networks or not. We also analyze the topology of existing synthetic networks. Results show that those networks are not topologically consistent with the European transmission power networks. The straightforward interpretation of the twelve dimensions allows for the improvement of synthetic-network-generation algorithms.
- Chapter 5 proposes a new algorithm to generate synthetic spatial power grids. The proposed algorithm mimics the historical evolution of power systems by taking into account economic and technical factors. The algorithm is articulated in two steps, the first step is focused on economic efficiency to meet demand, and the second one is targeted at increasing network robustness while achieving some topological attributes. We generate a synthetic network for the Portuguese, Spanish and French 400-kV transmission networks. Those networks are shown to be topologically consistent, according to the metrics presented in Chapter 2 and Chapter 3, with real ones. The

parametrical nature of the proposed model allows for the generation of different instances of consistent power networks, an exciting feature for grid generation. The content related to the generation of synthetic power grids has been published as:

- R. Espejo, S. Lumbreras, and A. Ramos, "A Complex-Network Approach to the Generation of Synthetic Power Transmission Networks," IEEE Systems Journal, pp. 1–4, 2018.

It has also been presented in the Windfarms 2017 conference:

- R. Espejo, S. Lumbreras, and A. Ramos, "Generating statistically consistent synthetic power networks for testing renewable integration models," Windfarms 2017, Madrid, Spain, Jun 2017.

- Chapter 6 introduces to the assessment of power-network vulnerability with complex-network metrics. Based on prior work, we show that pure topological metrics do not give conclusive results in vulnerability analyses. However, extended topological metrics, which endow topological metrics with electrical considerations, provide satisfactory results with lower computational requirements. This chapter proposes a new extended metric, the electrical line centrality, that can be applied to ranking lines according to the impact of line failure in the network. The proposed metric is based on the idea of betweenness centrality, and it considers parameters related to power demand, generation, and transmission lines. Simulations confirm the improvement of results concerning prior works. The proposal of the line electrical centrality has been published as:
  - R. Espejo, S. Lumbreras, A. Ramos, T. Huang, and E. Bompard, "An extended metric for the analysis of power-network vulnerability: the line electrical centrality", PowerTech 2019, Milan, Italy, Jun. 2019.
- Chapter 7 extracts conclusions and summarizes the main contributions of this thesis. Finally, it outlines further research.

2

A TRADITIONAL APPROACH TO POWER-NETWORK TOPOLOGY

2.1. Introduction to power-network topology

The analysis of network topology is the previous step to the generation of synthetic power grids and the analysis of the power-network vulnerability. In the complex-network field, several studies have tried to characterize network topology by finding common structures or patterns in different networks. Albert et al. presented the case of scale-free networks, those networks in which the distribution of node degree (the number of lines attached to each node) follows a power law [37]. Scale-free networks are robust against random failures. However, they are incredibly vulnerable in case of deliberate attacks, since the loss of some prominent nodes or links has the potential to disrupt the whole network. Erdös et al. examined random graphs, which are characterized by a low network average clustering coefficient (the probability that the neighbors of one node are also connected among them) and short distances among nodes. Random networks are vulnerable under both random -accidental- and deliberate attacks [43]. Small-world networks have low characteristic path length, but their network average global clustering coefficient is higher than in the case of random networks [44].

The analysis of power-network topology is, therefore, of particular interest in the application of complex-network techniques in research into power systems. Existent works have studied whether power-network topology fits the models above or not [38]. Most of these works have focused on specific national power grids such as the Iranian, South Korean or North American power grids [45], [46], [39]. However, there is not a homogenous conclusion when defining power-grid topology, e.g. whether power grids are small-world networks or not, or what probabilistic function fits degree distribution better. This lack of consensus may lie on the heterogeneous data used in prior analyses, for instance by comparing networks with different voltage levels. We find it is necessary to present a consistent topological analysis of different power grids based on comparable data in order to obtain definite conclusions about power-network topology. This would allow us to make a comparison among countries, extracting information about topological metrics and analyzing how complex-network metrics scale with

network size. Consequently, our work provides information about the metrics that better describe transmission power grids and their properties.

This chapter introduces the analysis of the power grid as a complex network by presenting different ways of characterizing and modeling power grids as graphs in Section 2. Section 3 analyzes fifteen European transmission networks from a topological point of view; it focuses on how complex-network metrics scale with network size. Section 4 discusses the implications of these results. Section 5 analyzes the impact of the analysis on the generation of synthetic power grids. Finally, Section 6 presents chapter takeaways.

The analysis of power network topology is the prior step to the generation of synthetic power grids and the assessment of power-network vulnerability. Although several studies have been proposed to characterize the topology of power networks, results are not consistent. Divergent results may be the consequence of using different voltage levels, with different topological properties, or different model assumptions. This chapter focuses on transmission power networks.

2.2. Modeling power grids as complex networks

Complex systems are large sets of individual units that are highly interconnected among them [32]. Power networks are large infrastructure networks composed of power lines that interconnect demand with power generation plants (both demand nodes and generators can be understood as nodes or substations). Accordingly, power networks are complex networks, and they can be modeled as graphs. In this case, graphs, $G(N, L)$, are set of vertices or substations, N, that are linked through edges or transmission lines, L.

Power networks, like other infrastructure networks such as roads, can be modeled as weighted graphs. In the case of power networks, edge weight may represent the maximum power that can flow between vertices (i.e., transmission capacity) or be used to characterize other electrical properties (such as line impedance) [47]. Pagani and Aiello make a thorough review of existing papers in which power grids are modeled as weighted or unweighted graphs [38]. In addition to weight, lines may be endowed with direction. Directed networks can be used to represent how power flows through the network in a specific scenario of demand and generation. A directed network is not always an accurate model since power can flow both ways. Therefore, in order to be general, power grids should be modeled as simple or non-directed networks.

Based on the previous considerations, power grids can be represented mathematically by an adjacency matrix, a matrix where non-zero elements reveal the existence of lines linking two nodes and their impedance if applicable. The adjacency matrix is symmetrical for non-directional graphs.

As in the case of lines, nodes can be endowed with weights. Node weight may help to reflect the importance of a node in the system, leading to more accurate results [48]. In the case of power systems, weight can represent the amount of energy that is injected or withdrawn. In addition to weights, colors can be used to differentiate between demand or generation nodes or to classify generation nodes based on generation technology or cost.

Power grids, as an example of transportation networks, are spatial networks (nodes are embedded in a geographical space). The location of nodes will directly affect the growth of the network and system dynamics. Several works have analyzed the implications of spatial embeddedness: node degree is limited by the physical space to be connected, the distance-dependent cost of lines limits the probability of linking two distant nodes, and there is no correlation between clustering coefficient and node degree (based on the power grid of the Western United States) [49]. Finally, Barthélemy states that power grids are planar (they can be drawn in a two-dimensional space in such a way that edges do not cross each other) [50]. However, if we consider that power-networks are embedded, we cannot state that they are planar graphs.

Finally, power grids are a clear example of interdependent systems. Their correct functioning depends on other networks, e.g., other power networks (higher or lower voltage networks) or other types of networks such as gas networks or communication networks [1], [51], [52]. In this case, every single network is represented as a layer of the whole system. If we model power grids as multilayer networks in which each voltage level is a different layer, the dependency between networks may be represented by a set of edges that connect different layers. These edges would represent the transformer impedance. Mathematically, each layer is represented by independent adjacency matrixes.

A **graph** $G(N, L)$ is a set of vertices N that are linked through edges L.

Regarding **edges**, graphs can be classified as:

- **Weighted / Unweighted**
- **Directed / Non-directed (or simple)**

Nodes may be endowed with some features:

- **Weight**
- **Color**

In **spatial** graphs **nodes** are **embedded** in the geographical space.

2.3. Global statistics and power grids

This section carries out a topological analysis of power grids by applying complex-network metrics to fifteen European transmission networks. The analyzed transmission networks are composed of two voltage levels, 400 kV, and 220 kV. This investigation analyzes both networks

as independent grids and as a single one. When modeling both voltage levels as a whole, the model does not omit transformers: transmission lines connected to the primary and secondary windings are connected to different nodes, as shown in Figure 2-1. Therefore, the number of nodes, in that case, is equal to the number of buses in the 400-kV network plus the number of buses in the 220-kV network. Multiple lines connecting two buses are modeled as one single edge to enhance the use of complex-network techniques to describe power-network topology. This model is unweighted and non-directed. Data are obtained from ENTSO-e [12]. These data provide information about how nodes are connected irrespective of node location. Accordingly, it disregards the spatial nature of power networks. One of the critical points of this work is the use of comparable data that allows us the comparison among countries. As previously mentioned, prior studies were based on heterogeneous data; voltage levels included in those studies were not always clear and varied depending on the work. This made comparisons untrustworthy and made it difficult to draw definite conclusions. This section analyzes how the metrics scale with network size, intending to generalize power-grid topological properties.

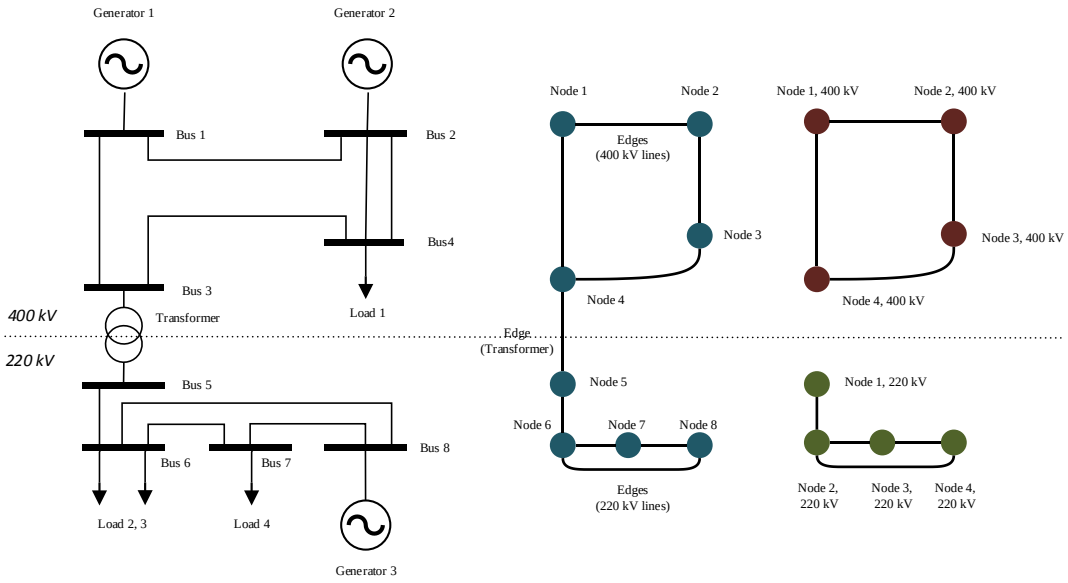


Figure 2-1. Graph models for power networks.
The 400 kV and 220 kV components as a single graph (center) or as independent layers (right)

2.3.1. Network size

Network size is the most basic metric when describing network structure. Network size is defined by the number of nodes, N, (number of substations) and the total number of edges, L, (transmission lines, considering only connections between two nodes regardless of the specific number of circuits). As shown in Table 2-1, the size of the European power grids that have been analyzed varies significantly among countries. This ranges from power grids with just 50 nodes as in the case of Hungary to 1,659 substations in France. Conspicuous factors that determine the number of nodes in each country were not found. Besides this, when considering different voltage levels, there is no correlation between network size and voltage level (400 kV or 220

kV).

Table 2-1. Network size of European transmission power networks.

Country	N	N_{400}	N_{220}	L	L_{400}	L_{220}
Hungary	50	28	22	80	38	22
Netherlands	55	35	20	63	40	18
Greece	57	57	0	80	80	0
Bulgaria	63	21	42	82	27	49
Serbia	84	35	49	107	37	61
Belgium	88	58	30	105	67	32
Austria	89	31	58	119	40	69
Romania	117	46	71	160	64	79
Switzerland	158	37	121	221	46	157
Portugal	159	57	102	237	79	148
Poland	163	59	104	247	82	138
Italy	634	262	372	812	321	437
Germany	782	480	302	1090	671	341
Spain	798	201	597	1115	284	731
France	1659	386	1273	2160	477	1479

N is the number of nodes, L is the number of transmission lines. The indices 400 and 220 show network voltage level, in case of no index, the network is the combination of 400 kV and 220 kV.

In the case of transmission lines, the number of lines in each country scales linearly with the number of nodes, as shown in Figure 2-2. This relation is valid in the three cases that have been analyzed, $L \propto 1.32N$, ($R^2 = 0.998$), $L_{400} \propto 1.33N_{400}$, ($R^2_{400} = 0.993$) and $L_{220} \propto 1.17N_{220}$, ($R^2_{220} = 0.999$). As a direct consequence of the linear correlation between the number of nodes and number of lines, network connectivity (the number of existing connections divided by all possible combinations of lines in the graph) inversely scales with the number of nodes (following a power law), $NCon. \propto 2.48N^{-0.985}$, ($R^2 = 0.996$).

- **Network size** is defined by number of nodes N and the number of edges L. In this work the number of nodes is equal to the number of substations.
- In the European transmission power networks, size highly varies among countries.
- The number of lines installed scales linearly with the number of nodes. The 400-kV network has a higher number of lines per node than the 220-kV network.

2.3.2. Degree distribution

Not only the total number of lines in a graph but the number of lines attached to each node also determines the dynamical behavior of networks. Node degree is defined as the number of lines connected to each node. The local nature of node degree makes it a non-manageable metric in large networks.

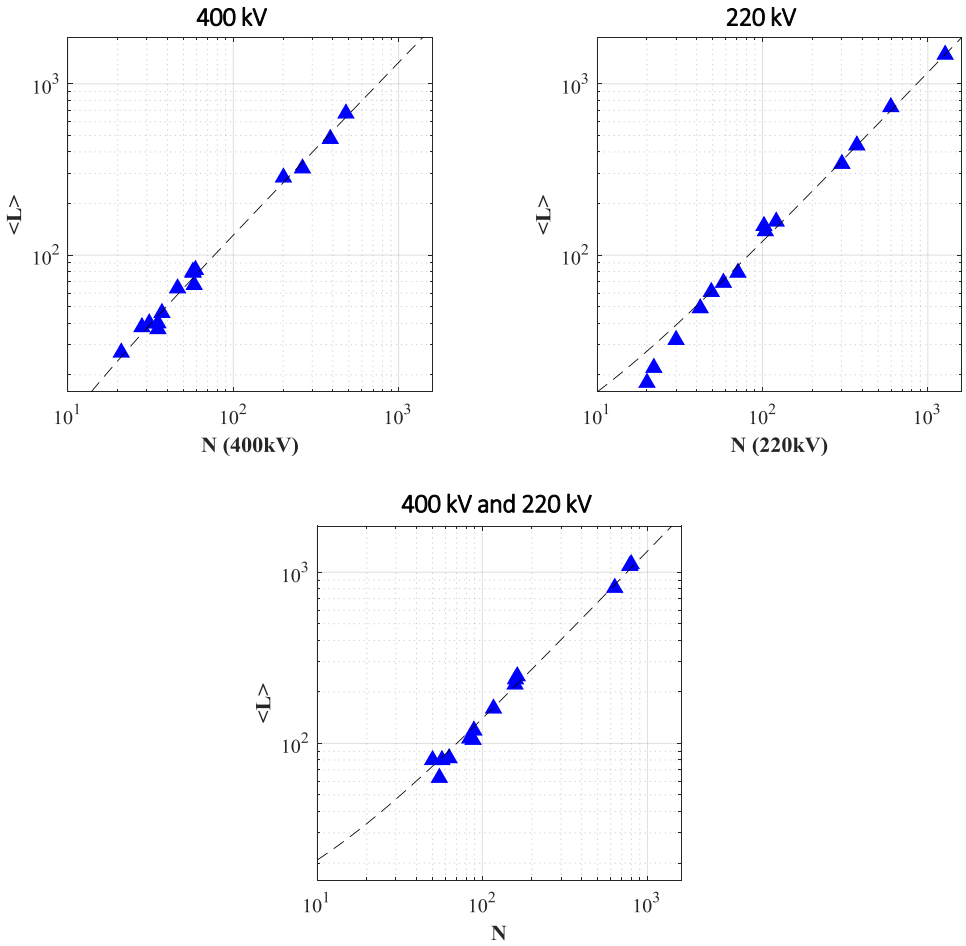


Figure 2-2. Relation between the number of nodes and the number of edges in the European transmission power networks.

Degree distribution (that is the probability of a node to have k lines attached to it) provides a better approach to explain network topology. Barabási and Albert firstly pointed out that the degree distribution of real complex networks often follows a power law ($\gamma \cdot x^{-\alpha}$, the variable x is the node degree), with a α of 4 in the case of power grids [37], several studies have discussed if the degree distribution in power grids follows a power law or an exponential function.

Networks where degree distribution follows a power law are also called scale-free networks. One of the main properties of scale-free networks is scale invariance. That is, the degree distribution is always characterized by the same α, irrespective of sample size. In terms of vulnerability, scale-free networks are robust against random failure and vulnerable when suffering deliberate attacks [37]. However, the degree distribution of several real networks, such as the worldwide transportation network or mail network, was found to follow an exponential function ($\alpha \cdot e^{\beta x}$) [53].

Exponential-degree distributions are characterized by having a faster decay to zero than power laws. Accordingly, the probability of having nodes with a high degree is slightly larger in scale-free networks. Recent studies show that the degree distribution of power networks is best approximated by an exponential function [38]. In particular, we observe that when considering a single grid (200 kV + 400 kV), the exponential fitting provides better results than the power law. If power grids are analyzed considering voltage layers independently, a power-law function fits best in some countries. As shown in Table 2-2, fitting empirical data to ($\alpha \cdot e^{\beta x}$) results in values of β, the exponential decay ratio, that range -0.66 and -0.30. That means that the smaller the β, the faster the decay and therefore the probability of finding nodes with high a degree is lower, as Figure 2-3 shows.

Finally, neither a power-law function nor an exponential function is a precise fit for the unweighted degree distribution of European transmission power networks from a graphical point of view (Figure 2-3). Representation of the degree distribution in a log-log plot would be a straight line in case of following a power law. Similarly, in the case of an exponential function, the degree distribution would be like a straight line in the linear-log plot. However, in the European transmission power networks, there are some divergences with those two patterns.

There is no mathematical relation between β and network size. Furthermore, the mode of the degree distribution also provides information when comparing different power-network topologies. The distribution mode varies between 1 or 2 (it is three just in the case of the 220 kV- Portuguese power network) (see Table 2-2).

Table 2-2. Degree properties of European transmission power networks.

Country	$\langle k \rangle$	$\langle k_{400} \rangle$	$\langle k_{220} \rangle$	$Mo(k)$	$Mo(k_{400})$	$Mo(k_{220})$	Assortativity coefficient	β
Hungary	2.22	2.45	1.91	1	2	1	-0.521	-0.56
Netherlands	2.10	2.00	1.80	2	2	1	-0.930	-0.61
Greece	2.81	2.81	-	2	2	0	-0.050	-0.66
Bulgaria	2.60	2.57	2.33	1	2	1	-0.367	-0.31
Serbia	2.55	2.11	2.49	1	1	1	-0.228	-0.49
Belgium	2.33	2.20	2.21	1	1	2	-0.240	-0.37
Austria	2.67	2.58	2.31	2	2	2	-0.208	-0.60
Romania	2.74	2.78	2.23	2	2	2	-0.144	-0.50
Switzerland	2.78	2.36	2.60	2	2	2	-0.016	-0.66
Portugal	2.98	2.77	2.90	3	1	3	-0.105	-0.30
Poland	2.99	2.78	2.60	2	2	2	-0.118	-0.49
Italy	2.53	2.4	2.35	1	1	1	-0.187	-0.37
Germany	2.58	2.57	2.12	2	2	2	-0.159	-0.62
Spain	2.79	2.83	2.45	2	2	2	-0.061	-0.64
France	2.59	2.42	2.32	1	1	1	-0.215	-0.44

$\langle k \rangle$ is average degree, $Mo(k)$ is the mode of degree distribution and β is the coefficient of the exponential adjustment of degree distribution ($\alpha \cdot e^{\beta x}$). The indices 400 and 220

show network voltage level, in case of no index, the network is the combination of 400 kV and 220 kV.

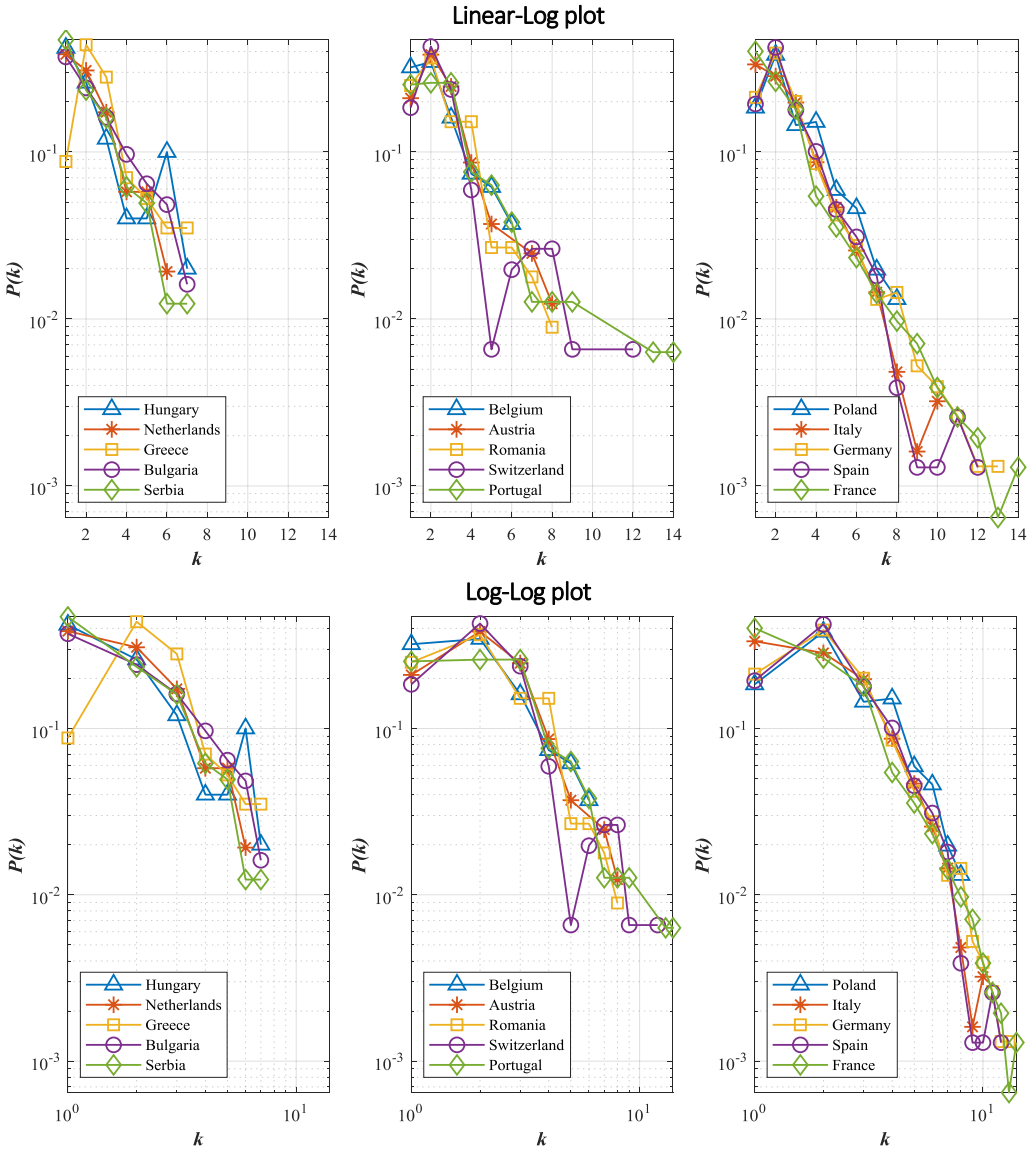


Figure 2-3. Unweighted degree distribution of European transmission power networks.

The average node degree, $\langle k \rangle$, is reasonably constant among European transmission networks since the number of lines per node scales linearly with network size. Accordingly, the average degree of the 400-kV network is higher than the 220-kV network. When analyzing both networks together the average node degree is higher since they have a higher number of connections (they include transformers) for the same set of nodes.

Finally, we analyze whether power networks are assortative or disassortative networks by calculating the network-assortativity coefficient r (it ranges from -1 to 1). The network-assortativity coefficient is the Pearson correlation coefficient of the degree at either end nodes of edges. This is calculated by equation (2-1), where j_i and k_i are the degree of the vertices at

the ends of the i-th edge [54]. Assortative networks (with a positive coefficient) are those in which nodes with high degree (also called hubs) tend to link to other hubs. Disassortative networks (with a negative coefficient) may present star-like features [36].

$$r = \frac{L^{-1}\sum_i j_i k_i - \left[L^{-1}\sum_i \frac{1}{2}(j_i + k_i)\right]^2}{L^{-1}\sum_i \frac{1}{2}(j_i^2 + k_i^2) - \left[L^{-1}\sum_i \frac{1}{2}(j_i + k_i)\right]^2} \quad (2\text{-}1)$$

European power networks tend to be disassortative: hubs tend to be connected to nodes with low degrees. However, assortativity coefficients are low, and no definite conclusions can be obtained about network topology. As in the case of β, the network-assortativity coefficient does not scale with the total number of nodes.

- The number of edges attached to each node is the **node degree**.
- Power-network degree distribution depends on network location. The average node degree is fairly constant in the European transmission networks.
- European transmission power networks are disassortative.

2.3.3. Shortest-path length

In addition to degree distribution, distances among nodes also condition the dynamic behavior of transportation and communication networks, since the shortest path among nodes provides an optimal path for transmitting system units between two nodes [32]. Characteristic path length (the average shortest path between any two nodes) and network diameter (maximum shortest path) characterize distances among nodes in a network.

In most analyzed power grids, the distribution of shortest-path lengths follows a quasi-normal distribution. However, in some cases, distances spread out to larger values, with positive skewness. This indicates that while some nodes are relatively well-connected (lower values of shortest path) there is a set of nodes that are far from the core of the network. This might be explained by the existence of highly meshed cores weakly connected among them. It might also represent the existence of a big hub, which is the center of peripheral nodes. Results show topological differences among countries. For instance, while the French network size is twice the Italian one, the Italian network has a larger diameter for a similar characteristic path length (as shown in Table 2-3). This might be explained by country geography: in the case of Italy, two main corridors connect the north and the south of the country, which has a relatively long and narrow shape.

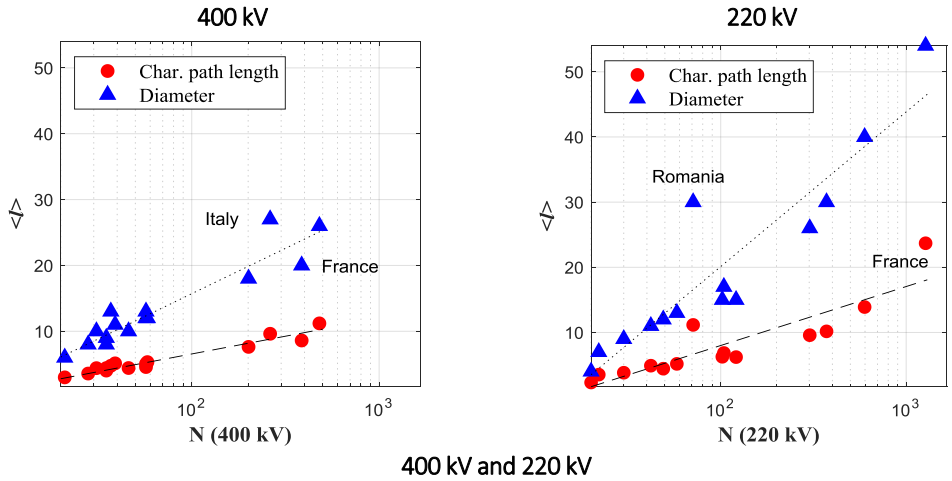
In terms of scalability, although geographical properties may condition the shortest-path distribution, both characteristic path length, and diameter scale logarithmically with the total number of nodes in all our studied cases, $\langle l \rangle \propto 2.48 \log n$, $(R^2 = 0.891)$; $\langle D \rangle \propto 6.24 \log n$, $(R^2 = 0.791)$; $\langle l_{400} \rangle \propto 2.35 \log n$, $(R^2 = 0.951)$; $\langle D_{400} \rangle \propto 5.98 \log n$, $(R^2 = 0.881)$;

$\langle l_{220} \rangle \propto 3.93 \log n$, ($R^2 = 0.788$); and $\langle D_{220} \rangle \propto 10.33 \log n$, ($R^2 = 0.850$); as shown in Figure 2-4. This is also the case for random networks, where both metrics also scale with $\log n$. However, in scale-free networks, characteristic path length and diameter scale with $\log \log n$ [55].

Table 2-3. Distance-based properties in European transmission power networks.

Country	$\langle l \rangle$	$\langle l_{400} \rangle$	$\langle l_{220} \rangle$	d	d_{400}	d_{220}	γ_1	$\gamma_{1,400}$	$\gamma_{1,220}$
Hungary	3.56	3.58	3.53	8	8	7	0.359	0.4	0.291
Netherlands	5.92	4.03	2.32	15	8	4	0.399	0.082	0.181
Greece	4.57	4.57	-	12	12	-	0.358	0.358	-
Bulgaria	4.65	3	4.9	10	6	11	0.001	0.071	0.238
Serbia	4.9	4.41	4.44	11	9	12	0.244	0.144	0.732
Belgium	7.04	5.32	3.82	20	12	9	0.765	0.348	0.328
Austria	5.92	4.4	5.18	14	10	13	0.245	0.319	0.434
Romania	5.82	4.42	11.16	11	10	30	-0.159	0.215	0.413
Switzerland	6.02	4.76	6.22	15	13	15	0.354	0.731	0.27
Portugal	6.10	5.05	6.27	13	13	15	0.074	0.557	0.201
Poland	6.24	5.13	6.85	15	11	17	0.172	0.132	0.248
Italy	11.98	9.62	10.17	32	27	30	0.437	0.463	0.369
Germany	12.19	11.2	9.58	29	26	26	-0.018	0.032	0.561
Spain	10.45	7.63	13.9	24	18	40	0.001	0.167	0.664
France	12.17	8.86	23.69	30	20	54	0.016	0.13	-0.007

$\langle l \rangle$ is the characteristic path length, d is network diameter, γ_1 is skewness of distance distribution. The indices 400 and 220 show network voltage level, in case of no index, the network is the combination of 400 kV and 220 kV.



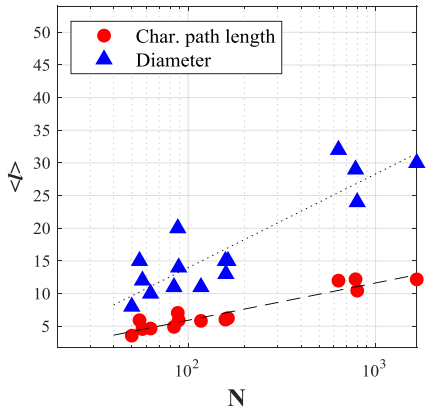


Figure 2-4. Characteristic path length and network diameter versus network size.

- The **characteristic path length** of a network is the average shortest path among all pairs of nodes. **Network diameter** is the maximum distance (shortest path) among all pairs of nodes in a network.
- In the European transmission power networks, characteristic path length and network diameter scale logarithmically with network size. Distances are slightly shorter in the 400-kV network.

2.3.4. Betweenness centrality

Betweenness centrality measures the centrality of a node in a network by counting the number of times a node or a line appears in the shortest path between other two nodes. In this chapter, betweenness centrality $B(u)$ refers to node betweenness centrality and it is defined by equation (2-2), where $n_{s,t}(u)$ is the number of shortest paths from s to t through node u and $N_{s,t}$ is the number of shortest paths from s to t. Since we are modeling power networks as undirected graphs, shortest paths from s to t and t to s count as one path.

$$B(u) = \frac{1}{2}\sum_{s,t\neq u}\frac{n_{s,t}(u)}{N_{s,t}} \tag{2-2}$$

Betweenness centrality may be used as a vulnerability metric in power networks. However, a clear relation between the betweenness centrality and dynamical behavior is not easy to infer since power does not follow the shortest path in terms of links -power flows are determined by Kirchhoff's laws-. Several works have modified the definition of betweenness centrality by the inclusion of electrical information (see Chapter 6).

Prior studies showed that betweenness-centrality distribution follows a power law in transmission networks [38]. Accordingly, most of the nodes are not in the shortest paths. In the case of the European networks, we observe that the percentage of nodes with a value of betweenness centrality that is equal to zero (they might be expected to have a negligible effect

on overall vulnerability) highly varies among countries and voltage levels. In the 400-kV network, it ranges from 13% in the case of Romania to 52% in France and the 220-kV network, 68% in Hungary and 23% in Poland. Usually, this percentage is higher in 220 kV networks (38% of nodes on average) compared to 400 kV networks (30% of nodes), except in the case of France, Portugal, and Serbia. This is related to how transmission power networks mesh. If the betweenness centrality of a node in a certain voltage level is zero, the node is connected just to one line in that voltage level. However, if we analyze both voltage layers together (400 kV and 220 kV), we observe that most of those nodes are connected to other lines. Therefore, the variation in the percentage of nodes with a betweenness centrality that is equal to zero shows differences in network structure; network mesh can be built in the same or at a lower voltage level.

The percentages above do not scale with network size. However, the mean and maximum value of betweenness centrality in power networks scale with the total number of nodes (see Table 2-4). As shown in Figure 2-5, mean betweenness centrality and maximum betweenness centrality may be characterized by a power law. When considering the 400 kV and 220 kV voltage levels together, the relationship is the following: $\langle BC \rangle \propto 0.28\ N^{1.43}$, ($R^2 = 0.965$) and $\max(BC) \propto 0.20\ N^{1.96}$, ($R^2 = 0.979$). In this network, the mean betweenness centrality may be also fitted with a linear regression $\langle BC \rangle \propto 5.73\ N$, ($R^2 = 0.995$). However, if we compare with the 400-kV layer and the 220-kV layer a linear regression is not an accurate fitting. The best regression for the 400 kV and 220 kV layers are the following: $\langle BC_{400} \rangle \propto 0.26\ N^{1.48}$, ($R^2 = 0.993$), $\max(BC_{400}) \propto 0.31\ N^{1.88}$, ($R^2 = 0.989$), $\langle BC_{220} \rangle \propto 0.21\ N^{1.46}$, ($R^2 = 0.887$) and $\max(BC_{220}) \propto 0.27\ N^{1.85}$, ($R^2 = 0.937$).

- The **betweenness centrality** of a node is the number of times that node is in the shortest path between all pairs of nodes in the network.
- The mean and maximum value of betweenness centrality scale with network size in the European transmission networks. They follow a power law.

Table 2-4. Betweenness centrality in European transmission power networks

Country	$\langle BC \rangle$	$\langle BC_{400} \rangle$	$\langle BC_{220} \rangle$	$max\ (BC)$	$max(BC_{400})$	$max\ (BC_{220})$
Hungary	3.12×10^{1}	3.48×10^{1}	2.66×10^{1}	1.81×10^{2}	1.81×10^{2}	1.37×10^{2}
Netherlands	1.33×10^{2}	4.08×10^{1}	4.80×10^{0}	5.58×10^{2}	1.71×10^{2}	1.90×10^{1}
Greece	1.00×10^{2}	1.00×10^{2}	-	7.64×10^{2}	7.64×10^{2}	-
Bulgaria	1.13×10^{2}	2.00×10^{1}	7.99×10^{1}	5.15×10^{2}	9.35×10^{1}	4.51×10^{2}
Serbia	1.62×10^{2}	5.79×10^{1}	8.26×10^{1}	1.33×10^{3}	3.40×10^{2}	7.31×10^{3}
Belgium	2.51×10^{2}	1.23×10^{2}	2.20×10^{1}	1.37×10^{3}	6.27×10^{2}	1.12×10^{2}
Austria	2.02×10^{2}	4.46×10^{1}	6.64×10^{1}	1.62×10^{3}	2.09×10^{2}	3.75×10^{2}
Romania	2.79×10^{2}	7.69×10^{1}	3.26×10^{2}	2.63×10^{3}	2.98×10^{2}	1.12×10^{3}
Switzerland	3.94×10^{2}	6.76×10^{1}	2.67×10^{2}	4.93×10^{3}	2.53×10^{2}	3.16×10^{3}
Portugal	4.03×10^{2}	1.13×10^{2}	2.66×10^{2}	3.30×10^{3}	7.07×10^{2}	2.55×10^{3}
Poland	4.24×10^{2}	1.20×10^{2}	2.87×10^{2}	4.55×10^{3}	7.96×10^{2}	1.39×10^{3}

Italy	3.48×10^{3}	1.12×10^{3}	6.57×10^{2}	5.00×10^{4}	8.81×10^{3}	1.33×10^{4}
Germany	4.37×10^{3}	2.44×10^{3}	3.26×10^{2}	1.05×10^{5}	4.25×10^{4}	4.70×10^{3}
Spain	3.76×10^{3}	6.63×10^{2}	1.55×10^{3}	1.16×10^{5}	7.14×10^{3}	2.19×10^{4}
France	9.23×10^{3}	1.51×10^{3}	1.35×10^{4}	3.32×10^{5}	1.99×10^{4}	2.49×10^{5}

$\langle BC \rangle$ is the mean value of betweenness centrality, $max(BC)$ is the maximum value of betweenness centrality. The indices 400 and 220 show network voltage level, in case of no index, the network is the combination of 400 kV and 220 kV.

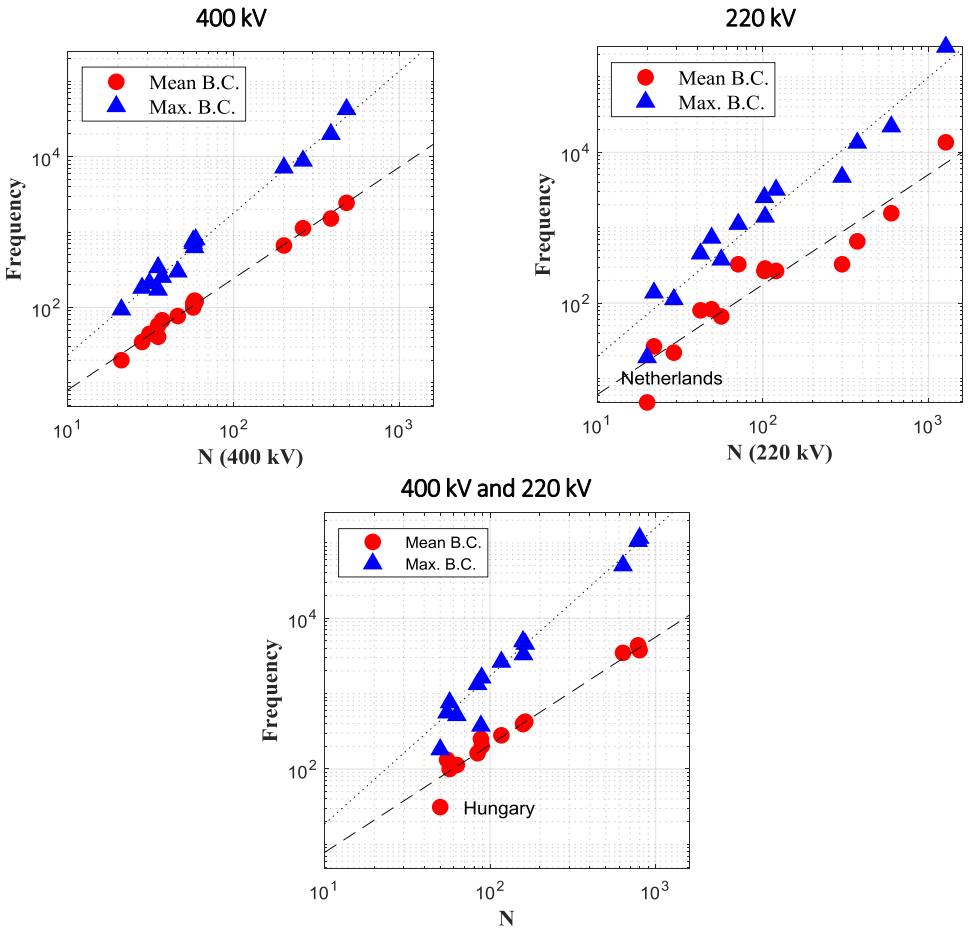


Figure 2-5. Maximum betweenness centrality and mean betweenness centrality versus network size

2.3.5. Network average clustering coefficient

In the previous sections, the analysis of distances and centrality might indicate the existence of highly clustered hubs. Similarly, the network average clustering coefficient may help to explain if there is a tendency to make clusters in power networks. The network average

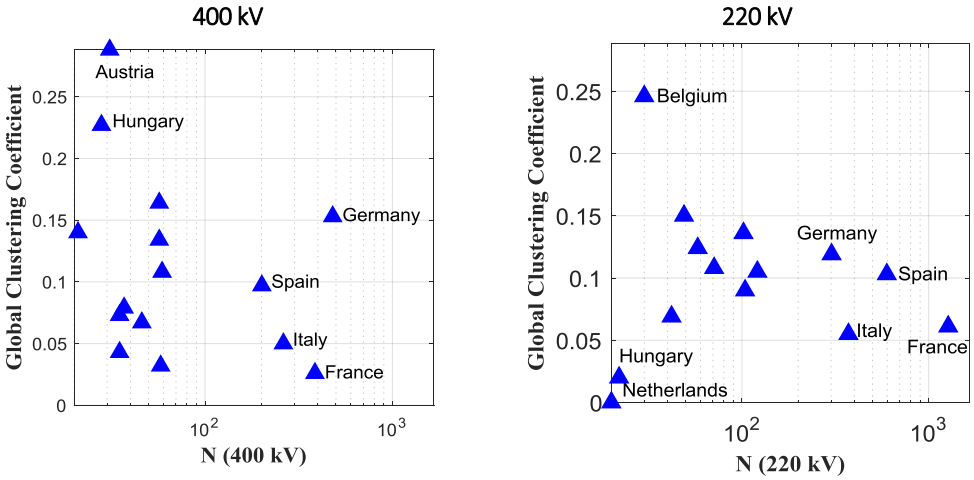
clustering coefficient $\langle cc \rangle$ shows the probability that the neighbors of one node are also connected among them. The network average clustering coefficient (it is the average value of node clustering coefficient) is calculated using expression (2-3), where T_i is the number of triangles in which node i is a vertex.

$$\langle cc \rangle = \frac{1}{N} \sum_{i=1}^{N} \frac{2\,T_i}{k_i(k_i - 1)} \tag{2-3}$$

This metric may help to understand why France and Germany have similar characteristic path lengths when the total number of nodes in France is twice larger than Germany. As shown in Figure 2-6, the network average clustering coefficient in Germany is three times larger than the one in France. That difference might explain the similarity in terms of distances: there is a higher tendency in Germany to form local clusters and therefore transmission lines reinforce those clusters in short distance rather than medium or long distances as in the case of France.

Overall, there is a low tendency to form clusters in power grids. Most nodes have no lines connecting their neighbors (the node clustering coefficient is zero) and, only in two countries, the percentage of nodes with all their neighbors connected is above 10%.

Finally, the network average clustering coefficient does not follow any relation with the total number of nodes, as shown in Figure 2-6. This might be a key indicator when comparing power grids. The network average clustering coefficient ranges between 0.05 and 0.15 when considering 400 kV and 220 kV layers together. By comparing node clustering coefficient and node degree, the node-clustering coefficient decreases with the degree, having a value of one only in nodes that are linked just to two or three neighbors.



400 kV and 220 kV

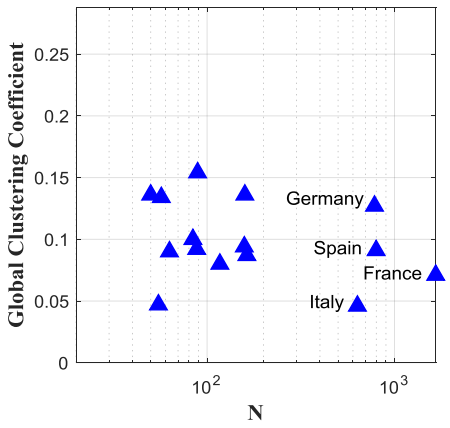


Figure 2-6. Network average clustering coefficient versus network size.

- The **network average clustering coefficient** shows the tendency to make clusters (triangles) in the network.
- In the European transmission power networks, the network average clustering coefficient highly varies among countries and voltage levels. Although it does not follow a pattern, the tendency to make clusters is low.

2.3.6. Is the power grid a small-world network?

Small-world networks are networks where most nodes are not neighbors among them, but they can be reached from other nodes by traversing only a few edges. In these networks, the characteristic path length grows logarithmically with network size $\langle l \rangle \propto \log n$. This property is similar to random graphs. However, small-world networks are characterized by having a higher network average clustering coefficient than random graphs. Based on previous considerations, Watts and Strogatz proposed an algorithm for generating random graphs with small-world properties [44].

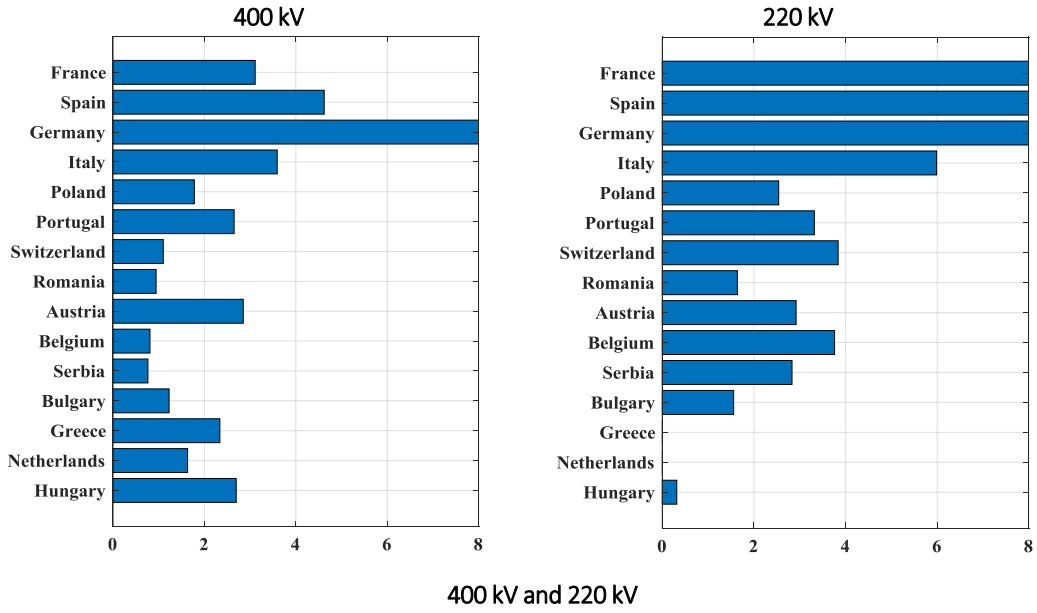
A network can be considered a small-world network if it has a similar characteristic path length than a random network and its network average clustering coefficient is much higher than the network average clustering coefficient of a random network, that is $\langle l \rangle \sim \langle l_{rand} \rangle$ and $\langle cc \rangle \gg \langle cc_{rand} \rangle$, where the network average clustering coefficient of a random network is defined by: $\langle cc_{rand} \rangle \sim \langle k \rangle / N$ and the characteristic path length of a random network by: $\langle l_{rand} \rangle \sim \ln(N) / \ln(\langle k \rangle)$. The small-world index S compares the previous ratios to determine if networks are small-world (2-4) [56]. If S is bigger than one, the network can be considered a small-world network. Values of S for the European Transmission Networks are shown in Table 2-5.

$$S = \frac{\frac{\langle l \rangle}{\langle l_{rand} \rangle}}{\frac{\langle cc \rangle}{\langle cc_{rand} \rangle}} \tag{2-4}$$

In prior studies, there was not a clear answer when analyzing whether power networks are small-world networks or not [38].

According to previous research, low and medium-voltage networks do not appear to be small-world networks [57]. However, most high-voltage networks are small-world networks. Results obtained in this work show that if we consider the 400 kV and 200 kV levels together, all networks have a small-world index bigger than one, as shown in Figure 2-7. However, when analyzing both layers independently, there are some cases in which S is under one, and therefore those networks cannot be considered small-world networks. Those cases are Belgium (400 kV), Romania (400 kV), Serbia (400 kV), Hungary (220 kV) and Netherlands (220 kV). In the case of Belgium, Serbia, Hungary, and the Netherlands, their global clustering coefficient is small in comparison to a random graph.

- **Small-world networks** are networks where most nodes are not neighbors among them but can they be reached from other nodes by traversing only a few edges.
- Not all European transmission power networks display the characteristic structure of small-world networks.



400 kV and 220 kV

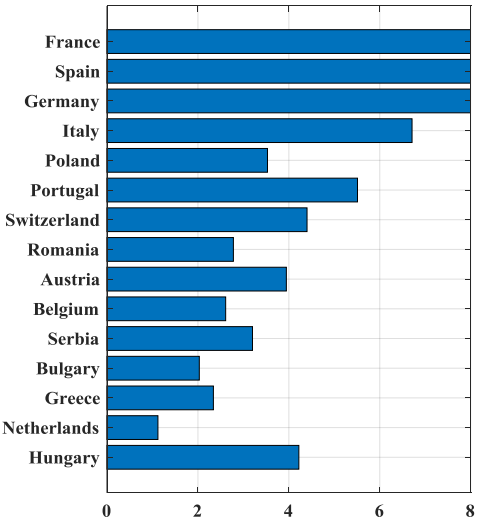


Figure 2-7. Small-world network index in the European transmission power networks

Table 2-5. Small-world properties of the European transmission power networks.

Country	$\langle l_{rand} \rangle$	$\langle l_{400,rand} \rangle$	$\langle l_{220,rand} \rangle$	$\langle cc \rangle$	$\langle cc_{400} \rangle$	$\langle cc_{220} \rangle$	$\langle cc_{rand} \rangle$	$\langle cc_{400,rand} \rangle$	$\langle cc_{220,rand} \rangle$	S	S_{400}	S_{220}
Hungary	4.90	3.72	4.77	0.13	0.22	0.02	0.04	0.08	0.08	4.2	2.7	0.3
Netherlands	5.40	5.13	5.1	0.04	0.07	0.00	0.03	0.05	0.09	1.1	1.6	0.0
Greece	3.92	3.92	-	0.13	0.13	-	0.04	0.04	-	2.3	2.3	-
Bulgaria	4.33	3.22	4.41	0.09	0.14	0.06	0.04	0.12	0.05	2.0	1.2	1.5
Serbia	4.74	4.75	4.27	0.1	0.04	0.15	0.03	0.06	0.05	3.2	0.7	2.8
Belgium	5.28	5.16	4.3	0.09	0.03	0.24	0.02	0.03	0.07	2.6	0.8	3.7
Austria	4.56	3.62	4.85	0.15	0.28	0.12	0.03	0.08	0.04	3.9	2.8	2.9
Romania	4.73	3.74	5.33	0.08	0.06	0.1	0.02	0.06	0.03	2.7	0.9	1.6
Switzerland	4.95	4.21	4.99	0.09	0.07	0.1	0.01	0.06	0.02	4.4	1.1	3.8
Portugal	4.64	3.97	4.34	0.13	0.16	0.13	0.01	0.04	0.02	5.5	2.6	3.3
Poland	4.65	3.99	4.85	0.08	0.1	0.09	0.01	0.04	0.02	3.5	1.7	2.5
Italy	6.94	6.35	6.97	0.04	0.05	0.05	0.00	0.00	0.00	6.7	3.5	6.0
Germany	7.02	6.54	7.61	0.12	0.15	0.11	0.00	0.00	0.00	22.1	16.6	13.4
Spain	6.5	5.11	7.14	0.09	0.09	0.10	0.00	0.01	0.00	16.1	4.6	12.9
France	7.79	6.75	8.48	0.07	0.02	0.06	0.00	0.00	0.00	29.2	3.1	11.9

$\langle l_{rand} \rangle$ is the characteristic path length of a random network with the same number of nodes, $\langle cc \rangle$ is the network average clustering coefficient, $\langle cc_{rand} \rangle$ is the global clustering coefficient of a random network with the same number of nodes, S is the small-world index. The indices 400 and 220 show network voltage level, in case of no index, the network is the combination of 400 kV and 220 kV.

2.4. Inferring the topology of power grids

2.4.1. Power networks, a matter of size

Several global statistics traditionally used in complex networks have been applied in Section 3 to fifteen transmission power networks to analyze network topology. Based on this analysis, we can differentiate two main groups of power networks when considering network size. The first group: Germany, Spain, France, and Italy with more than 600 nodes. The rest of the countries belong to a group in which networks have less than 160 substations. Those countries with more than 600 nodes are also the most extensive in terms of electricity consumption. However, the order in both lists is not the same. Besides, these four countries are also in the group of the biggest European countries in terms of area. However, the area of Poland is slightly larger than the area of Italy, and the number of substations in Italy is four times the number of substations in Poland; network size is more related to electrical consumption than geographical area.

Furthermore, when analyzing the number of nodes in each voltage level, we observe that the size of the 220-kV layer and the 400-kV layer do not scale with network size. For example, although Spanish and German power grids are similar in terms of network size, the percentage of nodes belonging to the 400-kV layer and 200 kV layer is the opposite. In ten out of the fifteen countries, the 220-kV layer is larger than the 400-kV layer. Therefore, in Europe, we differentiate two main groups of transmission power networks in terms of network size. However, several factors such as electrical consumption or country area lead to topological differences among countries in both groups.

As explained above, there is a linear correlation between the total number of lines and the total number of nodes in power networks. Results showed that the number of lines per node in 400-kV networks is around 13% larger than in 220-kV networks. That might be explained in terms of vulnerability, in power systems, the higher the voltage level, the higher the level of network reliability since the amount of energy that is transmitted in a power network grows with voltage level. Therefore, the 400-kV network has more lines per node; this means that network connectivity is higher in 400-kV networks, and they are more meshed than the 220-kV networks. Similarly, prior studies that analyzed medium and low voltage networks showed that the number of lines per node in those voltage levels is much lower, 1.09 and 1.03 respectively [57].

These results are in line with the ones obtained for other transmission-power networks such as the Iranian power network or North American power networks, where the number of lines per node is similar to the obtained for European countries [45], [39]. However, they differ from the results obtained for the South Korean power network. There, we observe three high-voltage levels: 765 kV, 345 kV and 154 kV in which the ratios between the number of links and the number of nodes are 3.81, 3.22 and 1.64 [46]. Those ratios are significantly larger than the previous cases. We should point out that differences among studies may be explained by how power networks are modeled (model assumptions), for instance, whether several lines

connecting the same pair of nodes are considered as one edge or not. Furthermore, the topography may also condition network topology as discussed below.

- Network size does not depend on country area.
- We differentiate two groups of transmission power networks in Europe based on network size: large networks (> 600 nodes) and small networks (< 150 nodes).
- The proportion of nodes in the 220 kV or in the 400 kV varies with location.
- The 400-kV network has a more meshed structure than the 220-kV network.

2.4.2. The meshed structure of transmission power networks

A. Average node degree

Since the number of lines can be approximated by the expressions given in Section 3.2.1 that linearly scales with total number of nodes, and the average degree is defined as the ratio between total number of lines and total number of nodes, the expected average degree in a power network can be approximated in 2.64 for transmission networks (400 kV and 220 kV) and 2.66 and 2.34 if considered the 400-kV and 220-kV layers independently. However, these results contrast with some test cases such as the PEGASE-89 in which the average degree is 4.72. The PEGASE-89 case study represents a fraction of the European high-voltage power network (400 kV, 220 kV, and 150 kV). Although other networks that are shown in reference [38] present differences concerning this study, those differences are not so significant as in the case of the South Korean power network that is mentioned above. As previously mentioned, those differences may be explained by how power networks are modeled as complex networks. On equal terms, the average degree distribution would be one of the first conditions to fit when validating a synthetic or test network since there is a clear correlation between the total number of nodes and the total number of lines.

B. Degree distribution

Although the number of lines correlates linearly with network size, how lines are distributed (degree distribution) varies among countries. Beyond the discussion, if degree distribution follows a power-law or an exponential function, we observe that the coefficients that characterize both functions vary by country. However, in all cases, we find a right-skew distribution where the maximum degree is lower than ten in most countries. Concerning the maximum degree, results obtained from this study are far from the maximum degree of the South Korean power grid, which is 18, or the U.S. Eastern Interconnect, U.S. Western Interconnect, and U.S. Texas Interconnect presented in which the maximum value of the node degree is 29 [39]. This might be explained by the inclusion of lower voltage layers.

Similarly, β (i.e., exponential decay ratio) provides information about network structure. While Spain and Germany have a similar β, France has a lower one. This means that the decay of the degree distribution is faster in the case of France, and therefore, the number of nodes with a high degree is lower. This might have consequences regarding vulnerability: networks with highly connected hubs might be more vulnerable to specific attacks to these hubs. Although lines are differently distributed, all power networks are disassortative, nodes with a high degree tend to be attached to nodes with low degree. Therefore, in power networks, most nodes have a low degree (one or two connections) and only a few nodes have a higher degree. This might be explained by the capital-intensive nature of power networks.

C. Network distances

In the case of network distances among nodes, the 220-kV layer has higher values of network diameter and characteristic path length. This difference is substantial in France, where characteristic path length and diameter are almost three times larger in the 220-kV network. As previously mentioned, 220-kV networks are generally more extensive in terms of network size and less meshed. As explained in Section 2.3, the location does not determine distance distribution, as can be seen for instance in the cases of France and Italy. Although Italy has a smaller geographical area and number of nodes, it has a higher diameter (220 kV and 400 kV) than France does. In this case, the skewness index describes the difference between both countries: the distance distribution in France is more normally distributed than in Italy. In Italy we have two main cores (North and South) connected mainly through two corridors.

Consequently, the analysis of network distances supports that the 220-kV network is less meshed than the 400-kV. Based on power-system considerations, the more critical nature of the 400-kV network leads to a more meshed structure. Similarly, 220-kV lines are built to connect lower distances than 400-kV lines. Therefore, this explains that the diameter of 220-kV networks is larger than the 400-kV network and that the 220-kV network is larger in terms of size.

Regarding distance analyses, results obtained in this work are consistent with the analyses of the Iranian and South Korean power networks. However, they are far from the results obtained in another study about North American power networks (Eastern, Western and Texas Interconnected power networks) where the values of network diameter are 94, 61, 37 respectively [39]. Once again, it is necessary to clearly define network models in order to distinguish structural topological differences from differences that are introduced because of different model assumptions (e.g., voltage levels, corridors vs. circuits, transformers included or not). This will support the justification of some questions (for instance, the impact of geography in network topology), that are ambivalently answered in prior works, as shown in

reference [38].

- Average node degree is similar in the European transmission power networks. However, we find inconsistencies with prior works.
- Degree distribution is conditioned by network location. Most nodes have a low number of connections. The presence of hubs is scarce, and they do not tend to be connected among them. They are disassortative networks.
- Network characteristic path length and network diameter do not depend on country area. Distances are larger in the 400-kV network.
- The analysis of the European transmission networks differs from prior works that analyzed the North American power network. Differences may lie on voltage levels and model assumptions.

2.4.3. About the small-world nature of power networks

Network average clustering coefficient highly varies among countries and voltage level. It does not scale with network size. However, our study provides a range of values where the network average clustering coefficient of power networks can be expected to lie. We observe that the values of the network average clustering coefficient are low, which shows that there is not a tendency to make clusters in power networks. We also observe that in relatively large countries such as France, Spain, and Germany, the network average clustering coefficient is 30 times larger than in the case of random networks with the same number of nodes. However, in small power networks, network average clustering coefficient values are similar to ones in random networks. This, therefore, conditions the categorization of power networks as small-world networks. France, Spain, and Germany are small-world networks (S>20). However, in the rest of the countries, the small-world index is significantly lower. For example, in the Netherlands, the network average clustering coefficient is quite similar to the theoretical network average clustering coefficient of a random network. This might reveal differing dynamics concerning network size: only as a power system grows does it make sense to build hubs or clusters that ensure the efficient exploitation of the system as a whole, which needs of shorter distances among any pairs of nodes overall. Smaller networks might have a more local structure, which much larger distances between nodes – which is less efficient but, on the flip side, can make them more resistant to attacks or failures. We can conclude that, in terms of clustering, power networks do not follow a similar pattern. However, this work provides a realistic range in which the network average clustering coefficient of synthetic power grids should lie.

Furthermore, the values of the network average clustering coefficient in power networks are more extensive than in random networks. Prior studies have pointed out to several different answers when questioning whether power networks are small-world networks or not. While an analysis based on North American power grids states that power grids are not small-

world networks [39], other previous works confirm that power networks belong to that group [46]. However, as was discussed by Pagani and Aiello, the answer to that question is case dependent and influenced by the voltage level considered [38]. Results provided in this work show that with a homogenous treatment of data, transmission power networks could be characterized as small-world networks only if 400-kV and 220-kV layers are considered together.

- There is no consensus in the literature about the characterization of the power grid as small-world networks.
- Based on the analysis carried out in this chapter, we cannot state that power grids are always small-world networks.

2.5. Topological consistency of synthetic power grids

The analysis presented in this chapter might be used in the generation of synthetic power grids from a double perspective. On the one hand, the topology of resulting synthetic power grids should be validated by the comparison with real topologies. This chapter proposed several metrics in order to check the accuracy of synthetic power grids regardless of network size. As shown previously, while some properties scale with network size there are other properties where it is not possible to estimate those parameters based on the number of substations, as in the case of network average clustering coefficient. However, this study provides a reasonable range for those metrics. Synthetic power grids should meet those objectives before being considered as case studies. This topological validation is something that was missing in most of the prior work in which several algorithms were described to generate synthetic power grids [22], [21], [58]. Networks obtained with those algorithms should be therefore tested with the topological metrics used in this chapter. Those metrics should be used beyond other considerations related to the minimum spanning tree or the Delaunay triangulation as done in reference [59]. By using just those last two considerations, we cannot provide an accurate and complete validation of network topology.

On the other hand, the conclusions obtained in this work show that network properties vary by country, so flexible algorithms are needed. Although some metrics scale with network size, we also observe a certain deviation level that depends on the country. Even when generating synthetic networks with the same number of substations, algorithms should be flexible to generate different topologies.

Global statistics can be used to validate synthetic power grids. However, we should be cautious since not all statistics scale with network size.

Models to generate synthetic power grids should be flexible enough to build network with different topologies.

2.6. The drawbacks of global statistics

The use of global statistic allows for the characterization of power-network structure. However, there are questions that those metrics cannot solve. For example, we cannot clearly explain why the French 220-kV and the German 220-kV networks have similar values of network diameter if the French network is 4.31 times larger in terms of nodes than the German one.

Furthermore, to compare network topologies, global statistics use average values (average node degree, characteristic path length, network average clustering coefficient, average betweenness centrality) or maximum values (network diameter, maximum betweenness centrality). Those values may guide to misleading results since they do not consider the shape of the distribution. Different distribution functions might have similar average values. This is discussed in Chapter 4.

Although most of the global statistics used in this chapter scale with network size, there is a certain deviation concerning the regression lines. Sound analysis is necessary to explain that deviation and to compare networks with different sizes. Although we have found some patterns in the European transmission power networks, there are inconsistencies regarding prior works. Furthermore, those patterns might change when including in the analysis of other power networks such as the North American transmission power network. Accordingly, we need a new method to compare networks regardless of network size.

Global statistics might be insufficient to describe power network topology and to compare network structure. The main drawbacks are:

- The use of average or maximum values might be misleading.
- Metrics might not scale with network size. That hinders the comparison among networks of different size.

2.7. Takeaways

This chapter introduces to the topological analysis of fifteen European transmission power networks. The two main voltage levels, 400-kV, and 220 kV are included independently and as a whole (which leads to a total of 45 networks). Our results show that network size (number of nodes) varies with countries and it is not determined by conspicuous factors. The number of lines scales linearly with the total number of nodes. Therefore, average degree distribution might be approximated as a constant in power networks.

Degree distribution varies across countries. However, all networks studied are disassortative (widely connected hubs tend to connect to poorly connected nodes). This means

that power grids tend to present star-like features.

In terms of distances, both characteristic path length and network diameter grow logarithmically with the number of nodes in all cases. The analysis of distances is completed with the skewness index, which shows whether distances are normally distributed or not. Most networks show a positive skew. This indicates that while some nodes are relatively well-connected (lower values of shortest path) there are set of nodes that are far from the core of the network, which might describe the presence of big hubs that are the center of peripheral nodes.

Regarding betweenness centrality, most nodes have low values (they do not tend to appear in shortest paths among nodes), which means the network is not vulnerable to losing them. Maximum and average values of betweenness centrality follow a power law with respect to the number of nodes.

Finally, the network average clustering coefficient highly varies across countries and voltage levels, and it presents larger values in power networks compared to random graphs. When considering transmission networks as 400-kV and 220-kV voltage levels together, all countries have a small-world index above one, and they can be therefore considered small-world networks. This means that the shortest path between nodes is relatively low when compared to random networks, which points to efficiency in their design. However, not all networks are small-world networks if voltage levels are considered independently. When analyzing both layers independently we observe that 400-kV networks have a higher average degree (they have more lines per node and distances are lower). This points out to a more meshed structure in 400-kV networks. Although both layers are considered transmission power networks, they display differences from a topological point of view.

3

AN INNOVATIVE TOOL TO DESCRIBE NETWORK TOPOLOGY

3.1. From global statistics to local descriptors

The analysis of complex-network topology can support the understanding of the principles that guide network evolution and condition network behavior [32]. Although most works have described network topology with global statistics, like the ones used in Chapter 2, local statistics have been also used to explain network structure [36], [60]. While global statistics, such as characteristic path length or betweenness centrality, considered the topology of the network as a whole, local properties only consider the connections of each node and its closest neighbors' connections. Some global statistics result from local descriptors. For example, the network average clustering coefficient is the average value of the node clustering coefficient, which measures the tendency of each node to make clusters.

Both global and local metrics complement each other, since different communities may coexist in the same network with different topological properties (what is known as structural subunits) [61]. Global metrics, such as the degree distribution, provide a panoramic view of networks that may have implications on their dynamics. For instance, the particular degree distribution of computing networks, they are scale-free networks, makes them relatively resistant to accidental failures but vulnerable to targeted attacks [62]. However, global metrics disregard the complexity of local structures that might be crucial to understand the behavior of networks, as it has been shown for the case of the internet [63]. Furthermore, local processes condition the development of network topology [64]. Consequently, topological analyses should include the use of local statistics that focus on the local structure of complex networks.

This chapter improves the characterization and understanding of network topology by proposing a twelve-dimensional metric, the GHuST framework, that is based on network local structures. Advantages of this novel framework are:

- Enhanced topological description: the twelve dimensions fully describe the structure of networks, covering most relevant aspects of local and global topology systematically.

- Size independence: the proposed framework explains network properties regardless of network size. This enables comparisons among networks with a different number of nodes and edges.
- Computational simplicity: this new statistic only considers 2-node and 3-node graphlets and they follow easily from the adjacency matrix. It reduces computational complexity with respect to prior analyses that require counting higher node graphlets.

The application of the novel metric to a set of five real networks demonstrates the accuracy of the framework to explain network topology. Furthermore, this new framework enhances network classification, and it can be used as a tool to confirm the topological accuracy of synthetic networks. This validation is usually missing in the generation of synthetic power grids, where there is a weak topological validation, or it is done only by a few global statistics [65]. Therefore, this tool can be introduced to compare the topology of both real and synthetic networks systematically.

The rest of the chapter is organized as follows: Section 2 introduces the use of local descriptors. Section 3 presents the GHuST, a novel framework to analyze network topology from graphlet decomposition. Section 4 applies the proposed framework to explain the topology of five real networks of different natures, and it compares results with other metrics traditionally used. Section 5 uses dimensionality reduction methods to evaluate the performance of the proposed framework when it is applied to a large sample of networks.

- **Global statistics** analyze network topology as a whole.
- **Local descriptors** only consider the connections of each node and its closest neighbors' connections.
- A novel framework, GHuST, is proposed to analyze the structure of complex networks. This framework is based on graphlet decomposition, a local descriptor.
- The main strengths of the framework are: full topological description, size independence and computational simplicity.

3.2. An introduction to local descriptors

An example of a local-topology statistic is the motif distribution. Motifs are recurring subgraph patterns that appear more often in a given network than in a random one (the base case against which the network under study is compared to is known as the null model). Motifs were proposed to understand the evolutionary design principles of complex networks from a local perspective [66]. They search for critical local structures that determine network behavior. However, the choice of the null model is often problematic [67]. Furthermore, motifs are partial subgraphs (they do not necessarily include all the connections between nodes), which leads to a loss of information that may be compelling to understand network structure [68].

Unlike motifs, graphlets allow for network decomposition in small subgraphs that preserve all connections among nodes. Graphlets are small connected induced (they preserve all edges among the subset of nodes) subgraphs of an extensive network [69]. The presence of graphlets in a network is not conditioned by a null model; they appear at any frequency. Although graphlets may be comprised of an arbitrarily large number of nodes, the most commonly studied graphlets are 2- to 5-node subgraphs, given that higher degrees are more difficult to calculate and interpret.

The automorphism orbit of a graphlet is defined as the set of nodes that are topologically symmetric in the graphlet [70]. Orbits, therefore, define the relative position of nodes concerning the rest of the nodes in the graphlet. A node can appear in more than one orbit in the network. When a node is in orbit n, it is said that node touches O_n.

In the case of G_4 (a four-node graphlet in which three nodes are connected to a central one) the node with three connections (green node) is in (touches) O_7. The three nodes (blue nodes) with only one connection are in (touch) O_6, as shown in Figure 3-1. Accordingly, the three nodes that are in O_6 have the same relative position in the graphlet (they have the same topological properties) and they are in the same orbit. However, they are topologically different from the central node (there is only one node in that orbit). Consequently, nodes that belong to G_4 can be in O_6 or in O_7. We can only differentiate two different orbits or positions inside that graphlet.

The main drawback of using graphlets to describe networks is that counting them is computationally intensive; recent works have proposed more efficient algorithms for graphlet counting though [71]–[75]. Figure 3-1 shows all 2- to 5-node graphlets and their automorphism orbits. The description of network topology is therefore limited by graphlet size. Although larger graphlets may complete the description of network topology, this would be unmanageable from a computational point of view.

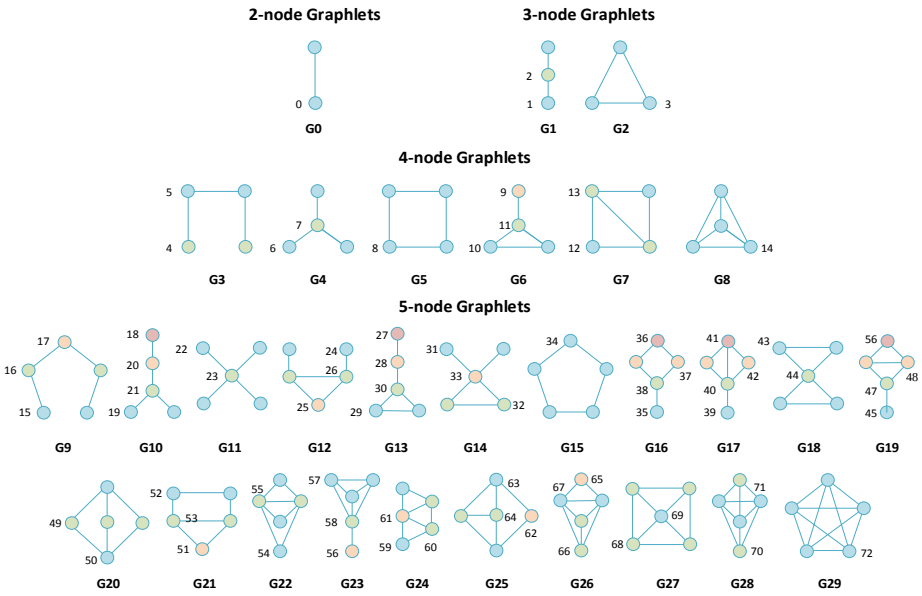


Figure 3-1. All 2- to 5-node graphlets and their automorphism orbits.

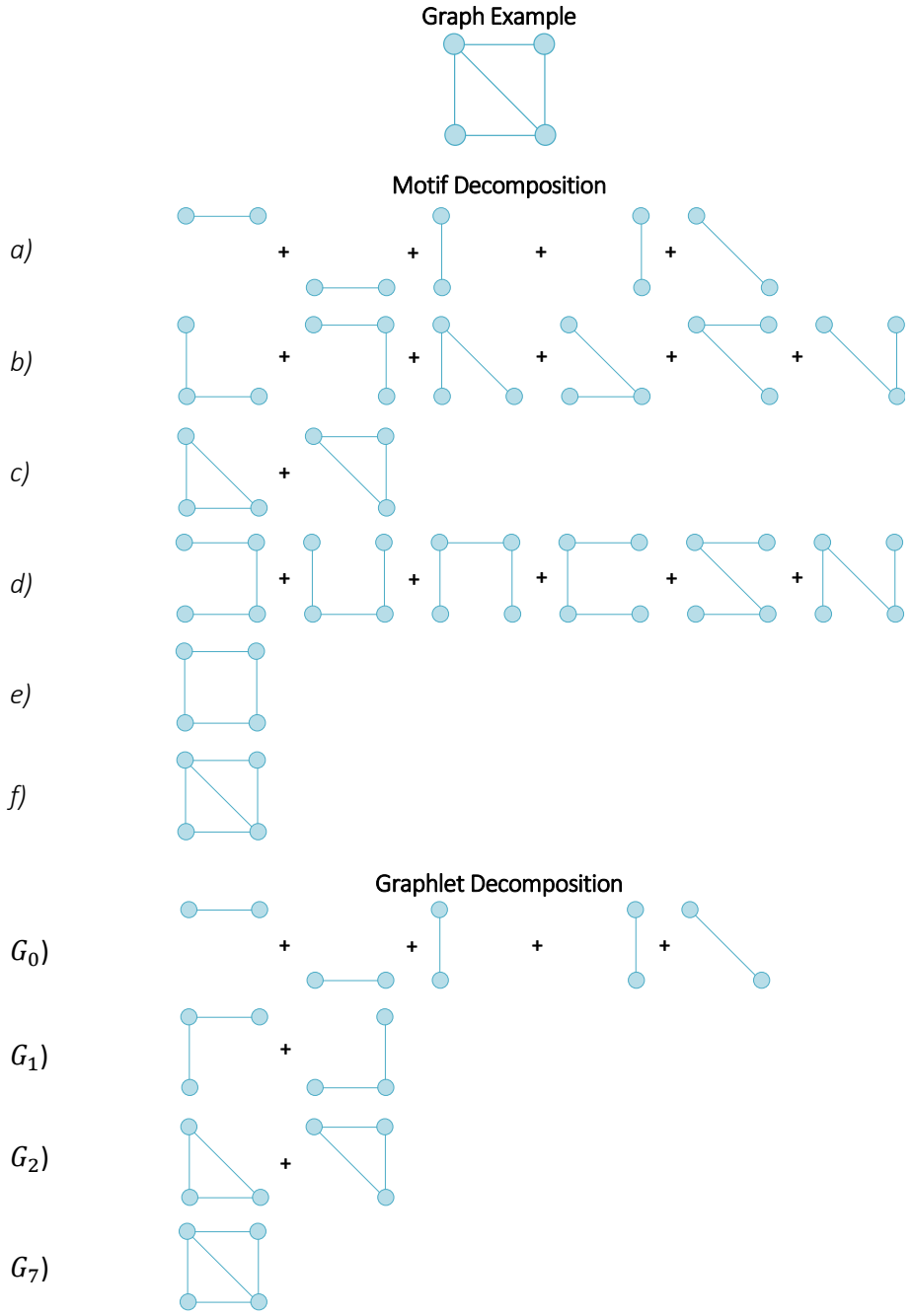


Figure 3-2. An example of the motif and graphlet decomposition.

Figure 3-2 shows an example of the motif and graphlet analyses. The graph used as an example is formed by 4 nodes and five edges. In the case of motif decomposition, the graph can be divided into six different subgraphs. Those subgraphs are not necessarily induced. Subgraphs *a*, *c*, and *f* hold all the connections among the subset of nodes taken (they are induced). However, subgraphs *b*, *d*, and *e* omit one of the existing connections among nodes in the original graph. Therefore, they do not preserve all the edges from the real network, and they are not induced subgraphs.

In the case of graphlets, they should preserve all edges connecting nodes. Consequently, we can only find four types of graphlets in the example.

In motif analysis, beyond the decomposition of the graph in smaller structures, we need to analyze whether those subgraphs are statistically significant or not. For example, to state that there is an overabundance of triangles, we have to compare the number of triangles concerning a null model. We may generate a random 4-node graph to compare its structure with the graph used as an example. There are several algorithms to generate random graphs (with different model assumptions) that generate different topologies. Accordingly, results will be conditioned by the algorithm used to generate the null model.

Graphlet analyses only define the frequency of each graphlet in the network. The frequency of each graphlet is not compared with a null model (a comparison with a null model would be needed to enhance the understanding of results, as discussed in Section 3.4.1). Accordingly, the graph used as an example results from the combination of the following graphlets: G_0 (5), G_1 (2), G_2 (2), G_7 (1) (as shown in Figure 3-2).

Furthermore, we know the automorphism orbits that each node touches. Nodes that have two connections touch O_0 (2), O_1 (2), O_3 (1), O_{12} (1) (they are part of G_0, G_1, and $G_{7)}$. Nodes with three connections touch O_0 (3), O_2 (1), O_3 (2), O_{13} (1). The frequency of each graphlet in a network results easily from the frequency of the automorphism orbits. For instance, the number of triangles in a network is equal to the frequency of O_3 in the network divided by three. In this example, the total frequency of O_3 is 6 (2 for nodes with three connections and 1 for nodes with two connections), and the number of triangles, G_2, is 2.

Graphlet decomposition considers all possible induced topologies for a subset of nodes. This is a strength with respect to motif decomposition, where the user should define the structure that should be found. In the prior example, users should define the subgraphs to be identified in the real network. However, in the case of graphlets decomposition, those subgraphs are already defined. Users only have to compare the real case with those predefined structures. The number of non-induced topologies highly increases with the subgraph size. In both motif analyses and graphlets analyses, the main drawback is the need for computationally intensive models.

Several models developed for the network alignment problem prove the adequacy of graphlets as a local topological descriptor [76]–[79]. The network alignment problem aims to find corresponding nodes between different networks. Nodes that play a similar role in both networks from a topological point of view. In this field, graphlet decomposition has been revealed as a crucial tool to solve the problem. The basis of those models is the degree signature of a graphlet [70]. The degree signature of a graphlet is an extension of the node degree that quantifies the number of times each node in the network appears (touches) in an orbit. Consequently, graphlets provide a complete description of local network topology (the orbits each node touches) that enhances the solution of the network alignment problem. Similarly, graphlets might support the comparison among networks or the study of the role played by nodes in the network [80], [81]. Despite being a good descriptor of local properties,

the use of graphlet distribution (or graphlet degree signature) is not enough to have an insight into the global topological properties of networks. Yaveroğlu et al. propose the analysis of orbit correlation (whether there are graphlets that tend to appear together) to characterize network structure and to ease the interpretation and implications of topological properties in real applications [82].

- **Motifs** are recurring subgraphs that are statistically relevant in a network. To determine the presence of motifs in a network is necessary a null model to compare with.
- **Graphlets** are small connected induced subgraphs of a large network. A network may be described by the frequency of graphlets (independently of a null model).
- Motif and graphlet decompositions need of computationally intensive models.

3.3. Understanding network structure from local properties

As explained above, graphlets can be a convenient tool for explaining the local structure of networks. Unfortunately, graphlet decomposition does not consider any interaction between graphlets. Besides, in large networks, counting graphlets is computationally intensive. It also supplies a substantial number of dimensions that are difficult to interpret (30 graphlets and 73 orbits in the case of using from 2- to 5-node graphlets). Motivated by this desire to simplify and improve topological analyses through graphlet decomposition, this section proposes a novel method that reduces the topological analysis of networks to a twelve-dimensional metric, the GHuST framework. This metric can be calculated in any non-directed and unweighted network.

3.3.1. The GHuST framework

The twelve dimensions are obtained from the decomposition of networks in 2-node and 3-node graphlets, comprising three graphlets (G_0, G_1 and G_2) and four orbits (O_0, O_1, O_2, O_3). The adjacency matrix succinctly reveals the number of times a node touches those orbits. In non-directed networks, the adjacency matrix is symmetric, and the sum of the elements in the ith row (or ith column) is, therefore, the degree of a node or $O_{0,i}$ (3-1).

$$O_{0,i} = \sum_j Adj_{i,j}, \quad \forall\, j \tag{3-1}$$

The number of times a node i touches O_1 is equal to the number of nodes j that are connected to node i by a two-edge path (through node k) (3-2). If a node j can be reached from node i through one or two edges simultaneously, nodes i and j are vertices of a triangle, and they touch O_3. Alternatively, the non-zero elements of Adj^2 show the number of two-edge paths that connect two nodes. However, this matrix does not consider if those nodes are vertices of a triangle or not.

$$O_{1,i} = \sum_j \sum_k \left(Adj_{i,j} Adj_{j,k}\right)\left(1 - Adj_{i,k}\right), \ \forall j,k \neq i \tag{3-2}$$

$O_{2,i}$ is the binomial coefficient $\binom{n}{2}$ where n is the number of edges attached to a node that is not connected among them attached to node i. $O_{2,i}$ can also be obtained from (3-3).

$$O_{2,i} = \sum_j \sum_k \left(Adj_{i,j} Adj_{i,k}\right)\left(1 - Adj_{j,k}\right), \ \forall j,k \neq i \tag{3-3}$$

As an extension of O_1, a node i touches O_3 when it is the vertex of a triangle (3-4). In this case, the number of times node i is a vertex of a triangle is also equal to $\frac{1}{2} Adj_{i,i}^3$.

$$O_{3,i} = \sum_j \sum_k (Adj_{i,j} Adj_{j,k})(Adj_{i,k}), \ \forall j,k \neq i \tag{3-4}$$

Besides, for the four orbits, $P_{t,i}$ is a binary variable that is 1 if node i is at least once in orbit t or 0 otherwise (3-5).

$$P_{t,i} = \begin{cases} 1, & O_{t,i} > 0 \\ 0, & O_{t,i} = 0 \end{cases} \tag{3-5}$$

To enhance readability, the twelve dimensions are classified into four categories: ***G****lobal connectivity*, ***Hu****bs*, ***S****trings*, and ***T****riangles*. Those categories cover different aspects of network structure that might condition network behavior. Furthermore, these categories allow for an intuitive interpretation of topology implications in real-world applications. For instance, in power networks, the higher presence of strings might mean a lower level of network robustness (higher probability of having energy not supplied in the network in case of line failure, given that when there is a failure in a string all the nodes that are downstream will be affected). Similarly, the presence of large strings in an email graph (nodes stands for community members and edges connect the people who send an email with the people who receive the email) will show that the community may follow a clearly defined hierarchical structure.

To enhance network comparison, it is desirable that the twelve dimensions of the metric range between 0 and 1. In cases where a dimension does not do it, we propose a scaling factor. The twelve dimensions are defined as follows:

A. Global connectivity

Line-surplus coefficient, $\boldsymbol{\rho_1}$. It stands for the surplus of lines in the network with respect to the minimum number of lines needed to build a connected graph (3-6). Given a set of nodes, N, the minimum number of lines, L_0, to have a connected graph is $L_0 = N - 1$, in case of large networks $L_0 \approx N$. As we only consider connected graphs, $N = \sum_i P_{0,i}$. The number of lines installed in a network is $\frac{\sum_i O_{0,i}}{2}$. This dimension is therefore related to the average node degree,

and it supplies information about line density in a network. In networks with a radial structure (trees), ρ_1 tends to zero. The higher the value of ρ_1 the more meshed a network is.

$$\rho_1 = \frac{1}{2}\frac{\sum_i O_{0,i}}{\sum_i P_{0,i}} - 1 \tag{3-6}$$

We define ρ'_1 (3-7) to scale ρ_1 between 0 and 1. Networks with ρ'_1 close to 1 have a highly meshed structure.

$$\rho'_1 = 1 - \frac{1}{\rho_1 + 1} \tag{3-7}$$

Leaf rate, $\boldsymbol{\rho_2}$. This ratio compares the proportion of nodes with just one connection, known as leaf nodes, to the rest of nodes in the network that are not vertices of a triangle. This ratio discerns between networks in which edges may form a homogenous mesh that touches most nodes and networks characterized by the presence of hubs connecting low-degree nodes. This metric is calculated as the complementary of the ratio between the number of nodes that touches O_1 but does not touch O_3 and the number of nodes that touches O_2 but does not touch O_3 (3-8).

All sets of three-connected nodes are either in graphlets G_1 or G_2. For those nodes that belong to G_2 and they are not part of G_3, they may touch O_1, O_2 or both simultaneously. A node is only in O_2 if it is the center of an isolated star, that is, the rest of the network nodes are connected to it. By assuming that networks have a more complex structure, no nodes can touch exclusively O_2. However, a node can touch exclusively O_1. This occurs in cases where nodes have only one connection, or they are the non-common vertex of two triangles that share one or two vertices. Accordingly, leaf nodes are defined by: $P_{1,i} = 1$, $P_{2,i} = 0$ and $P_{3,i} = 0$. Nodes that are not leaf nodes or vertices of a triangle are defined by: $P_{1,i} = 1$, $P_{2,i} = 1$ and $P_{3,i} = 0$. When ρ_2 is close to one, the presence of leaf nodes is high. The lower this coefficient, the lower the number of nodes that have just one connection; this is characteristic of star graphs.

$$\rho_2 = 1 - \frac{\sum_i P_{2,i}(1 - P_{3,i})}{\sum_i P_{1,i}(1 - P_{3,i})} \tag{3-8}$$

Leaf-base strength, $\boldsymbol{\rho_3}$. This ratio analyses if leaf nodes are connected to either hubs or low-degree nodes. This is the average number of times leaf nodes touch O_1 (3-9). The value of O_1 for leaf nodes is equal to the degree of its neighbor. Thus, the higher the value of O_1, the higher the degree of the node to which they are connected. Large values of ρ_3 may signal the presence of hubs in the network.

$$\rho_3 = \frac{\sum_i O_{1,i} P_{1,i}(1 - P_{2,i})(1 - P_{3,i})}{\sum_i P_{1,i}(1 - P_{2,i})(1 - P_{3,i})} \tag{3-9}$$

This dimension might be scaled with the maximum value of node degree, $\max(O_{0,i})$, in the network (3-10). If ρ_3 tends to zero, leaf nodes are connected to low-degree nodes. They may be the end nodes of node strings.

$$\rho'_3 = \frac{\rho_3}{max\ (O_{0,i})} \tag{3-10}$$

B. Hubs

Hub coefficient, $\boldsymbol{\rho_4}$. This dimension studies whether there is a tendency to form hubs in the network or not. It measures the average number of times nodes touch O_2 (3-11). All nodes touch O_2 except for leaf nodes and nodes that are only in G_2(they are only vertices of triangles). The larger the number of connections of a node, the larger the value of $O_{2,i}$. Large values of ρ_4 therefore shows there is a tendency to make hubs in the network. Unlike ρ_3, the hub coefficient does not linearly correlate with node degree; $O_{2,i}$ is given by the binomial coefficient $\binom{n}{2}$ where n is the number of non-connected edges attached to node i when the $O_{0,i}$ is greater than 2. If two networks have similar values of ρ_1, but different values of ρ_4, there is a higher tendency to make hubs in one network than in the other.

$$\rho_4 = \frac{\sum_i O_{2,i}}{\sum_i P_{2,i}} \tag{3-11}$$

To range between 0 and 1, ρ_4 can scale with the maximum value of $O_{2,i}$ in the network (3-12).

$$\rho'_4 = \frac{\rho_4}{\max\ (O_{2,i})} \tag{3-12}$$

Hub-connectivity coefficient, $\boldsymbol{\rho_5}$. It analyzes if hubs tend to connect among them. This dimension is defined by the Spearman's rank correlation between $O_1$and O_2, (3-13) where $cov(rg_{O_1}, rg_{O_2})$ is the covariance of the rank variables of $O_1$and O_2 and $\sigma_{rg_{O_1}}$, $\sigma_{rg_{O_2}}$ are the standard deviation of both rank variables. This is one of the correlations proposed by Yaveroğlu et al.[82]. If ρ_5 tends to one means that nodes with high O_2 are also nodes with high values of O_1. The number of times a node touches $O_{1,i}$ increases with the degree of a node and its neighbors' degree. However, the value of $O_{2,i}$ only depends on node degree; the higher the number of connections of a node, the higher the value of $O_{2,i}$. Consequently, nodes with a high value for O_1 and O_2 have a high node degree, they are hubs, and they are connected to other hubs. Therefore, a value close to 1 means that hubs tend to connect among them.

$$\rho_5 = \frac{cov(rg_{O_1}, rg_{O_2})}{\sigma_{rg_{O_1}} \sigma_{rg_{O_2}}} \tag{3-13}$$

This dimension is also scaled to range from 0 to 1 (3-14).

$$\rho'_5 = \frac{\rho_5}{2} + 0.5 \tag{3-14}$$

C. Strings

String coefficient, ρ_6. This coefficient measures the proportion of nodes in the network that are in the middle of a string. A string is formed by two end nodes (one or both nodes are linked to the rest of the network and there is no edge connecting them) and a set of intermediate nodes that are connected consecutively and have no links with the rest of the network. Consequently, a node is in the middle of a string if it has two connections, it touches $O_{2,i}$ only once ($U_{2,i} = 1$) and it is not a vertex of a triangle ($U_{3,i} = 1$). Therefore, ρ_6 is the ratio between the number of nodes that are in the middle of a node string and the total number of nodes that touch O_2 (3-15). Not all degree-two nodes touch O_2 once (triangle vertices do not touch O_2). In addition, not all nodes that touch O_2 once are in the middle of a node string. A node might touch O_2 only once if it is a shared vertex of a triangle ($O_{3,i} > 0$ and $U_{3,i} = 0$), so the node is not part of a string.

$$\rho_6 = \frac{\sum_i U_{2,i} U_{3,i}}{\sum_i P_{2,i}} \tag{3-15}$$

$$U_{2,i} = \begin{cases} 1, & O_{2,i} = 1 \\ 0, & O_{2,i} \neq 1 \end{cases} \tag{3-16}$$

$$U_{3,i} = \begin{cases} 1, & O_{3,i} = 0 \\ 0, & O_{3,i} \neq 0 \end{cases} \tag{3-17}$$

Characteristic string length, ρ_7. This dimension is the average length of node strings (considering only middle nodes and disregarding the end nodes of the string) in the network as shown in (3-18), where n is the number of node strings in the network. If ρ_7 is equal to one, it means that all node strings have two end nodes and only one middle node.

$$\rho_7 = \frac{\sum_i U_{2,i} U_{3,i}}{n} \tag{3-18}$$

To enhance network comparison, ρ_7 is scaled as its inverse (3-19)

$$\rho'_7 = \frac{n}{\sum_i U_{2,i} U_{3,i}} \tag{3-19}$$

D. Triangles

Triangle rate, ρ_8. This coefficient studies whether there is a tendency to make triangles in the network or not. It measures the proportion of triangles (G_2) in a network with respect to the total three-node graphlets (3-20). The number of G_2 in the network is equal to $\frac{\sum_i O_{3,i}}{3}$ and the number of G_1 is equal to $\sum_i O_{2,i}$. This ratio is similar to the global clustering coefficient. However, many works in the literature use the network average clustering coefficient to analyze network properties. The network average clustering coefficient weights more nodes with a low degree (as discussed in Section 3.4.2). Thus, it is not a correct measure to analyze

network with a non-homogenous degree distribution. The average network clustering coefficient, therefore, differs from the value of ρ_8 which considers the whole topology of the network.

$$\rho_8 = \frac{\sum_i O_{3,i}}{3\sum_i O_{2,i} + \sum_i O_{3,i}} \tag{3-20}$$

Triangle concentration, $\boldsymbol{\rho_9}$. This coefficient shows if triangles tend to be concentrated in networks. Triangles are concentrated when there are nodes that are vertices of two or more triangles. The dimension ρ_9 is complementary to the ratio between the number of nodes that are vertices of triangles and the number of triangles in the network (3-21). The higher the number of triangles that share some vertices the lower the value of ρ_9. If triangles have no shared vertices, the maximum value of $O_{3,i}$ is 1, and $O_{3,i} = P_{3,i}$. Therefore, the number of nodes that are in a triangle is three times the number of G_2 in the network ($3\,G_2 = \sum_i O_{3,i} = \sum_i P_{3,i}$). However, if triangles share vertices, $\sum_i P_{3,i} < 3\,G_2$. As ρ_9 converges to 0, the number of graphlets of type G_7, G_8, G_{17}, G_{19}, G_{22}, G_{23}, G_{24}, G_{25}, G_{26}, G_{27}, G_{28} and G_{29} (graphlets composed of triangles with shared vertices) converges to 0 too.

$$\rho_9 = 1 - \frac{\sum_i P_{3,i}}{\sum_i O_{3,i}} \tag{3-21}$$

Triangle pervasiveness, $\boldsymbol{\rho_{10}}$. This dimension analyzes if triangles tend to cover the whole network or if they are concentrated around a few nodes. It measures the proportion of nodes in the network that are vertices of triangles (3-22). If a node is a vertex of a triangle, $P_{3,i} = 1$. As explained, in connected graphs, the number of nodes in a network is $\sum_i P_{0,i}$. This coefficient compliments ρ_8 and ρ_9, since it sheds light whether triangles form a mesh that comprises most nodes in a network or not. A high value of ρ_8 might be a consequence of networks in which triangles are connected to hubs and low-degree nodes have a non-meshed structure or networks in which all nodes are connected by a triangle mesh. Therefore, ρ_{10} allows for the discernment between those types of networks, this coefficient would have a low value in the first case, and it would be close to one in the second network.

$$\rho_{10} = \frac{\sum_i P_{3,i}}{\sum_i P_{0,i}} \tag{3-22}$$

Triangle connectivity, $\boldsymbol{\rho_{11}}$. It measures if triangles are isolated in the network or they are part of a highly meshed structure. A triangle is isolated if one or two of its vertices are not connected to the rest of the network. Consequently, those vertices have only two connections, they touch $O_{1,i}$ and $O_{3,i}$ and they do not touch $O_{2,i}$. Thus, ρ_{11} is the ratio between the number of triangle vertices that are not connected to other nodes ($U_{2,i}$=1) and the total number of nodes that are vertices of triangles ($\sum_i P_{3,i}$) (3-23). The lower the value of ρ_{11}, the lower the number of isolated triangles in the network.

$$\rho_{11} = \frac{\sum_i P_{3,i}\, U_{2,i}}{\sum_i P_{3,i}} \tag{3-23}$$

Triangle degree, $\boldsymbol{\rho_{12}}$. This dimension shows if triangles tend to be connected to hubs or to low-degree nodes. It is the average degree of triangle vertices (3-24). That is the mean value of $O_{0,i}$ for those nodes that are in a triangle ($P_{3,i} = 1$). High values of ρ_{12} mean that triangles are connected to hubs. The lower the value of ρ_{12}, the lower the average node degree of triangle vertices.

$$\rho_{12} = \frac{\sum_i O_{0,i}\, P_{3,i}}{\sum_i P_{3,i}} \tag{3-24}$$

To range between 0 and 1 ρ_{12} is scaled with the maximum value of node degree (3-25).

$$\rho'_{12} = \frac{\rho_{12}}{\max\,(O_{0,i})} \tag{3-25}$$

Table 3-1 summarizes the definition and interpretation of the twelve dimensions of the GHuST framework. All dimensions, except ρ_1, ρ_5 and ρ_8, are new indices proposed in this thesis. As explained, ρ_1 is related to the average node degree, ρ_5 was proposed by Yaveroğlu et al.[82], and ρ_8 is the global clustering coefficient.

The new metric, **the GHuST framework**, is defined by twelve dimensions that cover four aspects of network topology: **G**lobal connectivity, **Hu**bs, **S**trings and **T**riangles.

To enhance network comparison, we propose a set of scale factors. Accordingly, all dimensions range between 0 and 1.

3.4. Explaining the topology of real networks

This section applies the twelve-dimensional metric to a set of five real networks to prove the usefulness of the proposed framework. It aims to prove if the information provided by the GHuST framework is consistent with the global-topology statistics usually used to describe network structure.

The set of five networks includes: two infrastructure networks the Minnesota road network and a power grid that represents the Western States Power Grid of the United States [44], [83], two social networks: an extract of Facebook and the email interchanges among members of a Spanish university [83], [84], and a network that represents the metabolic reaction of the E.coli bacteria [85]. For this analysis, all networks are modeled as unweighted and undirected graphs.

Table 3-1. Name, definition, and values of GHuST dimensions.

Name:	Definition:	Values:
Line-surplus coefficient	$\rho_1 = 1 - \frac{2\ \sum_i P_{0,i}}{\sum_i O_{0,i}}$	$\rho_1 \to 1$: highly meshed structure $\rho_1 \to 0$: no meshed structure
Leaf rate	$\rho_2 = 1 - \frac{\sum_i P_{2,i}(1 - P_{3,i})}{\sum_i P_{1,i}\,(1 - P_{3,i})}$	$\rho_2 \to 1$: large presence of leaf nodes $\rho_2 \to 0$: low presence of leaf nodes
Leaf-base strength	$\rho_3 = \frac{\sum_i O_{1,i}\ P_{1,i}\ (1 - P_{2,i})(1 - P_{3,i})}{\sum_i P_{1,i}\ (1 - P_{2,i})\ (1 - P_{3,i})} \frac{1}{\max(O_0)}$	$\rho_3 \to 1$: leaf nodes connected to high-degree nodes $\rho_3 \to 0$: leaf nodes connected to low-degree nodes
Hub coefficient	$\rho_4 = \frac{\sum_i O_{2,i}}{\sum_i P_{2,i}} \frac{1}{\max(O_2)}$	$\rho_4 \to 1$: presence of hub nodes $\rho_4 \to 0$: no presence of hub nodes
Hub-connectivity	$\rho_5 = \frac{1}{2} \frac{cov(rg_{O_1}, rg_{O_2})}{\sigma_{rg_{O_1}} \sigma_{rg_{O_2}}} + \frac{1}{2}$	$\rho_5 \to 1$: hubs tend to connect to other hubs $\rho_5 \to 0$: hubs do not tend to connect to other hubs
String coefficient	$\rho_6 = \frac{\sum_i U_{2,i} U_{3,i}}{\sum_i P_{2,i}}$	$\rho_6 \to 1$: high presence of strings $\rho_6 \to 0$: low presence of strings
Characteristic string length	$\rho_7 = 1 - \frac{n}{\sum_i U_{2,i} U_{3,i}}$	$\rho_7 \to 1$: long strings $\rho_7 \to 0$: short strings
Triangle rate	$\rho_8 = \frac{\sum_i O_{3,i}}{3\ \sum_i O_{2,i} + \sum_i O_{3,i}}$	$\rho_8 \to 1$: high presence of triangles $\rho_8 \to 0$: low presence of triangles
Triangle concentration	$\rho_9 = 1 - \frac{\sum_i P_{3,i}}{\sum_i O_{3,i}}$	$\rho_9 \to 1$: triangles tend to share vertices $\rho_9 \to 0$: triangles do not tend to share vertices
Triangle pervasiveness	$\rho_{10} = \frac{\sum_i P_{3,i}}{\sum_i P_{0,i}}$	$\rho_{10} \to 1$: most nodes are part of a triangle $\rho_{10} \to 0$: most nodes are not part of a triangle
Triangle connectivity	$\rho_{11} = \frac{\sum_i P_{3,i}\ U_{2,i}}{\sum_i P_{3,i}}$	$\rho_{11} \to 1$: triangle vertices tend to be unconnected to the rest of network nodes $\rho_{11} \to 0$: triangle vertices tend to be connected to the network
Triangle degree	$\rho_{12} = \frac{\sum_i O_{0,i}\ P_{3,i}}{\sum_i P_{3,i}} \frac{1}{\max(O_0)}$	$\rho_{12} \to 1$: triangle vertices are high-degree nodes $\rho_{12} \to 0$: triangle vertices are low-degree nodes

These five networks have different sizes and display completely different structures, as shown in Figure 3-3. The two social networks and the metabolic network are in the range of 1,000 to 1,500 nodes, and the two infrastructure networks are two and five times larger, respectively. However, the number of edges is much higher in the social networks; in the case of the Facebook network, the number of edges is twenty times larger than in the road networks. Differences in network size obscure the comparison among networks with global statistics. In some cases, as in distance-based metrics, it is not always possible to infer if there is a change in a variable because of network size or network structure.

3.4.1. Graphlets a matter of interaction

Scalability is one of the problems when using graphlets to describe network topology. The number of graphlets that a node touches depends on its degree and its neighbors' degree, but it also depends on the entire structure of the network. Subsequently, two networks with the same size (same number of nodes and edges) may have a different number of total graphlets. To make a comparison among networks regarding graphlets, we scale the frequency of a graphlet concerning the frequency of all same-size graphlets, as shown in Figure 3-4.

Considering 3-node graphlets, the distribution shows that the percentage of triangles G_2 looks extremely low in the five networks; in the metabolic and road network, the percentage of triangles is under 1.5%. Only in the network representing Facebook friendships does it reach 10%. However, a null model is necessary to compare results. Unless a network is formed exclusively by triangles, that is a network in which all nodes are connected among them, the frequency of G_1 in the network is not zero. Therefore, the value of G_2 has an upper bound.

About 4- and 5-node graphlets, the frequency distribution shows that a few frequencies prevail over the rest. In the case of 4-node graphlets, G_5 to G_8 account for 9% of the power-grid and 2.5% of the road-network distribution. Similarly, in the metabolic network and in the road networks, G_{12} to G_{29} account for less than 7.4% of 5-node graphlets. The presence of more connected graphlets (G_{12} to G_{29}) is only relevant to Facebook, where they represent 38% of 5-node graphlets.

Two graphlets dominate the metabolic network: G_4 (92% of 4-node graphlets) and G_{11} (87% of 5-node graphlets). That distribution of frequencies contrasts with the other networks in which predominant frequencies are G_3 (4-node graphlets) and G_9 and G_{10} (5-node graphlets). The number of times a node is in G_4 and G_{11} is the binomial coefficient $\binom{n}{k}$ where n is the number of non-connected edges attached to node i and k is three or four respectively. Therefore, those frequencies rapidly increase with the presence of hubs. The largest value of node degree in the metabolic network is 638 and the average node degree is 9.13, indicating a network with a few hubs connected to low-degree nodes. The predominance of those frequencies makes it impossible to infer a sound description of the metabolic network topology based on graphlet distribution will be limited to relatively low degrees.

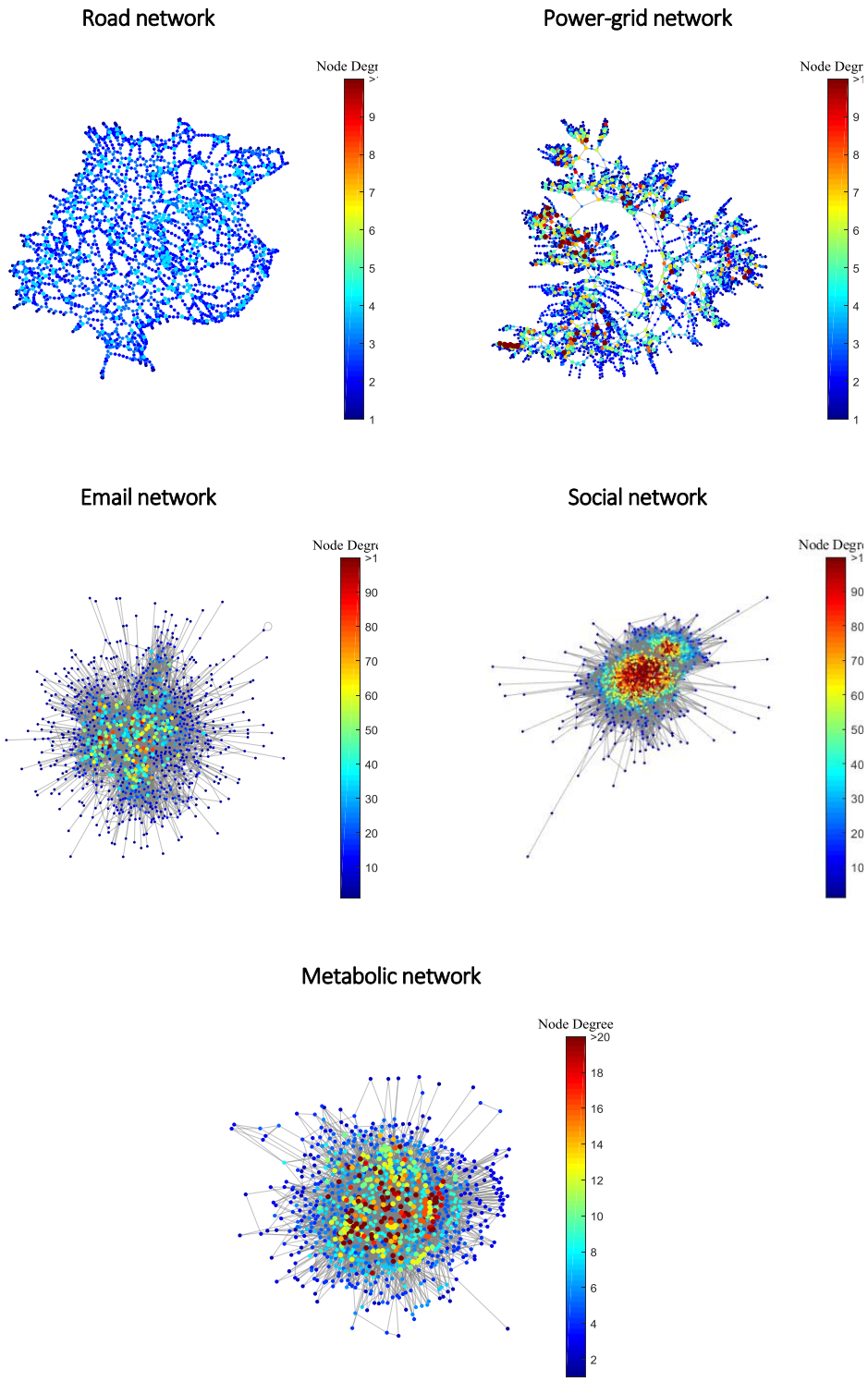


Figure 3-3. Graph representation of five real networks.

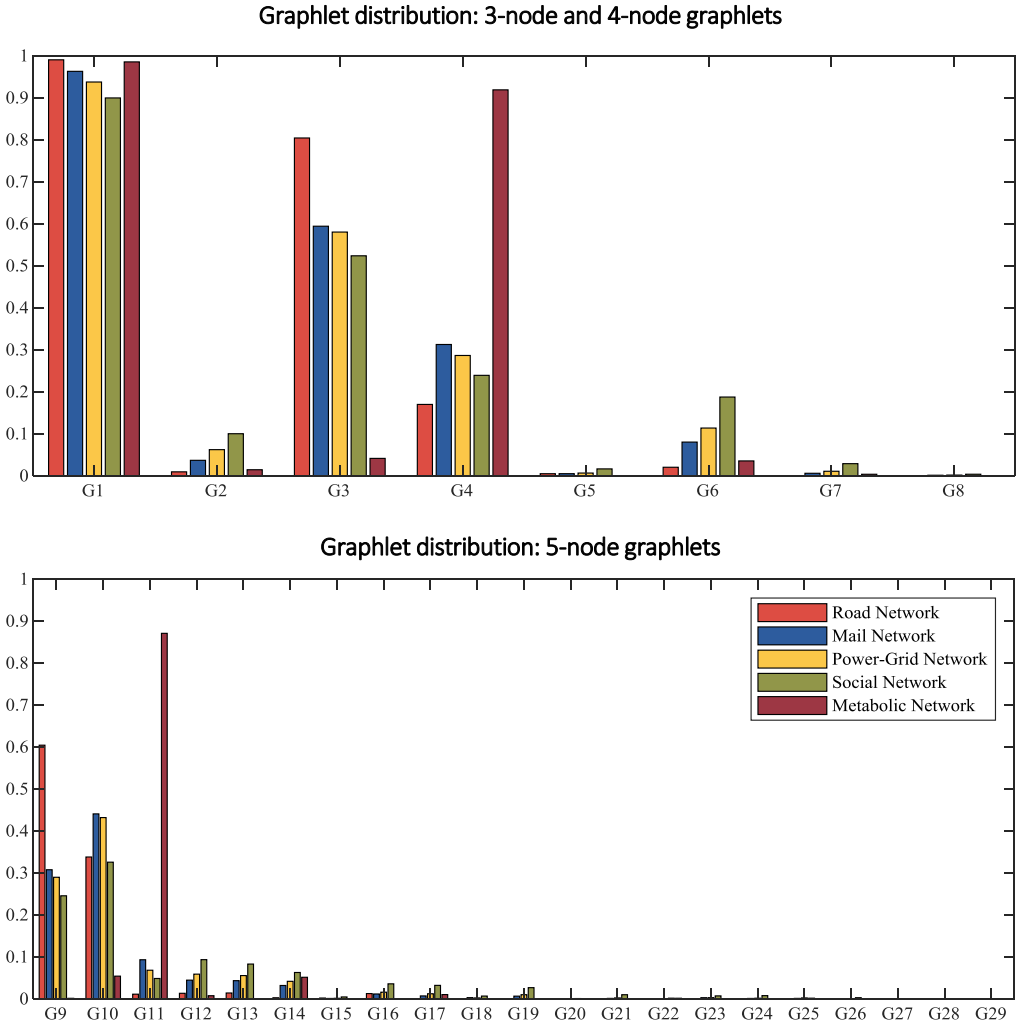


Figure 3-4. 3- to 5-node graphlet distribution of five real networks.

In infrastructure networks, connections are cost-intensive and highly connected subgraphs are not as frequent as in social networks. If we compare the two infrastructure networks, we see that there is a lower tendency to make triangles in the road network (3-node graphlets). However, the number of triangles in the power grid is not necessarily larger than in the road network, since the total number of graphlets depends on network structure. In the power grid, the value of G_4 is twice larger than the road network. There are nodes with a higher degree than in the road network. Indeed, the global statistics show that the maximum degree is four times higher in the power-grid case. When analyzing 5-node graphlets, G_9 and G_{10} explain 95% and 75% of power-grid and road-network graphlet distribution, respectively. As in the prior case, the main conclusion is that the average node degree is higher in the power grid and local structures tend to be more connected than in the road networks (since highly connected graphlets have a slightly higher frequency). However, this information is not enough to characterize network topology accurately, and it might be misleading.

In the case of Facebook, it is not possible to infer if the values of G_9 and G_{10} are because of the presence of hubs or not. This also requires a more in-depth analysis with a correct null model. When comparing the two social networks, the email network looks to have a less connected structure, since the frequency of highly connected graphlets in the email network is much lower than in the Facebook network. If we compare the mean absolute error ($MAE = \sum_i \frac{|G_{i,a} - G_{i,b}|}{n}$), the most similar networks in terms of graphlets frequencies are the power grid and the email network (MAE = 0.008). The MAE between Facebook and email network is 0.021. However, when analyzing the global topological statistics (see Table 3-2), we see that the power grid and the email networks display entirely different structures.

Based on prior results, the use of graphlet distribution cannot infer the topological characteristics of complex networks. Other statistics should complete that topological analysis.

The analysis of graphlet distribution shows that in some networks, as in the case of infrastructure networks, only a few graphlets characterize network structure, so that calculating higher orders does not bring much additional information. Our method only uses G_0, G_1 and G_2. This reduces the complexity of measuring 30 graphlets and 72 orbits.

Graphlet distribution is not an accurate tool to infer the topological characteristics of complex networks. The use of graphlets provide an incomplete description of network structure.

3.4.2. Spinning edges to connect nodes

The proposed method overcomes the limitations of graphlet distributions to explain network topology by a twelve-dimensional metric. To analyze results, Table 3-2 shows a set of global statistics used to analyze the five real networks, and Table 3-3 shows the value of the GHuST framework for those networks. In Table 3-3 values are not scaled. Figure 3-5 shows the scaled values of GHuST dimensions.

Table 3-2. Global topological properties of five real networks.

	N	L	D	$\langle k \rangle$	$\max(k)$	Ass. Coeff.	$\langle l \rangle$	d	$\langle BC \rangle$	$\max(BC)$	$\langle cc \rangle$
Road	2,642	3,303	0.02 %	2.5	5	-0.187	35.35	99	4.52×10^{4}	6.95×10^{5}	0.016
Power-grid	4,941	6,594	0.03 %	2.7	19	0.004	18.98	46	4.44×10^{4}	3.51×10^{6}	0.080
Mail	1,133	5,451	0.43 %	9.6	71	0.078	7.21	8	1.47×10^{3}	2.52×10^{4}	0.220
Social	1,446	59,589	2.85 %	82.5	375	0.067	2.22	6	887	1.88×10^{4}	0.323
Metabolic	1,039	4,741	0.44 %	9.13	638	-0.251	2.47	6	766	2.46×10^{5}	0.377

A. Global connectivity

The first dimension, ρ_1, relates the number of nodes and edges. This dimension scales linearly with network size. This is a strength concerning other metrics such as edge density. While the number of edges to have a complete graph increases with $\Delta N(\Delta N - 1)N_0$, where ΔN is the increase in nodes and N_0 the first set of nodes, the minimum number of lines to have a connected graph increases with ΔN. In the five real networks used, there is no discrepancy in the order provided by edge density, D, and ρ_1. However, there would have been discrepancies in the comparison of the following two networks: a network with 1,000 nodes and 2,000 edges and another network with 1,100 nodes and 2,200 nodes. The number of edges per node is the same in both networks. They have twice the number of edges needed by the minimum spanning tree, and there is no variation in ρ_1. However, the edge density of the second network is lower than the edge density of the first (0.20% and 0.18% respectively). Therefore, results provided by ρ_1 give a better understanding of the relation between the number of nodes and edges.

Table 3-3. Values of GHuST dimensions for a set of five real networks

	ρ_1	ρ_2	ρ_3	ρ_4	ρ_5	ρ_6	ρ_7	ρ_8	ρ_9	ρ_{10}	ρ_{11}	ρ_{12}
Road	0.25	0.04	1.85	2.18	0.72	0.57	1.55	0.01	0.03	0.05	0.01	3.42
Power grid	0.34	0.31	3.520	4.86	0.78	0.42	1.53	0.04	0.51	0.19	0.23	4.43
Email	3.81	0.51	16.25	85.97	0.96	0.08	1.04	0.06	0.95	0.74	0.06	12.34
Social	40.21	0.61	93.06	3965.03	0.98	0.01	1.00	0.10	1.00	0.98	0.01	84.02
Metabolic	3.56	0.04	321.00	472.05	0.80	0.06	1.06	0.01	0.96	0.84	0.05	10.31

ρ_i is the dimension i of the GHuST framework.

For the infrastructure networks, ρ_2 is lower in the power grid than in the road network. Indeed, the number of leaf nodes is 3.67% in the road network and 24.81% in the power grid. Therefore, we can infer that the power grid has nodes with a higher degree than the road network since the number of nodes per line and the percentage of nodes with only one connection is higher. The global statistic, maximum value of node degree, confirms that hypothesis.

In the case of the social networks, ρ_2 is lower than in the other networks (0.607 in the Facebook network and 0.515 in the email network). However, the percentages of nodes with just one connection are 1.17% and 13.23%. The dimensions ρ_8 and ρ_9 explain this inconsistency. Both social networks have a significant presence of triangles concerning other networks (friends of friends tend to be friends themselves). Indeed, only 1.93% of Facebook nodes are not part of a triangle and 25.86% of nodes in the email network. Therefore, ρ_2 only applies to those nodes that are not vertices of triangles. Accordingly, most nodes that are not vertices of a triangle are nodes with one connection. Similarly, in the metabolic network, 15% of nodes are not vertices of triangles, and the number of nodes with only one connection is scarce (0.5% of total nodes). Consequently, the value of ρ_2 is 0.036 in the metabolic network.

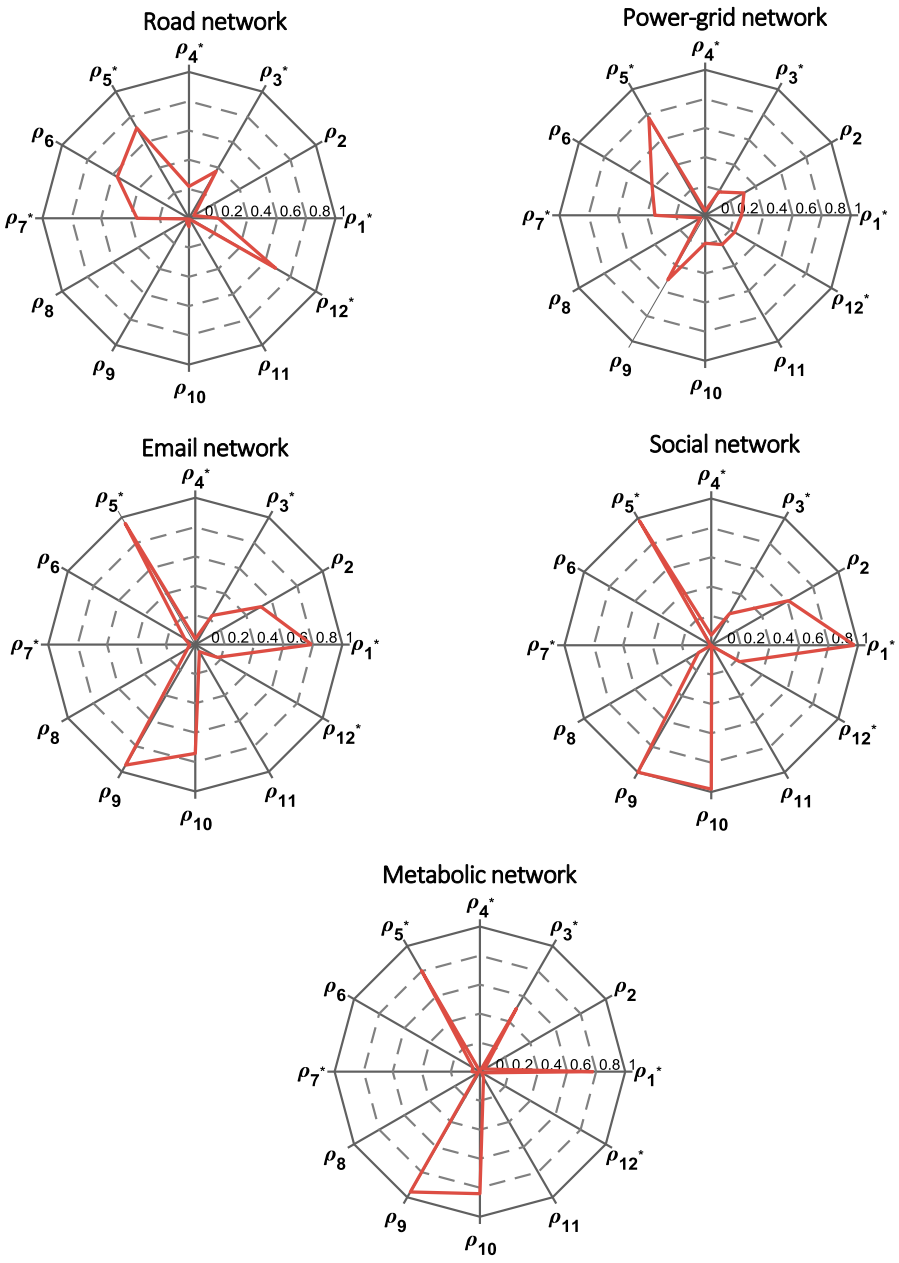


Figure 3-5. Graphical representation of the GHuST framework for a set of five real networks

The value of ρ_3 shows that in the power grid the neighbors of nodes with one connection have a higher node degree than in the road network. Therefore, we infer that the road network has a more homogenous mesh, and the range of node-degree distribution is smaller than in the power grid which tends to create hubs, as shown in Figure 3-3. The presence of hubs is also a characteristic of the metabolic network, where we see that ρ_3 is 321. This value is considerably larger than the Facebook network with a greater number of edges per node. This leads to the existence of a small number of hubs that concentrate most connections. The ratio between maximum node degree and average node degree is 70, a huge value in comparison with the other four networks.

B. Hubs

The tendency of a network to make hubs is supported by ρ_4. As in the case of ρ_3, the value of ρ_4 is 5.5 times larger in the metabolic network than in the email network (both networks have a similar number of nodes and edges). The maximum degree is 638 in the metabolic and 71 in the email network. Both networks have a similar number of edges per node (ρ_1), the percentage of nodes with only one connection is lower in the metabolic network (ρ_2), and ρ_3 is extremely large, so we may confirm the prior hypothesis that the high value of ρ_4 is the consequence of a few nodes with a high node degree. Accordingly, the third quantile of the degree distribution is 8 in the metabolic network and 13 in the email network. Therefore, although the hubs in the metabolic network have more connections on average, the number of nodes with a high degree is higher in the email network. In the case of the Facebook network, ρ'_4 is slightly larger than in the metabolic and in the mail network. We infer that in Facebook, the difference between high-degree nodes and low-degree nodes is not so big as in the other two networks.

In the case of the two infrastructure networks, the higher the value of ρ_4 in the power grid reinforces our first insight about their network topological properties. Furthermore, ρ'_4 shows that in the road network the mean value of O_2 is closer to the maximum value. That means that the road network has a more homogeneous mesh than the power grid where there should be a few nodes with large values of $O_{2,i}$.

Finally, ρ_5, shows if hubs tend to connect to other hubs. That correlation is clear in the social networks. If we choose the 50 nodes with the highest degree in the Facebook network, we see those nodes have connections to 50% (average value) of those nodes. However, in the infrastructure networks and in the metabolic network we cannot state whether hubs tend to connect among them or not, ρ_5 are around 0.75 in the range [0,1]. The values of ρ_5 diverge from the network assortativity coefficient, which has values close to zero (as shown in Table 3-2). Therefore, based on the network assortativity coefficient, nodes tend to connect high-degree nodes and low-degree nodes indifferently. The network assortativity coefficient measures if the degree of a node is correlated with its neighbors' degree. A positive correlation means that high degree nodes have connections with other hubs. Furthermore, low degree nodes are connected to nodes with a low number of connections. By contrast, in a network with a negative correlation, low degree nodes are only connected to high degree nodes. Moreover, hubs are not connected among them. In large networks, this coefficient might be misleading. We cannot state if hubs tend to connect to other hubs considering the network assortativity coefficient since it is conditioned to the way in which low degree nodes are also connected. Because of network size, hubs might be connected to other hubs and low-degree nodes at the same time. Therefore, the network assortativity coefficient would be close to zero (there is not a linear correlation between node degree and its neighbors' degree) and we will not obtain accurate information about the connection of hubs among them. The dimension ρ_5 overcomes this limitation.

C. Strings

Based on ρ_5, we may think that in networks in which hubs connect among them, distances will be smaller. Accordingly, the characteristic path length of the Facebook network is 2, and the diameter is 6. Here, it is difficult to compare changes in the characteristic path length and the diameter since it does not scale linearly with network size. We see that the diameter is lower in networks with hubs (social and metabolic) than in the infrastructure networks. To reinforce this analysis, ρ_6 shows that in the two infrastructure networks there are many nodes that are part of strings. The presence of those strings on Facebook is almost zero, and the length of those strings, ρ_7, is 1. Similarly, the closer the value of ρ'_7 to 0, the shorter the node string. In the email network and in the metabolic network, there are a few more strings, but their length is also close to 0. However, in the infrastructure networks, there is a higher presence of strings. In the case of the road network, the average length of node strings is similar to the case of the power-grid network.

D. Triangles

The dimensions $\rho_8, \rho_9, \rho_{10}, \rho_{11}, \rho_{12}$ supply detailed information about network clustering. As previously mentioned, the network average clustering coefficient, $\langle cc \rangle$, places more weight on low-degree nodes. In the case of the metabolic network, $\langle cc \rangle$ is 0.377. That might lead to the conclusion that nearly 40% of each node's neighbors form a complete connected subgraph. However, this contrasts with ρ_8 that shows the metabolic network as the one with the lowest numbers of triangles. As shown in the graphlet distribution, less than 1.4% of three-connected nodes are triangles. The high value of $\langle cc \rangle$ concerning ρ_8 shows that triangles in the metabolic network are connections of low-degree nodes. This is something that can be easily checked with ρ_{10}. In the metabolic network, the average degree of triangle vertices is 10.3, this value is close to the average node degree and far from the maximum degree in the network, 638. The number of edges needed by a node whose degree is 638 to have a value of local clustering coefficient equal to 1 is 215,644. Furthermore, ρ_9 shows that 96% of triangles share vertices, which reinforces the idea of low-degree nodes whose neighbors tend to form clusters. Those three dimensions explain network clustering, and they improve the information provided by the traditionally used network average clustering coefficient $\langle cc \rangle$.

The road network has a similar value of ρ_8. However, we see that more triangles do not share vertices; the average vertex degree is 3.41 and based on the first metric dimensions, we can conclude that the total number of triangles in the road network is lower (159 and 1,998 respectively). The total number of G_1 and G_2 in the road network is lower and therefore ρ_8 has similar values. To support this, we see that in the road network only 5% of nodes are vertices of a triangle, ρ_9. However, in the metabolic network, 84% of nodes are part of at least one triangle. Comparing the two infrastructure networks, the power grid has a higher number of triangles (651) and ρ_8 is larger. Unlike the road network, 26% of triangle vertices are not connected to the rest of the network. However, since in the power grid the maximum node degree is much higher than in the road network, the value of ρ'_{12} is higher in the road network.

In the power grid, 50% of triangles share one of their vertices, there are more lines per node than in the road network (ρ_1) and more isolated nodes (ρ_2), this might lead to the conclusion that in the power grid there are more triangle vertices that are not connected to other nodes, (that is whose node degree is two). In the power grid, 20% of triangle vertices have degree equals to 2, in the road network that percentage is 0.6%. This is something that we see in ρ_{11}, 23% of nodes that are vertices of triangles have no more connections in the power grid.

Regarding the two social networks, both have a large number of triangles. In the case of Facebook, 10% of 3-node graphlets are triangles; this is a high value considering the presence of hubs, ρ_4, which increases the number of total 3-node graphlets. Indeed, 98% of Facebook nodes are part of a triangle, as shown in ρ_{10}. Furthermore, almost all triangles share their vertices since ρ_9 is close 0. In the case of the email network, the presence of triangles in the network is 6%; this value is high in comparison with the network average clustering coefficient of another email network [86]. Only 5% of triangle vertices in the email network are not part of two or more triangles. Finally, if we compare the email network with the metabolic network, we can observe that in both networks ρ_{12} is similar. In the metabolic network, it looks like triangles are not part of hubs, since ρ_{12} is much lower than the maximum node degree (low value of $\rho_{12}{}'$).

Table 3-4. Information provided by the GHuST model for 5 real networks

	Global connectivity	Hubs	Strings	Triangles
Road	Low number of lines per node Scarce presence of leaf nodes	Low number of hubs	Presence of strings formed by several nodes	Low presence of triangles that do not share vertices
Power grid	Low number of lines per node Presence of leaf nodes	Low number of hubs	Presence of strings formed by several nodes	Low presence of triangles
Email	High number of lines per node Presence of leaf nodes	Low presence of hubs with respect to social networks. Hubs are connected among them	Scarce presence of strings	High presence of triangles that tend to share vertices and cover the whole network
Social	Highly meshed structure Leaf nodes connected to hubs	High presence of hubs that are connected among them	Scarce presence of strings	High presence of triangles that tend to share vertices and cover the whole network
Metabolic	High number of lines per node Scarce presence of leaf nodes	High presence of hubs, tendency to be connected lower than in social networks	Scarce presence of strings	High presence of triangles that tend to share vertices and cover the whole network

The description provided by the GHuST framework fully describes the topology of real networks.

Result are consistent with global statistics traditionally used in complex networks. Furthermore, this framework overcomes the main drawbacks of global statistics.

3.5. A panoramic view offered by local properties

The previous section has illustrated the application of the proposed metric as a tool for summarizing the main topological features of complex networks. This section aims at evaluating the performance of this technique using a large sample, 1404 graphs, of real networks from different domains: Autonomous Systems, Enzymes, Facebook, Power Network, Retweet, Roads, and Web.

The autonomous-systems set stands for 733 daily instances of graphs of routers comprising the internet [87]. The enzymes, Facebook, retweet, roads, web, and some power-network graphs are obtained from an open-access network repository [83]. The enzyme dataset includes 476 samples (the analysis only considers graphs with more than 20 nodes). The Facebook set consists of 108 networks of friendship connections. The power-network graphs comprise the transmission (220 kV and 400 kV) power networks of fifteen European countries, and a set of power networks (7 graphs) obtained from the open-access repository (voltages levels are not specified) [12], [83]. The retweet networks form a set of 32 graphs. The road set includes 16 instances. Finally, 17 networks are part of the web graphs.

Once we compute the twelve-dimensional metric for each network, a Principal Component Analysis (PCA) is used to reduce the dimensionality of the proposed statistic. It enables visual inspection of our data. PCA is a statistical technique that seeks to obtain a linear combination of the original variables in such a way that the maximum variance is explained. This allows us to obtain a low-dimensional representation of the data that captures most of the original information. Varimax rotation was applied to improve the understanding of PCA analysis. However, results obtained with varimax rotation did not improve the results shown in this section. Furthermore, any network with unusual topological properties will be highlighted in our analysis, providing a tool for detecting outliers.

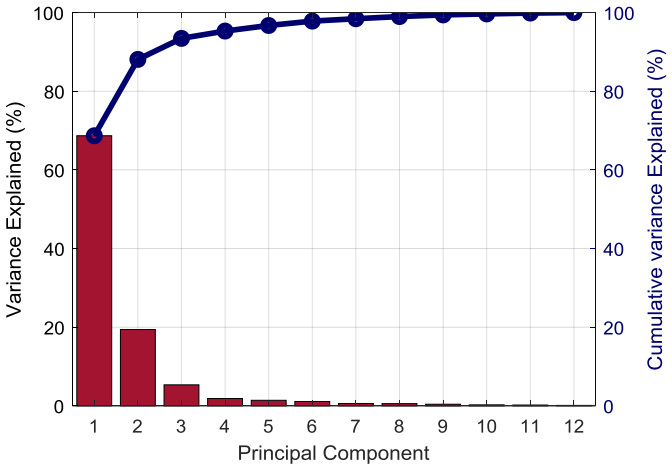


Figure 3-6. Variance explained and cumulative variance explained by each of the principal components resulting from the PCA analysis to a set of 1404 networks.

Figure 3-6 shows the proportion of variance explained by the principal components. By selecting the first three components, we are able to capture 93.4% of the variance of the original data, allowing us to obtain a low-dimensional view of the distribution of our data. The weights of the twelve dimensions of our metric for each component are shown in Figure 3-7, and they can be used to obtain an interpretation of each component. The first component (68.7% of variance), accounts for a positive contribution of ρ'_4, ρ_{10}, ρ'_{12} and a negative contribution of ρ_2, ρ_6 and ρ_{11}. Therefore, the main topological differences among the networks analyzed lie on the proportion of leaf-nodes, presence of hubs and strings, as well as the triangle pervasiveness and connectivity coefficients and triangle degree. A similar interpretation can be obtained for the second component (19.4% of variance) and the third component (5.3% of variance) based on Figure 3-7.

By projecting the coordinates of our twelve-dimensional data on the space spanned by the first 3 principal components, we can visualize the distribution of the metric for each network in this new axis system. As seen in Figure 3-8 and in Figure 3-9, networks from different processes tend to have similar topological properties, hence showing clear groupings in the principal-component space.

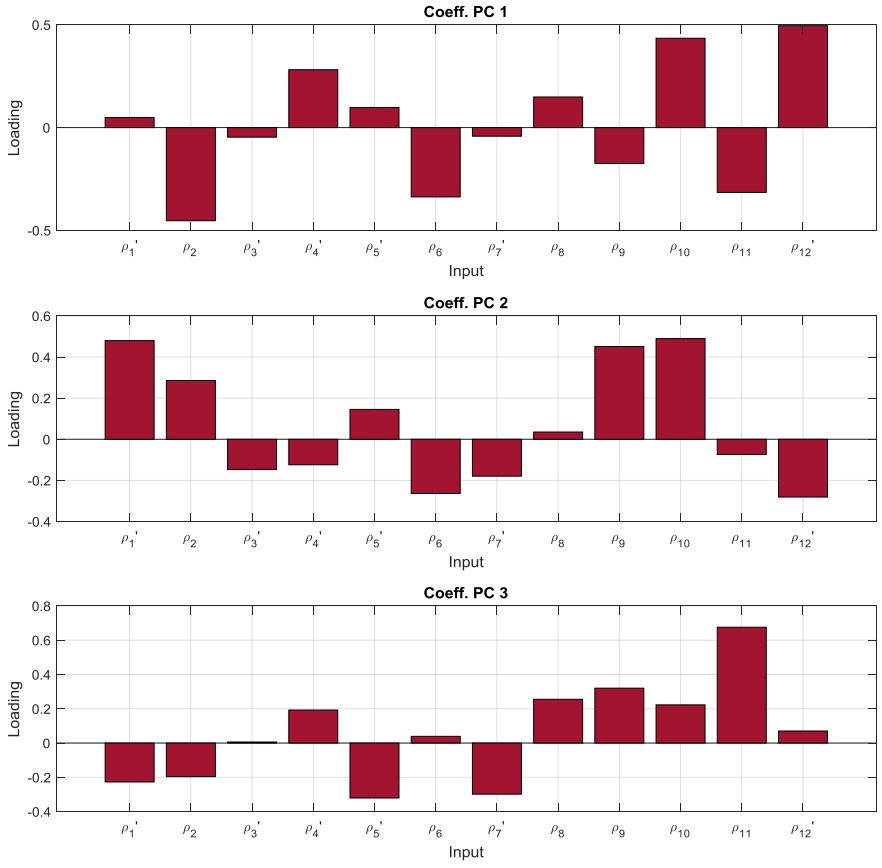


Figure 3-7. Contribution of each dimension of GHuST to the three first principal components.

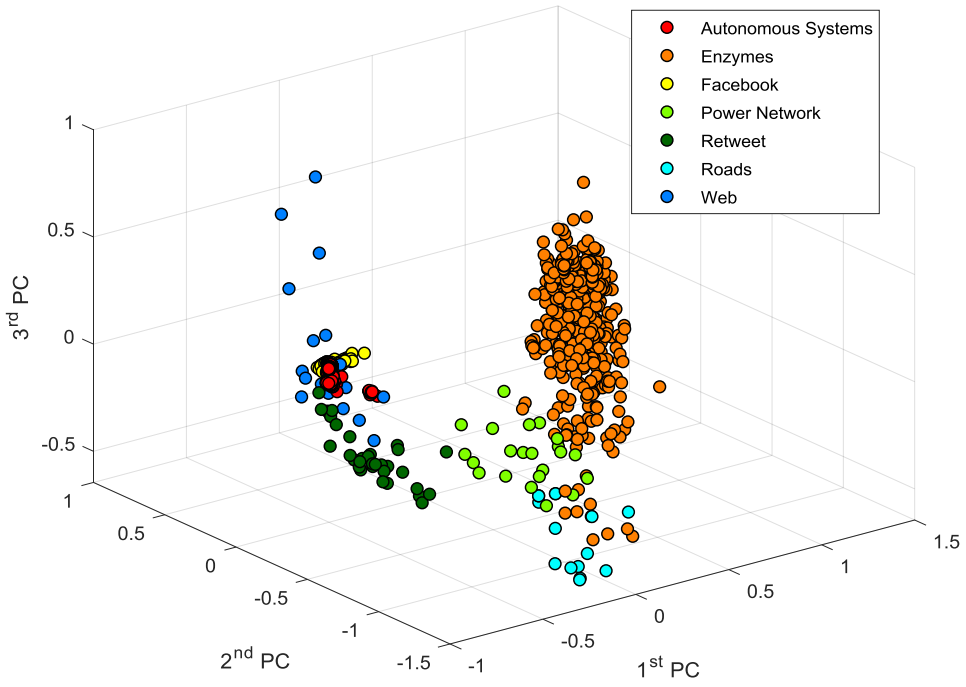


Figure 3-8. Graphical representation of 1,404 networks in the 3D space defined by the three first principal components

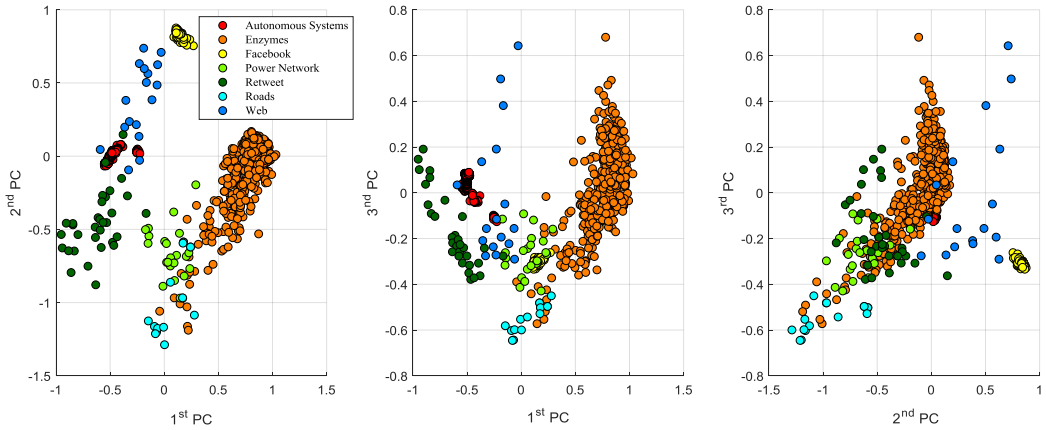


Figure 3-9. 2D projections of the 1,404 networks in the space defined by the three first principal components

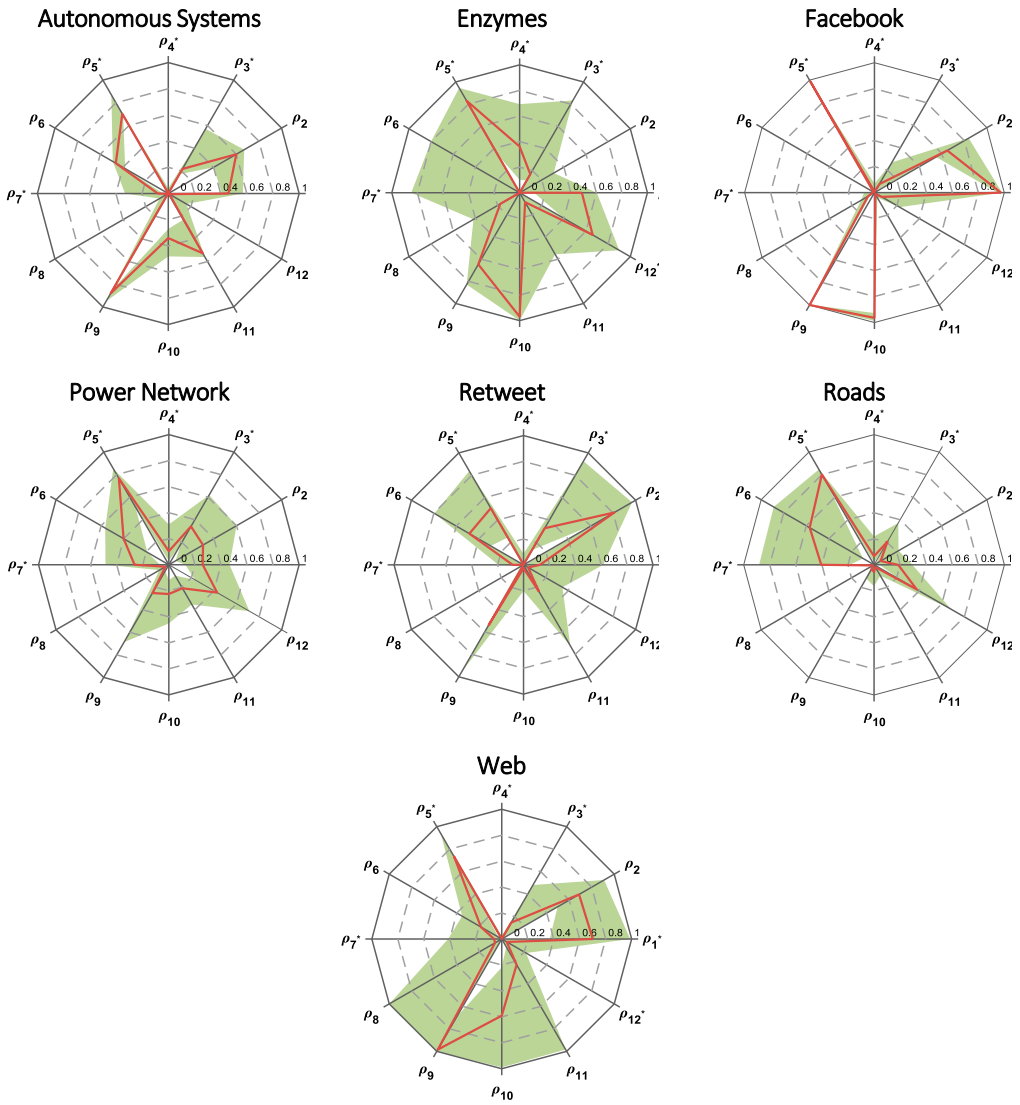


Figure 3-10. Range of variation and median value of each metric dimension for the seven sets of networks analyzed.

The autonomous-system and Facebook networks form two bounded clusters in the three-first principal-component space. Despite being the category with more instances, all the autonomous-system instances are close to -0.5 in the first component and to 0 in the second and third components. Since in the first principal component, ρ_i have positive and negative loadings, we cannot state if those values close to zero are the consequence of low values of all components, or they are the consequence of the balance between positive and negative loadings. Figure 3-10 shows the range in which the twelve dimensions vary. We see that in the autonomous systems, the value of the second component is the balance between positive loadings (ρ'_1, ρ_9 and ρ_{10}) and negative loadings (ρ'_3, ρ_6 and ρ'_7); the other dimensions are close to zero. Similarly, the analysis of ranges for each type of network allows for the classification of graphs. In the case of Facebook graphs, most analyzed instances have values of

$\rho'_3, \rho'_4, \rho_6, \rho'_7, \rho_8, \rho_{11}, \rho'_{12}$ that are close to zero and the values of $\rho'_1, \rho'_5, \rho_9, \rho_{10}$ are close to one. In Exhibit A, the reader can find a detailed explanation of ρ_i distribution for each type of network.

Regarding the two infrastructure networks, roads and power networks comprise two independent clusters. Although some road networks are close to some power grids in the space defined by the first and second principal components, they are clearly delimited in the other two projections of the three first principal components.

Both roads and power networks have low values for the second component, that is low values of ρ'_1, ρ_9 and ρ_{10}. Accordingly, the number of connections in comparison with the minimum spanning tree is low, there is a low number of triangles in the network, and they do not tend to share vertices. The instances of roads and power networks that have similar values for the second principal component have a similar number of edges per node. They are the power networks with the lowest number of lines per node concerning other power networks and the roads with a higher number of lines per node in their category.

Unlike social networks, connections in infrastructure networks are cost-intensive, and they are conditioned by topological, morphological, technical, economical, permitting, environmental, managerial and political factors [88]. Consequently, the influence of all those factors may lead to different topological properties depending on regions. Furthermore, in the case of power networks, graphs may include different voltage levels, or they may be the result of different model assumptions [89]. This uncertainty leads to a lack of consensus about some of the topological properties of power networks [38].

The cluster with the most variation among its members belongs to the enzymes group. This shows that a network cannot be classified in the enzyme group, such as Facebook networks. The green area that shows the range in Fig. 6 almost covers all the dodecagon. The topological properties of enzymes are case dependent.

Finally, we can also see two clusters considering the web and retweet group. In the case of web networks, there is a significant variation in the third component. It ranges from -0.3 to 0.7. This variation is caused by the significant difference in ρ_{11} (triangle-connectivity coefficient). Although the median of the analyzed instances has a low value, this coefficient ranges from 0 to 1. In the web case, we also see that although most instances have a triangle coefficient (ρ_8) close to zero, there is an instance in which ρ_8 tends to 1 (the network is mainly formed by triangles). This coefficient is coherent with the network average clustering coefficient [83]. Accordingly, this framework also supports the quick detection of potential outliers.

PCA analysis can be used for each set of networks independently. Therefore, the dimensions with more significant loadings for the first components are the ones that exhibit the most variance in each original set; hence those dimensions will provide information about the topological differences between networks of the same set. Dimensions that have similar values for all networks in the set will have a low contribution to the first components as they are characteristics of those networks.

The explained variance for each principal component and the coefficient that shape the first component are shown in Figure 3-11 and Figure 3-12, respectively. Low-Dimensional representation of the projections of the metrics in the three first principal components for each set of networks can be seen in Exhibit B. In the case of the road networks; the first component explains 88.5% of the variance. This component is mainly defined by ρ_6 and ρ'_7. Therefore, the difference among roads networks lies on the number of nodes that are part of a node string in the network and the length of those strings.

When analyzing power networks, we observe that the first component only explains 44% of the variation. Consequently, the number of coefficients to describe and to explain differences among power networks is larger. The first component is mainly described by ρ'_1, $\rho_2, \rho'_3, \rho'_4, \rho_9, \rho'_{12}$. It is necessary to include five principal components, to explain 95% of the variance of the data. This increases the number of metric dimensions required to have a deep understanding of power-network topology. In the case of Facebook, the first component explains 72% of the variance. Consequently, the main differences lie in the leaf coefficient, leaf-connection degree, and triangle degree.

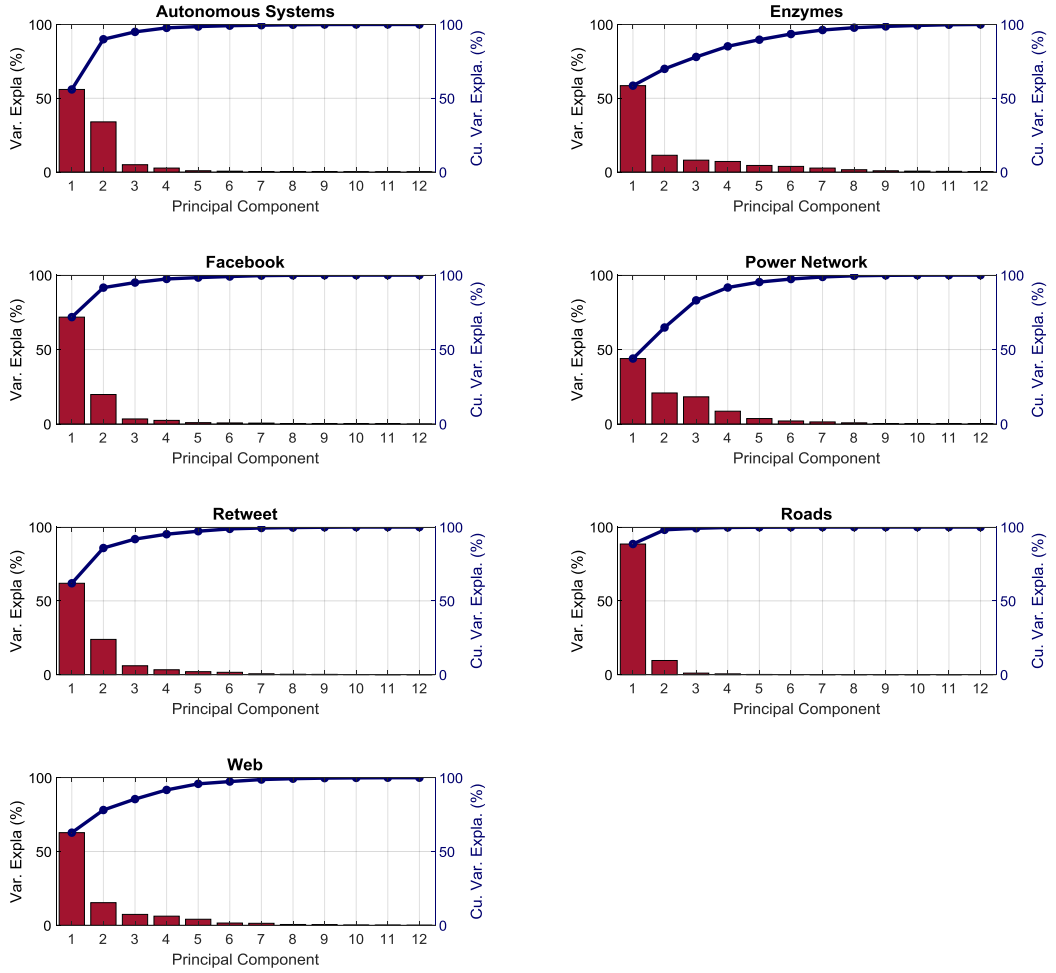


Figure 3-11. Variance explained and cumulative variance explained by each of the principal components of the resulting PCA applied independently to each type of network analyzed.

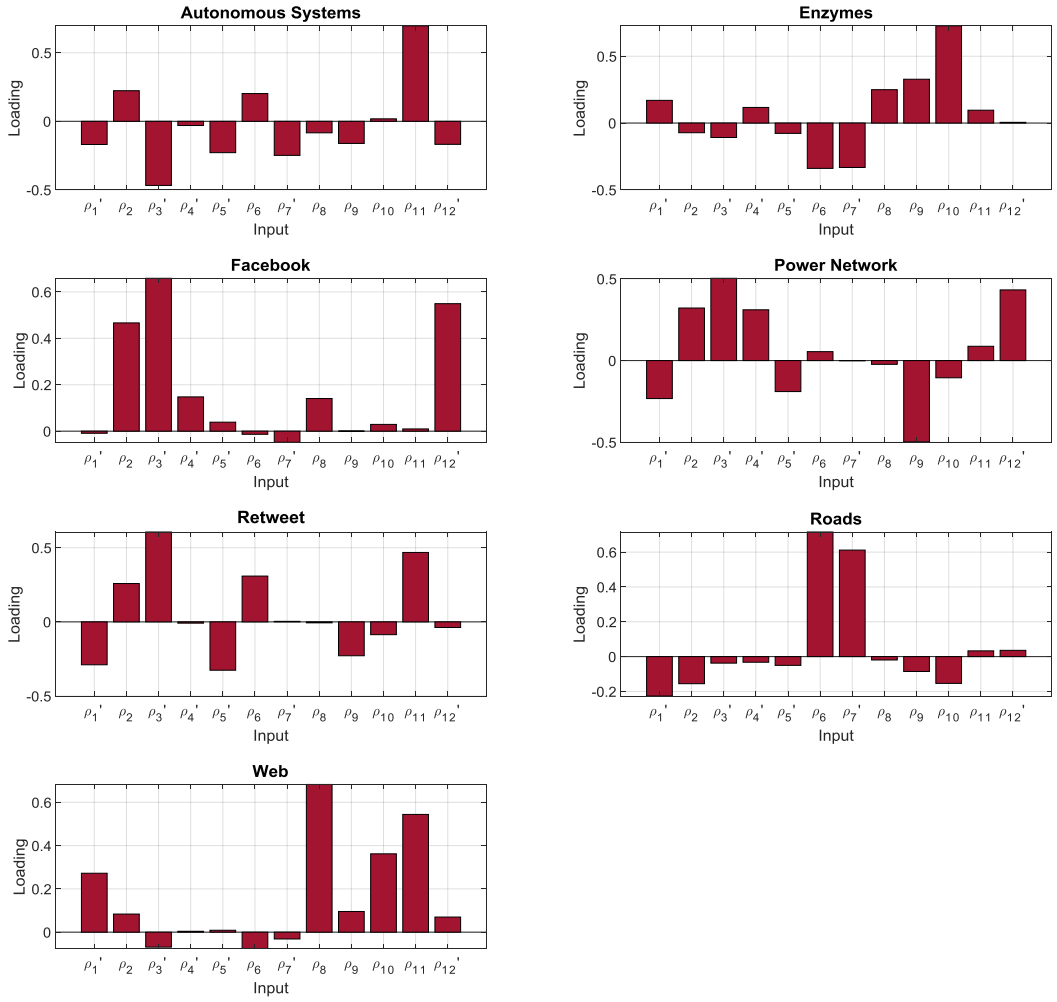


Figure 3-12. Contributions of each dimension of the GHuST framework to the first principal component obtained for each set of networks analyzed.

Results show the strengths of the proposed method to compare networks of different nature and to find the topological differences among same-nature networks.

The use of PCA to reduce the dimensions of the GHuST framework allows for graphical representation of networks in the three-dimensional space.

Networks form different processes tend to have similar topological properties, hence showing clear groupings in the principal-component space.

3.6. Takeaways

The analysis of network graphlets, a local-topological statistic, gives rise to a new description of the global topology of complex networks. This thesis introduces an innovative method that analyzes the interaction among graphlets to explain and characterize network topology. This method is based on 2- and 3-node graphlets (three graphlets and four orbits) that are easily derived from the adjacency matrix. Therefore, it overcomes the limitation of counting high degree graphlets that might be cost-intensive for large networks.

The application of the novel framework to five real networks shows that the proposed method is consistent with the global statistics traditionally used to characterize network structure. Furthermore, it overcomes two of their main drawbacks: the use of metrics based on average values and the application of metrics that do not scale linearly with network size. Accordingly, the comparison among networks of different sizes does not require any analysis of metric scalability.

The proposed method has been also validated with a large sample study of networks that arise in different fields. Results prove that the information provided by this novel metric can be used to identify the underlying topological features of the networks and even to provide us with a visual tool to distinguish networks with different properties.

Consequently, this method might explain the evolution in both local and global properties of networks in which growth affects the whole structure. It can also be used to compare networks where network growth does not necessarily imply a change in local properties. This is common in infrastructure networks.

Finally, this work sets up a systematic analysis consisting of a twelve-dimensional metric to explain the properties of the network structure. Moreover, the proposed method allows for the translation of topological properties into other scientific dimensional languages. This is possible because global properties are explained from local structures that are easily interpretable.

4

THE ROLE OF TOPOLOGY IN SYNTHETIC POWER GRIDS

4.1. Applying the GHuST framework to power networks

The use of global statistics, such as network average clustering coefficient, characteristic path length, or network diameter, are not enough to provide a sound explanation of power-network topology, as explained in Chapter 2 and Chapter 3. For instance, based on those global statistics, we cannot explain why the French 220-kV and the German 220-kV networks have similar values of network diameter if the French network is 4.3 times bigger than the German one. Furthermore, the use of global statistics may lead to misleading conclusions, as discussed in previous chapters with the network average clustering coefficient. Moreover, global statistics do not always scale with network size. This hinders the correct characterization of power-network topology and the comparison among networks.

Not only may the use of global statistics give an incomplete description of network topology, but they can also condition the topological validation of synthetic power grids. As explained previously, synthetic power networks are non-real, albeit realistic, power networks that are topologically and electrically consistent with real networks. Accordingly, it is necessary to define a transparent methodology to validate the structure of synthetic networks. This method should compare the topology of synthetic and real networks regardless of network size.

This chapter proposes the use of the GHuST framework to complete the topological description of power networks and to analyze the topological consistency of synthetic power grids. First, the use of this novel framework will allow us to have a better understanding of network topology, solving those questions about network topology that cannot be answered in Chapter 2. Second, the GHuST framework will set up a complete method to analyze the topological consistency of synthetic networks.

The rest of the chapter is organized as follows; Section 2 compares the European transmission power networks based on GHuST dimensions. Section 3 proposes the use of the GHuST framework for the validation of synthetic power grids. Section 4 analyzes the topological consistency of published synthetic power networks. Finally, Section 5 discusses the results.

The GHuST framework overcomes the limitations of global statistics traditionally used in complex networks. The main applications of the GHuST framework in power networks are:

- Full topology description of power networks.
- Topological validation of synthetic power grids.

4.2. Completing the topological description of power networks

The introduction of the GHuST framework showed that the topology of power networks is different from the topology of other networks such as social or road networks. In Chapter 3, we could observe that those networks were clearly differentiated in the topological space. We also saw that the variance in the power-network cluster was higher than in the Facebook or autonomous-system clusters. Accordingly, topological differences among power-network instances were higher. We stated that the lack of information about voltage level or network location was leading to that dispersion. We, therefore, assumed that we might find some sub-clusters inside the power-network region.

The lack of a large set of real power networks makes it impossible to carry out statistical analysis to verify the existence of sub-clusters based on the power-network structure. However, a more in-depth analysis of power-network topology with the GHuST framework may complete the description of power-network topologies given in Chapter 2. We apply the GHuST framework to the fifteen European transmission power networks presented in Chapter 2 (we also consider the 400-kV and the 200-kV networks both as a single network and as independent networks). This allows for a comparative analysis based on location (country) and voltage level.

4.2.1. Relating voltage level and topology

To analyze if voltage level conditions the topology of the European transmission power networks, we apply the GHuST framework to both voltage levels independently (400 and 220 kV) and to the network that includes both voltage levels linked by transformers. Figure 4-1 shows the range, the first, second, and third quartile for each dimension of GHuST.

The twelve dimensions show that the 220-kV network has a slightly less meshed structure than the 400-kV network. Based on the difference between quartiles, in the 220-kV network, the number of lines installed is lower, there is a higher number of leaf nodes, the average degree of leaf connections is lower, and there is a lower tendency to make hubs. Regarding node strings, the number of node strings and the length of those strings is slightly higher in the 220-kV network. Similarly, the presence of triangles is lower in the 220-kV network; they cover a lower number of vertices, and the percentage of shared vertices is also lower. When analyzing both voltage levels together, the structure has a higher number of lines per node, since it includes transformers. Beyond that change, there is no remarkable difference that can be obtained from GHuST concerning the analysis of both voltage levels independently.

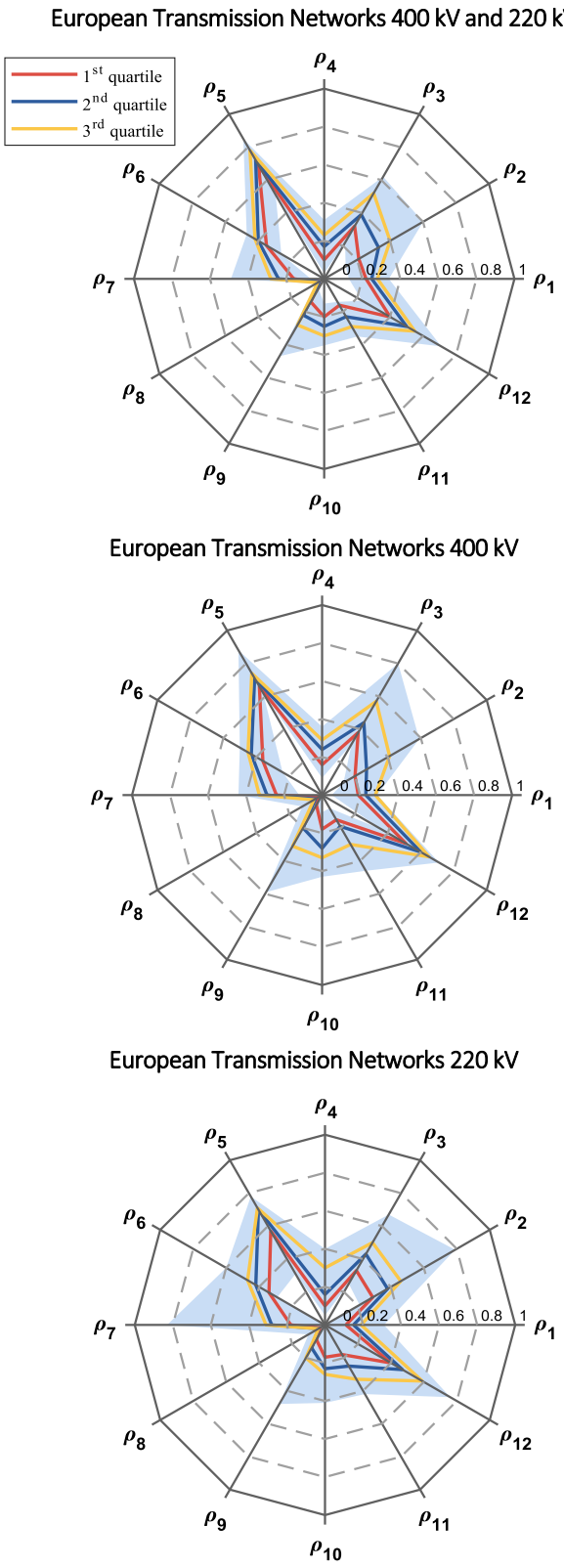


Figure 4-1. Range of variation for each dimension of the GHuST framework for the European transmission power networks

The differences seen between the 220-kV and 400-kV network are in line with the results showed in Chapter 2. The less meshed structure leads to higher distances among nodes (higher values of characteristic path length and network diameter) in the 220-kV network. Moreover, the maximum and mean betweenness centrality are higher in the 220-kV network since the number of alternative paths to go from one node to another is lower.

Voltage level slightly conditions the topology of transmission power networks. The 400-kV network has a more meshed structure with a higher presence of hubs and triangles. The number of strings is lower than in the 220-kV network.

4.2.2. Countries define network structure

Chapter 2 showed that European transmission power networks followed some topological patterns. We found the number of lines, characteristic path length, network diameter and betweenness centrality (mean and maximum values) scale with network size. However, most countries have a certain deviation regarding the regression line. This could not be explained by global statistics. The topological information provided by the GHuST framework sheds light on the analysis done in Chapter 2. Values of GHuST for the 220-kV network, the 400-kV network, and 200-kV and 400-kV network together are shown in Table 4-1 to Table 4-3.

Results show that networks with equivalent size may display completely different topologies. This is the case of the Portuguese and the Swiss 400-kV and 200-kV networks. Both have similar size, 159 and 158 nodes respectively. However, the Portuguese network has more lines installed and a higher presence of leaf nodes, hubs, and triangles. Those triangles share a higher number of vertices in the Portuguese network. Furthermore, the number of node strings and their length are larger in the Swiss network. It shows that the Swiss network has a more homogeneous structure with lower complex local structures. This explains the topological differences between both countries regarding characteristic path length, network diameter, or betweenness centrality that are not so intuitive. The mean value of betweenness centrality is higher in the case of Portugal since the presence of hubs and leaf nodes may lead to the presence of network components with higher values of centrality in the network. The presence of hubs may also lead to a lower network diameter. However, a more homogeneous mesh with a smaller number of leaf nodes may reduce the characteristic path length.

The tendency to form hubs in the Portuguese 400-kV network is exceeded only by France. While in France ρ_4 is twice the mean value for the European countries, the number of lines installed per node is one of the lowest values in the European networks. We also see that the percentages of node strings in the French networks are low and the characteristic string lengths are also the lowest values among the fifteen countries.

Table 4-1. GHuST values for 400-kV and 220-kV European transmission networks.

Country	N	ρ_1	ρ_2	ρ_3	ρ_4	ρ_5	ρ_6	ρ_7	ρ_8	ρ_9	ρ_{10}	ρ_{11}	ρ_{12}
Hungary	(50)	0.17	0.60	0.62	0.29	0.51	0.33	0.13	0.04	0.17	0.30	0.33	0.56
Netherlands	(55)	0.13	0.33	0.61	0.29	0.67	0.47	0.39	0.03	0.22	0.13	0.14	0.71
Greece	(57)	0.29	0.13	0.34	0.17	0.67	0.45	0.32	0.05	0.33	0.32	0.17	0.53
Bulgaria	(63)	0.23	0.43	0.61	0.31	0.70	0.28	0.09	0.03	0.06	0.27	0.18	0.64
Serbia	(84)	0.21	0.49	0.45	0.09	0.72	0.33	0.07	0.03	0.43	0.20	0.35	0.32
Belgium	(88)	0.16	0.38	0.54	0.25	0.66	0.40	0.23	0.03	0.14	0.20	0.28	0.59
Austria	(89)	0.25	0.23	0.39	0.20	0.72	0.42	0.26	0.05	0.31	0.30	0.33	0.50
Romania	(117)	0.27	0.22	0.50	0.20	0.85	0.40	0.49	0.03	0.13	0.22	0.15	0.53
Switzerland	(158)	0.29	0.20	0.22	0.09	0.78	0.43	0.19	0.02	0.14	0.23	0.31	0.40
Portugal	(159)	0.33	0.37	0.32	0.12	0.74	0.26	0.17	0.05	0.47	0.35	0.16	0.39
Poland	(163)	0.34	0.16	0.40	0.17	0.82	0.38	0.16	0.03	0.25	0.26	0.19	0.54
Italy	(634)	0.22	0.36	0.38	0.10	0.77	0.37	0.27	0.02	0.18	0.15	0.14	0.44
Germany	(782)	0.28	0.26	0.31	0.08	0.72	0.40	0.27	0.04	0.26	0.3	0.25	0.33
Spain	(798)	0.28	0.20	0.32	0.10	0.79	0.44	0.29	0.03	0.13	0.25	0.23	0.41
France	(1659)	0.23	0.41	0.23	0.03	0.73	0.41	0.24	0.02	0.22	0.19	0.25	0.23

ρ_i is the dimension i of the GHuST framework.

Table 4-2. GHuST values for 400-kV European transmission networks.

Country	N	ρ_1	ρ_2	ρ_3	ρ_4	ρ_5	ρ_6	ρ_7	ρ_8	ρ_9	ρ_{10}	ρ_{11}	ρ_{12}
Bulgaria	(21)	0.22	0.27	0.8	0.34	0.72	0.33	0.2	0.04	0.00	0.29	0.33	0.63
Hungary	(28)	0.26	0.38	0.58	0.36	0.62	0.35	0.17	0.06	0.20	0.43	0.42	0.60
Austria	(31)	0.23	0.13	0.42	0.22	0.65	0.43	0.44	0.12	0.46	0.42	0.46	0.53
Serbia	(33)	0.05	0.53	0.62	0.29	0.67	0.41	0.29	0.01	0.00	0.09	0.33	0.50
Netherlands	(35)	0.13	0.33	0.54	0.28	0.64	0.50	0.33	0.04	0.22	0.20	0.14	0.71
Switzerland	(37)	0.20	0.25	0.31	0.22	0.88	0.50	0.43	0.03	0.00	0.24	0.22	0.67
Romania	(46)	0.28	0.14	0.60	0.24	0.75	0.43	0.29	0.03	0.17	0.22	0.10	0.63
Greece	(57)	0.29	0.13	0.34	0.17	0.67	0.45	0.32	0.05	0.33	0.32	0.17	0.53
Portugal	(57)	0.28	0.56	0.45	0.32	0.71	0.24	0.13	0.09	0.59	0.37	0.19	0.62
Belgium	(58)	0.13	0.42	0.56	0.27	0.65	0.43	0.33	0.01	0.00	0.10	0.17	0.67
Poland	(59)	0.28	0.20	0.44	0.28	0.83	0.43	0.33	0.06	0.37	0.32	0.11	0.66
Spain	(201)	0.29	0.19	0.41	0.16	0.78	0.47	0.32	0.03	0.14	0.28	0.21	0.52
Italy	(262)	0.18	0.39	0.39	0.10	0.71	0.38	0.28	0.02	0.11	0.15	0.15	0.39
France	(386)	0.19	0.59	0.37	0.11	0.72	0.32	0.12	0.01	0.21	0.13	0.08	0.42
Germany	(480)	0.28	0.21	0.34	0.09	0.69	0.45	0.29	0.04	0.29	0.34	0.27	0.36

ρ_i is the dimension i of the GHuST framework.

Table 4-3. GHuST values for 220-kV European transmission networks.

Country	N	ρ_1	ρ_2	ρ_3	ρ_4	ρ_5	ρ_6	ρ_7	ρ_8	ρ_9	ρ_{10}	ρ_{11}	ρ_{12}
Greece	(0)	-	-	-	-	-	-	-	-	-	-	-	-
Netherlands	(20)	0.00	0.44	0.66	0.40	0.45	0.60	0.83	0.00	0.00	0.00	0.00	0.50
Hungary	(22)	0.00	0.79	0.67	0.40	0.31	0.29	0.00	0.02	0.00	0.14	0.00	0.76
Belgium	(30)	0.09	0.47	0.43	0.32	0.63	0.40	0.17	0.10	0.20	0.41	0.42	0.62
Bulgaria	(42)	0.14	0.52	0.51	0.29	0.56	0.33	0.13	0.04	0.08	0.26	0.18	0.61
Serbia	(49)	0.20	0.54	0.54	0.11	0.71	0.32	0.25	0.04	0.48	0.29	0.36	0.34
Austria	(58)	0.16	0.32	0.49	0.31	0.52	0.38	0.29	0.04	0.07	0.25	0.36	0.64
Romania	(71)	0.10	0.31	0.44	0.19	0.71	0.55	0.56	0.05	0.11	0.23	0.25	0.52
Portugal	(102)	0.31	0.28	0.36	0.12	0.70	0.30	0.17	0.04	0.35	0.34	0.17	0.40
Poland	(104)	0.25	0.21	0.48	0.19	0.78	0.43	0.26	0.03	0.11	0.23	0.25	0.55
Switzerland	(121)	0.24	0.21	0.25	0.10	0.69	0.47	0.19	0.03	0.18	0.23	0.33	0.40
Germany	(302)	0.12	0.39	0.36	0.11	0.61	0.47	0.31	0.04	0.15	0.23	0.33	0.40
Italy	(372)	0.15	0.38	0.32	0.09	0.75	0.42	0.31	0.03	0.23	0.16	0.19	0.43
Spain	(597)	0.19	0.24	0.33	0.09	0.72	0.52	0.39	0.03	0.12	0.23	0.31	0.40
France	(1273)	0.14	0.44	0.26	0.05	0.70	0.43	0.31	0.02	0.18	0.15	0.24	0.27

ρ_i is the dimension i of the GHuST framework.

That may explain why France has a similar characteristic-path length and network diameter than Germany in the 400 kV and 220 kV network, but the number of nodes in France is more than twice the number of nodes in the German 400-kV and 220-kV network.

Results also show divergences concerning the network average clustering coefficient showed in Chapter 2. The Italian 220-kV network has a lower network average clustering coefficient than the French 220-kV network. However, the number of triangles is lower in the French network. Additionally, the average degree of those triangles is also lower in the French network. Consequently, this leads to a higher value of network average clustering coefficient as discussed in Chapter 3. The comparison of the twelve dimensions shows that the Portuguese 400-kV network is one of the fifteen countries with large values of lines per node. However, a large number of lines installed does not necessarily imply a lower number of leaf nodes. Portugal is one of the countries with a higher number of leaf nodes (ρ_2 is 0.56). Therefore, lines are concentrated in some areas. Indeed, the Portuguese 400-kV network is the second network with a higher tendency to make hubs. Furthermore, we find a high presence of triangles in the network, and they tend to share vertices. Consequently, the Portuguese network has highly complex local structures that cannot be explained with global statistics. We also may find networks with similar structures. The Spanish and German 400-kV networks have similar values for global statistics. Those countries also have similar values for GHuST. The slight differences between them lie on the percentage of shared vertices and vertices degree.

Network topology displays substantial differences depending on location. We have not found electrical considerations that explain those differences. Consequently, the generation of synthetic power grids should be flexible enough to adapt resulting in networks to the complexity of the country they stand for. Furthermore, the analysis of network topology with global statistic has been revealed insufficient to give a sound explanation of network topology.

The GHuST framework shows that the idiosyncrasy of a country highly conditions the structure of transmission power networks. The twelve dimensions show topological differences among countries that were not explained by global statistics.

4.3. Topological validation of test cases

Based on the topological description given by the GHuST framework along with results in Chapter 2, we can conclude that although there are structural similarities among power networks, each country displays different topologies. Those differences may impact on network operation or network robustness. Consequently, the topology of synthetic power networks should be tested to analyze the topological consistency concerning real networks.

Several models have been proposed for the generation of synthetic power grids, both spatial and non-spatial networks [21], [22], [65], [90]. However, some of those works did not give enough attention to topological validation [65]. That validation is usually based on global statistics, that as explained in this chapter, have two main drawbacks: the use of average values that might be misleading and the use of distance-based metrics that do not scale linearly with the number of nodes. Accordingly, the use of global statistics may lead to wrong validation.

We propose the use of the GHuST framework to validate the topology of synthetic power networks. The twelve dimensions are complementary, and they should not lead to biased conclusions. Furthermore, they are size-independent, and it is not necessary to analyze metric scalability.

The twelve dimensions of GHuST in a synthetic network should be similar to the dimensions of GHuST in the real power network it represents. In case real data are not available, the synthetic network can be compared with the results obtained for the European transmission power networks. These data, provided by ENTSO-e, show the real topology of fifteen transmission networks. However, the topology of other real networks might display different topologies. Consequently, based on the comparison with those networks, we can only conclude whether the topology of the synthetic network is consistent (from a statistical point of view) with the European power network topologies or not. We cannot conclude that a network with different values of GHuST is incorrect.

Since values in the second and third quartiles are close (see Figure 4-1), the higher the difference in a dimension, the lower the probability of finding that topology in a real network. In this thesis, we consider that a synthetic network is topologically consistent if the twelve dimensions of GHuST are in the range defined by the European transmission networks. Because of the wide range for some dimensions, we will point out if a dimension is in the first or fourth quartile. This does not mean that the topology of the synthetic network is inaccurate; it is just a sign of caution. The use of a higher number of real power-network instances would increase

the statistical consistency of the ranges defined for each dimension of the GHuST framework (the methodology to validate the topology of synthetic power grids would be the same).

This chapter only focuses on the topological validation of synthetic power grids. A complete validation of synthetic power grids should also include electrical considerations. However, if a synthetic power grid is not topologically consistent with real networks, we can conclude that the synthetic network is not an accurate network model. Once synthetic power grids are validated from a topological point of view, they should be validated considering electrical considerations such as network operation.

The GHuST framework allows for a sound topological comparison between real and synthetic network topologies. It compares network structures regardless of network size. A synthetic network is consistent from a topological point of view if the twelve dimensions of GHuST are similar to the real network it stands for.

4.4. Analyzing the topology of synthetic networks

This section analyzes the topology of four sets of non-real power networks: ACTIVSg, Columbia University synthetic network, PEGASE, and SDET. Those network models are available in the open-access repository DR Power [91].

The ACTIVSg and the Columbia University synthetic networks are spatial networks; their nodes are geographically distributed. Both sets of networks result from two novel algorithms proposed to generate spatial synthetic power grids (those algorithms are fully described in Chapter 5). Consequently, an in-depth topological analysis would also help to understand the behavior and effectiveness of both algorithms to generate realistic synthetic power networks. The PEGASE and the SDET networks are non-spatial, and there is no information about the generation process that has been followed to create them. Accordingly, the GHuST framework would highlight exclusively the topological differences or similarities between those networks and the European transmission networks used as a reference.

4.4.1. ACTIVSg

This set of six synthetic grids stands for some parts of the North American power grid. Network size ranges from 200 nodes to 70,000 nodes. Those networks have been developed in the context of the US ARPA-E Grid Data research project [20]. Both topological and electrical considerations drive the generation process of those networks [59].

A. ACTIVSg 200

This 200-node network is located in the central part of Illinois (US), and it includes two voltage levels: 230 kV and 115 kV. The 230-kV network has 17 nodes, and the 115-kV network has 134 nodes. The rest of the nodes are low-voltage nodes (13.8 kV) connected to the transmission network through transformers; they are leaf nodes. To analyze the structure of this network, we compare the ACTIVSg-200 network with the 220-kV European networks. Since transformers are not included in the reference networks, we also analyze the topology of the ACTIVSg 200 after removing low-voltage nodes. The values of GHuST are shown in Table 4-4.

We see that in the analysis of the entire network (including low-voltage nodes), the twelve dimensions are consistent with the reference. Although some dimensions are in the first or third quartile, all of them are in the reference range. However, after removing low-voltage nodes, the reduction in the number of leaf nodes leads (ρ_2) out of the range. Moreover, the tendency of hubs to be connected among them (ρ_5) is also extremely high. This might be caused by a lower value of maximum node degree in the synthetic network (differences in the degree distribution of the synthetic and the reference networks).

From the analysis of each voltage level independently, we can conclude that the 230-kV network has an extraordinarily large number of triangles. The value of ρ_8 (0.12) is three times larger than the mean value for the fifteen European countries. Similarly, the number of nodes that is part of a triangle, ρ_{10}, (47%) is also far from the mean value of the European 220-kV power networks (26%).

The number of lines per node that is an input of the algorithm is consistent with the reference in all cases. The 115-kV layer has a lower number of lines per node than the 220-kV network. The properties of the 115 kV are expected to be slightly different, and we have no reference to compare them. However, all values are in the range defined by the 220-kV reference networks.

Although values of GHuST are in the range in most cases, the presented differences may lead to infer that the algorithm generates a more homogenous structure than the observed in the European case. With a similar number of lines per node, this mesh has a lower number of nodes with one connection and a higher number of triangles. This might be caused by the use of the Delaunay triangulation to build the network, as explained in the next chapter.

Table 4-4. GHuST values for the ACTIVSg 200 network

	ρ_1	ρ_2	ρ_3	ρ_4	ρ_5	ρ_6	ρ_7	ρ_8	ρ_9	ρ_{10}	ρ_{11}	ρ_{12}
Entire network	0.18	0.43	0.48	0.09	0.71	0.42	0.45	0.02	0.21	0.16	0.06	0.40
Excluding low-volt. Nodes	0.23	0.19	0.53	0.20	0.82	0.47	0.49	0.03	0.21	0.21	0.06	0.66
115 kV network	0.15	0.25	0.48	0.17	0.77	0.53	0.51	0.03	0.15	0.17	0.09	0.61
230 kV network	0.19	0.33	0.40	0.35	0.70	0.42	0.40	0.12	0.33	0.47	0.25	0.65

ρ_i is the dimension i of the GHuST framework.
Green values are in the second or in the third quartile, orange values are in the first or in the fourth quartile, red values are out of the range.

B. ACTIVSg 500

The ACTIVSg-500 network is located in the northwestern part of South Carolina, and it comprises two voltage levels: 345 kV and 138 kV. As in the prior case, this synthetic network also includes 13.8 kV nodes. Before applying the GHuST framework, we focus on network size. The number of nodes in the 138-kV network is more than twelve times larger than in the 345-kV network. In the reference networks, the maximum ratio between the number of nodes in both voltage levels is lower than 4. This difference may lead to misleading results when analyzing the network as a whole. In this case, GHuST values for the entire network are compared with the 400-kV and 220-kV reference networks. We compare the 345-kV synthetic network with the 400-kV reference networks and the 138-kV synthetic network with the 220-kV reference networks. The values of GHuST are shown in Table 4-5.

In the case of considering the 345-kV and 138-kV networks together, we find several inconsistencies. The number of nodes that are part of a string, ρ_6, is low. The average length of those strings, ρ_7, is shorter than the reference. Moreover, the number of triangles, ρ_8, the number of nodes that are vertices of a triangle, ρ_{10}, and the percentage of isolated triangles, ρ_{11}, are out of the range. The rest of the dimensions are in the third or fourth quartile in the case of the entire network.

As in the prior case, in the 138-kV network, we observe that the number of nodes in a string is low. The rest of the values are in range. For the 345-kV network, the only inconsistency is related to string length, ρ_7.

According to prior results, the large number of 138-kV nodes with low values of ρ_8, ρ_{10}, ρ_{11} leads to low values of those dimensions in the entire network. As in the ACTIVSg-200 network, we observe that the algorithm tends to make a large number of triangles for the 345-kV network (ρ_8 and ρ_{10} are in the fourth quartile), and the percentage of shared vertices is low. However, the 135-kV network is featured by an extremely low number of triangles.

Table 4-5. GHuST values for the ACTIVSg 500 network

	ρ_1	ρ_2	ρ_3	ρ_4	ρ_5	ρ_6	ρ_7	ρ_8	ρ_9	ρ_{10}	ρ_{11}	ρ_{12}
Entire network	0.14	0.57	0.30	0.07	0.68	0.09	0.00	0.01	0.11	0.08	0.02	0.33
Excluding low-volt. nodes	0.17	0.46	0.30	0.14	0.68	0.13	0.07	0.01	0.11	0.10	0.02	0.47
138 kV network	0.11	0.49	0.29	0.13	0.67	0.18	0.11	0.01	0.12	0.08	0.07	0.46
345 kV network	0.18	0.25	0.48	0.31	0.84	0.45	0.50	0.06	0.08	0.35	0.36	0.64

ρ_i is the dimension i of the GHuST framework.
Green values are in the second or in the third quartile, orange values are in the first or in the fourth quartile, red values are out of the range.

C. ACTIVSg 2000

This network is located in the State of Texas, and the voltage levels included are 500 kV, 230 kV, 161 kV, and 115 kV. As in prior cases, it also includes low-voltage nodes (24 kV, 22 kV, 20 kV, 28 kV, 13.8 kV, and 13 kV). However, in the real ERCOT system, there is no 500-kV nor 230-

kV power network. The results of the topological study are shown in Table 4-6.

In the analysis of all voltage levels together (including low-voltage nodes), the tendency of hubs to be connected among them, ρ_5, is out of the range, and the percentage of triangles in the network (ρ_8) is almost null. Indeed, the percentage of nodes that are vertices of a triangle (ρ_{10}) is 2% (the minimum value of this dimension in the reference networks is 12%). Moreover, the percentage of shared vertices (ρ_9) and the percentage of isolated vertices (ρ_{11}) are out of the range.

In the case of excluding the low-voltage nodes, ρ_2 is close to zero. That means that most nodes have at least two connections. Consequently, the presence of leaf nodes is scarce after removing transformers; only 7 nodes out of 1,151 were connected to the network through one transmission line. This is unusual in the analyzed networks, where the mean value of ρ_2 is 0.32. Additionally, ρ_3 and ρ_4 are low (first quartile) and ρ_5 is high (fourth quartile). This might reflect differences in the degree distribution of the synthetic network in comparison with the reference networks. Furthermore, triangle properties are not consistent with the reference.

When considering the 500-kV network independently, only ρ_3 and ρ_{12} are also out of range. This reinforces the idea about the presence of inconsistencies related to the degree distribution. High values of ρ_3 and ρ_{12} might reflect that the maximum node degree of that network is low in comparison with the reference. This network has the same problem with triangles than the prior cases. In the 115-kV and 161-kV networks, the value of ρ_2 (leaf nodes) is 0.11 below the minimum value in the 220-kV reference networks. It is expected that the lower the voltage level, the lower the mesh and the higher the number of leaf nodes. Accordingly, the lack of triangles in this subnetwork might not be so relevant.

As in prior synthetic networks, there is a significant difference between the number of nodes in the 500-kV network (120 nodes) and the 230-kV, 161-kV, 115-kV networks (1,431 nodes). The topology of this network reinforces the idea that the algorithm used may lead to homogenous structures with a low number of leaf nodes. Thus, it cannot replicate the clustering (triangulation) of real networks.

Table 4-6. GHuST values for the ACTIVSg 2,000 network

	ρ_1	ρ_2	ρ_3	ρ_4	ρ_5	ρ_6	ρ_7	ρ_8	ρ_9	ρ_{10}	ρ_{11}	ρ_{12}
Entire network	0.25	0.23	0.46	0.04	0.86	0.39	0.33	0.00	0.00	0.02	0.03	0.26
Excluding low-volt. nodes	0.30	0.00	0.24	0.09	0.88	0.42	0.35	0.00	0.00	0.02	0.03	0.40
115 kV network	0.20	0.09	0.26	0.16	0.85	0.52	0.43	0.00	0.00	0.03	0.05	0.54
161 kV network	0.18	0.07	0.30	0.22	0.83	0.53	0.43	0.00	0.00	0.01	0.17	0.63
230 kV network	0.20	0.15	0.24	0.11	0.79	0.49	0.30	0.00	0.00	0.02	0.00	0.46
500 kV network	0.22	0.12	0.41	0.28	0.87	0.45	0.40	0.01	0.00	0.05	0.17	0.70

ρ_i is the dimension i of the GHuST framework.
Green values are in the second or in the third quartile, orange values are in the first or in the fourth quartile, red values are out of the range.

D. ACTIVSg 10k

This network stands for the US part of the Western Electricity Coordinating Council (WECC) system. The ACTIVSg-10k network comprises seven voltage levels: 115 kV, 138 kV, 161 kV, 230 kV, 345 kV, 500 kV and 765 kV as well as distribution nodes (1 kV, 13.2 kV, 13.8 kV, 18 kV, 20 kV, 22 kV and 24 kV). However, there is no 765-kV network in the actual WECC transmission power network. As in prior cases, the GHuST framework is calculated for the entire network, and for each voltage level independently, results are shown in Table 4-7.

In the case of the entire network, we observe that ρ_3, ρ_8, ρ_{10}, and ρ_{11} are out of the reference range. The degree of leaf-node connections, the number of triangles, the percentage of nodes that are part of a triangle, and the percentage of isolated triangles are not consistent with the European networks. There is only 6% of nodes that are vertices of triangles, and the number of isolated triangles is null. Furthermore, the low values of ρ_4 (tendency to make hubs) and the high value of ρ_5 (hubs tend to be connected among them) question the degree distribution of the synthetic network.

In the analysis of each voltage level independently, we see that the number of leaf nodes is higher in the 138-kV, 161-kV network than in 345-kV, 500-kV, and 765-kV networks. As mentioned, it is expected that the higher the voltage level, the higher the mesh and the lower the number of nodes with one connection. In this test case, triangulation in the 765-kV and 500-kV networks is consistent with the reference networks.

As in the case of ρ_1, the number of leaf nodes is lower in the 115-kV, 138-kV, and 161-kV networks. Those values are in the fourth quartile of the 220-kV reference network.

Although these networks have a lower number of evident inconsistencies (red cells), it continues showing differences that require a more in-depth analysis before concluding the accuracy of the network.

Table 4-7. GHuST values for the ACTIVSg 10,000 network

	ρ_1	ρ_2	ρ_3	ρ_4	ρ_5	ρ_6	ρ_7	ρ_8	ρ_9	ρ_{10}	ρ_{11}	ρ_{12}
Entire network	0.18	0.36	0.20	0.03	0.78	0.32	0.17	0.01	0.18	0.06	0.07	0.26
Excluding low-volt. Nodes	0.20	0.26	0.17	0.05	0.79	0.38	0.20	0.01	0.18	0.07	0.10	0.31
115 kV network	0.17	0.29	0.27	0.11	0.77	0.38	0.23	0.02	0.25	0.10	0.13	0.47
138 kV network	0.13	0.29	0.30	0.14	0.75	0.43	0.24	0.01	0.07	0.05	0.15	0.51
161 kV network	0.13	0.27	0.34	0.20	0.73	0.45	0.22	0.01	0.00	0.06	0.14	0.59
230 kV network	0.15	0.39	0.30	0.17	0.80	0.41	0.21	0.05	0.33	0.23	0.18	0.54
345 kV network	0.14	0.37	0.20	0.07	0.78	0.40	0.27	0.02	0.21	0.13	0.14	0.35
500 kV network	0.15	0.39	0.37	0.16	0.78	0.35	0.29	0.04	0.22	0.20	0.11	0.51
765 kV network	0.16	0.39	0.40	0.26	0.77	0.35	0.29	0.03	0.23	0.17	0.13	0.65

ρ_i is the dimension i of the GHuST framework.
Green values are in the second or in the third quartile, orange values are in the first or in the fourth quartile, red values are out of the range.

E. ACTIVSg 25k

This synthetic network has been referred to as the Northeast and Mid-Atlantic regions of the US. It includes 69-kV, 100-kV, 115-kV, 138- kV, 161-kV, 230-kV, 345-kV, 500-kV and 765-kV networks. For the topological analysis, we only consider the 230-kV network that is compared with the 220-kV reference network and the 345-kV, 500-kV and 765-kV networks that are compared with the 400-kV reference network. We do not include the analysis of the entire network due to its large size and the inclusion of low voltage levels that might lead to misleading results. In this case, several components (disconnected graphs) may appear for each voltage level; for instance, the 100-kV network is formed by 13 components. The 230-kV network has three components, one of them is formed but a few lines, and it is not considered in the analysis. Table 4-8 shows the value of GHuST for those networks.

We see that the 230-kV network has a low number of lines per node; its structure is close to the minimum spanning tree. The number of leaf nodes is consequently high, and the average length of strings is also low (high value of ρ_7). In the case of ρ_3 the value of the first component (1,444 nodes) is out of range and ρ_4 is low. This directly leads to question the degree distribution. Furthermore, the percentage of triangles and the number of nodes that are part of a triangle are low. Indeed, only 6% of nodes are vertices of a triangle (in the reference network the mean value is 22%).

In the first component of the 765-kV network (218 nodes), the number of lines per node is extremely low ($\rho_1 = 0.05$), this value is similar to the 230-kV network. The number of nodes with one connection is high (they are in range). As in prior cases, main inconsistencies come from low values of ρ_3 and ρ_4 and high values of ρ_5 as well as triangle-related dimensions. The low values of ρ_1 may lead to low-connected structures. This contrasts with the topology of prior synthetic networks with meshed structures in which the number of leaf nodes was low.

Table 4-8. GHuST values for the ACTIVSg 25,000 network

	ρ_1	ρ_2	ρ_3	ρ_4	ρ_5	ρ_6	ρ_7	ρ_8	ρ_9	ρ_{10}	ρ_{11}	ρ_{12}
230 kV network Comp. 1	0.06	0.45	0.21	0.10	0.85	0.41	0.16	0.01	0.13	0.06	0.06	0.47
230 kV network Comp. 2	0.06	0.46	0.38	0.24	0.69	0.32	0.00	0.01	0.00	0.07	0.00	0.67
345 kV network	0.13	0.49	0.30	0.19	0.87	0.34	0.06	0.03	0.25	0.15	0.02	0.63
500 kV network	0.09	0.45	0.21	0.10	0.86	0.39	0.13	0.02	0.29	0.08	0.01	0.49
765 kV network Comp. 1	0.05	0.46	0.33	0.20	0.84	0.40	0.12	0.02	0.08	0.10	0.18	0.56
765 kV network Comp. 2	0.18	0.47	0.28	0.19	0.88	0.36	0.13	0.07	0.74	0.15	0.00	0.67

ρ_i is the dimension i of the GHuST framework.
Green values are in the second or in the third quartile, orange values are in the first or in the fourth quartile, red values are out of the range.

F. ACTIVSg 70k

This network has been referred to as the Eastern US. The voltage levels included are the same that in the ACTIVSg-25k network (69-kV, 100-kV, 115-kV, 138-kV, 161-kV, 230-kV, 345-kV, 500-kV and 765-kV). In this synthetic network the number of components for each voltage

level is higher: 4 components in the 230-kV network, 3 components in the 345-kV network, 4 components in the 500-kV networks and 2 components in the 765-kV network. As in the ACTIVSg 25k network, components with a low number of nodes are not included in the analysis that is shown in Table 4-9.

The topological inconsistencies found are similar to the inconsistencies displayed by the ACTIVSg-25k case. The number of lines per node is low in all the subnetworks. This ratio is higher in the 138-kV and 115-kV subnetworks. This contrasts with the European networks. Similarly, although values of ρ_3, ρ_4 and ρ_5 might be in range in some cases, they are close to the limits. The degree of leaf-node connections is low in most cases, the tendency to make hubs is high in the 500-kV networks, and the tendency to connect hubs among them is extremely high. Consequently, degree distribution may differ concerning the reference networks.

Additionally, triangle properties continue to be incoherent. For instance, in the 230-kV network, the number of triangles is considerably low. The percentage of nodes that are vertices of a triangle is just 7% (the mean value in the reference network is 22%). If we compare the 230-kV network with the entire 220-kV European network (including all countries as one network), the percentage of nodes that are vertices of a triangle is 18%.

The 365-kV, 500-kV, and 765-kV networks have high values of leaf-nodes. Three components of the 500-kV network have a structure with the minimum number of lines to have a connected graph. Although those values are in the reference range, they are far from the median value. In those networks, values of ρ_5 are high and ρ_{10} is inconsistent. Finally, the number of triangles is only consistent in the first component of the 765-kV network.

Table 4-9. GHuST values for the ACTIVSg 70,000 network

	ρ_1	ρ_2	ρ_3	ρ_4	ρ_5	ρ_6	ρ_7	ρ_8	ρ_9	ρ_{10}	ρ_{11}	ρ_{12}
230 kV network Comp. 1	0.06	0.45	0.16	0.06	0.89	0.42	0.12	0.02	0.30	0.07	0.02	0.37
230 kV network Comp. 2	0.09	0.41	0.23	0.10	0.84	0.41	0.22	0.01	0.16	0.07	0.08	0.46
345 kV network	0.10	0.45	0.24	0.15	0.92	0.41	0.13	0.02	0.37	0.09	0.01	0.56
500 kV network Comp. 1	0.07	0.42	0.25	0.16	0.90	0.45	0.15	0.02	0.27	0.08	0.02	0.60
500 kV network Comp. 2	0.00	0.33	0.33	0.44	0.71	0.83	0.40	0.00	0.00	0.00	0.00	0.67
500 kV network Comp. 3	0.00	0.38	0.33	0.47	0.78	0.80	0.25	0.00	0.00	0.00	0.00	0.67
500 kV network Comp. 4	0.00	0.39	0.29	0.26	0.91	0.64	0.33	0.00	0.00	0.00	0.00	0.40
765 kV network Comp. 1	0.11	0.48	0.32	0.19	0.91	0.36	0.15	0.04	0.33	0.17	0.06	0.59
765 kV network Comp. 2	0.03	0.50	0.35	0.23	0.86	0.47	0.18	0.02	0.00	0.09	0.17	0.64

ρ_i is the dimension i of the GHuST framework.
Green values are in the second or in the third quartile, orange values are in the first or in the fourth quartile, red values are out of the range.

4.4.2. Key points about the topology of ACTIVSg networks

The application of the GHuST framework to the ACTIVSg networks shows that in most cases there are inconsistencies with respect to the European transmission networks. On the one hand, those inconsistencies may be a consequence of the model used to generate the

networks. On the other hand, North American power networks may display different topologies concerning European networks. Consequently, it would be necessary to apply the GHuST framework to the real North American power grid to have a better comparison with synthetic networks. However, data for real North American power grids are not available. With the information available, we can only compare synthetic North American power grids with real European power grids.

Regarding results, we observed that some networks display a homogenous structure with a low number of leaf nodes. This contrasts with the analysis of the European transmission power networks. Furthermore, we detect inconsistencies regarding the tendency to make hubs and the tendency to connect hubs among them. This might reflect that the algorithm is not able to fit a realistic degree distribution. This is crucial because of the impact of node degree on network vulnerability, as discussed in Chapter 2.

The algorithm is not able to build topologies in which triangles are similar to the reference networks. This is related to the complexity of local structures that might condition the operation and robustness of networks.

As previously explained, those inconsistencies do not mean that the topology of those synthetic networks is inaccurate. However, further studies would be required to use the proposed model to generate European synthetic power grids.

The ACTIVSg synthetic networks are not topologically consistent with respect to the real European networks. Number of leaf nodes, tendency to make hubs and to connect among them as well as triangulation are the main inconsistencies found in most cases.

4.4.3. Columbia University synthetic power grid with geographical coordinates

The Network Imitating Method Based on Learning (NIMBLE) is used to generate a synthetic network based on the properties of the North American and Mexican power networks (this algorithm is explained in Chapter 5) [58], [90]. The resulting network stands for the Western Interconnection (WI) power network that includes the Western Electricity Coordinating Council in the United States and the Western Electricity Coordinating Council in Canada [92].

Regarding connections, the dataset includes information about the end nodes of connections and line/transformer impedance. However, it does not differentiate between lines and transformers, and there is no information about the voltage level. According to the ACTIVSg-10k network that stands for the US part of the WECC system (it does not include the Canadian portion), the Columbia University synthetic power grid may include 765-kV, 500-kV, 345-kV, 230-kV, 161-kV, 138-kV, and 115-kV transmission networks.

The lack of voltage information hinders the topological validation of the network since we cannot compare the topology by voltage level. The values of GHuST are shown in Table 4-10; the network is compared with the 400-kV and 220-kV reference network.

Table 4-10. GHuST values for the Columbia University synthetic network

	N	ρ_1	ρ_2	ρ_3	ρ_4	ρ_5	ρ_6	ρ_7	ρ_8	ρ_9	ρ_{10}	ρ_{11}	ρ_{12}
Columbia U.	14430	0.14	0.37	0.16	0.03	0.77	0.45	0.28	0.01	0.10	0.10	0.31	0.23

ρ_i is the dimension i of the GHuST framework.
Green values are in the second or in the third quartile, orange values are in the first or in the fourth quartile, red values are out of the range.

Our analysis shows that the number of lines installed, ρ_1, is close to the minimum value in the European transmission networks. This might be a consequence of the inclusion of low voltage networks, with a low meshed structure. The value of ρ_2 is close to the median. Consequently, the number of leaf nodes is consistent with the reference. However, we might expect a higher number of leaf nodes, when considering lower voltage levels. The value of ρ_3 and ρ_4 is extremely low. The degree of leaf-node connections is low, and there is no tendency to make hubs in the networks. Those low values might be a consequence of the maximum value of node-degree. The highest number of connections in the synthetic network is 16, and the average maximum node degree in the reference networks is 9.4. Consequently, low values of ρ_3 and ρ_4 show inconsistencies regarding degree distribution. Moreover, the tendency of hubs to be connected among them, ρ_5, is high (fourth quartile). Regarding strings, ρ_6, ρ_7 are consistent with the reference.

In the paper presenting this synthetic grid, the authors show that the network average clustering coefficient is similar to the real network one. However, ρ_8 is low. We introduced in Chapter 3 that the use of the network average clustering coefficient might be misleading. Furthermore, the average degree of triangles is lower than in the ACTIVSg (both networks are supposed to stand for the same real power grid), we may think that in real network triangles may be located in higher degree nodes. Thus, this would lead to having similar values of network average clustering coefficient with a lower number of triangles in the network. We confirm this hypothesis with the absolute value of ρ_{12} (without being scaled). This is 3.9 in the Columbia Synthetic network and 4.48 in the ACTIVGs case.

The validated network average clustering coefficient for this synthetic network (WI system) is 0.048. However, Cotilla et al. state that the network average clustering coefficient of the WI is 0.073 [39]. Characteristic path length also presents divergent values in both works. Assumptions used to model the networks as a graph or voltage levels included may be the cause of this difference. The authors should look through it in order to do an accurate validation of the synthetic network. As in the ACTIVSg-10k case, triangle properties are not consistent with the European reference. In the Columbia network, only 10% of nodes are vertices of a triangle (ρ_{10}), and only 10% of vertices are shared among triangles (ρ_9). Furthermore, the number of isolated triangles (ρ_{11}) is high.

The NIMBLE model only uses a topological criterion to generate synthetic networks. Beyond the electrical features that might be consistent with real networks, the topology is not realistic in comparison with the European transmission networks. As in the prior case, it would be necessary to check if those inconsistencies lie on the model itself or in the topology of the North American Power grid.

The Columbia-University synthetic network displays topological inconsistencies with respect to the European transmission networks, and to another synthetic network that stands for the same real network. It also diverges from other published studies.

4.4.4. PEGASE

A set of networks were designed to represent the European transmission network. In the context of a European Commission project, the Pan European Grid Advanced Simulation and State Estimation (PEGASE) aims to work in the field of real-time control and operation planning of the pan-European network [93]. Five fictitious, albeit realistic, networks form the PEGASE test cases [94]. The voltage levels included, and the number of components for each voltage level is the following:

- PEGASE 89: 150 kV (2 components), 220 kV (2 components), 380 kV (1 component).
- PEGASE 1354: 220 kV (30 components), 380 kV (2 components).
- PEGASE 2869: 110 kV (2 components), 150 kV (17 components), 220 kV (34 components), 380 kV (2 components).
- PEGASE 9241: 110 kV (47 components), 120 kV (7 components), 150 kV (24 components), 154 kV (14 components), 220 kV (53 components), 380 kV (4 components), 400 kV (1 component), 750 kV (1 component).
- PEGASE 13659: 110 kV (47 components), 120 kV (7 components), 150 kV (154 components), 220 kV (53 components), 380 kV (4 components).

Since PEGASE networks stand for the Pan-European network and node location is not given (we cannot split up the network by country), we compare the PEGASE networks with the values of GHuST obtained for the network that includes the 220-kV and 400-kV network of the fifteen countries. The values of GHuST for the ENTSO-e network are shown in Table 4-11 and for PEGASE networks in Table 4-12. The ENTSO-e network is compounded of the 24 countries included in the 2016 TYNDP [12].

Table 4-11. GHuST values for the Continental Europe 220-kV and 400 kV network

	N	ρ_1	ρ_2	ρ_3	ρ_4	ρ_5	ρ_6	ρ_7	ρ_8	ρ_9	ρ_{10}	ρ_{11}	ρ_{12}
ENTSO-e	5757	0.28	0.32	0.22	0.03	0.76	0.41	0.26	0.03	0.25	0.29	0.29	0.21

ρ_i is the dimension i of the GHuST framework

Table 4-12. GHuST values for the PEGASE networks

	ρ_1	ρ_2	ρ_3	ρ_4	ρ_5	ρ_6	ρ_7	ρ_8	ρ_9	ρ_{10}	ρ_{11}	ρ_{12}
PEGASE 89	0.57	0.26	0.42	0.18	0.88	0.32	0.13	0.22	0.95	0.40	0.08	0.54
PEGASE 1354	0.21	0.46	0.34	0.08	0.76	0.37	0.21	0.02	0.17	0.16	0.20	0.34
PEGASE 2869	0.28	0.34	0.30	0.07	0.79	0.40	0.26	0.04	0.55	0.23	0.21	0.30
PEGASE 9241	0.35	0.22	0.10	0.02	0.80	0.48	0.37	0.17	0.92	0.24	0.21	0.13
PEGASE 13659	0.27	0.48	0.16	0.02	0.76	0.42	0.35	0.12	0.92	0.16	0.19	0.15

ρ_i is the dimension i of the GHuST framework.
Green values are in the second or in the third quartile, orange values are in the first or in the fourth quartile, red values are out of the range.

The PEGASE-89 case is an example of network reduction. Accordingly, topology is not expected to fit the properties of the real network. Indeed, we observe that the number of lines installed per node (0.57) is much larger than the maximum value of ρ_1 for the European countries (0.34). Furthermore, as PEGASE 89 includes the 110-kV network, we might expect that the total number of lines per node would be lower than in the 220-kV and 400-kV. In this test case, we also see the presence of more complex local structures. Indeed, there is a high tendency to make clusters. The proportion of triangles is seven times larger than in the reference network. Moreover, those triangles tend to share vertices among them (95% of triangle vertices are part of two or more triangles).

In the rest of PEGASE networks, the number of lines installed per node is consistent with the reference. Although we observe some small differences, for example, in the PEGASE-13,659 network the number of leaf nodes is huge (this might be the consequence of including 110-, 120- and 150-kV network), the main inconsistencies are related to triangles. In the PEGASE-1,354 network and in the PEGASE-2,869 network, the proportion of triangles is according to the reference. However, in the PEGASE-9,241 network and PEGASE-13,659 network, where the presence of triangles is expected to be low, ρ_8 is really high. This significant presence of triangles contrasts with the percentages of nodes that are part of a triangle, ρ_{10}, which is lower than the reference. So, those networks do not represent the complexity of local structures.

Finally, there are topological differences concerning the network provided by ENTSO-E. Although PEGASE networks include lower voltage levels, they display a more complex structure than the ENTSO-E network. Since location is not provided, we cannot detect if those inconsistencies are presented in all countries or not. Furthermore, ENTSO-E classifies networks according to Continental Europe, Baltics, and Great Britain. The networks used as reference are countries that are part of Continental Europe. PEGASE networks might include other countries. This might introduce some deviation concerning the reference. Finally, since all PEGASE networks stand for the same real network, but they have different sizes; they might be the result of network reduction or clustering techniques.

There are slight differences between the PEGASE networks and the European transmission networks. They might be the consequence of network reduction techniques or the voltage levels and countries included in the analysis.

4.4.5. Sustainable Data Evolution Technology (SDET)

The Sustainable Data Evolutionary Technology project aims to develop: "Evolvable open-access large-scale datasets to accelerate the development of next-generation power grid optimization" [95]. They introduce the concept of Data Evolvability in order to "disrupt the current ad hoc cycles of static dataset generation". However, there is no explanation about the implications of this novel concept in the generation of synthetic power networks.

Although four synthetic networks have been published in an open-access repository (SDET 500, SDET 2000, SDET 3000, SDET 4000), there is no information about the methodology followed in the synthetic network generation. There is only a succinct presentation about the project, where authors state that the generation of synthetic networks would be based on real-data anonymization and algorithms that are based on graph theory [96]. We assume that these networks are preliminary results, since their objective is to generate networks with more than 100,000 nodes, and the size of those four networks ranges from 588 to 4,661 nodes.

The voltage levels and the number of components of each voltage level are the following:

- SDET 500: 500 kV (2 components), 345 kV (4 components), 230 kV (1 component), 161 kV (1 component), 138 kV (9 components), 69 kV (5 components).
- SDET 2000: 500 kV (5 components), 345 kV (11 components), 138 kV (67 components), 69 kV (14 components), 66 kV (13 components).
- SDET 3000: 500 kV (3 components), 345 kV (12 components), 138 kV (40 components), 115 kV (3 components), 110 kV (4 components), 66 kV (2 components).
- SDET 4000: 500 kV (5 components), 345 (11 components), 138 kV (150 components), 66 kV (14 components).

From the analysis of voltage levels and components, we observe a high number of components for each voltage level. In the SDET 3000, the number of components of the 345-kV network is 4 times larger than the 115-kV network. It contrasts with the European network, where the 220-kV network (around 3400 nodes) has 47 components, and the 400-kV network (around 2100 nodes) has 5 components. Thus, the assignment of voltage level should be checked, since a large number of components might have a direct consequence on network operation.

We have applied the GHuST framework to the 500-kV, 345-kV, 230-kV components with more than five nodes. Results are shown in Table 4-13 to Table 4-16.

Table 4-13. GHuST values for the SDET 500 network

	ρ_1	ρ_2	ρ_3	ρ_4	ρ_5	ρ_6	ρ_7	ρ_8	ρ_9	ρ_{10}	ρ_{11}	ρ_{12}
Entire network	0.13	0.25	0.29	0.08	0.76	0.58	0.47	0.01	0.00	0.04	0.14	0.45
230 kV network	0.00	0.23	0.37	0.18	0.49	0.80	0.56	0.00	0.00	0.00	0.00	0.40
345 kV network Comp. 1	0.05	0.24	0.36	0.23	0.48	0.81	0.46	0.00	0.00	0.00	0.00	0.40
345 kV network Comp. 2	0.09	0.31	0.44	0.40	0.81	0.45	0.30	0.02	0.00	0.09	0.33	0.67
500 kV network Comp. 1	0.17	0.10	0.75	0.39	0.64	0.67	0.33	0.00	0.00	0.00	0.00	0.50
500 kV network Comp. 2	0.05	0.21	0.44	0.31	0.75	0.67	0.50	0.00	0.00	0.00	0.00	0.50

ρ_i is the dimension i of the GHuST framework.
Green values are in the second or in the third quartile, orange values are in the first or in the fourth quartile, red values are out of the range.

Table 4-14. GHuST values for the SDET 2,000 network

	ρ_1	ρ_2	ρ_3	ρ_4	ρ_5	ρ_6	ρ_7	ρ_8	ρ_9	ρ_{10}	ρ_{11}	ρ_{12}
Entire network	0.18	0.27	0.32	0.05	0.79	0.55	0.45	0.01	0.19	0.05	0.16	0.33
345 kV network Comp. 1	0.14	0.24	0.26	0.05	0.76	0.60	0.44	0.01	0.17	0.06	0.13	0.41
345 kV network Comp. 2	0.00	0.50	0.50	1.00	0.50	1.00	0.50	0.00	0.00	0.00	0.00	1.00
345 kV network Comp. 3	0.00	0.71	0.68	0.55	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.40
345 kV network Comp. 4	0.16	0.24	0.40	0.16	0.67	0.55	0.29	0.01	0.00	0.07	0.33	0.57
345 kV network Comp. 5	0.00	0.81	0.82	0.60	0.07	0.33	0.00	0.00	0.00	0.00	0.00	0.25
345 kV network Comp. 6	0.00	0.71	0.68	0.55	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.40
500 kV network Comp. 1	0.00	0.75	0.67	0.65	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.40
500 kV network Comp. 2	0.13	0.50	0.47	0.08	0.77	0.48	0.35	0.00	0.00	0.04	0.17	0.40
500 kV network Comp. 3	0.13	0.54	0.53	0.09	0.73	0.37	0.13	0.00	0.00	0.05	0.00	0.33
500 kV network Comp. 4	0.06	0.33	0.55	0.42	0.61	0.40	0.25	0.00	0.00	0.00	0.00	0.50

ρ_i is the dimension i of the GHuST framework.
Green values are in the second or in the third quartile, orange values are in the first or in the fourth quartile, red values are out of the range.

Table 4-15. GHuST values for the SDET 3,000 network

	ρ_1	ρ_2	ρ_3	ρ_4	ρ_5	ρ_6	ρ_7	ρ_8	ρ_9	ρ_{10}	ρ_{11}	ρ_{12}
Entire network	0.22	0.38	0.21	0.01	0.77	0.50	0.39	0.06	0.84	0.11	0.21	0.19
230 kV network Comp. 1	0.00	0.45	0.53	0.67	0.50	0.50	0.00	0.00	0.00	0.00	0.00	0.67
230 kV network Comp. 2	0.00	0.33	0.44	0.44	0.50	0.83	0.60	0.00	0.00	0.00	0.00	0.67
345 kV network Comp. 1	0.19	0.32	0.38	0.11	0.77	0.48	0.30	0.05	0.52	0.20	0.29	0.45
345 kV network Comp. 2	0.00	0.11	0.33	0.42	0.68	0.88	0.71	0.00	0.00	0.00	0.00	0.67
345 kV network Comp. 3	0.00	0.60	0.56	0.67	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.67
345 kV network Comp. 4	0.00	0.40	0.50	1.00	0.50	1.00	0.67	0.00	0.00	0.00	0.00	1.00
345 kV network Comp. 5	0.33	0.00	0.80	0.67	0.00	0.00	0.00	0.38	0.67	0.83	0.60	0.68
500 kV network	0.35	0.23	0.24	0.05	0.75	0.48	0.34	0.14	0.87	0.25	0.12	0.33

ρ_i is the dimension i of the GHuST framework.
Green values are in the second or in the third quartile, orange values are in the first or in the fourth quartile, red values are out of the range.

Table 4-16. GHuST values for the SDET 5,000 network

	ρ_1	ρ_2	ρ_3	ρ_4	ρ_5	ρ_6	ρ_7	ρ_8	ρ_9	ρ_{10}	ρ_{11}	ρ_{12}
Entire network	0.19	0.23	0.27	0.05	0.80	0.56	0.47	0.01	0.11	0.05	0.20	0.33
345 kV network Comp. 1	0.18	0.21	0.25	0.05	0.80	0.51	0.44	0.00	0.13	0.03	0.10	0.39
345 kV network Comp. 2	0.00	0.56	0.60	0.63	0.21	0.00	0.00	0.00	0.00	0.00	0.00	0.50
345 kV network Comp. 3	0.00	0.81	0.82	0.60	0.07	0.33	0.00	0.00	0.00	0.00	0.00	0.25
345 kV network Comp. 4	0.07	0.33	0.47	0.38	0.71	0.53	0.30	0.02	0.00	0.11	0.00	0.83
345 kV network Comp. 5	0.00	0.60	0.56	0.67	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.67
345 kV network Comp. 6	0.00	0.39	0.42	0.35	0.82	0.57	0.63	0.00	0.00	0.00	0.00	0.50
345 kV network Comp. 7	0.00	0.50	0.50	1.00	0.50	1.00	0.50	0.00	0.00	0.00	0.00	0.00
500 kV network	0.16	0.38	0.39	0.10	0.78	0.40	0.32	0.01	0.00	0.06	0.19	0.41

ρ_i is the dimension i of the GHuST framework.
Green values are in the second or in the third quartile, orange values are in the first or in the fourth quartile, red values are out of the range.

The values of GHuST show that SDET networks are not topologically consistent with the European transmission networks. Indeed, most dimensions are out of the range or in the third and fourth quartile.

Although the SDET networks do not provide the location, we might assume that they stand for the North American power network. The values of GHuST are also far from the values of the ACTIVSg or the NIMBLE networks. Results might lead to think that these networks have been published without being tested. Since there is no information about the power networks they stand for, or the methodology followed to generate those networks, we cannot provide a sound analysis of topological inconsistencies.

The SDET networks are completely inconsistent with the topology of the European transmission power network.

4.5. Takeaways

The application of the GHuST framework enhances the topological characterization of power networks. The analysis of the European transmission networks shows that differences in the structure of the 400-kV and the 220-kV networks are low. Nevertheless, the network location clearly conditions the topology of power grids. Those differences are apparent in the twelve dimensions of the GHuST framework. Moreover, results are consistent with the global statistics used in Chapter 2.

Furthermore, the GHuST framework allows for the validation of synthetic power grids. This chapter has analyzed the topology of four sets of synthetic networks: ACTIVSg, Columbia-University synthetic network, PEGASE, and SDET. The ACTIVSg and the Columbia-University synthetic network stand for North American power networks, and they display topological inconsistencies concerning the European networks. We cannot state if those differences are a

consequence of the model used to generate the networks or if the real North American power network has a different structure. However, we observe that both sets of networks have topological differences among them. Consequently, it is highly likely that the models used to generate them cannot replicate the topology of real networks accurately. We find that those algorithms have difficulties to replicate the complexity of local structures, that is, triangles. In most cases, dimensions related to triangles are out of the reference ranges. Furthermore, the analysis reveals problems that might be related to the degree distribution of those networks (e.g., the number of leaf nodes or the tendency to make hubs in the network).

The PEGASE networks also present some differences concerning the European transmission networks. They have a large number of triangles, and they are not well distributed through the network. However, since all PEGASE networks stand for the same real network (but they have different sizes) we may infer that the use of network-reduction techniques may lead to topological differences concerning the real network. Finally, the SDET networks are entirely inconsistent with the topology of the European transmission networks. The lack of information about the real network they represent, or the algorithm used to generate those networks makes it impossible to determine the cause of those inconsistencies.

The GHuST framework is a useful tool to analyze the topology of synthetic networks since it allows for the detection of inconsistencies. Furthermore, those inconsistencies are easily interpretable, so this analysis may support the introduction of changes in the existing algorithms to improve results.

5

A NOVEL ALGORITHM TO GENERATE SYNTHETIC POWER GRIDS

5.1. What are synthetic grids?

The generation of synthetic power grids is a practical alternative to the lack of public network models. Synthetic networks are non-real, albeit realistic, power grids that are topologically and electrically consistent with a real network. Accordingly, the operation and control of synthetic networks are similar to real networks.

Chapter 4 analyzes the topology of existing synthetic power networks. It shows that published synthetic-network algorithms are not topologically consistent with the European transmission power networks. Indeed, we found some topological properties of real networks that were not replicated in synthetic networks (existing algorithms cannot imitate the complexity of local structures in power networks). Furthermore, some of them lack node location, which is essential in applications such as Transmission Expansion Problem.

This chapter makes a review of the existing algorithms to generate synthetic networks, and it proposes a novel algorithm for the generation of synthetic spatial power grids (synthetic networks in which nodes are endowed with geographical location). The algorithm is articulated in two steps:

1. The first step focused on building a basic network to meet generation and demand.
2. The second step targeted at increasing network robustness whilst achieving topological attributes.

We also showed that different power networks might have different topologies, so any synthetic generation procedure must be adjustable in order to generate representative grids. The proposed algorithm has adjustable parameters that enable it to generate synthetic power grids with different topological properties.

The rest of this chapter is organized as follows. Section 2 reviews existing works. We introduce a new algorithm to generate synthetic power grids in section 3. Section 4 presents a case study. Finally, section 5 summarizes results.

Synthetic power grids are non-real, albeit realistic, power grids that are topologically and electrically consistent with real power grids. They are a practical alternative to the lack of public power-network models.

A novel algorithm is proposed to generate spatial synthetic power grids.

The model is flexible enough to adapt results to different topologies.

5.2. State-of-the-art review

The generation of synthetic networks has attracted the attention of several studies in the complex-network field. Those works focus on the generation of networks (graphs) that fit with some topological properties. Based on their nature, we differentiate between purely topological algorithms and hybrid models. The pure topological algorithms connect nodes disregarding network nature (in this case, the electrical nature of power grids). Those networks lack electrical parameters. Hybrid models combine complex-network techniques with power-system methods to generate synthetic networks that are endowed with electrical information.

5.2.1. Purely topological algorithms

Based on the idea "the rich get richer", Barabási and Albert presented their preferential attachment model [37]. In this model, nodes are consecutively added to the network and linked to existing ones. The probability of being linked to a node is correlated with the degree of existing nodes so that already well-connected nodes have a higher probability of being selected for new links. The preferential attachment model generates scale-free networks, networks in which node degree follows a power-law. As discussed in Chapter 2, the power grid cannot be considered a scale-free network.

Wats and Strogatz presented a method to generate small-world networks [44]. These networks are characterized by having a high network average clustering coefficient concerning random networks and small characteristic path length. As in the case of the scale-free network, Chapter 2 questions the systematic characterization of transmission power networks as small-world networks.

Other models to generate pure topological networks include the Erdős–Rényi model that generates random networks [43]. However, the topological properties of power grids do not fit with the structure of random networks. Moreover, several versions or prior algorithms have been proposed to generate synthetic networks [36]. However, the lack of electrical considerations makes them an inaccurate tool to build synthetic power networks. All prior models generate networks without considering their electrical nature. They may lead to evident inconsistencies such as demand nodes that cannot satisfy demand because of a lack of transmission capacity.

Purely topological algorithms generate synthetic networks disregarding the electrical nature of power grids. This may lead to inconsistencies regarding network operation or control.

5.2.2. Hybrid models

Hybrid models combine complex-network theory with power-system principles; they may be classified concerning the network location. We differentiate between algorithms that endow nodes with geographical location and algorithms that do not consider the spatial nature of power grids.

A. Non-spatial models

The *RT-nested-Smallworld* is an algorithm that generates synthetic power grids based on purely statistical information [21]. Based on the assumption that power networks are small-world networks, it generates synthetic networks with electrical features such as line impedance. The algorithm was improved by the introduction of an electrical classification of nodes into load, generation, or connection nodes [97], [98]. However, based on prior results, power networks cannot be always described as small-world networks [38], [89].

The *cluster-and-connect* model generates synthetic networks based on purely topological information and can potentially fit any degree distribution [99]. Nevertheless, it takes an existing network as a starting point and merely reshuffles its connections. Scaling to new sizes is not possible with this algorithm. Although the resulting networks have the same target degree distribution, other topological properties are not tested. This may make them completely different from a topological point of view.

Despite not considering the geographical location of nodes, other approaches have introduced the distance between pairs of nodes as a design parameter, the probability of linking two nodes decreases as distances increases [100], [101].

The prior models, as well as most pure topological algorithms, disregard the geographical location of nodes. Therefore, the resulting synthetic grids do not bear any geographical significance. Node location is a crucial factor in applications such as Transmission Expansion Planning.

B. Spatial models

Based on the distance among nodes, Patania et al. and Wang et al. go one step beyond, and they propose algorithms to effectively generate spatial networks [102], [103]. In the first case, the *Epsilon-disc* model connects nodes if the distance between them is below a specific limit. In the second case, lines are distributed following a length distribution that should be introduced as an input. As distance is correlated with the cost of installing a line, the decision

process resorts to a soft economic approach. The Epsilon-disc model also includes the electrical characterization of resulting synthetic grids. However, as they show in their paper, the topological properties of resulting networks do not fit well with the properties of real networks. The idea of considering distance as an economic criterion has been used in other studies such as in the model proposed by Deka and Vishwanath [104]. The degree distribution of resulting networks tends to follow an exponential function.

In an attempt to generate a synthetic spatial network similar to the US power network in properties, a model was proposed to make an in-depth electrical characterization of lines and nodes [105]. Based on the North American Eastern Interconnect grid, lines are linked using the Delaunay triangulation. Although the network could fit a realistic degree distribution, the authors do not check if it fits other topological metrics. A further improvement of this work was proposed where an iterative process decides which line should be added to the system [22], [59]. The algorithm chooses the set of lines that better contribute to the performance of the resulting grid in terms of power flow. The set of candidate lines is also based on the Delaunay triangulation. Besides, this model only considers the average degree as a topological input. The ACTIVSg networks showed in Chapter 4 are generated with this algorithm. As shown in that chapter, the topology of those networks is not consistent with the properties of the European transmission networks. The use of an average node degree as the unique topological input is not sufficient to ensure the topological consistency of resulting networks. Moreover, this model always generates the same network for the same set of nodes. Consequently, this model cannot adapt to the heterogeneity of power-network topologies shown in Chapter 4.

Schultz et al. [106] present an algorithm that first generates a minimum spanning tree. Second, it adds new lines to connect nearby nodes. However, these assumptions are not consistent with the historical evolution of power networks leading to unrealistic topologies. A similar approach is the base of the *NIMBLE* algorithm [90]. First, the algorithm adds nodes and connect them to their closest nodes to form a connected graph (the resulting network is not necessarily a minimum spanning tree). Second, new lines are added to the network based on degree distribution is similar to scale-free networks (there are differences in nodes with one connection and highly connected nodes), line length is limited and the higher the density of the area the higher the node degree. This algorithm, therefore, adds line regarding degree distribution. As a result of the algorithm, a synthetic network is published. The analysis carried out in Chapter 4 shows that the network is inconsistent with the topological properties of the European transmission power networks. Furthermore, its structure differs from the topology of other synthetic power grid that stands for the same real network and from prior works.

Despite the introduction of an electrical characterization, existing models for generating synthetic power grids do not provide results that are consistent with the topology of real networks, apart from matching a degree distribution in some particular cases. Considering this, we propose a new algorithm that mimics the evolution of real power networks to generate synthetic spatial power networks. Moreover, existing algorithms are not parametrizable, and they cannot adapt to the structure of the resulting synthetic network to different topologies, as explained in Chapter 4.

Hybrid models generate networks combining complex-network techniques and power-systems theory. Hybrid models may generate spatial or non-spatial networks. This work focuses on spatial networks.

Synthetic power networks generated with existing spatial algorithms are not topologically consistent with real networks. Furthermore, those algorithms are not flexible to adapt network structure to different topologies.

5.3. Algorithm description

This section proposes a new algorithm to generate spatial synthetic power grids. Although node generation is described, the novelty of the algorithm lies in the wiring process (how nodes are connected). Furthermore, this section highlights the need for a parametrical algorithm to generate networks with different topologies.

5.3.1. The need for a parametrical algorithm

The growth of power networks, as a case of infrastructure networks, is conditioned by topological, morphological, technical, economical, permitting, environmental, managerial or political factors [88]. These factors are not aligned with the optimal power flow method, and different network topologies may appear depending on countries or regional areas (as shown in Chapter 2 and Chapter 4). Besides, because of network evolution, power plants that were built decades ago may become underused, and new generation investments are allocated. Accordingly, the generation of synthetic power grids cannot be tackled with optimization models as performed in transmission expansion planning. Following this, a non-parametrical algorithm cannot be the solution for generating synthetic power grids. Even when we have the same set of nodes, we might generate different topologies due to geography, political decisions or electricity generation mix. We, therefore, propose a parametrical algorithm that is flexible enough to adopt different topologies depending on economic and technical factors.

This new algorithm considers the economic and technical dimensions as the most relevant factors that explain the structure of power networks. Those factors guide the construction of a base network in which demand is supplied. Accordingly, we generate networks that meet with the design target. However, their structure may differ from real networks since other relevant factors such as environmental constraints were not considered. Once the algorithm has provided a network that can meet demand, and it is robust in case of some component failures, new lines are added to achieve topological consistency.

The heterogeneous topology of transmission power networks leads to the need for flexible algorithms to generate synthetic power grids. Accordingly, algorithms should be able to generate networks with different topologies.

The proposed algorithm considers technological and economic considerations as the most relevant factors that guide network design.

5.3.2. Node Generation

The algorithm starts with node creation; it distributes substations where power is demanded, or generators are connected. The algorithm considers two types of nodes: demand nodes (which include interconnection substations) and generation nodes. If there are demand and generation connected to a node, this is classified as demand or generation node based on the balance between power demand and generation capacity.

Since power networks are spatial, the algorithm can assign node location according to a spatial probability distribution function. Geographical characteristics, the availability of the fuel or renewable resources, as well as transportation infrastructures (e.g., gas pipelines), might determine the location of nodes. Probability functions must be introduced as an input.

After locating nodes, the algorithm endows them with electrical properties. In the case of demand, the algorithm requires the total power demanded by the network and how it is distributed (for instance, with an exponential distribution function [102]). Demand can also be distributed based on economic considerations such as GDP per region [59]. It should be noted that, although magnitudes such as GDP per region or renewable resource availability might be chosen to represent a specific region, this is not necessary for the algorithm –nodes can be generated randomly based on any distribution defined by the user.

Once demand is set, the algorithm addresses generation features. Total generation capacity and electricity generation mix are parameters of the model; it will distribute generators in the area where the synthetic network is generated under those parameters.

As explained previously, this thesis does not introduce a novel methodology regarding the random generation of nodes and their characteristics; it assumes the prior work described in the literature. The novelty of the proposed methodology lies in the creation of the network also called the wiring process.

5.3.3. Building a connected graph

Once nodes are set, the algorithm links them with a basic network in which demand is met. In this first step, the algorithm tries to minimize network costs while preserving the physical principles that govern power networks. Accordingly, the decision to install a line is based on

installation cost (that is assumed to be proportional to the distance between two nodes) and line contribution to demand-supply (transmission-line-capacity constraints). If line cost linearly increases with distance, a minimum spanning tree would be the cheapest solution to have a connected graph. However, this is only valid if all lines installed have similar properties. In the case of considering a set of different transmission lines, e.g. different transmission capacity, the minimum spanning tree neglects cost differences. Furthermore, a minimum spanning tree does not consider how power flows through lines. This might lead to electrical inconsistencies such as line overloads. We, therefore, propose the combination of an economic criterion (installation cost) with electrical considerations.

The inclusion of different transmission line capacities (with different cost), rapidly increases the complexity (number of variables and constraints) of the minimum spanning tree problem. Accordingly, it cannot be tackled with classical optimization techniques. We apply the divide-and-conquer scheme to build a connected graph. The algorithm divides the network into small subnetworks that are connected afterward. The solution to this might not be the optimal solution to the problem. Nevertheless, as previously mentioned, we are not trying to build an optimal network but a reasonable one, one that has properties that are similar to the real network. The construction of the connected graph is divided into three stages: node clustering, intra-cluster connection, and inter-cluster connection (as shown in Figure 5-1).

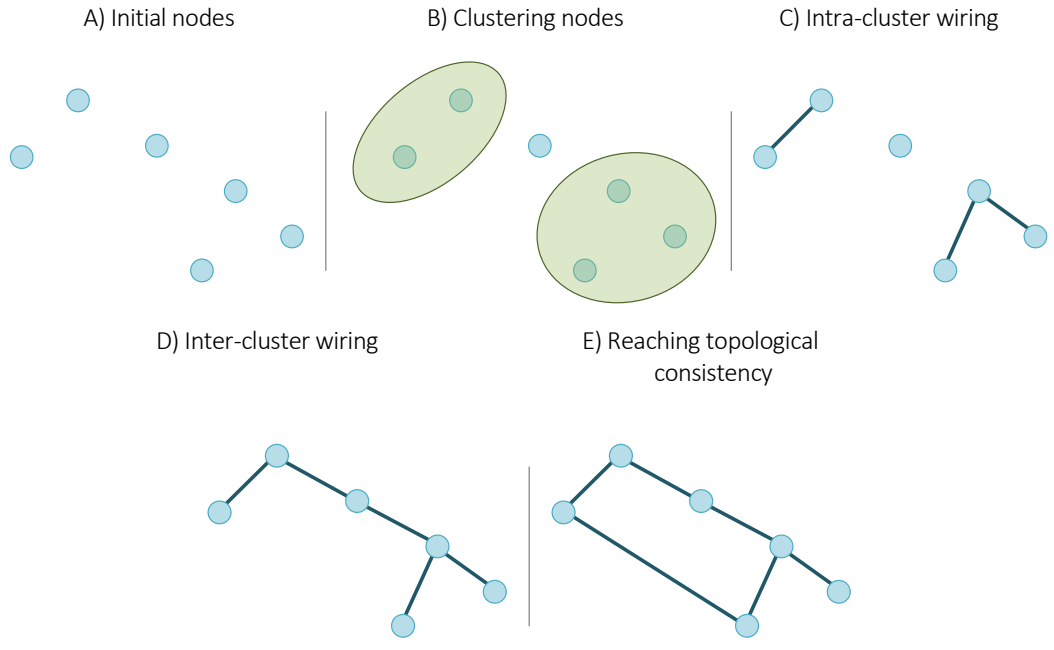


Figure 5-1. Steps followed by the model to build a synthetic power network.

A. Clustering nodes

To reduce problem size, the model groups demand nodes with the closest generator that is able to supply its demand. Each generation node is defined as the center of a cluster.

Consequently, the number of clusters in the network is equal to the number of generation nodes. The number of generation nodes may differ from the number of generators, since more than one generator may be connected to the same node.

To define the nodes that belong to each cluster, the algorithm minimizes the sum of distances $D_{i,j}$ between each demand node (N_D) and generators (N_G) (5-1). The variable $\alpha_{i,j}$ determines whether a demand node i is connected to a cluster (generator) j or not.

$$\min \sum_{i \in N_D} \sum_{j \in N_G} D_{i,j} \alpha_{i,j} \tag{5-1}$$

Each cluster should be able to supply the demand of its nodes. Consequently, the sum of node demand (PD_i) must be lower than the generation capacity of that cluster (G_j) (5-2).

$$\sum_{i \in N_D} \alpha_{i,j} PD_i \leq G_j \qquad \forall\, j \tag{5-2}$$

Finally, a demand node may be connected to several generators. Accordingly, $\alpha_{i,j}$ is an integer variable that represents the proportion of demand that is satisfied by each generator (5-3).

$$\sum_{j \in N_G} \alpha_{i,j} = 1 \qquad \forall\, i \tag{5-3}$$

This local approach tries to mimic the origin of power networks, in which power networks were built relatively close to demand nodes.

This step clusters demand nodes around generators. Each cluster has enough generation capacity to supply the demand of its nodes. Demand nodes might be assigned to more than one cluster.

B. Intra-cluster wiring

The algorithm connects the nodes of each cluster with the minimum-cost network that is able to supply demand (5-4). Accordingly, the decision ($\beta_{i,j,k}$) of connecting nodes i and j (belonging to cluster N_C) with a line of transmission capacity k is conditioned by the distance among nodes $D_{i,j}$, and the cost of the different transmission line considered C_k. To build the network, the model chooses the type of lines to be installed based on a line catalog introduced as an input. Different transmission capacities are considered.

$$\min \sum_{i \in N_C} \sum_{j \in N_C} \sum_{k \in K} \beta_{i,j,k} D_{i,j} C_k \tag{5-4}$$

Because of network model assumptions explained in Chapter 2, there is only one line that can connect nodes i and j (5-5).

$$\sum_{k \in K} \beta_{i,j,k} \leq 1 \qquad \forall\, i,j \tag{5-5}$$

To ensure that demand is met, we estimate power flows $f_{i,j,k}$. Because of the non-meshed structure expected as a result of this step, the power flow through a line is calculated considering only the first Kirchhoff's law (5-6). Accordingly, the effect of the second Kirchhoff's law is neglected. This simplifies the problem solution. To estimate the flow through lines, PD_i is the demand of each node and PG_i is the power injected by each generator ($PG_i = \sum_{j \in N_C} PD_j$). OL and IL are the set of outgoing lines and incident lines connected to node i.

$$\sum_{j \in OL} \sum_{k \in K} f_{j,i,k} + PG_i = \sum_{j \in IL} \sum_{k \in K} f_{i,j,k} + PD_i \quad \forall\, i \tag{5-6}$$

Furthermore, the flow through a line has an upper and lower bound fixed by the transmission line capacity L_k (5-7).

$$\underline{L_k}\, \beta_{i,j,k} \leq f_{i,j,k} \leq \overline{L_k}\, \beta_{i,j,k} \qquad \forall\, i,j,k \tag{5-7}$$

To ensure that all clusters are connected subgraphs, interconnection nodes are assigned with a small demand. Since there is only one generator per cluster and equation (5-6) ensures that demand is supplied, all nodes are therefore connected.

This process is repeated for all the clusters defined in the prior stage. Finally, the network is formed by a set of connected subnetworks that are disconnected among them. Since demand nodes may be connected to more than one generator, the number of connected subnetworks may differ from the number of initial clusters.

Although several clusters might be connected after this step, there is no guarantee that the resulting network is a connected graph.

Each cluster is connected with the minimum-cost network that is able to satisfy demand. The design of that network includes power-flow considerations.

Although some clusters may be linked, the resulting network is in general a disconnected graph.

C. Inter-cluster wiring

To build a connected graph, the connection of disconnected clusters (those clusters that do not share any demand node) is the last step. Reliability considerations, as well as economic criteria, will lead to the installation of new lines.

After linking demand and generation nodes with a basic network (demand is met under normal operating conditions), the algorithm tries to increase network robustness. The algorithm tries to find alternative sources to supply demand in case of generator failures. First, each cluster tries to find a backup cluster (or generator). Second, if there are no close generators with enough capacity to supply demand in case of failure, the algorithm installs the lines that connect clusters at the minimum cost.

The first stage is inspired by the generation N-1 criterion. This reliability principle ensures that demand will be met in case of generator failure. Consequently, if one of the generators fails, there should be another available generator to supply demand. This new generator should be able to inject the same amount of power as it was being injected by the failed generator. The process to find a backup generator starts with the estimation of cluster reserve margin CRM (5-8). This is the difference between the generation capacity of each cluster and the total demand in that cluster. Accordingly, the cluster reserve margin is the ability of each cluster to supply other clusters in case of failure.

$$CRM_i = \sum_{i \in N_C} PD_i \, \alpha_{i,j} - G_i \tag{5-8}$$

In the case of connected clusters in the prior stage (intra-clustering wiring), that reserve is first assigned to the cluster to which they are connected. Consequently, clusters that share a demand node have backup generators that meet the total or partial needs of power in case of failure. Subsequently, each cluster tries to find the backup cluster iteratively. It starts from the largest clusters in terms of power demand (sum of node demand of a cluster). It assumes that the larger the cluster size the larger the bargaining power of that cluster. If no cluster can supply the total demand, it finds the cluster that can supply the maximum amount of power. The flowchart for this process is shown in Figure 5-2.

The connection between clusters is limited by cluster distance (maximum length of transmission lines). The distance among clusters is the distance among their closest nodes.

Once the connection of clusters is fixed, the model installs the line with the lowest installation cost that links both clusters. In this stage, the connection among clusters disregards power-flow constraints. The capacity of the new transmission line is fixed based on the expected power flow from the backup cluster in case of failure.

If there is not a close cluster that can supply demand in case of failure, or the resulting network is not a connected graph, clusters that remain disconnected C are linked with the lowest investment option. Consequently, the model minimizes the investment cost of new lines to have a connected graph (5-9). Since power flows are not considered, cost correlates with the distance between nodes $CD_{i,j}$.

$$\min \sum_{i \in C} \sum_{j \in C} \gamma_{i,j} \, CD_{i,j} \tag{5-9}$$

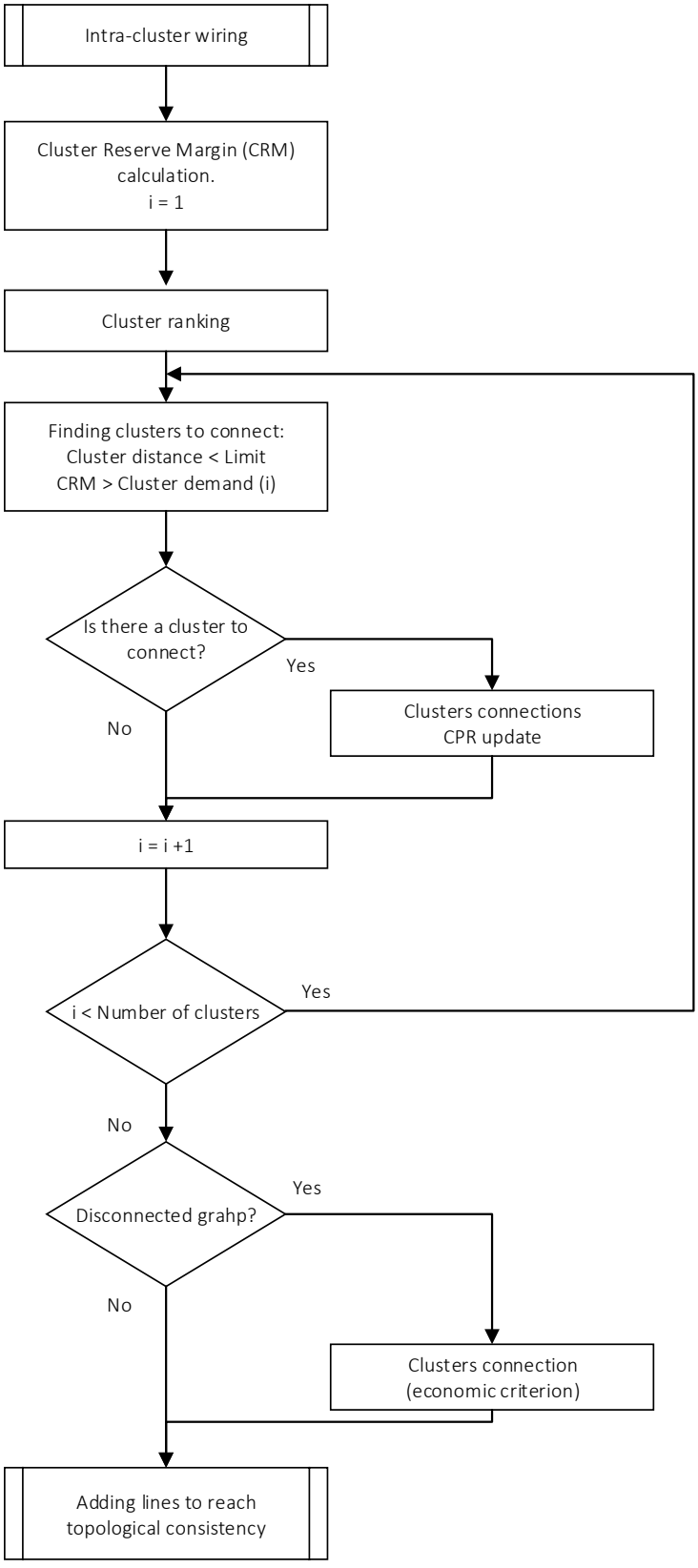


Figure 5-2. Flowchart of the inter-cluster-wiring stage.

To ensure a connected graph, the problem has a similar constraint as that used in the prior stage with the first Kirchhoff's law. In a connected graph, the demand of all nodes could be satisfied with a single node (disregarding capacity constraints). Accordingly, the algorithm solves the theoretical flow $u_{i,j}$ through the connections among clusters (5-10). It defines a power withdrawal vector W in which all clusters (except the generation cluster) have a small demand (for instance one unit of power), and the injection vector I, in which only one node satisfies all the theoretical demand (this is equal to the number of nodes minus one). Then, the decision to install a line is conditioned by the maximum length of transmission lines (5-11).

$$\sum_{j \in C} u_{j,i} + I_i = \sum_{j \in C} u_{i,j} + W_i \qquad \forall\, i \tag{5-10}$$

$$\gamma_{i,j}\, CD_{i,j} < L \qquad \forall\, i,j \tag{5-11}$$

Finally, the model allows for the iterative inclusion of generators based on technology. This attempts to mimic the historical development of power systems. The historical evolution of generators was articulated around periods characterized by the installation of single dominant generation technology. First, thermal plants of different technologies (coal, nuclear or natural gas) were installed followed by renewable plants. Accordingly, this step allows the user to introduce nodes iteratively. In each iteration, new generation nodes are connected.

Disconnected clusters are linked based on reliability considerations. Clusters find a backup cluster to supply demand in case of generator failure.

If the resulting network is not a connected graph, the model installs lines that make a connected graph at the minimum cost.

5.3.4. Adding lines to reach topological consistency

Once the algorithm has provided a basic network where demand is satisfied at low investment cost, the model focuses on increasing network robustness by adding new lines. The model adds new connections trying to reach a target in terms of topology. Previously, we have shown that the multiple factors that guide the real evolution of power networks are not effectively replicated by existing models. They are based on power-system considerations and soft topological criteria. We propose the introduction of sophisticated topological criteria as well as power-flow considerations to guide synthetic-network generation. Accordingly, during this step, new lines are added with a double objective. On the one hand, they try to improve network robustness as well as network operation. On the other hand, the installation of a line is conditioned by the expected topology of the synthetic network defined by the GHuST framework.

The inputs of the model used to guide the topological evolution of synthetic power

networks are the expected degree distribution and the expected values of GHuST dimensions for the synthetic network. Chapter 2 pointed out the importance of degree distribution due to its implications in terms of network vulnerability. We also showed that the use of a global statistic such as degree distribution might result in misleading. However, the twelve dimensions of the GHuST framework provide a complete description of network topology. Accordingly, by including the GHuST framework and the degree distribution together, we will be able to reach a topological consistency of resulting synthetic power networks.

Furthermore, to overcome the drawbacks of pure topological methodologies explained in section 5-2, power-flow considerations guide the installation of those lines. Accordingly, resulting networks are expected to be both topologically and electrically consistent with real networks.

This step is divided into three stages:

1. *Preventing islands:* the algorithm tries to install lines that reduced network vulnerability.
2. *Guiding node degree:* lines are added individually to improve network operation.
3. *Reaching GHuST consistency:* lines are added to mimic those aspects of network evolution that cannot be introduced in the algorithm.

A. Preventing islands

Networks resulting from the first step, *building a connected graph*, are expected to have a poorly meshed structure. Accordingly, those networks are highly vulnerable in case of line failure. Indeed, line removal might split the network into two components. Beyond the consequences in terms of network dynamics, this division is critical if there is a deficit of generation in one of those components. Therefore, generation and demand will not meet.

As in the *inter-cluster wiring*, this stage relies on the N-1 criterion. Unlike the inter-cluster wiring that considers generator failure, this stage only focuses on line failures. The algorithm tries to build alternative paths to supply demand in case of connections failure. Accordingly, the installation of new lines will prevent the formation of islands (disconnected components). Figure 5-3 shows the flowchart of this stage.

The complete fulfillment of the N-1 criterion would lead to a network in which all nodes have at least two connections. However, as shown in Chapter 4, the number of leaf nodes in power networks cannot be neglected. Accordingly, the model will try to reach the N-1 approach while preserving topological consistency. The installation of line reinforcements is therefore based on power-system considerations, but it is conditioned by the contribution of that line to network topology. In this stage, the topological contribution is measured through the degree distribution. A line contributes to the degree distribution if the installation of that line helps to reduce the difference between the degree distribution of the synthetic network and the target one. For instance, if the number of leaf nodes in the synthetic network is equal to the number

of leaf nodes in the target distribution, no more lines will be connected to nodes with only one connection. If new lines would be linked to nodes connected to only one transmission line, the number of leaf nodes in the synthetic network will be lower than in the reference network. This error cannot be corrected with the installation of new lines in a further step, and the topology of the synthetic network will not be consistent with real power networks. Consequently, the desired degree distribution conditions the installation of new lines in this stage.

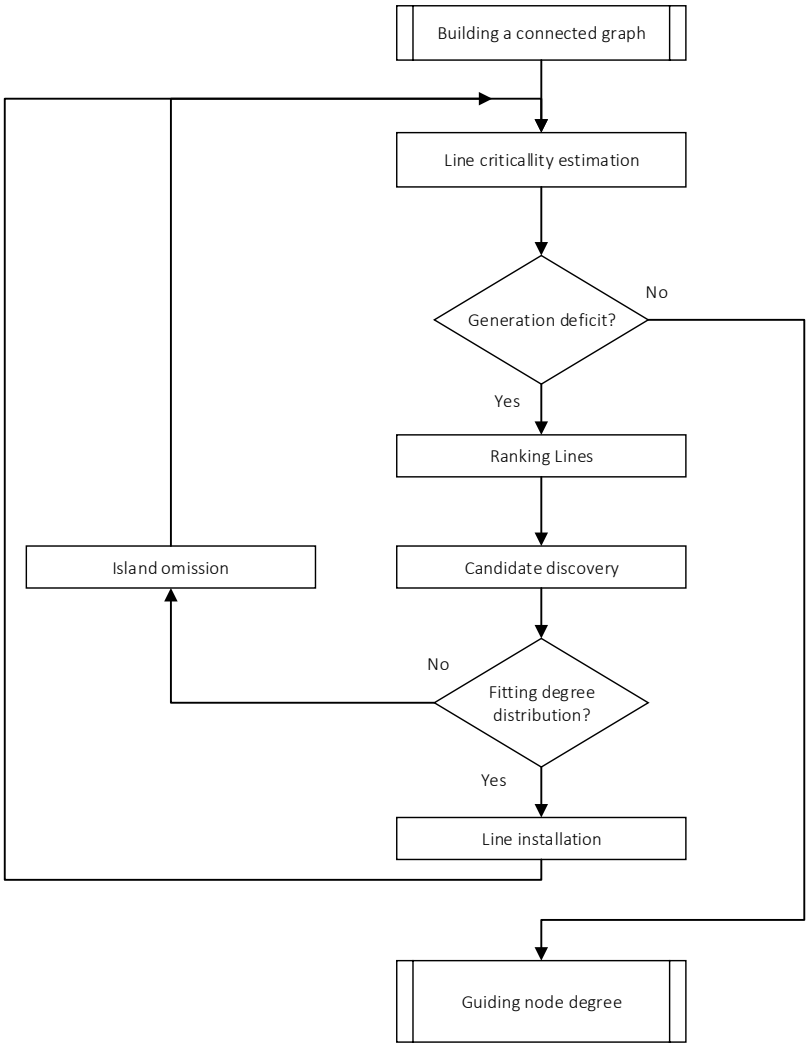


Figure 5-3. Flowchart of the preventing-island stage

Since not all nodes can be connected with two lines, and disconnected components may appear in case of line failure, it is necessary to define a ranking to prioritize line reinforcements. This ranking is based on the impact the failure of an existing line has on the network. We calculate the impact of each line removal in the system with the *line criticality index*. This index is based on the Loss of load index (later explained in Chapter 6) [107]. The line criticality index of a line LC_i is the maximum value of Power Not Served in the disconnected components PNS_j that appear after the failure of that line (5-12). Generation deficit is the difference between the generation capacity G of nodes that belong to component C and the total demand of those

nodes PD (5-13).

$$LC_i = \begin{cases} \max(PNS_j), & \max(PNS_j) \geq 0 \\ 0, & \max(PNS_j) < 0 \end{cases} \quad (5\text{-}12)$$

$$GD_i = \sum_{k \in C} PD_k - \sum_{k \in C} G_k \quad (5\text{-}13)$$

The algorithm only simulates the failure of one line. Accordingly, the maximum number of disconnected components is two. Furthermore, as generation capacity in the network should be higher than total demand, generation deficit may appear in only one of those components.

After ranking existing lines based on their impact on demand-supply, the algorithm looks for the new lines that can mitigate that deficit of generation in case of line failure. Any line that connects both components will avoid the formation of islands. However, the algorithm tries to minimize the impact of new lines in the network. Consequently, a new line should minimize the number of cases in which line failures lead to a disconnected graph.

To decide what is the line that should be installed, the algorithm finds a set of potential candidate lines that would contribute to increasing network reliability. Subsequently, it evaluates those candidates, and it chooses the line that better fits with network requirements.

For each line that leads to a disconnected component with a deficit of generation, the algorithm first figures out the nodes that belong to that component (the origin of candidate lines should be one of those nodes). To maximize the number of cases in which the candidate line mitigates the effect of a disconnected graph, the algorithm chooses as the origin of the candidate lines all leaf nodes of the subgraph. The leaf node that was connected to the line that has caused the disconnected graph is excluded. Figure 5-4 illustrates the process followed. In the case of line failure, a string of four nodes would be disconnected. However, if we connect the leaf node (red node) with the network, we avoid the formation of a disconnected graph in case of removing the three lines of the string (blue lines). Accordingly, the best option to be the origin of candidate lines is the leaf node. However, to meet the constraint related to the number of leaf nodes in the network, we also consider as the origin of the candidate line the connection of the leaf node (yellow node).

Once the origin of the candidate lines is fixed, the algorithm looks for the end of those candidate lines. It is clear that the end would be located in the component where there is not a generation deficit. All those nodes whose distance to the initial node is lower than the maximum line length are considered the end nodes of potential candidate lines. Subsequently, the algorithm analyzes if those lines contribute to the degree distribution or not. This is repeated every time a line is installed. Furthermore, to avoid the connection with close nodes, and to encourage the connection with other network areas, the algorithm imposes a constraint in terms of network distance. This contributes to increase network reliability since the lower the distances in the network the higher network robustness [108]. The pair of nodes that would be directly connected in case of installing a candidate line should be separated by a minimum

distance. That distance is a minimum number of edges regarding the shortest path matrix. Finally, the model chooses the line that connects the furthest nodes. In the case of two lines with a similar contribution, the cheapest line is installed.

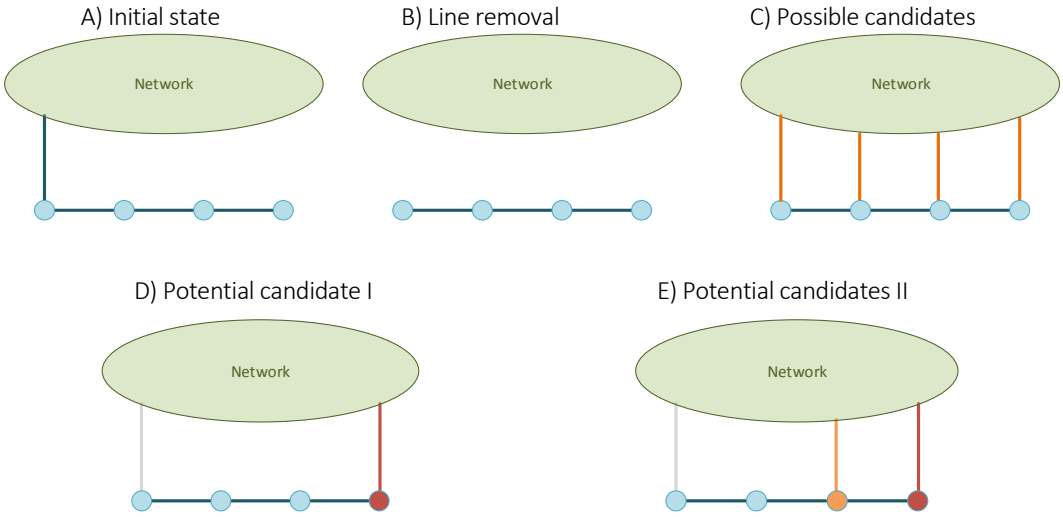


Figure 5-4. An example of the candidate line proposal in the preventing-island stage

New lines are added to increase network reliability. The algorithm builds alternative paths to supply demand in case of line failure.

This stage focuses on avoiding the formation of disconnected components in which there would be a deficit of generation and demand cannot be therefore met.

The installation of new lines is conditioned by the contribution of that line to fit with the desired degree distribution.

B. Guiding node degree

Although the previous step introduced topological considerations, it only analyses the contribution of new lines to the degree distribution of the synthetic network. As explained in Chapter 3, two networks with similar degree distribution may still display significant topological differences. To ensure topological consistency, we propose the introduction of the GHuST framework to guide network generation. However, we do not have a sound understanding of the marginal contribution of an individual line to reach a GHuST target (considering the twelve dimensions). Although the contribution to each dimension looks intuitive, the correlation among GHuST dimensions needs for further research. Furthermore, unlike degree distribution, in the GHuST framework, we do not know how the installation of a line conditions error mitigation in the future. We have no certainty if the error in the GHuST dimensions, which is

introduced by the installation of a new line, can be corrected with further lines. Therefore, the algorithm can only compare the topological consistency of the synthetic network with the GHuST framework if all lines are installed. The algorithm, therefore, evaluates sets of candidate lines to be installed together (instead of a single line).

Chapter 2 showed the average number of lines per node in the European transmission networks, $L \approx 1.33N$. To have a connected graph, the minimum spanning tree installs $N-1$ lines. Although this algorithm does not use the minimum spanning tree, we might expect that the number of lines installed at the building-a-connected-graph step will be similar (or slightly higher). Consequently, the algorithm should add approximately $0.33N$ lines. Despite the number of lines installed in the prior stage (*preventing islands*), the number of new lines and the number of candidate lines (nodes that can be connected and that line contributes to the degree distribution) may result in an extremely large, and an unmanageable number of potential networks in which evaluate the GHuST framework.

To reduce the number of candidate lines as well as the total number of lines to be installed, this stage focuses on installing lines in low-degree nodes. Those lines are installed analyzing its contribution to the degree distribution and base on power-flow considerations. Consequently, this stage disregards the GHuST framework, but it reduces the number of candidate lines in further stages. If the number of leaf nodes in the synthetic network and the target network is the same, all candidate lines that include a leaf node should be omitted. Accordingly, this step starts by installing lines considering leaf nodes and nodes with two connections. The higher the number of lines installed, the lower the number of candidates. However, it also reduces the possibilities to reach the reference in terms of GHuST.

The process followed is described in Figure 5-5. If the number of nodes with a specific node degree X is lower in the synthetic network than in the target, new lines should be installed. Accordingly, all nodes whose node degree is $X-1$ are potential nodes to be the origin/end of candidate lines. Once that candidate lines are found (following the criteria explained in the prior section, contribution to degree distribution and line length), those candidates are evaluated in terms of power flow to decide the line to be installed. This is done iteratively until there is no difference between the synthetic network and the target network.

The algorithm installs the line that makes a reduction in terms of power flow per distance E, which can be considered as a proxy of the optimal transmission expansion decision. It estimates the sum of the product between power flow through a line f_i and the length of that line LL_i (5-14). This measure analyzes the contribution of the line to network operation. In case two lines have the same impact, the cheapest line is installed.

$$E = \sum_{i \in L} f_i \, LL_i \tag{5-14}$$

Transmission line capacities are fixed based on the expected power flow in the optimal economic dispatch.

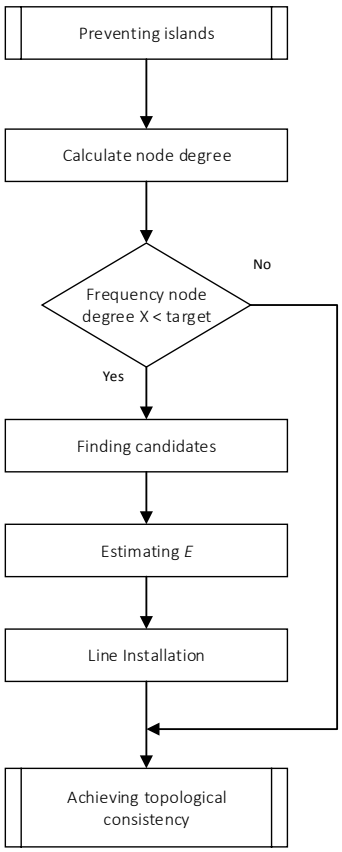


Figure 5-5. Flowchart of the guiding-degree stage

To reduce the number of candidate lines as well as the number of lines to be installed, this stage focuses on the connection of low-degree nodes. Accordingly, it tries to reduce the difference between the degree distribution of the synthetic network and the target.

C. Achieving GHuST consistency

As explained in Chapter 4, existing synthetic networks cannot replicate the complexity of the real power-network structure. Indeed, in most cases, those synthetic networks do not capture the complexity of local structures. Although this model has conditioned the installation of new lines to their contribution to the degree distribution, this is not enough to create networks that are topologically consistent. Accordingly, we have introduced the use of the GHuST framework to guide network generation.

As previously mentioned, new lines should be evaluated together. Accordingly, we propose the generation of a large set of candidate networks that are evaluated afterward.

Those sets of networks are generated with a random process. Since the number of

combinations is not manageable, the algorithm generates a large number of candidate sets to be evaluated. Then, it filters the networks whose topology is consistent with the target. This random process is guided based on electrical and topological criteria. The process is illustrated in Figure 5-6. First, the model determines the number of lines to be installed in the synthetic network (this is obtained from the target degree distribution). It also finds the candidate lines that can be installed based on degree distribution and line length. Subsequently, the model decides randomly the lines that are installed. This is done iteratively until meeting the target degree distribution. The decision to install a line depends on a probability function. Every time a line is installed, the list of candidate lines and probabilities are updated.

The probability of installing a line is conditioned by 3 aspects:

- Estimated power flow, since running a power flow in each iteration would need large computational requirements; we estimate the power through a candidate line based on the state of the synthetic network. Based on the DC power flow definition, power flows through network lines are defined by equation (5-15), where X is the diagonal matrix of line reactance, A the incidence matrix (reduced to the slack bus), P the vector of power injections at each node, and θ are the system voltage angles.

$$F = X^{-1}\, A^T\, \theta \tag{5-15}$$

$$\theta = [A\, X^{-1}\, A^T]^{-1} P \tag{5-16}$$

If we consider that changes in θ after the installation of a new line are small, we can estimate the power through a new line as (5-17).

$$P_{i,j}^{est} = {X_{i,j}}^{-1}\, (\theta_j - \theta_i) \tag{5-17}$$

The probability of installing a line based on the estimated power flow $P_{i,j}^{est}$, is proportional to the maximum power flow estimation for a candidate line (5-18).

$$pf_{i,j} = \frac{P_{i,j}^{est}}{\max(P^{est})} \tag{5-18}$$

- Line length, the probability of installing a line is also conditioned by the length of the line. The model assumes that the line length is the distance between two nodes. The probability of installing large lines may differ from the probability of finding a short line in the network. Those probabilities might vary depending on the country as introduced by Espejo et al. [65].

$$pl_{i,j} = \begin{cases} \alpha, & D_{i,j} \le l_1 \\ \beta, & l_1 < D_{i,j} \le l_2 \\ \gamma, & l_2 < D_{i,j} \le l_3 \\ & \dots \end{cases} \tag{5-19}$$

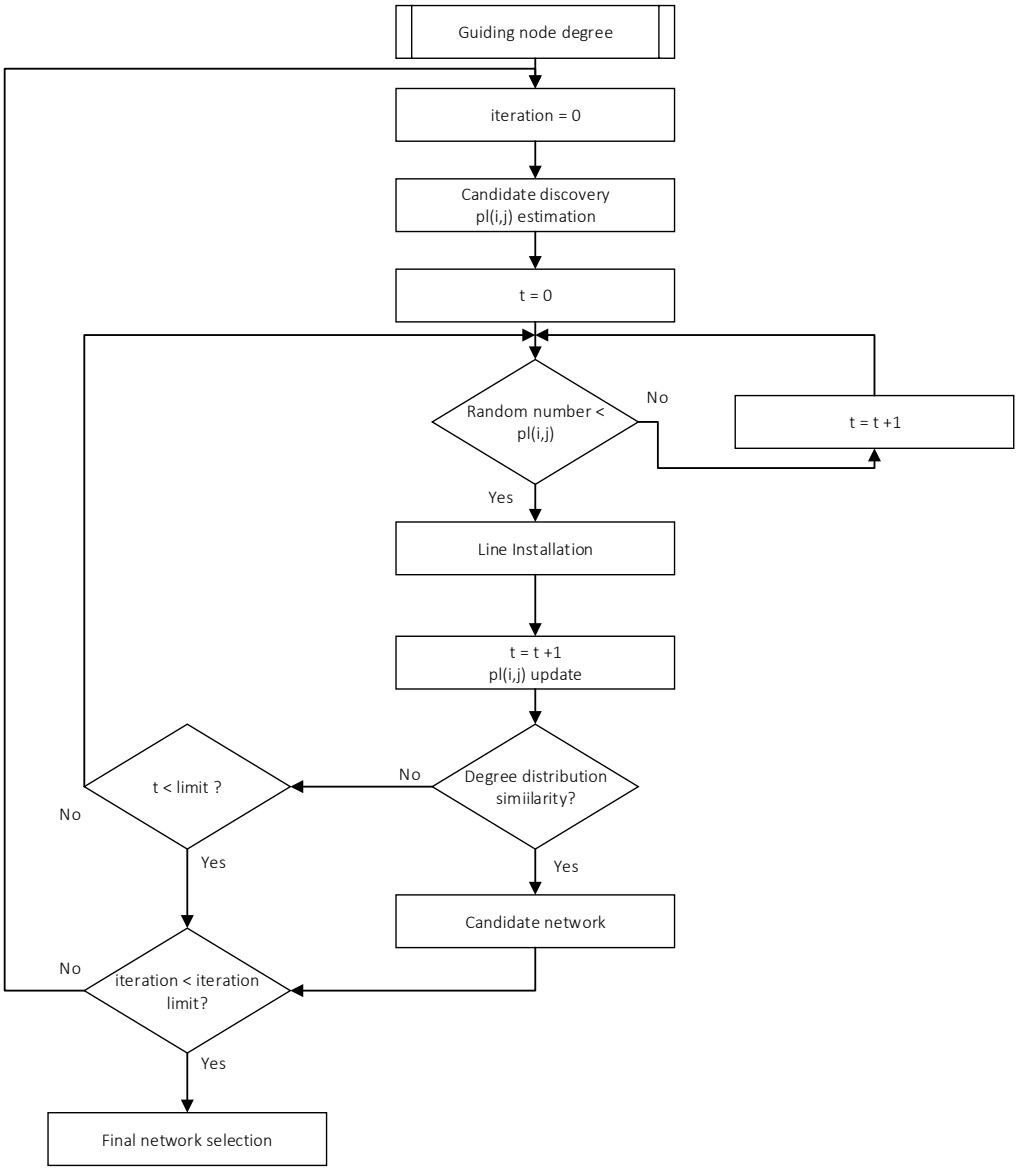


Figure 5-6. Flowchart of the achieving-GHuST-consistency stage

- **Graph distances**, we pointed out the importance of graph distances (minimum number of edges to go from one node to another) concerning network vulnerability. Furthermore, we observe a lack of triangles in existing synthetic networks. Accordingly, the probability of installing a line based on graph distances differentiates two cases. First, if the distance is two edges, the addition of that line would create a triangle. The probability of installing a triangle is fixed by the user with the parameter δ. Finally, if the distance is higher than 2, the higher the distance, the higher the probability of installing that line.

$$pd_{i,j} = \begin{cases} \delta, & d_{i,j} = 2 \\ 1 - \frac{1}{d_{i,j}}, & d_{i,j} > 2 \end{cases} \quad (5\text{-}20)$$

Accordingly, the probability of installing a new line depends on power-flow estimations, the geographical distance and the graph distance between nodes (5-21).

$$p_{i,j} = pf_{i,j}\, pl_{i,j}\, pd_{i,j} \tag{5-21}$$

This process is repeated until the degree distribution of the synthetic network meets with the target. If the degree distribution meets the target, that network is a candidate network to be evaluated. The algorithm also fixes a time limit since there is no guarantee that the synthetic network will meet the target. Following this process iteratively, the algorithm is able to generate a large set of networks with the same degree distribution. However, their topological properties may differ from the target networks.

The model applies the GHuST framework and filters all those networks in which the relative error (concerning the reference network) is below a limit in the twelve dimensions. Subsequently, it chooses the network that minimizes the mean relative error. Accordingly, this process ensures that final networks are topologically consistent with real transmission power networks.

Finally, the algorithm presents enough flexibility to generate synthetic power grids with different topologies. By introducing an electricity generation mix or a line catalog, different topologies may appear during the first step of the model. The result of the second step is conditioned by degree distribution, and the values of GHuST introduced as input. This flexibility is crucial to replicate the historical evolution of power grids as shown previously.

The achieving-GHuST-consistency stage follows an iterative process to generate a set of synthetic networks that have the same degree distribution than the target network.

The GHuST framework is applied to choose the candidate network that better fits with the target topology.

5.4. A synthetic network for Spain, Portugal, and France

The objective of this section is to test the ability of the proposed algorithm to generate synthetic power grids that are consistent with the topology of real power networks. We use three cases: the Spanish 400-kV network, the Portuguese 400-kV network, and the French 400-kV network. The Spanish network is composed of 235 nodes and 334 lines, the Portuguese network has 69 nodes and 93 lines, and the French network is formed by 217 nodes and 283 lines. The Spanish and Portuguese networks are obtained from the 2014 TYNDP (2030 scenario) of ENTSOE-e [12]. The French case is obtained from RTE [13].

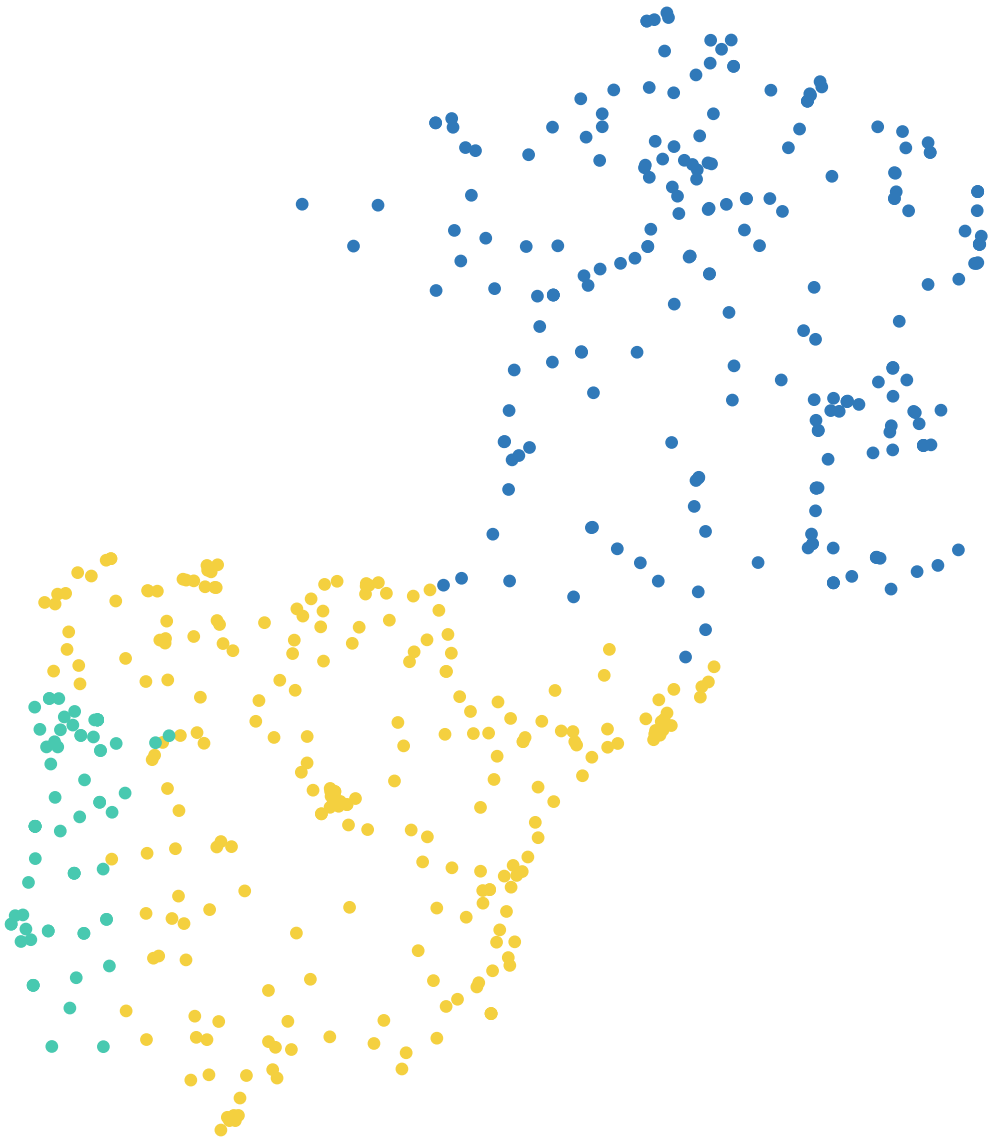


Figure 5-7. Node location in the Spain-Portugal-France synthetic network.

Node location and node attributes are introduced as inputs of the case study. As previously mentioned, we only focus on testing the ability of the proposed algorithm to link those nodes. The node location used as input is the real location of nodes for the three countries. Figure 5-7 shows the geographical distribution of nodes for the three countries. The target degree distribution, as well as the values of GHuST for each country, are fixed considering the real degree distribution and the GHuST values of real networks.

The model starts clustering demand nodes around generators. The clusters obtained for each country are shown in Figure 5-8. The number of clusters is 73 clusters in Spain, 21 clusters in Portugal and 60 clusters in France. Accordingly, the average cluster size is 3.2, 3.3 and 3.6 nodes, respectively. The number of clusters with no demand nodes (generators that are not connected to other nodes) is 9 clusters in Spain, 6 clusters in Portugal and 15 clusters in France.

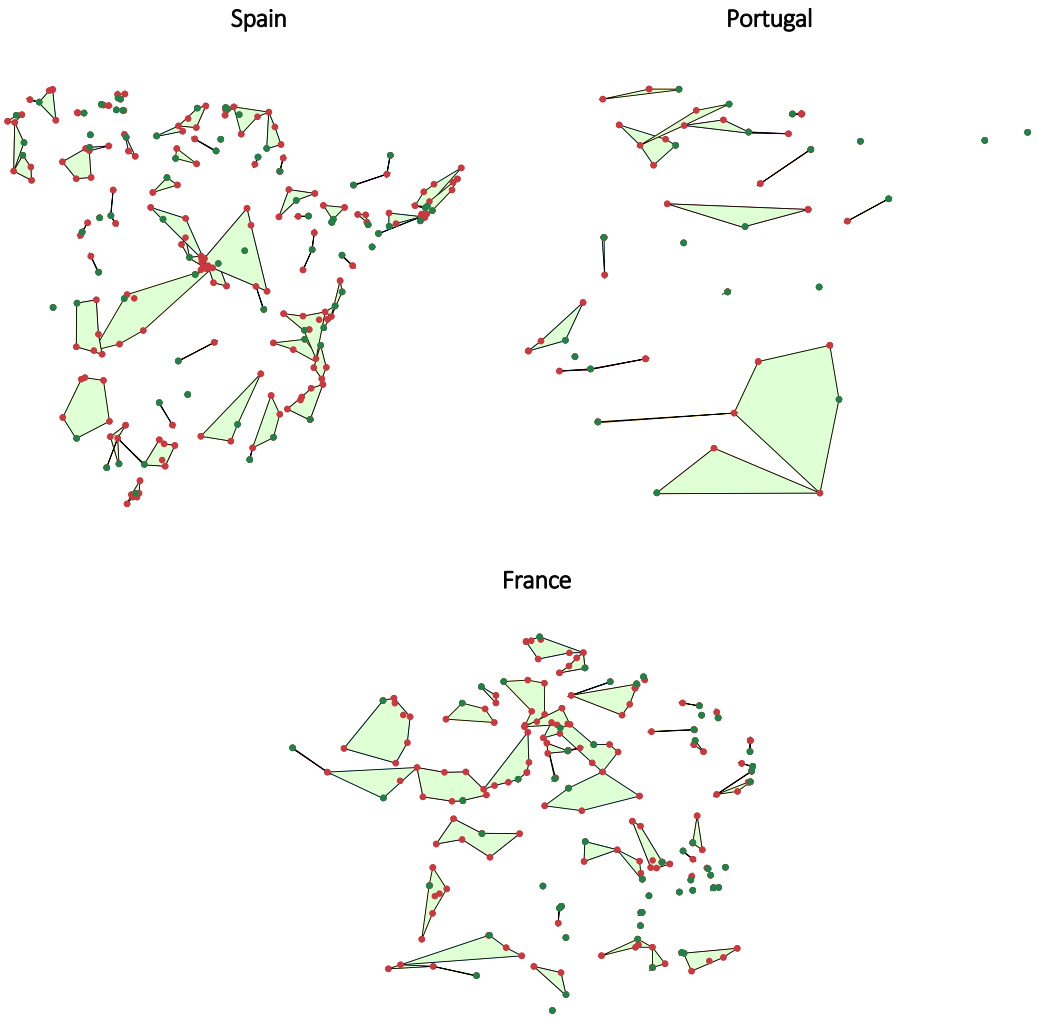


Figure 5-8. Clusters formed in the Spain-Portugal-France synthetic network.
Red points are demand nodes and green points are generators. The green area represents the clusters formed and the nodes that belong to those clusters.

Since the number of renewable power plants that are directly connected to transmission networks is small, and we have no detailed information about the year of construction of power plants, we only consider one iteration. Accordingly, all generators are introduced at the same time.

To connect clusters, the intra-cluster-wiring stage considers three types of lines. The values of the transmission capacities are 700 MVA, 1,500 MVA, and 2,000 MVA. Those values are the most representative frequencies of the thermal-rating distribution for those countries [109]. The number of lines installed in this stage is 411 (191 lines in Spain, 52 lines in Portugal and 168 lines in France). Lines added are shown in Figure 5-9.

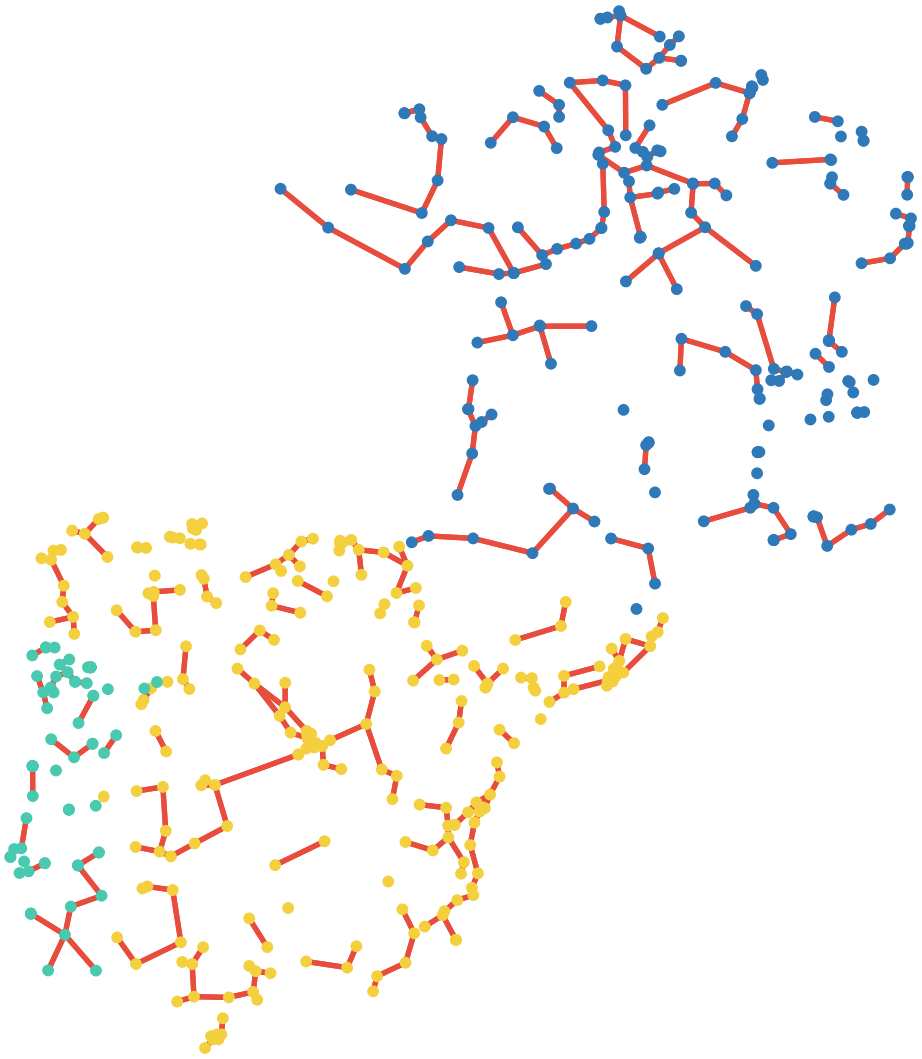


Figure 5-9. Intra-cluster-wiring stage in the Spain-Portugal-France synthetic network.

Regarding those demand nodes that belong to more than one cluster, the system is divided into 45 disconnected clusters in Spain, 17 disconnected clusters in Portugal, and 48 disconnected clusters in France. To obtain a connected graph, those disconnected clusters are linked to find a backup cluster. In case no backup cluster is found the algorithm minimizes network cost to have a connected graph.

Figure 5-10 shows the clusters made in each country and how they are connected. Each point represents a disconnected cluster. The location of the point is the mean value of the latitude and longitude of all nodes that belong to each cluster. Grey edges stand for the connection candidates that have been considered. Candidate connections are proposed based on the shortest geodesic distance among clusters (geographical distance between the closest nodes of different clusters). Finally, red lines represent the connections installed to have a connected graph.

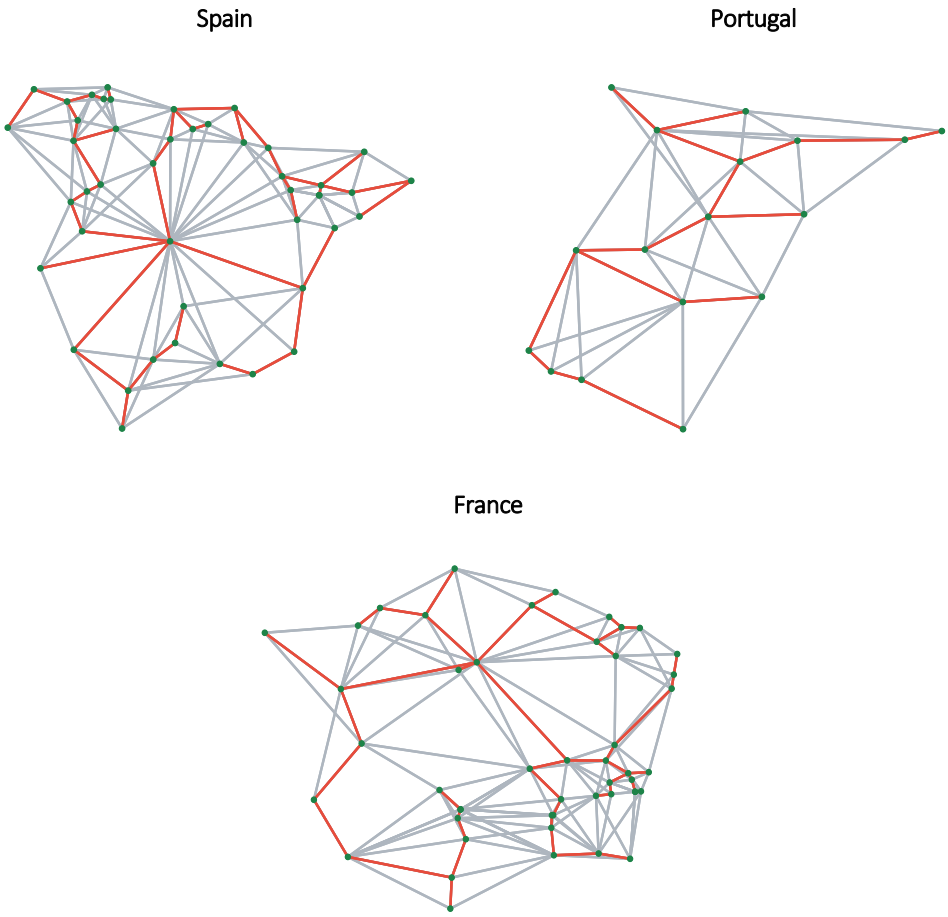


Figure 5-10. Connection of disconnected clusters in the Spain-Portugal-France synthetic network.

Connections among clusters should be translated into node connections. The algorithm defines the connection between a pair of nodes of different clusters based on geographical distance. Furthermore, international connections (Spain – Portugal, and Spain – France) are introduced manually at this stage. Lines installed (red lines) are shown in Figure 5-11.

The number of lines installed in each country is 236 lines in Spain, 68 lines in Portugal and 216 lines in France. The percentage of the remaining lines to be installed in further stages is 29.4% in Spain, 26.9% in Portugal and 23.67% in France.

To increase network reliability, the model tries to reduce the potential formation of islands. As we can see in Figure 5-11, the removal of a large number of lines will divide the system into two components. The algorithm verifies if there is a generation deficit in each of those potential disconnected components. In the case of deficit, the model analyses if the installation of a new line would avoid the formation of islands.

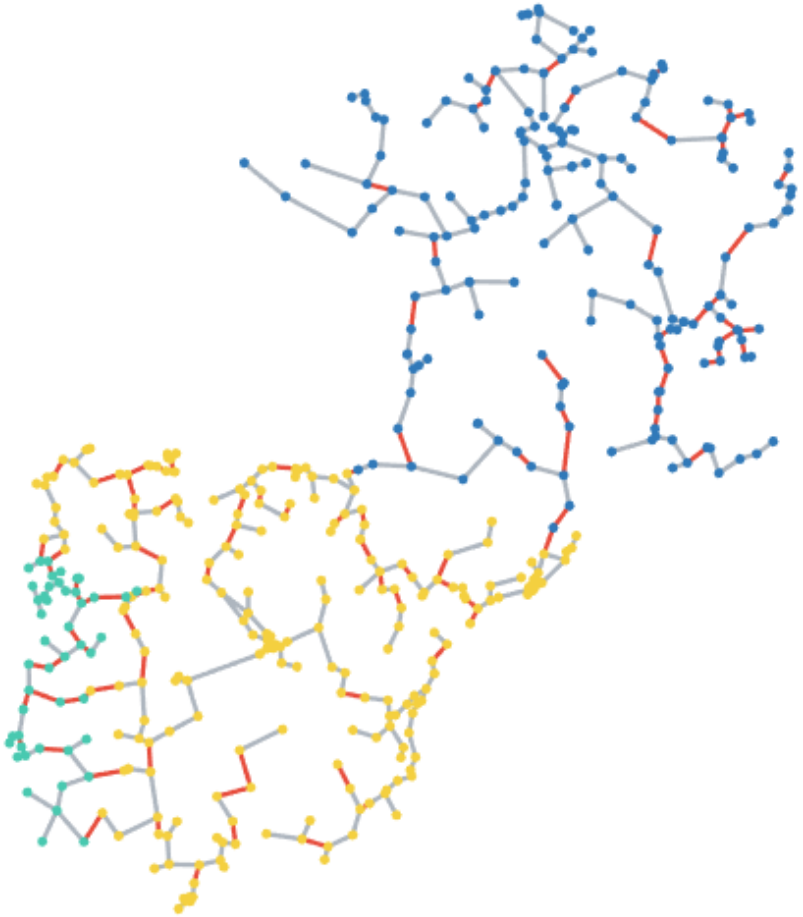


Figure 5-11. Inter-cluster-wiring stage in the Spain-Portugal-France synthetic network.

The preventing-island stage is repeated in two iterations. In both iterations, candidate lines are chosen based on the transmission-line length and its contribution to degree distribution. Furthermore, in the first iteration, candidate lines should connect nodes that are separated by nine or more edges. In the second iteration, that limit is lower, the minimum distance is four edges. This constraint tries to avoid the formation of really meshed local structures, in which island formation is avoided with close nodes. Accordingly, the number of lines required to increase network reliability would be extremely high.

The number of lines installed in the first iteration (node distance for candidate lines higher than 9 edges) is 11 lines in Spain, 2 lines in Portugal, 9 lines in France. The percentage of new lines to be installed is 26.1% in Spain, 24.7% in Portugal and 20.5% in France.

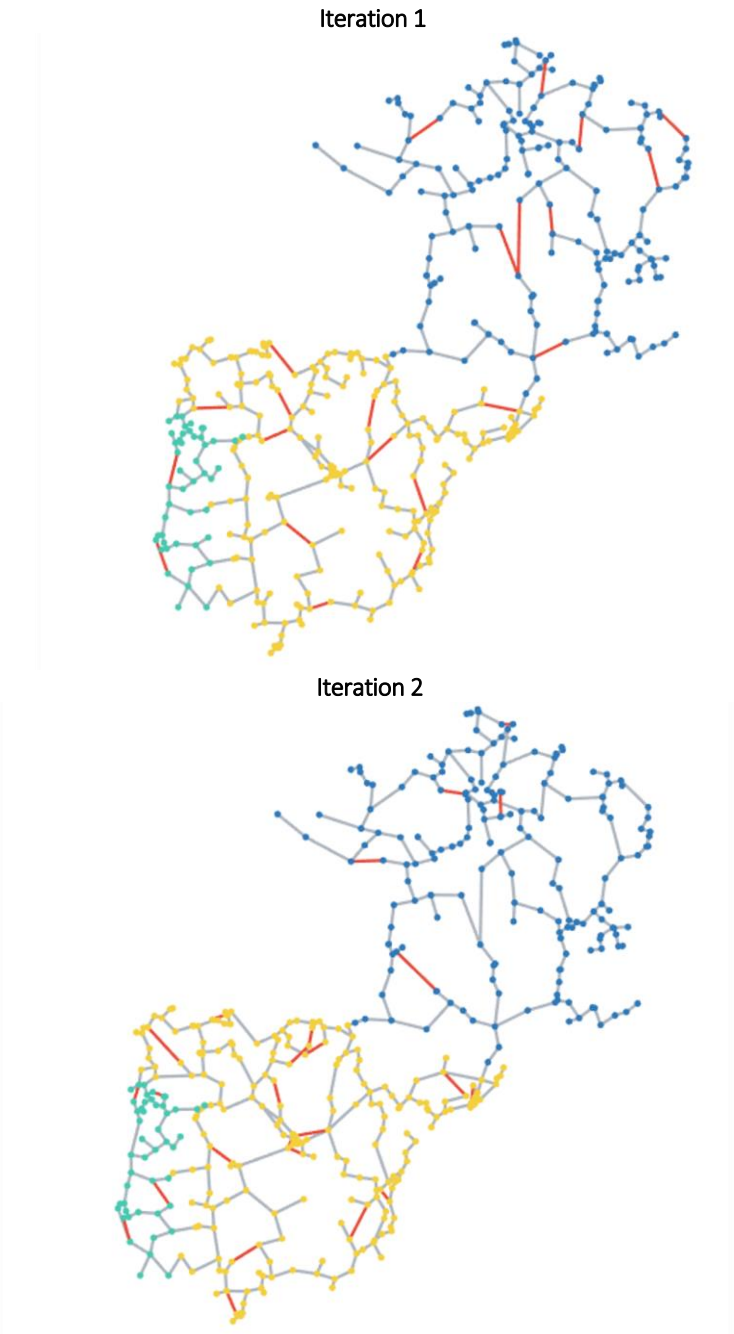


Figure 5-12. Preventing-islands stage in the Spain-Portugal-France synthetic network.

In the second iteration (node distance for candidate lines higher than 4 edges), the number of lines installed is 16 lines in Spain, 4 lines in Portugal, 5 lines in France. Finally, the percentage of remaining lines is 21.3% in Spain, 20.4% in Portugal and 18.7% in France. Lines installed in both iterations are shown in Figure 5-12.

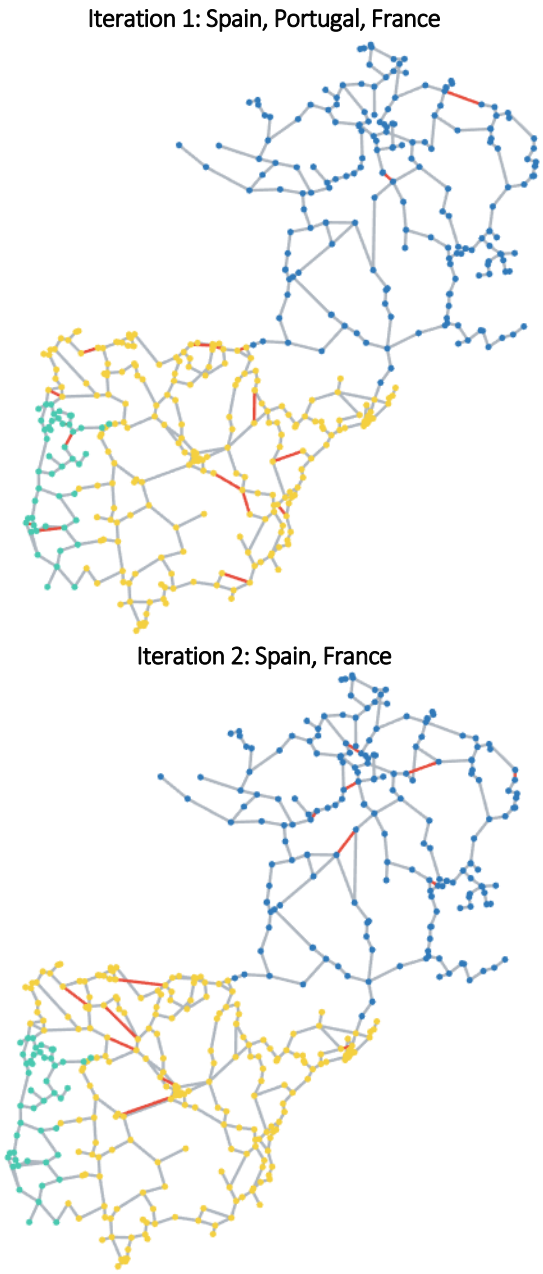


Figure 5-13. Guiding-node-degree stage in the Spain-Portugal-France synthetic network.

The average percentage of lines to be installed by the algorithm at the end of the preventing-island stage is 20.13%. This would lead to a high number of combinations to achieve consistency in terms of GHuST. To reduce the number of line candidates and possible combinations, the guiding-node-degree stage installs new lines following power-flow considerations. Those lines are added in order to reduce the error between the degree distribution of the synthetic network and the target degree distribution. It focuses on nodes with one and two connections. Because of the small size of Portugal (number of nodes) in comparison with Spain and France, the algorithm only installs new lines attached to leaf nodes in that country.

This stage is therefore divided into two iterations, as shown in Figure 5-13. In the first iteration, only nodes with one connection are considered. The number of lines installed is 15 lines in Spain, 6 lines in Portugal and 4 lines in France. The percentage of lines to be installed is 16.8% in Spain, 13.9% in Portugal and 17.3% in France. In the second iteration (nodes with two connections), the number of lines installed and the percentage of lines to be installed are 7 lines and 14.7% in Spain and 7 lines and 14.8% in France.

Finally, the average percentage of new lines to be installed is 14.5% of the total lines in each country. Those lines are added based on a random process in the achieving-GHuST-consistency stage.

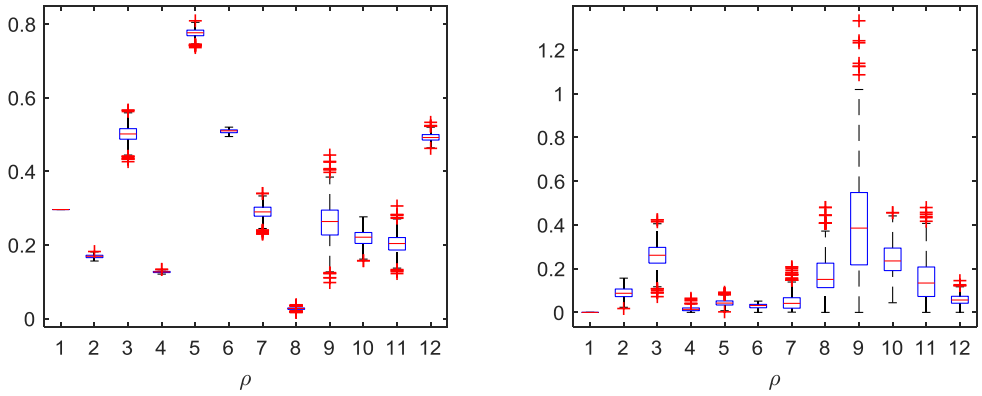
In the achieving-GHuST-consistency stage, the model performs 5,000 iterations for the Spanish and French networks and 3,500 iterations for the Portuguese network. The time limit is also lower in the Portuguese network.

In the case of Spain, the number of candidate lines is 935, and the number of lines to be installed is 56. The algorithm generates 1,295 networks that meet with the target degree distribution. In 3,705 iterations, the algorithm could not generate a valid network in terms of degree distribution. For the 1,295 networks that meet the degree distribution, the algorithm figures out the values of GHuST. Figure 5-14 shows the range of GHuST for each dimension as well as the relative error concerning the target. Since the number of installed lines is a consequence of meeting the degree distribution, the error of ρ_1 is 0% in all cases. In ρ_2, the maximum relative error is 15% and the median relative error is 8.7%. The value of the maximum relative error of ρ_4, ρ_5, and ρ_6 are always below 10% and the median relative errors are 1.4%, 4.2%, and 3.1% respectively. In the case of ρ_3, the median of the relative error is higher (26%), but the minimum error is 7%. Accordingly, most networks meet the distribution of GHuST regarding hubs and strings. Although there is a significative difference regarding, the degree of leaf-node connections (ρ_3), we can find networks in which the dimensions of GHuST regarding global connectivity are also consistent with the target.

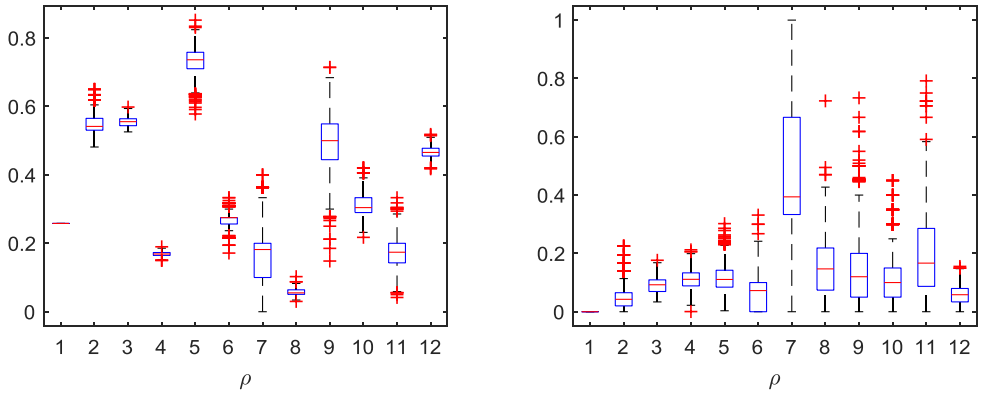
The higher deviation of GHuST is found in the dimensions related to triangles. We observe that the maximum value of the relative error is 1.33 in ρ_9. However, we also observe that there are instances in which the minimum relative error is 0%, 0%, 4%, 0% and 0% for ρ_8, ρ_9, ρ_{10}, ρ_{11}, and ρ_{12} respectively. Accordingly, although all networks have the same degree distribution, they display completely different topological properties. Indeed, the complexity of local structures varies. This reinforces the idea that the validation of synthetic power grids should go beyond global statistics and should use the GHuST framework proposed in the thesis.

Despite the variance in the values of GHuST related to triangles, there are networks in which those values are close to zero. The model will filter the networks with lower errors, as explained below. If no valid network is found, it is possible to rerun the algorithm increasing the number of iterations or reducing the number of lines installed in the previous stage.

Values of the GHuST framework (right) and relative error (left) for the Spanish synthetic power network



Values of the GHuST framework (right) and relative error (left) for the Portuguese synthetic power network



Values of the GHuST framework (right) and relative error (left) for the French synthetic power network

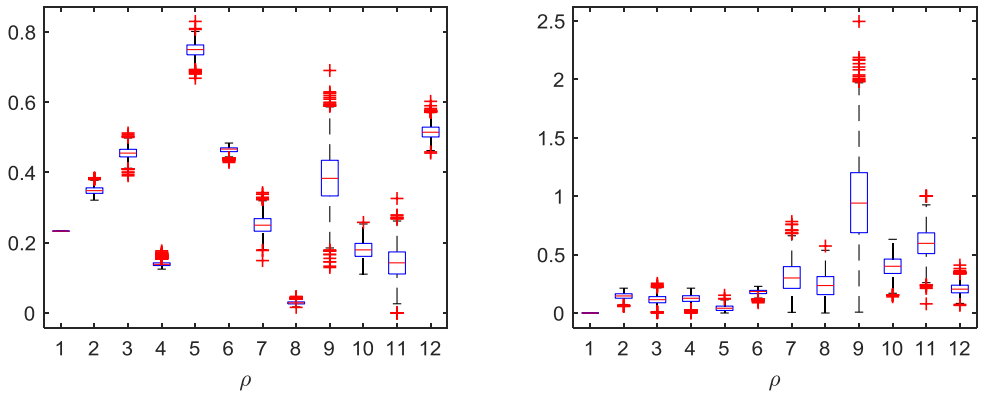


Figure 5-14. Values of the GHuST framework and relative error for the networks generated in the reaching-GHuST-consistency stage in the Spain-Portugal-France synthetic network.

In the Portuguese network, the number of lines to be installed is 13 lines. In this last stage, the model considers 167 candidates, and it generates 1,603 out of 3,500 networks that fit the degree distribution. In most dimensions of GHuST, (2, 3, 4, 5, 6, 8, 9, 10, and 12) the median of the relative error is below 15%. We only find a significative variance in ρ_7 (characteristic string length) and in ρ_9 (triangle concentration). There is a higher tendency to share vertices of triangles and strings tend to be longer in the synthetic networks. However, as in the case of Spain, we find some instances in which the relative error is close to 0% for those dimensions.

In the French network, although the number of lines to be installed is lower than in the Spanish case (43 lines), it considers 1,200 candidate lines, and it generates 2,145 networks that meet with the degree distribution. In those networks, global connectivity (ρ_1, ρ_2, and ρ_3) and hubs (ρ_4 and ρ_5) are consistent with the target (the median value of the relative error is below 15%). Although there is a higher relative error regarding string length, 25% of instances have a relative error below 20% considering strings (ρ_6 and ρ_7).

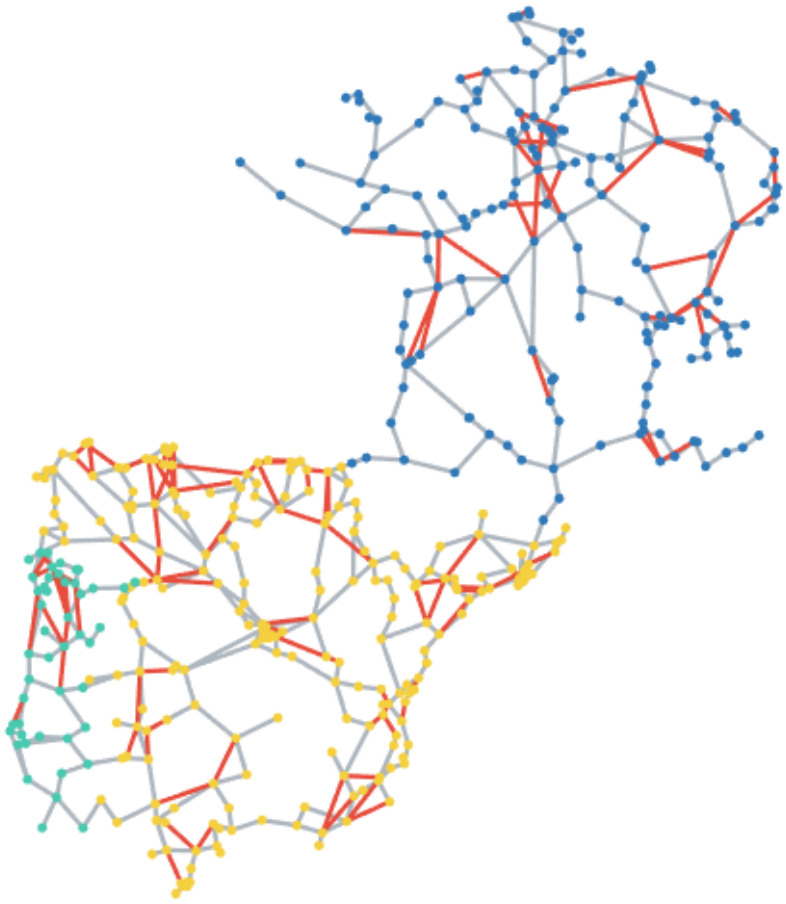


Figure 5-15. Reaching-GHuST-consistency stage in the Spain-Portugal-France synthetic network.

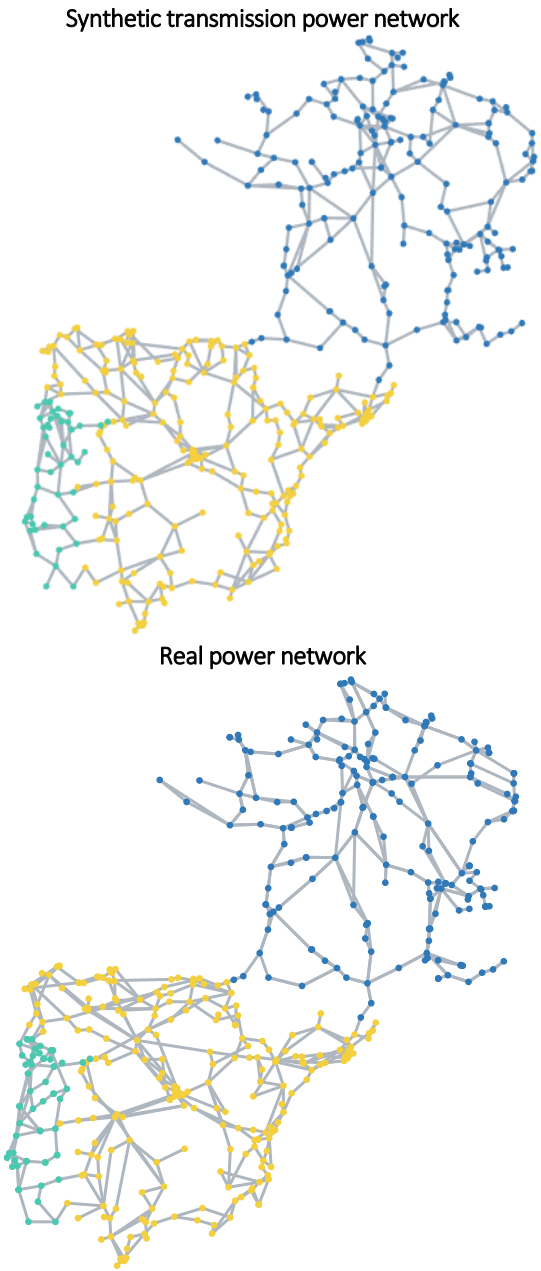


Figure 5-16. Spain-Portugal-France synthetic and real transmission power networks.

As in prior cases, higher variations appear in the dimensions related to triangles. In ρ_9 there are some outliers in which the difference regarding the target is higher than 200%. We can conclude that there is a higher tendency to share vertices. This might be a consequence of the degree-distribution constraint. However, in the case of triangles, we find instances in which GHuST dimensions have a relative error of 0%. This is the case of ρ_8, in which the relative error ranges from 0% to 57%. Besides, the minimum error is 13%, 7% and 6% for ρ_{10}, ρ_{11}, and ρ_{12}.

Finally, the instances that minimize the relative error for each country are chosen for the final synthetic network. The new lines installed (red lines) are shown in Figure 5-15.

Furthermore, Figure 5-16 illustrates a graphical comparison between the synthetic and the real networks. Further developments may condition the choice of the final network based on an electrical criterion, for instance, by comparing the distribution of power flows.

Table 5-1 summarizes the steps followed in the generation process of the synthetic power network. Table 5-2 shows the values of the GHuST framework for the three networks after the building-a-connected-graph step, the preventing-islands stage, and the guiding-node-degree stage.

Table 5-1. Steps followed and the percentage of lines installed in the generation process of the Spain-Portugal-France synthetic power network.

Step	Criterion	Percentage of lines installed	Result
Clustering nodes	Minimum distance	0% in Spain, 0% in Portugal, 0% in France	Nodes are grouped in clusters in which demand can be satisfied.
Intra-cluster wiring	Meeting demand at minimum cost	57.2% in Spain, 55.9% in Portugal, 59.4% in France	Demand is connected to generation.
Inter-cluster wiring	Cluster connection based on generation N-1 reliability criterion	70.6% in Spain, 73.1% in Portugal, 76.37% in France	Clusters are connected (connected graph).
Preventing islands	New lines installed based on line N-1 reliability criterion.	78.7% in Spain, 79.6% in Portugal, 82.3% in France	Network reliability is improved.
Guiding node degree	Degree-distribution consistency	85.3% in Spain, 85.2% in Portugal, 82.7% in France	The frequency of nodes with one or two connections is the same in the synthetic and in the real network.
Achieving consistency	GHuST consistency	100% in Spain, 100% in Portugal, 100% in France	The synthetic network is topologically consistent with the target.

Table 5-2. GHuST values for real and synthetic power networks.

		ρ_1	ρ_2	ρ_3	ρ_4	ρ_5	ρ_6	ρ_7	ρ_8	ρ_9	ρ_{10}	ρ_{11}	ρ_{12}
Generating a connected graph	Spain	0.00	0.28	0.37	0.18	0.65	0.65	0.50	0.00	0.00	0.00	0.00	0.40
	Portugal	0.00	0.49	0.44	0.13	0.48	0.46	0.31	0.00	0.00	0.00	0.00	0.25
	France	0.00	0.35	0.47	0.23	0.56	0.61	0.44	0.00	0.00	0.00	0.00	0.40
Preventing islands	Spain	0.11	0.15	0.41	0.20	0.68	0.59	0.44	0.00	0.00	0.00	0.00	0.40
	Portugal	0.07	0.38	0.52	0.12	0.54	0.53	0.35	0.00	0.00	0.00	0.00	0.25
	France	0.06	0.29	0.50	0.23	0.54	0.58	0.37	0.00	0.00	0.00	0.00	0.40
Guiding node degree	Spain	0.18	0.15	0.35	0.13	0.79	0.52	0.41	0.02	0.19	0.09	0.32	0.49
	Portugal	0.14	0.46	0.53	0.12	0.59	0.41	0.25	0.03	0.13	0.19	0.31	0.37
	France	0.10	0.31	0.44	0.17	0.60	0.48	0.31	0.01	0.07	0.06	0.00	0.52

The values of GHuST for the final synthetic and real transmission power networks are shown in Table 5-3. We observe that values for the synthetic and real networks are close in all the dimensions. In the Spanish synthetic network, the mean relative error is 6.0%. The highest difference is found in ρ_3 (17%) and in ρ_8 (11.3%). The relative error of the rest of the dimensions is below 10%. The mean relative error in the case of Portugal is 6.9% and the maximum relative errors are 10.9% in ρ_3 and 10% in ρ_6 and ρ_9. On the contrary, the error in ρ_2 is only 2%. In the third synthetic network, France, the mean relative error is 12%. This mean

error is higher because of the maximum error that is 23% in ρ_{10}. In France, triangulation has ha igher relative error than in other countries. However, the relative error of ρ_3 is only 5%.

Table 5-3. GHuST values for real and synthetic power networks.

	ρ_1	ρ_2	ρ_3	ρ_4	ρ_5	ρ_6	ρ_7	ρ_8	ρ_9	ρ_{10}	ρ_{11}	ρ_{12}
Real Spain	0.30	0.19	0.40	0.13	0.81	0.49	0.29	0.03	0.19	0.29	0.24	0.47
Synthetic Spain	0.30	0.18	0.47	0.13	0.79	0.50	0.31	0.03	0.20	0.25	0.21	0.47
Real Portugal	0.26	0.54	0.51	0.19	0.82	0.26	0.30	0.06	0.56	0.29	0.20	0.49
Synthetic Portugal	0.26	0.54	0.56	0.18	0.76	0.28	0.27	0.06	0.50	0.30	0.19	0.47
Real France	0.23	0.41	0.41	0.16	0.72	0.39	0.19	0.04	0.20	0.30	0.35	0.43
Synthetic France	0.23	0.37	0.43	0.15	0.74	0.44	0.20	0.03	0.23	0.23	0.30	0.44

Based on prior results, the algorithm generates networks with really similar topology to the real networks used as a reference. Furthermore, we cannot conclude that there is a dimension that the algorithm fails to replicate systematically. Although in Spain and Portugal ρ_3 has a higher deviation, the relative error is low in the case of France. All stages look to be significant. There is not a stage in which a low number of lines is installed.

As we pointed out in previous chapters, other authors have used some global statistics to prove the topological consistency of synthetic networks. Although we have shown the topological consistency of the synthetic network with the GHuST framework, we also compare the characteristic path length and network diameter. There is no error considering degree distribution and detailed validation of triangulation is provided with ρ_8 to ρ_{12}. Accordingly, the network average clustering coefficient is not compared.

The values of characteristic path length and network diameter for the synthetic and real networks are shown in Table 5-4. The relative error in the case of network diameter is 5% (Spain), 9% (Portugal), and 0% (France). The error is only one edge in both the Spanish and Portugal synthetic network. The relative error of the characteristic path length is 7% (Spain), 3.2% (Portugal), 5.9% (France).

Table 5-4. Characteristic path length and network diameter for the real and synthetic power networks

	Real Spain	Synthetic Spain	Real Portugal	Synthetic Portugal	Real France	Synthetic France
Network Diameter	19	20	11	12	21	21
Characteristic path length	8.06	8.63	4.98	5.14	8.27	8.76

Finally, we compare the synthetic networks that would have been generated for this set of nodes by the model proposed by Birchfield et al. [22]. Before running the referenced model, we modified the number of lines to be installed in the synthetic networks to be equal to the number of lines in the corresponding real networks, as opposed to considering the installation of 1.22 lines per node as done originally in their paper. This way, the mean node degree is equal in both the real and synthetic networks. However, the respective degree distributions obtained

from the synthetic and real networks were not in agreement. For instance, the maximum degree is 6 in the Spanish synthetic network, while in the real case, it is 9 connections per node. Regarding distance, the synthetic-network diameter is 35 edges, 15 edges, 33 edges in the case of Spain, Portugal and France respectively. The associated errors are 84.2%, 36.4% and 57.1% concerning the real networks. Similarly, the characteristic path lengths are 13.8 edges, 5.9 edges and 13.9 edges (relative error are 70.6%, 18.5% and 59.0%). Those values are also far from the real ones. Besides, we compare the GHuST framework. The mean relative errors are 44.6%, 27.7% and 28.0% for Spain, Portugal, and France respectively. Furthermore, in the Spanish synthetic network, the maximum relative errors are 161% for ρ_8, 68% for ρ_9 and 55% for ρ_{10}. In the case of Portugal highest relative errors are for ρ_9, ρ_{10}, and ρ_{11} (52.8%, 55.0% and 51.6%). The proposed model cannot replicate the local complexity (triangles) of the real networks. In the Portuguese synthetic network, there is also a significant error associated with strings, the relative error of ρ_6 is 46.8% and in the case of ρ_7 the relative error is 44.4%. In the French network, the number of leaf nodes (ρ_2) highly diverges with respect to the reference, the relative error is 49%. Highest relative errors are 86% (ρ_7) and 72% (ρ_{11}). This model has significant problems to replicate the topological aspects covered by the GHuST framework. Moreover, those results are in line with the conclusions obtained in Chapter 4 for the ACTIVSg networks generated with the same model.

Those errors are considerably larger than the obtained with the proposed algorithm. Therefore, while the referenced algorithm provides synthetic power networks in which AC power flow converges, these results show divergence between the synthetic networks and their corresponding real networks from a topological point of view. This divergence is insignificant when comparing the synthetic networks generated with the model proposed in this chapter.

These results show that the proposed algorithm is not only able to generate networks with the same degree distribution. Furthermore, results are also highly consistent with other topological metrics such as characteristic path length or network diameter. Furthermore, they are consistent with the GHuST framework.

The proposed algorithm is able to generate synthetic transmission power networks that are topologically consistent with real networks. It has been tested with the Spanish, Portuguese, and French 400-kV transmission power networks.

5.5. Takeaways

This chapter presents a new algorithm to generate synthetic power grids. Several models for this are proposed in the literature; however, they do not fit well with the properties of real power networks and they lack the flexibility necessary to generate different topologies according to the different factors that condition the evolution of power networks in different regions.

The proposed model considers economic and technical factors in order to mimic the topology of real power networks. The generation process is divided into two steps: *building a connected graph* and *adding lines to reach topological consistency*.

The objective of the first step is to design the simplest network that is able to supply demand at the lowest cost. The first power networks were non-meshed networks that supplied local areas. Once the demand is met, the planning process focuses on improving the quality of supply and network reliability, something that is achieved by adding new lines. Since this is a parametrical model, it can generate different network topologies for the same set of nodes. This is crucial in order to replicate the historical evolution of power systems, which depends on regional factors such as geography or the electricity generation mix, which has been widely discussed in prior chapters.

The wiring process was tested on the Spanish, Portuguese, and French 400-kV transmission network. In the three cases, we use the same set of nodes in the real grids (with the same electrical and geographical properties) in order to make network comparisons transparent. The resulting synthetic networks are consistent with the topology of their corresponding real networks. In the validation process, we have considered the GHuST framework and some global statistics. The algorithm is, therefore, able to generate networks that are topologically consistent with the real network. This is something that was missing in the literature as explained in Chapter 4.

The algorithm can, therefore, be used to generate case studies for power-network studies (such as the expansion of transmission power networks), where publicly available cases are scarce.

6

ASSESSING POWER-NETWORK VULNERABILITY

6.1. New challenges in network design

Power networks are interdependent infrastructures highly connected to other systems such as communication networks. Not only may a failure in power-networks lead to electric blackouts but a failure in other networks may also lead to power-network collapse. Power networks are therefore critical infrastructures that should be robust against failures. These failures may be the consequence of component breakdown or deliberate attacks. Deliberate attacks are targeted attacks that aim to collapse power networks, as the cyber-attack that caused the Ukraine blackout in 2015, affecting to 225,000 customers [27].

As introduced in Chapter 1, the power-network design usually includes N-1 or N-2 analyses. Accordingly, transmission networks are robust in the case of component failures. When a line fails, there are alternative routes to supply demand. Similarly, there are backup generators to avoid Power Not Supplied (PNS) in case of a committed-generator failure. However, those analyses are insufficient in the case of deliberate attacks.

Network design, therefore, rises to new challenges to build more robust networks. Consequently, it should include new criteria to protect networks against deliberate attacks. Prior work focuses on the detection of most vulnerable network components through an optimization problem [30], [31], [111]. Under the perspective of terrorists, the problem is to maximize the damage in the network with the lowest possible number of attacks. The detection of the most vulnerable element allows for the introduction of new measures to protect them. Although these formulations give an optimal solution, they involve computationally intensive models. This limits their use in the network design process since the large size of the network makes it impossible to run those algorithms. Alternative methods, such as the use of complex-network techniques, are needed to analyze the vulnerability of power networks including their interconnection with other types of networks. Thus, new methodologies should find a balance between computational requirements and result accuracy.

This chapter introduces a new metric, the *Electrical Line Centrality*, for the analysis of power-network vulnerability against deliberate attacks. Based on complex-network metrics, the electrical line centrality endows the betweenness centrality (presented in Chapter 2) with electrical parameters with a view to better resembling the physical principles that govern power networks. By applying this metric to power networks, we try to find the most critical lines in the network (line failure would have the highest impact on the network). This will support the development of new models to protect power networks in case of deliberate attacks.

The rest of this chapter is organized as follows. In Section 2, we introduce several vulnerability indices proposed in the literature. Section 3 presents the Electrical Line Centrality. Section 4 provides numerical analyses to prove the accuracy of the proposed metric. Finally, Section 5 extracts chapter conclusions.

This chapter introduces a **hybrid metric** to assess power network vulnerability: **Electrical Line Centrality**. Hybrid metrics combine topological metrics used in complex networks with the electrical features that characterize power networks.

6.2. Using complex networks to assess vulnerability

Recent work proposed to model power grids as complex networks to reduce the computational complexity in vulnerability analyses. As previously explained, complex-network methods focus on topology, which has been proved to play a crucial role in the propagation of cascade failures [112], [113].

Most of those works propose the analysis of power-network vulnerability through vulnerability indices. Those indices try to quantify the level of power-network vulnerability based on network topology. Existing indices have evolved from pure topological metrics to extended or hybrid metrics, i.e., topological metrics endowed with electrical information. Cuadra et al. did a comprehensive review of how complex-network concepts adapt to power-network-vulnerability analyses [40]. The main advantage of vulnerability indexes is the requirement of low computational resources. Accordingly, they may be effectively introduced in the network design problem.

6.2.1. Topological metrics

Topological indices, metrics that only consider the connection among nodes have been widely used to analyze power-network vulnerability. *The characteristic path length* and the *network average clustering coefficient* (explained in Chapter 2), were proposed to analyze the U.S. Western Systems Coordinating Council (WSCC) [108]. This work found that the power network is a small-world network. It concludes that power-network vulnerability increases when line removal leads to an increase of characteristic path length and a decrease in the small-world index (2-4). Accordingly, we can rank network components based on the impact they

would have on distance distribution.

Holmgren applied the *average node degree*, *characteristic path length,* and *network average clustering coefficient* in order to try to relate changes in network topology with network vulnerability by analyzing the Nordic power network and the power network of the western states of the U.S. [114]. The objective was to find if a change in a topological index (after component failure) may reveal an increase of network vulnerability. However, the author states that graph metrics are imprecise to study structural vulnerability.

Latora et al. introduce the *global efficiency* metric to measure the performance of networks [110]. This metric relates network efficiency with the distribution of distances among nodes (C-1). We can analyze the criticality of a network component, *damage*, by comparing the properties of the graph before and after the failure or attack (C-2) [115]. The application of this method to power networks allows for the identification of the most critical lines as well as the detection of the lines that should be installed to reduce network vulnerability. The application of this topological analysis to the Spanish, French, and Italian transmission networks shows that the removal (or failure) of only three edges would have severe consequences in those networks. Furthermore, the installation of just one line would lead to a remarkable increase in network efficiency [116]. Consequently, this should be taken into account when designing network expansion. The definition of global efficiency is adapted to power networks by considering only the paths that connect generators and demand nodes, which is *modified global efficiency* (C-3). This was applied to the analysis of North American in the Italian power grid [117], [118].

Rosas-Casals et al. find a correlation between *the cumulative degree distribution parameter* γ (a parameter that characterizes the degree probability distribution) and reliability indices (energy not supplied, total loss of power, restoration time and equivalent time of interruption) [119]. Accordingly, this parameter allows for the assessment of network vulnerability as a whole. This parameter cannot analyze the impact of each component in network vulnerability.

In order to analyze the goodness of *characteristic path length* and *connectivity loss*, those metrics were compared with the *blackout size model* [120]. The blackout size model uses electrical information to analyze adequacy by modeling cascading failures in power systems due to overloaded lines. The authors conclude that those metrics might mislead the evaluation of network vulnerability. Therefore, pure topological metrics are not an accurate tool to assess network vulnerability. Furthermore, they may lead to ambiguous results. Indeed, while Albert et al. state that the power network is highly vulnerable to the attack of high-load nodes or hubs; Wang and Rong conclude that attacks to low-load nodes may result is worse failures [121], [122]. Both prior analyses are based on pure topological methods.

Finally, Ouyang et al. also pointed out the lack of accuracy of purely topological metrics [123]. That conclusion results from the analysis of the correlations between topological metrics (characteristic path length, network efficiency, source-demand considered efficiency, network average clustering coefficient, connectivity level and size of the largest component) with energy not supplied in the network.

6.2.2. Hybrid metrics

To overcome the limitations of topological metrics, hybrid indices endow topological metrics with electrical parameters. Electrical features might be added into node description. For example, node information might include node demand or the power that can be injected in each substation by generators. This is the case of *loss of load* and *connectivity loss*.

Loss of load tries to estimate the ability of the network to meet demand after a node or edge failure [107]. If the removal of a line (or a node) splits up the system in two or more components, loss of load estimates the deficit of generation concerning demand in each component. Accordingly, the criticality of an element correlates with the total deficit of generation in the network after a failure (C-4).

Connectivity loss quantifies the decrease in the ability of a substation to receive power from generators. It is the average decrease in the number of generators that are connected to a node (directly or through a path) (C-6) [122]. The application of this metric to the North American power network reveals that power networks are very vulnerable to a failure in highly connected nodes (hubs). Furthermore, the authors state that power-network vulnerability is inherent to the structure of the grid.

Hybrid metrics endow edges with electrical properties. Node degree and betweenness centrality are modified by the inclusion of the value of power flow through lines [124]–[129], the resulting metrics are the *electrical degree centrality* (C-7) and the *electrical betweenness centrality* (C-8). The electrical degree centrality measures the importance of a node in the network regarding the power flows through the lines that are connected to it. Similarly, the electrical betweenness centrality assesses the centrality of a node based on the power that is injected or withdrawn in a node when lines operate at full capacity.

Further improvements try to introduce the physical principles that govern power networks. In power networks, power units flow based on Kirchhoff's' circuit laws. Consequently, they do not follow the shortest path. Accordingly, distances among nodes are measured by line impedance.

The *structural vulnerability index* (C-9) measures the ability of the network to supply demand [130]. It assumes that the contribution of a generator to a demand node decreases exponentially with electric distance.

Directed global efficiency modifies the global efficiency index to limit the exchange of power between generators and demand nodes and to include the electric distance instead of the shortest-path distance (C-10) [131]. Similarly, *net ability* (C-11) introduces line impedance as well as the maximum interchange of power between nodes to complete the definition of global efficiency [132], [133]. To assess the vulnerability of power networks, Wang et al. propose the analysis of changes in the *effective graph resistance* (A-12) [134], [135]. The electrical resistance between two nodes is the potential difference that appears when a unit of current flows from one node to another.

The *electrical centrality* (A-13) measures the importance of a node based on the electrical distance among nodes [136], [137]. The larger the electrical distance of a node with the rest of the nodes in the network, the lower the impact it has on network vulnerability. The *centrality index* considers that the criticality of a line is proportional to the sum of the maximum power flows that might be exchanged between all pairs of nodes in the network (C-15) [138], [139]. This is therefore conditioned by the power-transfer-distribution-factor PTDF matrix since the maximum power flow is limited by Kirchhoff's laws. The inclusion of line impedance is also included in other adaptations of the betweenness centrality such as the *hybrid flow betweenness*, or *the electrical betweenness* [107], [131].

Finally, the *extended betweenness* (C-16) proposes an adaptation of the betweenness centrality by including the PTDF matrix and transmission-line capacity [140]–[143]. This index is compared to the topological betweenness centrality and the assessment of network vulnerability with pure power-system techniques. While it improves results given by the topological index, results are worse than the ones provided by a pure electrical framework.

6.2.3. Other approaches

References [144]–[146] discuss the accuracy of vulnerability indices. They propose a novel method, which is based on the fault chain theory, to combine topological features and power-network operational characteristics. This method is based on the construction of a *correlation graph* that includes the structural features of the network as well as the operation status. The nodes of the correlation graph stand for the transmission lines of a network and edges represent the relationship between two transmission lines during fault propagation. The ranking of critical lines is done based on the topology of the correlation graph. Fault-Chain Theory is also used in a steady-state model to identify critical events that contribute to cascading-failure propagation [147].

Since the correlation graph considers network operation, it needs to run several DCOPF to build it. The number of power flows required might be unmanageable to introduce this method in the design of large interconnected power networks (because of computational requirements). Despite not considering network operation implicitly, we consider that vulnerability indices may be effectively incorporated in network design. As mentioned in Chapter 1, transmission expansion planning may benefit from these indices in two ways: by introducing them as a partial objective in the optimization function (it penalizes high values of vulnerability indices) or by including them as constraints (it establishes maximum values for the indices). Consequently, this is a strength of vulnerability indices concerning those methods based on Fault-Chain Theory.

Although hybrid metrics improve results given by topological metrics, new improvements should be introduced to reduce the gap between pure electrical considerations and complex network-based metrics. Furthermore, in order to introduce these metrics in the network design problem and to enhance the resolution of an optimization problem, vulnerability indices should be linear functions

Purely **topological metrics**, such as global statistics, are not suitable tools to assess network vulnerability.

Hybrid metrics improve results given by topological metrics. However, results are not good enough to replace electrical methods.

6.3. Electrical Line Centrality

This section proposes a hybrid metric to analyze power-network vulnerability. As previously explained, a hybrid or extended metric refers to the endowment of topological-based metrics with electrical considerations to replicate the physical behavior of power networks. The proposed metric, the Electrical Line Centrality ELC endows betweenness centrality with electrical properties. Betweenness centrality $BC(u)$ is defined as the number of times a node or a line is in the shortest path among all pair of nodes in a network (6-1), where $\sigma_{s,t}(u)$ is the number of the shortest path from s to t through node or line u and $\sigma_{s,t}$ is the number of shortest paths from s to t. In the case of modeling power networks as undirected graphs, shortest paths from s to t and t to s counts as one path. Therefore, in undirected networks, the earlier equation omits the coefficient 1/2.

$$BC(u) = \frac{1}{2}\sum_{i,j\neq u}\frac{\sigma_{i,j}(u)}{\sigma_{i,j}} \tag{6-1}$$

Unlike other networks in which network flow or information may move from node s to node t through the shortest path, Kirchhoff's laws govern power networks. Therefore, betweenness centrality cannot infer the dynamic behavior of power networks.

We propose the adaptation of betweenness centrality with the inclusion of electrical information about lines and nodes. ELC considers line reactance, the power demanded in each node and node generation capacity. ELC is the sum of power through a line for power interchanges among all pairs of generators and demand nodes in a network (6-2). ELC considers that power flows always go from generators to demand nodes, and the amount of power is proportional to the generation capacity and power demanded in each node.

$$ELC(l) = \sum \Delta F_{ij}\left(\Delta P_{r,}\Delta P_{s,}\right) \tag{6-2}$$

$\Delta F_{ij}\left(\Delta P_{r,}\Delta P_{s,}\right)$ is the incremental power that flows through a line that connects nodes i and j when there is a change in demand or generation capacity in nodes r and s.

In power networks, the DC Power Flow (DCPF) equations model power flow through lines (6-3). This model sets node voltages to 1 per unit and assumes that voltage angles differences among nodes are small. That means that $\cos\theta_{ij} \approx 1$ and $\sin\theta_{ij} \approx \theta_i - \theta_j$.

$$F_{i,j} = \frac{1}{X_{ij}}\left(\theta_i - \theta_j\right) \tag{6-3}$$

Considering that $\left(\theta_i - \theta_j\right) = \theta$ and $P = B\theta$ (where P is the vector of injected or withdrawal power and B the system susceptance matrix), power flows in a network can be written as (6-4), where A is the incidence matrix, X is the reactance matrix (diagonal matrix), and r is the subset of nodes without the slack bus.

$$F = ({X_r}^{-1}{A_r}^{T}{B_r}^{-1})\, P_r \tag{6-4}$$

The expression $({X_r}^{-1}{A_r}^{T}{B_r}^{-1})$ stands for the power-transfer-distribution-factor matrix S. It shows how power through lines changes when there is a new injection or withdrawal of power in one or more nodes of the system (6-5).

$$\Delta F = S\, \Delta P_r \tag{6-5}$$

In power networks, we can classify nodes as demand nodes, generator nodes, and interconnection nodes (nodes in which there is no demand or generator connected). By applying matrix block multiplication, we can rewrite (6-5) as (6-6):

$$\Delta F = [S_G \quad S_D \quad S_I] \begin{bmatrix} \Delta P_G \\ \Delta P_D \\ \Delta P_I \end{bmatrix} \tag{6-6}$$

where S_G, S_D and S_I are the blocks of the sensitivity matrix that shows how power through lines change when there is a change in generation, demand or interconnection nodes respectively. ΔP is the change of power in the generation (G), demand (D) or interconnection nodes (I).

By assuming that ΔP_G is always positive (generation nodes always inject power in the system), ΔP_D is always negative (demand nodes always withdraw power) and ΔP_I is equals to zero (there is no demand or generation connected to those nodes) (6-6) becomes:

$$\Delta F = S_G\, \Delta P_G - S_D\, \Delta P_D \tag{6-7}$$

As previously mentioned, ELC considers that power flows always go from generator nodes to demand nodes. By considering all combinations among generators and demand nodes, we can express (6-2) as follows:

$$ELC(l) = \sum_{j=1}^{j=N_G} S_{G\,i,j}\, w_g - \mathrm{C} \sum_{j=1}^{j=N_D} S_{D\,i,j}\, w_d \tag{6-8}$$

where w_g is the vector of generation capacity of each generation node, w_d the vector of the

power demand of each demand node and C is the generation capacity of the entire system, $C = \sum w_g$.

When calculating electrical line centrality of all lines of the system, (6-8) can be expressed as a product of matrices:

$$ELC = S_G\, w_g - C\, S_D\, w_d \tag{6-9}$$

As is the case of betweenness centrality, we can extend the concept of electrical node centrality to nodes. Electrical node centrality ENC is the sum of the ELC that are incident in that node (6-10).

$$ENC(n) = \sum_j ELC_{ij} \tag{6-10}$$

$$ENC = |A|\, ELC(l) \tag{6-11}$$

In (6-11), $|A|$ is the incidence matrix in which all elements are positive. On the contrary to ELC, flow direction does not directly condition ENC since it does not matter whether the flow is incident or outgoing.

This formulation bridges the gap between purely topological measures (fast and easy to calculate, but with limited use in power systems) and power-flow estimations. It complements betweenness centrality, with node and line information. It has the considerable advantage of having a compact matrix expression that can be efficiently calculated.

Finally, this new metric can be easily adapted to different operation scenarios by simply changing the vector of power demand w_d or the vector of nodal generation capacity w_g. Moreover, different scenarios might be considered by figuring out a weighted average ELC where weights represent the probability of each scenario.

Electrical Line Centrality endows topological betweenness centrality with electrical features.

ELC **is the sum of power through a line for power interchanges among all pair of generators and demand nodes in a network.**

It can be applied to assess the centrality of transmission lines and substations in the network.

6.4. Numerical Studies

In order to confirm the usefulness of ELC, we apply it to the IEEE 9-bust test system and to the IEEE 118-bus test system [148]. **We will try to determine the sequence of line attacks that will maximize network damage.** Accordingly, the higher the values of PNS in the network obtained with the vulnerability metric, the better the performance of the metric. As explained in Section 1, an optimization problem may give the exact solution to the analysis of power-network vulnerability against deliberate attacks. However, problem size leads to a computationally intensive analysis, which hinders problem resolution. We try to approach the optimal solution with an iterative method. We propose to select the most vulnerable lines in each iteration or attack. Accordingly, the method will choose the line whose removal leads to the largest PNS. Subsequently, it updates network topology and repeats the process to get a sequence of attacks. This is not the optimal solution since it ranks line impact regardless of the attack sequence. Unlike this procedure, the optimization problem would try to maximize the damage in the network by coordinating the order of attacks. However, given computational complexity, we use the iterative method as an approach to the maximum damage problem. We use it to provide a reference in terms of PNS that will be used as the target the proposed metric should reach. The estimation of PNS after a line attack is calculated with the DCOPF (implicitly, it assumes that the system is able to correctly respond to abrupt changes in demand and it will, therefore, be a lower bound to the real result).

6.4.1. IEEE 9-bus test system

The IEEE 9-bus test system includes 9 nodes, 6 lines, 3 transformers, 3 generators and 3 demand nodes (see Figure 6-1). Despite the small size of the case, there are 720 different sequences of line attacks or rankings. Eight of those rankings lead to the worst scenario under the PNS perspective (best attacks), and thirty rankings lead to the lowest values of PNS (worst attacks). Table 6-1 shows an example of both sequences of failure and the PNS obtained. This analysis allows us to confirm the accuracy of vulnerability indices. As shown in Table 6-1, values of PNS obtained with the iterative approach (target) are close to the optimal sequences.

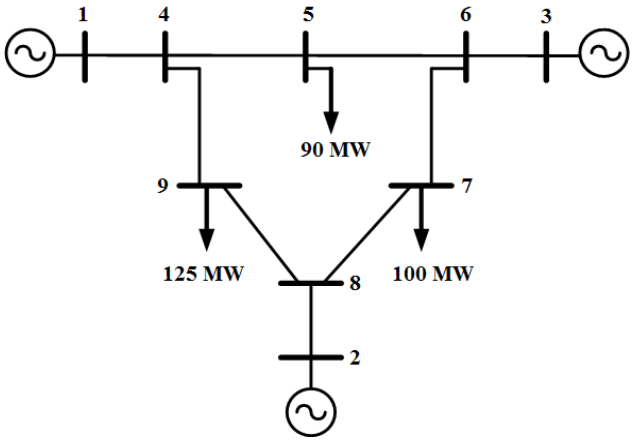


Figure 6-1. IEEE 9-buses test case

In the case of the proposed vulnerability index, ELC proposes the following ranking of lines: 8-9, 4-9, 5-6, 7-8, 6-7, 4-5. Corresponding values of PNS are: 0, 125, 125, 125, 225, 315. Those values are similar to the obtained for the best case. ELC improves results given by the target in the second and third attack.

We also analyze the ranking provided by the topological betweenness centrality. In this test case, all lines have the same value of betweenness centrality. This means that from a topological point of view, all lines would have the same impact on the network in case of failure. This clearly shows the limitations of purely topological indices. Regarding the extended betweenness centrality, values of PNS are: 0, 0, 125, 125, 225, 315. Therefore, the ELC provides a better solution than the obtained with prior metrics. It needs a lower number of attacks to have PNS in the network.

Table 6-1. Order of line failure and PNS in the IEEE 9-bus test system in the best case, the worst case, and the target.

Order	Best Attack		Worst Attack		Target	
	Line	PNS (MW)	Line	PNS (MW)	Line	PNS (MW)
1	8-9	0	4-5	0	6-7	0
2	4-9	125	6-7	0	7-8	100
3	7-8	125	8-9	0	4-9	100
4	6-7	225	5-6	90	8-9	225
5	5-6	225	7-8	190	4-5	225
6	4-5	315	4-9	315	5-6	315

In the IEEE 9-bus test system, the Electrical Line Centrality improves the results given by the Betweenness Centrality and by the Extended Betweenness Centrality.

Furthermore, results are very close to the optimal solution.

6.4.2. IEEE 118-bus test system

This system consists of 118 nodes, 177 lines, 9 transformers, and 56 generators. This chapter considers the model assumptions presented in Chapter 2. If there are two more electrical connections between two nodes, the model assumes them to be one single edge. The resulting graph has 118 nodes and 170 lines. Therefore, it considers that the largest number of attacks is 170.

We apply a similar procedure to assess metric performance in the IEEE 118-bus test system than in the IEEE 9-bus test system. The values of PNS reached by ELC are quite close to the target in first iterations, from attack number 0 to 21 (as shown in Figure 6-2). Both betweenness centrality and extended betweenness centrality are far from those values. Indeed, the number of attacks needed to cause PNS in the system with the ranking given by betweenness centrality

and by the electrical betweenness centrality is much larger than with ELC or the electrical procedure.

In the case of ELC, values of PNS highly diverge from the target values in the gap between 25 and 100, where extended betweenness gives better results (as shown in Figure 6-2). The sequence given is far from the approach used as a target. However, while that approach calculates the most vulnerable line in each iteration, the model only computes the three indices once. Accordingly, it does not consider changes in topology. After removing a line, network topology and system dynamic change and therefore line vulnerability changes too. Consequently, the model may improve results by updating metrics to changes in network topology because of prior failures.

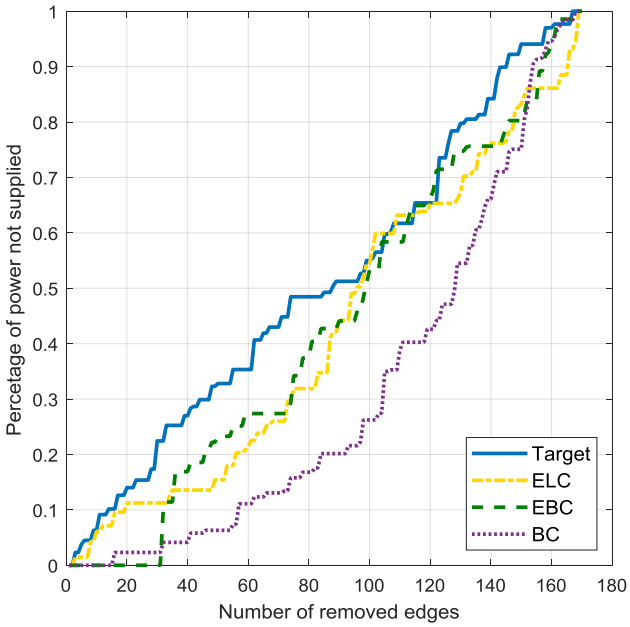


Figure 6-2. Percentage of PNS in the IEEE 118-bus test system after removing lines according to electrical considerations (target), electrical line centrality (ELC), extended betweenness centrality (EBC) and betweenness centrality (BC).

By updating rankings in each iteration, ELC continues to be a better approach to the target as shown in Figure 6-3. The difference between ELC and the target is smaller in the gap between 25 and 100 than in the first case. Furthermore, ELC also gives larger values of PNS than the target for a high number of attacks. This is possible since the procedure described above does not give the optimal solution.

To compare results, we use the mean absolute error MAE, both in MW and in percentage. In (6-12) x_i and y_i are the values of PNS obtained with different metrics and n the number of iterations or attacks.

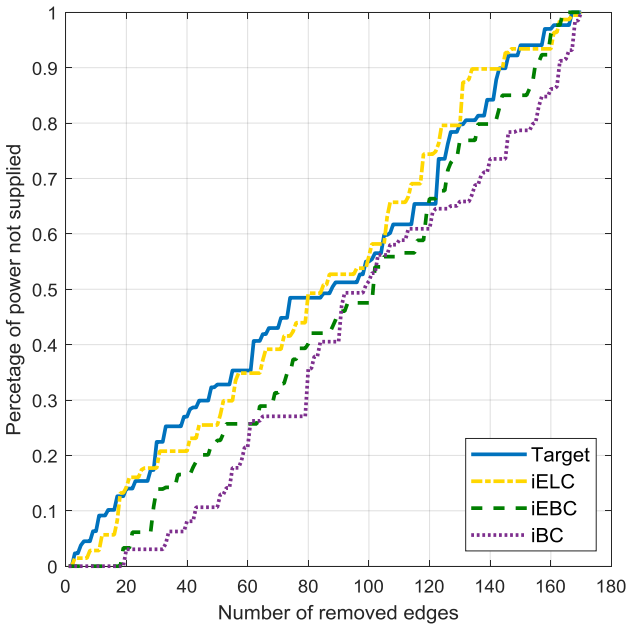


Figure 6-3. Percentage of PNS in the IEEE 118-bus test system after removing lines according to electrical considerations (target), iterative electrical line centrality (iELC), iterative extended betweenness centrality (iEBC) and iterative betweenness centrality (iBC).

$$MAE = \frac{\sum|x_i - y_i|}{n} \tag{6-12}$$

First, we compare the error of the topological and hybrid metrics concerning the target. Second, we define the error as the difference between values of PNS for each ranking and the largest value of PNS each attack (ELC provides higher values of PNS than the target in some iterations). Results show the accuracy of ELC over betweenness and extended betweenness centrality, as shown in Table 6-2. The error of betweenness centrality is more than twice the error given by the ELC. The proposed metric also reduces drastically the error of the extended betweenness centrality. Furthermore, when updating rankings iteratively, rankings are improved leading to a more vulnerable sequence of attacks. The reduction of error is above 50% in the case of ELC. The final error with respect to the electrical procedure is 10.89%.

While the improvement is also essential in the case of betweenness centrality, the error reduction is only around 4% in the case of the extended betweenness centrality. When considering the largest values of PNS, the error of ELC is 8.23%, this is much lower than the error obtained with the extended betweenness centrality and the topological betweenness centrality.

Table 6-2. Mean absolute errors of vulnerability indices with respect to electrical considerations and the larger values of PNS

Metric	MAE with respect to ELE		MAE with respect to the largest value of PNS	
	(MW)	(%)	(MW)	(%)
Target	-	-	23.69	2.59
ELC	129.14	23.33	149.24	25.24
iELC	47.19	10.89	25.23	8.23
EBC	118.49	30.66	138.66	32.30
iEBC	165.32	26.32	189.01	28.32
BC	292.16	54.07	315.47	55.44
iBC	100.32	36.98	122.34	38.95

Furthermore, the ELC is computationally simpler than the extended betweenness centrality. ELC and EBC are compared in terms of execution time. While ELC takes 0.018 seconds and $iELC$ 0.108 seconds in the IEEE 118-bus test case, the EBC and $iEBC$ take 0.170 seconds and 0.991 seconds respectively. Consequently, the new metric is almost ten times faster than the EBC. This time reduction is crucial in applications that require to assess network vulnerability a large number of times.

The accuracy of the Electrical Line Centrality is also proved in the IEEE 118-bus test case, where the ranking given by ELC is more vulnerable (critical) than the EBC.

To get results that are closer to electrical procedures, it is necessary to calculate ELC iteratively.

Moreover, it drastically reduces computational complexity with respect to the extended betweenness centrality.

6.5. Chapter takeaways

A new extended metric, the Electrical Line Centrality, is proposed in this chapter to assess power-network vulnerability against deliberate attack. This metric endows betweenness centrality with electrical information related to generation, demand, and transmission-line parameters. This allows us to overcome the limitations of purely topological metrics that do not consider the electrical nature of power networks. It also has the considerable advantage of having a compact matrix expression that can be efficiently calculated and it can be integrated into other models. The accuracy of the metric is tested with the IEEE 9-bus and IEEE 118-bus test systems. The Electrical Line Centrality supplies better results than existing hybrid metrics such as extended betweenness centrality with lower computational requirements.

7

CONCLUSIONS AND FURTHER RESEARCH

7.1. Conclusions

The increase of power-network complexity, as well as the interconnection with other networks, lead to the need for new models to effectively operate and control power networks. However, the lack of publicly available network models hinders research into power systems. Information about real power networks is scarce, and most of the existing test cases are old and do not reflect the current structure of the network.

We find two groups of initiatives that try to overcome the lack of public power-network models. On the one hand, OpenStreetMap-based initiatives proposed the construction of network models based on the real location of infrastructure components. However, those models do not include electrical information. Moreover, those initiatives mean the disclosure of the real location of power-network components. This might run into security issues. On the other hand, recent initiatives in the US propose the generation of synthetic power grids: non-real, albeit realistic power networks that are topologically and electrically consistent with real networks. This would allow for the publication of accurate network models in which network operation and control are similar to real power networks while preserving network security. Although the target is clear, the generation of synthetic power grids, it is necessary to develop new algorithms to generate these synthetic networks. The models traditionally used in power systems are not a suitable solution because of computational complexity.

The publication of network location is a risk that may increase network vulnerability. Besides, the interdependency connection with other networks also makes power networks more vulnerable. Traditionally, N-1 analyses have been used to design networks that are robust in the case of component failures. However, the large size of power networks, the higher degree of interconnection with other networks, and the risk of deliberate attacks require new methodologies to design more robust networks with manageable computational requirements.

The use of complex-network techniques may find a balance between the complexity of power systems and computational requirements. This thesis proposes a novel algorithm to generate synthetic power grids and a new metric to assess power-network vulnerability. In

both cases, complex-network analyses (that are based on network topology) are complemented with electrical information to capture the principles that govern power networks. As a previous step to the generation of synthetic power grids, and to the assessment of power-network vulnerability, the thesis focuses on the analysis of power-network topology. Global statistics traditionally used in complex-network are applied to a set of real high-voltage transmission power networks. Furthermore, a new framework is proposed to analyze the structure of complex networks.

7.1.1. Power-network topology

An in-depth description of power-network topology is crucial to generate synthetic power grids or to assess network vulnerability. Although several studies have tried to analyze and describe the topology of power networks, results were not consistent. They reached divergent conclusions. We pointed out that the use of heterogeneous data (inclusion of different voltage levels) and the use of different model assumptions (e.g., inclusion or not of transformers) are some of the causes of those inconsistencies. This work focuses on the analysis of high-voltage transmission power networks (400 kV and 220 kV).

A. Using global statistics

We study the transmission-power-network structure with a set of global statistics traditionally used in complex networks: *network size*, *degree distribution*, *characteristic path length*, *network diameter*, *betweenness centrality,* and *network average clustering coefficient*. We apply these metrics to fifteen European transmission power networks. The analysis focuses on the 400 kV and the 220 kV networks together as well as independent networks. We observe that there are topological differences among networks. In general, the 220-kV network has a less meshed structure than the 400-kV network. Furthermore, in the 400-kV network distances are lower, and the centrality of nodes is higher. Finally, the proportion of nodes that belong to the 400-kV or the 200-kV networks depends on the country.

We also analyze how the global statistics scale with network size. We observe that the number of lines correlates linearly with the number of nodes. Consequently, the average node degree of power networks can be approximated as a constant. However, the degree distribution varies among countries. Despite differences in the node degree distribution, all networks are disassortative -hubs tend to connect to poorly connected nodes- and they are not scale-free networks. The degree distribution fits better with an exponential function than with power law.

Characteristic path length and network diameter tend to grow logarithmically with the number of nodes in the European transmission power networks. Moreover, the skewness index shows that while some nodes are relatively well-connected, there is a set of nodes that are far from the core of the network. This might describe the presence of hubs, which are the center of peripherical nodes. Besides, the mean and maximum values of betweenness centrality follow a power-law concerning the number of nodes.

The network average clustering coefficient, also called the global clustering coefficient in the literature, highly varies with country and voltage level, and it does not scale with network size. The values of the network average clustering coefficient in power networks are higher than in random networks. Nevertheless, not all networks analyzed display a small-world network structure.

Although global statistics provide the first insight of power-network topology, they do not give a comprehensive characterization of the power-network structure. There are questions about network topology that are unsolved after this analysis (using global statistics). Furthermore, some of those metrics are based on average values, and they might be misleading. Moreover, the inclusion of new voltage levels or new network locations may vary metric scalability. Accordingly, this hinders the comparison among networks. Network comparison is a crucial step in order to validate the topological consistency of synthetic power grids.

B. A novel framework

To avoid the main drawbacks of global statistics, we propose an innovative tool, the GHuST framework, to analyze network topology systematically. This framework is based on graphlet decomposition (2- and 3- node graphlets). The main strengths are *full topology description*, *size independence,* and *computational simplicity*. Accordingly, this framework fully describes the structure of networks by covering the most relevant aspects of local and global properties. Furthermore, the framework explains network topology regardless of network size, and its characterized by its computational simplicity (it is calculated directly from the adjacency matrix).

The GHuST framework is defined by twelve dimensions that are grouped into four categories based on the topological aspects they cover: *global connectivity*, *hubs*, *strings,* and *triangles*. Finally, to enhance network comparison, the twelve metrics range between 0 and 1.

The application of the GHuST framework to five real networks (road, power-grid, email, social and metabolic) demonstrates that the information provided by the twelve dimensions is consistent with the global statistics traditionally used in complex networks. Furthermore, this method improves the results provided by graphlet decomposition that have been revealed insufficient. Finally, it allows for the comparison of the five networks disregarding network size.

Once the accuracy of the method is proved, we analyze a large set (1,404) of real networks of different nature (7 categories). The use of PCA to reduce the twelve dimensions allows for a graphical representation of the networks in a three-dimensional topological space defined by GHuST. There, we differentiate clusters of networks based on their topological properties. Those clusters can be identified with the seven types of networks used. Accordingly, the seven groups of networks analyzed (autonomous systems, enzymes, Facebook, power networks, retweets, roads, and webs) have different topologies. Consequently, this method enhances network classification and comparison. The twelve dimensions describe specific and intuitive aspects of network topology. That eases the interpretation of network topology and the

introduction of structural consideration in real-world applications.

Finally, the GHuST framework is applied to European transmission power networks. It completes the topological description of power networks given by global statistics. Furthermore, it allows for a straight comparison among network topologies regarding location and voltage level. As mentioned, differences are easy to understand and can be translated into the generation of synthetic power grids.

C. Contributions

The analysis of power network topology with global statistics has been published as:

- R. Espejo, S. Lumbreras, and A. Ramos, "Analysis of transmission-power-grid topology and scalability, the European case study," Physica A: Statistical Mechanics and its Applications, vol. 509, pp. 383–395, Nov. 2018.

The ρ framework has been presented in a working paper:

- R. Espejo, G. Mestre, F. Postigo, S. Lumbreras, A. Ramos, T. Huang, and E. Bompard "Exploiting graphlet-decomposition to explain the structure of complex networks."

7.1.2. Synthetic power grids

A. Topological analysis of existing synthetic power grids

Although several models were proposed in the literature to generate synthetic power grids, existing synthetic cases were not validated correctly from a topological point of view. In most cases, only a few global statistics, such as average node degree, were used. However, the use of global statistics may be insufficient to state that a synthetic power grid is topologically consistent with real power networks.

We propose the use of the GHuST framework to validate synthetic networks. We consider that synthetic networks are topologically consistent with real networks if they have similar values of GHuST for the twelve dimensions. In case there are no reference networks to compare with, we might use the range of GHuST dimensions given by the analysis of the European transmission power networks.

We applied the GHuST framework to a set of published network models: ACTIVSg, Columbia-University Synthetic Power Grid, PEGASE, and SDET networks. All those cases display topological inconsistencies concerning the European transmission power grids. The degree of those synthetic networks looks to be inconsistent with the reference network. Furthermore, those algorithms cannot replicate the local complexity of the real networks used as a reference. Since ACTIVSg, Columbia University synthetic network, and SDET stand for portions of the North American power grids, we only can conclude that there are topological inconsistencies regarding the reference. It is necessary to apply the GHuST framework to the North American power grid to discern if those differences are a consequence of the generation algorithms used

or if the North American power grid has different topology.

Beyond the inconsistencies found in the existing synthetic networks, the algorithms proposed in the literature are not flexible enough to adapt the topology of resulting synthetic networks to the structural differences found in the European transmission power networks.

B. A new model to generate synthetic power grids

This thesis proposes a novel algorithm to generate synthetic spatial power grids. Accordingly, nodes are endowed with geographical location. The algorithm uses technical and economic considerations as the most relevant factors that guide network design.

The algorithm tries to mimic the historical evolution of power networks in two steps: *building a connected graph* and *adding lines to reach topological consistency*.

The first step builds a connected graph to meet demand and generation at a minimum cost. This step is also divided into three stages to reduce the complexity of the problem: *clustering nodes*, *intra-cluster connection, and inter-cluster connection*. First, the algorithm clusters demand nodes around generators to connect them with the cheapest network that is able to supply demand (electrical considerations are included). Subsequently, clusters are connected based on reliability considerations.

The second step increases network robustness while achieving topological consistency. Consequently, the model adds new lines to ensure demand supply in case of line failure. This step is also divided into three stages: *preventing islands*, *guiding node degree* and *achieving GHuST consistency.* In the three stages, the installation of new lines is conditioned by electrical considerations and topological criteria. New lines are added only if they contribute to achieving a target degree distribution. Since networks with the same degree distribution may display different topological properties, the GHuST framework is used in the last stage to guide the generation process. Accordingly, the algorithm ensures that the resulting synthetic networks are topologically consistent with real ones.

The algorithm is tested on the Spanish, Portuguese, and French 400-kV transmission networks. The topology of the three networks (both global statistics and GHuST dimensions) is pretty similar to the topology of the real networks. Accordingly, this case proves the accuracy of the proposed algorithm to generate synthetic networks. Furthermore, the algorithm improves results given by existing algorithms.

Finally, the algorithm is flexible enough to generate networks with different topologies. This is crucial to adapt the structure of synthetic power networks to the heterogeneous topology found in the analysis of the European transmission power networks.

C. Contributions

Regarding the generation of synthetic power grids, two versions of the proposed algorithm

have been published as:

- R. Espejo, S. Lumbreras, and A. Ramos, "A Complex-Network Approach to the Generation of Synthetic Power Transmission Networks," IEEE Systems Journal, pp. 1–9, 2018.
- R. Espejo, S. Lumbreras, and A. Ramos, "Generating statistically consistent synthetic power networks for testing renewable integration models", Windfarms 2017, Madrid, Spain, Jun 2017.

7.1.3. Vulnerability assessment

A. A new hybrid metric

The use of complex networks in power systems can also support the assessment of power-network vulnerability. Several complex-network-based indices were proposed to rank the impact a line failure would have on a power network.

Although purely topological metrics were proved to provide non-accurate results because of their lack of electrical considerations, hybrid metrics find a balance between result accuracy and computational requirements.

We find several indices in the literature that combine global statistics traditionally used in complex networks with electrical considerations. However, in most cases, results were not tested, and some of them also need for computationally intensive models. We propose a new metric to assess power-network vulnerability: Electrical Line Centrality.

This hybrid metrics endows betweenness centrality with electrical information related to nodes (generation capacity and power demand) and transmission lines (line impedance).

The application of the new metric to the IEEE 9-bus test case and the IEEE118-bus test case shows that the Electrical Line Centrality supplies better results than pure topological metrics and prior hybrid metrics such as the extended betweenness centrality.

Furthermore, it drastically reduces computational requirements. In contrast with most of the prior metrics, the Electrical-Line-Centrality index is a linear function. The linearity of vulnerability indices is crucial to include them in the network design problem.

Finally, the ELC may also contribute to the increase of power-network resilience. It might support the design of power networks that are robust against deliberate attacks or cascade failures. Although the probability of those events might be low, they might cause network failures with severe consequences.

B. Contributions

The proposal of the line electrical centrality has been published as:

- R. Espejo, S. Lumbreras, A. Ramos, T. Huang, and E. Bompard, "An extended metric for the

analysis of power-network vulnerability: the line electrical centrality", PowerTech 2019, Milan, Italy, Jun. 2019.

7.2. Further research

This thesis proves the advantages of applying complex-network techniques to power systems and the insights that can be gained by using these techniques. Although the thesis contributes to the generation of synthetic power grids as well as in the assessment of power-network vulnerability; there are questions that require further research, which we summarize in the following lines.

7.2.1. Network topology

The GHuST framework might be the seed for the development of a novel method to classify complex networks based on topology. Furthermore, it would be the base of a new algorithm to generate graphs with predefined topological properties.

A. Network description and classification

The twelve dimensions of the GHuST framework allow for an in-depth description and classification of complex-network topology. The inclusion of new network instances may lead to the definition of topological standards or benchmarks (typical values of the twelve dimensions of GHuST for different types of networks). This would support network clustering and classification. The main drawback of this task is the lack of large network datasets that may condition the statistical significance of results. Further analyzes may also imply the definition of new GHuST dimensions or the modification of current dimensions to better capture other real networks properties.

The use of the GHuST framework to compare networks (as it is done in the synthetic network validation procedure) might be used to detect anomalies in the topology of complex networks. For instance, in social networks, the definition of typical values for each GHuST dimension might help to support the detection of "bots" used to increase the impact of news or marketing campaigns. The topology of bots' connections may differ from other accounts with a "normal" pattern of connections. Accordingly, bots might be located in a different place in the topological space defined by the GHuST framework. It would be necessary to define the ranges of expected values for different types of networks. As previously mentioned, this requires a broad set of networks to extract significant conclusions.

Furthermore, the GHuST framework provides a full description of network topology. This description needs to be effectively connected to network operational behavior and network design. Further research should determine how topology conditions network operation. For example, this might help to answer questions such as Do triangles increase network

robustness? It would help to discover the role of some structures in network operation. However, the main challenge is the bridge between global and local properties. Are nodal prices (local property) conditioned by the number of triangles is the network (global property)? However, there are other features that might be addressed with the GHuST framework, such as the risk of cascading failure in case of line failure.

In the case of power systems, topology would help to operate, design, and control power grids. New challenges such as the protection against deliberate attacks (N-X contingencies) or distributed generation (the same node might inject or withdraw power) may be faced with topology-based methods. As explained in this thesis, methodologies traditionally used in power systems may result in unmanageable tools for these problems. It, therefore, requires a sound analysis of the relation between network topology and electrical behavior.

B. Graph generation

As explained in Chapter 5, synthetic-network generation algorithms may disregard the specific nature of the networks under study. They may generate graphs defined by an adjacency matrix (only ones and zeros) with predefined topological properties. Accordingly, those models only consider the distribution of connections among nodes (e.g. the Preferential Attachment model). In the Preferential Attachment model, nodes are connected following a power law. Users can, therefore, predefine the degree distribution of the synthetic network. Similarly, a new model might allow for the generation of synthetic networks in which the users would define the topology of the resulting network (with the GHuST dimensions) in advance.

This might be useful to generate a broad set of networks that will support an empirical analysis of the relation among GHuST dimensions. In Chapter 5, we pointed out that the lack of understanding about the correlation of GHuST dimensions hinders the generation of synthetic power grids (the model could not analyze the contribution of an individual line to reach a target GHuST). Furthermore, it would help to determine if GHuST dimensions have lower or upper bounds. In Chapter 3, we explained that the number of triangles in a network might have an upper limit since an increase in G_3 (number of triangles) would also lead to an increase in G_2. Accordingly, a model to generate networks with specific topological properties would help to answer those questions.

7.2.2. Road to more realistic synthetic power grids

This thesis proposed an algorithm to generate synthetic power grids that are topologically consistent with real networks. However, to meet the complexity of real power networks, new research and improvements should be introduced.

A. Increasing the complexity of synthetic power grids

The proposed algorithm was tested with a single voltage layer, 400-kV networks. The

addition of new voltage levels (e.g., 200-kV) should be considered in further improvements. Although the algorithm is flexible to adapt the resulting topology to the properties of other voltage levels, it is necessary to determine other aspects such as the location of transformers. Furthermore, new improvements should consider the **inclusion in the wiring process of a wide range of connections**, for instance, double circuits.

More in-depth analyses are also needed to endow nodes with more realistic properties. Although this model focuses on the wiring process, and it assumes the hypothesis done by prior work regarding nodes, new improvements would lead to an accurate characterization of nodes (location and demand/generation properties). Furthermore, it would allow for the generation of multiple scenarios of generation and demand.

Finally, **the integration of geographical information** would also increase the realism of resulting networks. Geographical information might condition line installation by varying investment cost (e.g., it would increase line cost if the line goes through a natural park). Furthermore, it would give a better estimation of line length or geographical path.

B. Multilayer networks

We propose the **inclusion of new transmission voltage levels** to improve the accuracy of network models. However, as explained before, the higher degree of interconnection with other networks need to be captured in network models.

The **inclusion of distribution networks** (low-voltage and medium-voltage power networks) is crucial since they were proved to play a crucial role in the propagation of cascading failure. In this case, the main challenge is the connection of both networks since the large size of the resulting network may hamper the inclusion of electrical considerations such as power flows. The connection with other network models, such as gas or communication networks, would also be worthy. Those models would help to understand failure propagation among interdependent networks. Furthermore, it might support the increase of power network robustness by investments in other cheaper infrastructures, such as telecommunication networks. To reach that goal, the antecedent, and crucial step is the research into other network models.

7.2.3. Network vulnerability

Although the Electrical Line Centrality improves results given by prior works, new improvements are necessary to effectively introduced vulnerability indices in network design.

The objective of assessing power-network vulnerability is the designing of more robust networks. Consequently, **vulnerability metrics should be included in the network design process**. In Chapter 6 we pointed out that it might be considered by introducing vulnerability indices as a partial objective in the optimization function (it penalizes high values of vulnerability indices) or by including them as constraints (it establishes maximum values for the indices).

Consequently, it would be necessary to define a vulnerability cost or a maximum value of topological vulnerability to be included in the optimization problem. Furthermore, topological indices allow for the assessment of the impact component failures have on network vulnerability. Similarly, we can also assess the impact that new lines would have on the network. Accordingly, vulnerability indices might be used to propose candidates to be installed in the transmission expansion problem.

Complex-network techniques are a key tool in the generation of synthetic power grids and in the assessment of power network vulnerability. The inclusion of electrical considerations in complex-network theory find a balance between result accuracy and computational requirements.

This thesis proposes a novel algorithm to generate synthetic spatial power grids that generates networks that are topologically consistent with real power networks. Furthermore, it introduces a novel metric to assess power-network vulnerability. It improves results given by prior metrics and reduces computational complexity. Finally, in the analysis of network topology, this thesis contributes with a novel framework that allows for the description of network topology as well as the comparison among networks

Despite those advances, further research is needed to build more complex synthetic power grids or to introduce vulnerability indices in the network design problem.

References

[1] S. V. Buldyrev, R. Parshani, G. Paul, H. E. Stanley, and S. Havlin, “Catastrophic cascade of failures in interdependent networks,” *Nature*, vol. 464, no. 7291, p. 1025, Apr. 2010.

[2] T. Li, M. Eremia, and M. Shahidehpour, “Interdependency of Natural Gas Network and Power System Security,” *IEEE Transactions on Power Systems*, vol. 23, no. 4, pp. 1817–1824, Nov. 2008.

[3] A. G. Exposito, A. J. Conejo, and C. Canizares, Electric Energy Systems: Analysis and Operation. CRC Press, 2016.

[4] “About INSPIRE” [Online]. Available: https://inspire.ec.europa.eu/about-inspire/563. [Accessed: 01-May-2019].

[5] “Power Systems Test Case Archive - UWEE.” [Online]. Available: http://labs.ece.uw.edu/pstca/. [Accessed: 20-Apr-2019].

[6] “Power Systems Test Case Archive.” [Online]. Available: https://www.maths.ed.ac.uk/optenergy/NetworkData/. [Accessed: 20-Apr-2019].

[7] F. E. Postigo Marcos et al., “A Review of Power Distribution Test Feeders in the United States and the Need for Synthetic Representative Networks,” *Energies*, vol. 10, no. 11, p. 1896, Nov. 2017.

[8] R. D. Zimmerman, C. E. Murillo-Sanchez, and R. J. Thomas, “MATPOWER: Steady-State Operations, Planning, and Analysis Tools for Power Systems Research and Education,” *IEEE Transactions on Power Systems*, vol. 26, no. 1, pp. 12–19, Feb. 2011.

[9] Colffrin, C., Gordon, D., and Scott, P. "NESTA, the NICTA energy system test case archive,"2014.

[10] T. V. Jensen and P. Pinson, “RE-Europe, a large-scale dataset for modeling a highly renewable European electricity system,” *Scientific Data*, vol. 4, p. 170175, Nov. 2017.

[11] A. Xenophon and D. Hill, “Open grid model of Australia’s National Electricity Market allowing backtesting against historic data,” *Scientific Data*, vol. 5, p. 180203, Oct. 2018.

[12] “TYNDP Maps + Data.” [Online]. Available: https://www.entsoe.eu/major-projects/ten-year-network-development-plan/maps-and-data/Pages/default.aspx. [Accessed: 17-Jul-2017].

[13] “La carte du réseau, RTE France.” [Online]. Available: https://www.rte-france.com/fr/la-carte-du-reseau. [Accessed: 20-Apr-2019].

[14] M. Haklay and P. Weber, “OpenStreetMap: User-Generated Street Maps,” I*EEE Pervasive Computing*, vol. 7, no. 4, pp. 12–18, Oct. 2008.

[15] W. Medjroubi, C. Matke, and D. Kleinhans, “SciGRID – An Open Source Reference Model for the European Transmission Network,” 2015.

[16] Wupperinst, Open source German transmission grid model based on OpenStreetMap, 2019.

[17] J. Rivera, P. Nasirifard, J. Leimhofer, and H. Jacobsen, "Automatic Generation of Real Power Transmission Grid Models from Crowdsourced Data," *IEEE Transactions on Smart Grid*, p. 1, 2018.

[18] J. Rivera, C. Goebel, D. Sardari, and H.-A. Jacobsen, "OpenGridMap: An Open Platform for Inferring Power Grids with Crowdsourced Data," in *Energy Informatics*, 2015, pp. 179–191.

[19] W. Medjroubi, U. P. Müller, M. Scharf, C. Matke, and D. Kleinhans, "Open Data in Power Grid Modelling: New Approaches Towards Transparent Grid Models," *Energy Reports*, vol. 3, pp. 14–21, Nov. 2017.

[20] "ARPA-E | GRID DATA." [Online]. Available: https://arpa-e.energy.gov/?q=arpa-e-programs/grid-data. [Accessed: 09-Apr-2019].

[21] Z. Wang, A. Scaglione, and R. J. Thomas, "Generating Statistically Correct Random Topologies for Testing Smart Grid Communication and Control Networks," *IEEE Transactions on Smart Grid*, vol. 1, no. 1, pp. 28–39, Jun. 2010.

[22] A. B. Birchfield, K. M. Gegner, T. Xu, K. S. Shetye, and T. J. Overbye, "Statistical Considerations in the Creation of Realistic Synthetic Power Grids for Geomagnetic Disturbance Studies," *IEEE Transactions on Power Systems*, vol. 32, no. 2, pp. 1502–1510, Mar. 2017.

[23] C. M. Domingo, T. G. S. Roman, Á. Sanchez-Miralles, J. P. P. Gonzalez, and A. C. Martinez, "A Reference Network Model for Large-Scale Distribution Planning With Automatic Street Map Generation," *IEEE Transactions on Power Systems*, vol. 26, no. 1, pp. 190–197, Feb. 2011.

[24] L. L. Lai, H. T. Zhang, S. Mishra, D. Ramasubramanian, C. S. Lai, and F. Y. Xu, "Lessons learned from July 2012 Indian blackout," in *9th IET International Conference on Advances in Power System Control, Operation and Management (APSCOM 2012)*, 2012, pp. 1–6.

[25] NERC, "Glossary of Terms Used in NERC Reliability Standards," 2017.

[26] J. W. Bialek, "Why has it happened again? Comparison between the UCTE blackout in 2006 and the blackouts of 2003," in *2007 IEEE Lausanne Power Tech*, 2007, pp. 51–56.

[27] G. Liang, S. R. Weller, J. Zhao, F. Luo, and Z. Y. Dong, "The 2015 Ukraine Blackout: Implications for False Data Injection Attacks," *IEEE Transactions on Power Systems*, vol. 32, no. 4, pp. 3317–3318, Jul. 2017.

[28] Q. Chen and J. D. McCalley, "Identifying high risk N-k contingencies for online security assessment," *IEEE Transactions on Power Systems*, vol. 20, no. 2, pp. 823–834, May 2005.

[29] J. Salmeron, K. Wood, and R. Baldick, "Analysis of electric grid security under terrorist threat," *IEEE Transactions on Power Systems*, vol. 19, no. 2, pp. 905–912, May 2004.

[30] J. M. Arroyo and F. D. Galiana, "On the solution of the bilevel programming formulation

of the terrorist threat problem," *IEEE Transactions on Power Systems*, vol. 20, no. 2, pp. 789–797, May 2005.

[31] A. L. Motto, J. M. Arroyo, and F. D. Galiana, "A Mixed-Integer LP Procedure for the Analysis of Electric Grid Security Under Disruptive Threat," *IEEE Transactions on Power Systems*, vol. 20, no. 3, pp. 1357–1365, Aug. 2005.

[32] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang, "Complex networks: Structure and dynamics," *Physics Reports*, vol. 424, no. 4, pp. 175–308, Feb. 2006.

[33] L. da F. Costa et al., "Analyzing and modeling real-world phenomena with complex networks: a survey of applications," *Advances in Physics*, vol. 60, no. 3, pp. 329–412, Jun. 2011.

[34] Y. B. Kim et al., "Predicting the Currency Market in Online Gaming via Lexicon-Based Analysis on Its Online Forum," *Complexity*, 2017.

[35] H. Cetinay, F. A. Kuipers, and P. V. Mieghem, "A Topological Investigation of Power Flow," *IEEE Systems Journal*, vol. PP, no. 99, pp. 1–9, 2016.

[36] M. Newman, Networks: An Introduction. Oxford, New York: Oxford University Press, 2010.

[37] A.-L. Barabási and R. Albert, "Emergence of Scaling in Random Networks," *Science*, vol. 286, no. 5439, pp. 509–512, Oct. 1999.

[38] G. A. Pagani and M. Aiello, "The Power Grid as a complex network: A survey," *Physica A: Statistical Mechanics and its Applications*, vol. 392, no. 11, pp. 2688–2700, Jun. 2013.

[39] E. Cotilla-Sanchez, P. D. H. Hines, C. Barrows, and S. Blumsack, "Comparing the Topological and Electrical Structure of the North American Electric Power Infrastructure," *IEEE Systems Journal*, vol. 6, no. 4, pp. 616–626, Dec. 2012.

[40] L. Cuadra, S. Salcedo-Sanz, J. Del Ser, S. Jiménez-Fernández, and Z. W. Geem, "A Critical Review of Robustness in Power Grids Using Complex Networks Concepts," *Energies*, vol. 8, no. 9, pp. 9211–9265, Aug. 2015.

[41] S. Lumbreras and A. Ramos, "The new challenges to transmission expansion planning. Survey of recent practice and literature review," *Electric Power Systems Research*, vol. 134, pp. 19–29, May 2016.

[42] D. Citron et al., "Part III: Using Topological Information to Build More Robust Networks," unpublished.

[43] P. Erdos and A. Rényi, "On the evolution of random graphs," *Publ. Math. Inst. Hung. Acad. Sci*, vol. 5, no. 1, pp. 17–60, 1960.

[44] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, no. 6684, pp. 440–442, Jun. 1998.

[45] M. A. Saniee Monfared, M. Jalili, and Z. Alipour, "Topology and vulnerability of the Iranian power grid," *Physica A: Statistical Mechanics and its Applications*, vol. 406, no.

Supplement C, pp. 24–33, Jul. 2014.

[46] D. H. Kim, D. A. Eisenberg, Y. H. Chun, and J. Park, "Network topology and resilience analysis of South Korean power grid," *Physica A: Statistical Mechanics and its Applications*, vol. 465, no. Supplement C, pp. 13–24, Jan. 2017.

[47] M. E. J. Newman, "Analysis of weighted networks," *Physical Review E*, vol. 70, no. 5, Nov. 2004.

[48] M. Wiedermann, J. F. Donges, J. Heitzig, and J. Kurths, "Node-weighted interacting network measures improve the representation of real-world complex systems," *EPL (Europhysics Letters)*, vol. 102, no. 2, p. 28007, Apr. 2013.

[49] E. Ravasz and A.-L. Barabási, "Hierarchical organization in complex networks," *Phys. Rev. E*, vol. 67, no. 2, p. 026112, Feb. 2003.

[50] M. Barthélemy, "Spatial networks," *Physics Reports*, vol. 499, no. 1, pp. 1–101, Feb. 2011.

[51] S. Boccaletti et al., "The structure and dynamics of multilayer networks," *Physics Reports*, vol. 544, no. 1, pp. 1–122, Nov. 2014.

[52] M. Kivelä, A. Arenas, M. Barthelemy, J. P. Gleeson, Y. Moreno, and M. A. Porter, "Multilayer networks," *J Complex Netw*, vol. 2, no. 3, pp. 203–271, Sep. 2014.

[53] W. Deng, W. Li, X. Cai, and Q. A. Wang, "The exponential degree distribution in complex networks: Non-equilibrium network theory, numerical simulation and empirical data," *Physica A: Statistical Mechanics and its Applications*, vol. 390, no. 8, pp. 1481–1485, Apr. 2011.

[54] M. E. J. Newman, "Assortative Mixing in Networks," *Phys. Rev. Lett.*, vol. 89, no. 20, p. 208701, Oct. 2002.

[55] R. Cohen and S. Havlin, "Scale-Free Networks Are Ultrasmall," *Physical Review Letters*, vol. 90, no. 5, Feb. 2003.

[56] M. D. Humphries and K. Gurney, "Network 'Small-World-Ness': A Quantitative Method for Determining Canonical Network Equivalence," *PLoS ONE*, vol. 3, no. 4, Apr. 2008.

[57] G. A. Pagani and M. Aiello, "Towards Decentralization: A Topological Investigation of the Medium and Low Voltage Grids," *IEEE Transactions on Smart Grid*, vol. 2, no. 3, pp. 538–547, Sep. 2011.

[58] S. Soltan and G. Zussman, "Generation of synthetic spatially embedded power grid networks," in 2*016 IEEE Power and Energy Society General Meeting (PESGM)*, 2016, pp. 1–5.

[59] A. B. Birchfield, T. Xu, K. M. Gegner, K. S. Shetye, and T. J. Overbye, "Grid Structural Characteristics as Validation Criteria for Synthetic Networks," I*EEE Transactions on Power Systems*, vol. PP, no. 99, pp. 1–1, 2016.

[60] U. Alon, "Network motifs: theory and experimental approaches," *Nature Reviews Genetics*, vol. 8, no. 6, pp. 450–461, Jun. 2007.

[61] G. Palla, I. Derényi, I. Farkas, and T. Vicsek, “Uncovering the overlapping community structure of complex networks in nature and society,” *Nature*, vol. 435, no. 7043, pp. 814–818, Jun. 2005.

[62] “Scale-Free Networks,” Scientific American. [Online]. Available: https://www.scientificamerican.com/article/scale-free-networks/. [Accessed: 23-Mar-2019].

[63] HADDADI, Hamed, et al. “On the importance of local connectivity for Internet topology models" in *2009 21st International Teletraffic Congress*, 2009.

[64] R. Albert and A.-L. Barabási, “Topology of Evolving Networks: Local Events and Universality,” *Phys. Rev. Lett.*, vol. 85, no. 24, pp. 5234–5237, Dec. 2000.

[65] R. Espejo, S. Lumbreras, and A. Ramos, “A Complex-Network Approach to the Generation of Synthetic Power Transmission Networks,” *IEEE Systems Journal*, pp. 1–4, 2018.

[66] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, “Network motifs: simple building blocks of complex networks,” *Science*, vol. 298, no. 5594, pp. 824–827, Oct. 2002.

[67] Y. Artzy-Randrup, S. J. Fleishman, N. Ben-Tal, and L. Stone, “Comment on ‘Network Motifs: Simple Building Blocks of Complex Networks’ and ‘Superfamilies of Evolved and Designed Networks,’” *Science*, vol. 305, no. 5687, pp. 1107–1107, Aug. 2004.

[68] A. Mazurie, S. Bottani, and M. Vergassola, “An evolutionary and functional assessment of regulatory network motifs,” *Genome Biology*, vol. 6, no. 4, p. R35, Mar. 2005.

[69] N. Przulj, D. G. Corneil, and I. Jurisica, “Modeling interactome: scale-free or geometric?,” *Bioinformatics*, vol. 20, no. 18, pp. 3508–3515, Dec. 2004.

[70] T. Milenkoviæ and N. Pržulj, “Uncovering Biological Network Function via Graphlet Degree Signatures,” *Cancer Inform*, vol. 6, pp. 257–273, Apr. 2008.

[71] N. K. Ahmed, J. Neville, R. A. Rossi, N. G. Duffield, and T. L. Willke, “Graphlet decomposition: framework, algorithms, and applications,” *Knowl Inf Syst*, vol. 50, no. 3, pp. 689–722, Mar. 2017.

[72] D. Marcus and Y. Shavitt, “RAGE – A rapid graphlet enumerator for large networks,” *Computer Networks*, vol. 56, no. 2, pp. 810–819, Feb. 2012.

[73] S. Wernicke and F. Rasche, “FANMOD: a tool for fast network motif detection,” *Bioinformatics*, vol. 22, no. 9, pp. 1152–1153, May 2006.

[74] T. Hočevar and J. Demšar, “A combinatorial approach to graphlet counting,” *Bioinformatics*, vol. 30, no. 4, pp. 559–565, Feb. 2014.

[75] R. Itzhack, Y. Mogilevski, and Y. Louzoun, “An optimal algorithm for counting network motifs,” *Physica A: Statistical Mechanics and its Applications,* vol. 381, pp. 482–490, Jul. 2007.

[76] T. Milenković, W. L. Ng, W. Hayes, and N. Pržulj, “Optimal Network Alignment with

Graphlet Degree Vectors," *Cancer Inform*, vol. 9, pp. 121–137, Jun. 2010.

[77] J. Crawford and T. Milenković, "GREAT: GRaphlet Edge-based network AlignmenT," arXiv:1410.5103 [cs, q-bio], Oct. 2014.

[78] N. Malod-Dognin and N. Pržulj, "L-GRAAL: Lagrangian graphlet-based network aligner," *Bioinformatics*, vol. 31, no. 13, pp. 2182–2189, Jul. 2015.

[79] N. Malod-Dognin and N. Pržulj, "GR-Align: fast and flexible alignment of protein 3D structures using graphlet degree similarity," *Bioinformatics*, vol. 30, no. 9, pp. 1259–1265, May 2014.

[80] Hayes, W., Sun, K., & Pržulj, N. "Graphlet-based measures are suitable for biological network comparison," *Bioinformatics*, vol. 29, no. 4, pp. 483-491, Jan 2013

[81] Rossi, Ryan A., and Nesreen K. Ahmed, "Role Discovery in Networks" *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 4, pp. 1112-1131, Jan 2014.

[82] Ö. N. Yaveroğlu et al., "Revealing the Hidden Language of Complex Networks," *Scientific Reports*, vol. 4, p. 4547, Apr. 2014.

[83] R. Rossi and N. Ahmed, "The Network Data Repository with Interactive Graph Analytics and Visualization," in *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[84] R. Guimerà, L. Danon, A. Díaz-Guilera, F. Giralt, and A. Arenas, "The real communication network behind the formal chart: Community structure in organizations," J*ournal of Economic Behavior & Organization*, vol. 61, no. 4, pp. 653–667, Dec. 2006.

[85] J. Schellenberger, J. O. Park, T. M. Conrad, and B. Ø. Palsson, "BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions," *BMC Bioinformatics*, vol. 11, no. 1, p. 213, Apr. 2010.

[86] H. Ebel, L.-I. Mielsch, and S. Bornholdt, "Scale-free topology of e-mail networks," *Phys. Rev. E*, vol. 66, no. 3, p. 035103, Sep. 2002.

[87] "SNAP: Network datasets: Autonomous systems - AS-733." [Online]. Available: https://snap.stanford.edu/data/as-733.html. [Accessed: 23-Mar-2019].

[88] F. Xie and D. Levinson, "Modeling the Growth of Transportation Networks: A Comprehensive Review," *Netw Spat Econ*, vol. 9, no. 3, pp. 291–307, Sep. 2009.

[89] R. Espejo, S. Lumbreras, and A. Ramos, "Analysis of transmission-power-grid topology and scalability, the European case study," *Physica A: Statistical Mechanics and its Applications*, vol. 509, pp. 383–395, Nov. 2018.

[90] S. Soltan, A. Loh, and G. Zussman, "A Learning-Based Method for Generating Synthetic Power Grids," *IEEE Systems Journal*, vol. 13, no. 1, pp. 625–634, Mar. 2019.

[91] "DR POWER, Data Repository for Power system Open models With Evolving Resources." [Online]. Available: https://egriddata.org/. [Accessed: 10-Apr-2019].

[92] S. Soltan, A. Loh, and G. Zussman, "Columbia University Synthetic Power Grid with Geographical Coordinates." Data Repository for Power system Open models With

Evolving Resources (DR POWER); Department of Electrical Engineering at Columbia University, 2018.

[93] S. Fliscounakis, P. Panciatici, F. Capitanescu, and L. Wehenkel, "Contingency Ranking With Respect to Overloads in Very Large Power Systems Taking Into Account Uncertainty, Preventive, and Corrective Actions," *IEEE Transactions on Power Systems*, vol. 28, no. 4, pp. 4909–4917, Nov. 2013.

[94] C. Josz, S. Fliscounakis, J. Maeght, and P. Panciatici, "AC Power Flow Data in MATPOWER and QCQP Format: iTesla, RTE Snapshots, and PEGASE,"unpublished, Mar. 2016.

[95] R. Diao, "Sustainable Data Evolution Technology (SDET) for Power Grid Optimization," unpublished.

[96] "SDET Tool, DR POWER." [Online]. Available: https://egriddata.org/dataset/sdet-tool. [Accessed: 16-Apr-2019].

[97] Z. Wang and R. J. Thomas, "On Bus Type Assignments in Random Topology Power Grid Models," in *2015 48th Hawaii International Conference on System Sciences*, 2015, pp. 2671–2679.

[98] S. H. Elyas and Z. Wang, "A Multi-objective Optimization Algorithm for Bus Type Assignments in Random Topology Power Grid Model," in *2016 49th Hawaii International Conference on System Sciences (HICSS)*, 2016, pp. 2446–2455.

[99] J. Hu, L. Sankar, and D. J. Mir, "Cluster-and-Connect: An algorithmic approach to generating synthetic electric power network graphs," in *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 2015, pp. 223–230.

[100] S. S. Manna and P. Sen, "Modulated scale-free network in Euclidean space," *Physical Review E*, vol. 66, no. 6, Dec. 2002.

[101] B. M. Waxman, "Routing of multipoint connections," *IEEE Journal on Selected Areas in Communications*, vol. 6, no. 9, pp. 1617–1622, Dec. 1988.

[102] A. Patania et al., "Part I: Generating Random Networks that are Consistent with Power Transmission." unpublished.

[103] Z. Wang, R. J. Thomas, and A. Scaglione, "Generating Random Topology Power Grids," in *Proceedings of the 41st Annual Hawaii International Conference on System Sciences (HICSS 2008)*, 2008, pp. 183–183.

[104] D. Deka and S. Vishwanath, "Generative Growth Model for Power Grids," in *2013 International Conference on Signal-Image Technology Internet-Based Systems*, 2013, pp. 591–598.

[105] 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 *2016 IEEE Power and Energy Conference at Illinois (PECI)*, 2016, pp. 1–6.

[106] P. Schultz, J. Heitzig, and J. Kurths, "A random growth model for power grids and other spatially embedded infrastructure networks," *Eur. Phys. J. Spec. Top.*, vol. 223, no. 12, pp.

2593–2610, Oct. 2014.

[107] K. Wang, B. Zhang, Z. Zhang, X. Yin, and B. Wang, “An electrical betweenness approach for vulnerability assessment of power grids considering the capacity of generators and load,” *Physica A: Statistical Mechanics and its Applications*, vol. 390, no. 23–24, pp. 4692–4701, Nov. 2011.

[108] C. J. Kim and O. B. Obah, “Vulnerability Assessment of Power Grid Using Graph Topological Indices,” *International Journal of Emerging Electric Power Systems*, vol. 8, no. 6, 2007.

[109] L. Cuadrado Granda, “Validación y análisis de medidas de redes complejas aplicado a redes eléctricas de países europeos,” Universidad Pontificia Comillas, 2018.

[110] V. Latora and M. Marchiori, “Efficient Behavior of Small-World Networks,” P*hysical Review Letters,* vol. 87, no. 19, Oct. 2001.

[111] L. Che, X. Liu, Z. Shuai, Z. Li, and Y. Wen, “Cyber Cascades Screening Considering the Impacts of False Data Injection Attacks,” *IEEE Transactions on Power Systems,* pp. 1–1, 2018.

[112] Yang, Y., Nishikawa, T., & Motter, A. E. “Small vulnerable sets determine large network cascades in power grids" *Science*, vol. 358, no 6365, p. eaan3184, Jul 2018.

[113] P. Dey, R. Mehra, F. Kazi, S. Wagh, and N. M. Singh, “Impact of Topology on the Propagation of Cascading Failure in Power Grid,” *IEEE Transactions on Smart Grid*, vol. 7, no. 4, pp. 1970–1978, Jul. 2016.

[114] A. J. Holmgren, “Using Graph Models to Analyze the Vulnerability of Electric Power Networks,” *Risk Analysis*, vol. 26, no. 4, pp. 955–969, Aug. 2006.

[115] V. Latora and M. Marchiori, “Vulnerability and protection of infrastructure networks,” *Phys. Rev. E*, vol. 71, no. 1, p. 015103, Jan. 2005.

[116] P. Crucitti, V. Latora, and M. Marchiori, “Locating critical lines in high-voltage electrical power grids,” *Fluct. Noise Lett.*, vol. 05, no. 02, pp. L201–L208, Jun. 2005.

[117] P. Crucitti, V. Latora, and M. Marchiori, “A topological analysis of the Italian electric power grid,” *Physica A: Statistical Mechanics and its Applications*, vol. 338, no. 1–2, pp. 92–97, Jul. 2004.

[118] R. Kinney, P. Crucitti, R. Albert, and V. Latora, “Modeling cascading failures in the North American power grid,” *Eur. Phys. J. B*, vol. 46, no. 1, pp. 101–107, Jul. 2005.

[119] M. Rosas Casals and B. Corominas Murtra, “Assessing European power grid reliability by means of topological measures,” *WIT transactions on ecology and the environment,* vol. 121, pp. 527–537, 2009.

[120] P. Hines, E. Cotilla-Sanchez, and S. Blumsack, “Do topological models provide good information about electricity infrastructure vulnerability?,” *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 20, no. 3, p. 033122, Sep. 2010.

[121] J.-W. Wang and L.-L. Rong, "Cascade-based attack vulnerability on the US power grid," *Safety Science*, vol. 47, no. 10, pp. 1332–1336, Dec. 2009.

[122] R. Albert, I. Albert, and G. L. Nakarado, "Structural vulnerability of the North American power grid," *Physical Review E*, vol. 69, no. 2, Feb. 2004.

[123] M. Ouyang, Z. Pan, L. Hong, and L. Zhao, "Correlation analysis of different vulnerability metrics on power grids," *Physica A: Statistical Mechanics and its Applications*, vol. 396, pp. 204–211, Feb. 2014.

[124] A. B. M. Nasiruzzaman, H. R. Pota, and A. Anwar, "Comparative study of power grid centrality measures using complex network framework," in *2012 IEEE International Power Engineering and Optimization Conference Melaka, Malaysia*, 2012, pp. 176–181.

[125] A. B. M. Nasiruzzaman, H. R. Pota, and F. R. Islam, "Complex network framework based dependency matrix of electric power grid," *in AUPEC 2011*, 2011, pp. 1–6.

[126] A. B. M. Nasiruzzaman and H. R. Pota, "Transient stability assessment of smart power system using complex networks framework," in *2011 IEEE Power and Energy Society General Meeting*, 2011, pp. 1–7.

[127] A. B. M. Nasiruzzaman, H. R. Pota, and M. A. Barik, "Implementation of bidirectional power flow based centrality measure in bulk and smart power transmission systems," in *IEEE PES Innovative Smart Grid Technologies*, 2012, pp. 1–6.

[128] A. B. M. Nasiruzzaman, H. R. Pota, A. Anwar, and F. R. Islam, "Modified centrality measure based on bidirectional power flow for smart and bulk power transmission grid," *in 2012 IEEE International Power Engineering and Optimization Conference Melaka*, Malaysia, 2012, pp. 159–164.

[129] A. B. M. Nasiruzzaman, H. R. Pota, and M. A. Mahmud, "Application of centrality measures of complex network framework in power grid," in *IECON 2011 - 37th Annual Conference of the IEEE Industrial Electronics Society*, 2011, pp. 4660–4665.

[130] C. Liu, Q. Xu, Z. Chen, and C. L. Bak, "Vulnerability evaluation of power system integrated with large-scale distributed generation based on complex network theory," in *2012 47th International Universities Power Engineering Conference (UPEC)*, 2012, pp. 1–5.

[131] H. Bai and S. Miao, "Hybrid flow betweenness approach for identification of vulnerable line in power system," *Transmission Distribution IET Generation*, vol. 9, no. 12, pp. 1324–1331, 2015.

[132] E. Bompard, R. Napoli, and F. Xue, "Analysis of structural vulnerabilities in power transmission grids," *International Journal of Critical Infrastructure Protection*, vol. 2, no. 1–2, pp. 5–12, May 2009.

[133] S. Arianos, E. Bompard, A. Carbone, and F. Xue, "Power grid vulnerability: A complex network approach," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 19, no. 1, p. 013119, Mar. 2009.

[134] Wang, X., Koç, Y., Kooij, R. E., & Van Mieghem, P. "A network approach for power grid

robustness against cascading failures" 2*015 7th international workshop on reliable networks design and modeling (RNDM)*, pp. 208-214.

[135] Y. Koç, M. Warnier, R. Kooij, and F. Brazier, "Structural vulnerability assessment of electric power grids," in *Proceedings of the 11th IEEE International Conference on Networking, Sensing and Control*, 2014, pp. 386–391.

[136] P. Hines, S. Blumsack, E. C. Sanchez, and C. Barrows, "The Topological and Electrical Structure of Power Grids," in *2010 43rd Hawaii International Conference on System Sciences*, 2010, pp. 1–10.

[137] P. Hines and S. Blumsack, "A Centrality Measure for Electrical Networks," in *Proceedings of the 41st Annual Hawaii International Conference on System Sciences,* 2008, pp. 185–185.

[138] A. Dwivedi and X. Yu, "A Maximum-Flow-Based Complex Network Approach for Power System Vulnerability Analysis," *IEEE Transactions on Industrial Informatics*, vol. 9, no. 1, pp. 81–88, Feb. 2013.

[139] J. Fang, C. Su, Z. Chen, H. Sun, and P. Lund, "Power System Structural Vulnerability Assessment Based on an Improved Maximum Flow Approach," I*EEE Transactions on Smart Grid*, vol. 9, no. 2, pp. 777–785, Mar. 2018.

[140] E. Bompard, D. Wu, and F. Xue, "Structural vulnerability of power systems: A topological approach," *Electric Power Systems Research,* vol. 81, no. 7, pp. 1334–1340, Jul. 2011.

[141] E. Bompard, E. Pons, and D. Wu, "Extended Topological Metrics for the Analysis of Power Grid Vulnerability," *IEEE Systems Journal*, vol. 6, no. 3, pp. 481–487, Sep. 2012.

[142] E. Bompard, L. Luo, and E. Pons, "A perspective overview of topological approaches for vulnerability analysis of power transmission grids," I*nternational Journal of Critical Infrastructures*, vol. 11, no. 1, pp. 15–26, Jan. 2015.

[143] E. Bompard, D. Wu, and F. Xue, "The Concept of Betweenness in the Analysis of Power Grid Vulnerability," in 2*010 Complexity in Engineering*, 2010, pp. 52–54.

[144] X. Wei, J. Zhao, T. Huang, and E. Bompard, "A Novel Cascading Faults Graph Based Transmission Network Vulnerability Assessment Method," *IEEE Transactions on Power Systems*, vol. 33, no. 3, pp. 2995–3000, May 2018.

[145] X. Wei, S. Gao, T. Huang, E. Bompard, R. Pi, and T. Wang, "Complex Network Based Cascading Faults Graph for the Analysis of Transmission Network Vulnerability," *IEEE Transactions on Industrial Informatics*, pp. 1–1, 2018.

[146] X. Wei, S. Gao, T. Huang, T. Wang, and W. Fan, "Identification of Two Vulnerability Features: A New Framework for Electrical Networks Based on the Load Redistribution Mechanism of Complex Networks," *Complexity*, pp.1-14, 2019.

[147] A. Wang, Y. Luo, G. Tu, and P. Liu, "Vulnerability Assessment Scheme for Power System Transmission Networks Based on the Fault Chain Theory," *IEEE Transactions on Power Systems*, vol. 26, no. 1, pp. 442–450, Feb. 2011.

[148] “118 Bus Power Flow Test Case.” [Online]. Available: https://www2.ee.washington.edu/research/pstca/pf118/pg_tca118bus.htm.

Exibit A

Table A-1. GHuST values for autonomous-system graphs.

	$\rho_1{}'$	ρ_2	$\rho_3{}'$	$\rho_4{}'$	$\rho_5{}'$	ρ_6	$\rho_7{}'$	ρ_8	ρ_9	ρ_{10}	ρ_{11}	$\rho_{12}{}'$
Minimum	0.409	0.397	0.172	0.001	0.686	0.384	0.044	0.003	0.842	0.272	0.228	0.005
Quantile 1	0.459	0.577	0.191	0.001	0.699	0.456	0.144	0.004	0.917	0.318	0.496	0.007
Quantile 2	0.477	0.597	0.217	0.001	0.703	0.463	0.076	0.004	0.881	0.340	0.527	0.009
Quantile 3	0.510	0.619	0.257	0.002	0.762	0.431	0.087	0.010	0.884	0.369	0.430	0.016
Maximum	0.569	0.668	0.573	0.041	0.867	0.486	0.333	0.077	0.953	0.476	0.558	0.162

Table A-2. GHuST values for Enzymes.

	$\rho_1{}'$	ρ_2	$\rho_3{}'$	$\rho_4{}'$	$\rho_5{}'$	ρ_6	$\rho_7{}'$	ρ_8	ρ_9	ρ_{10}	ρ_{11}	$\rho_{12}{}'$
Minimum	0.113	0.000	0.111	0.101	0.342	0.000	0.000	0.000	0.000	0.000	0.000	0.400
Quantile 1	0.447	0.000	0.143	0.301	0.777	0.000	0.000	0.109	0.551	0.735	0.027	0.607
Quantile 2	0.487	0.000	0.167	0.361	0.831	0.000	0.000	0.175	0.647	0.969	0.085	0.660
Quantile 3	0.524	0.000	0.200	0.428	0.873	0.052	0.000	0.274	0.709	1.000	0.156	0.720
Maximum	0.604	0.308	0.833	0.693	0.946	0.800	0.848	0.417	0.829	1.000	0.556	0.900

Table A-3. GHuST values for Facebook graphs.

	$\rho_1{}'$	ρ_2	$\rho_3{}'$	$\rho_4{}'$	$\rho_5{}'$	ρ_6	$\rho_7{}'$	ρ_8	ρ_9	ρ_{10}	ρ_{11}	$\rho_{12}{}'$
Minimum	0.949	0.548	0.016	0.000	0.934	0.000	0.000	0.035	0.997	0.927	0.004	0.009
Quantile 1	0.970	0.623	0.051	0.002	0.982	0.006	0.000	0.050	0.998	0.955	0.010	0.039
Quantile 2	0.974	0.648	0.076	0.007	0.986	0.008	0.013	0.058	0.999	0.964	0.013	0.066
Quantile 3	0.977	0.681	0.122	0.017	0.987	0.010	0.027	0.071	0.999	0.972	0.016	0.110
Maximum	0.983	0.838	0.269	0.069	0.992	0.018	0.080	0.120	1.000	0.988	0.044	0.224

Table A-4. GHuST values for power-network graphs.

	$\rho_1{}'$	ρ_2	$\rho_3{}'$	$\rho_4{}'$	$\rho_5{}'$	ρ_6	$\rho_7{}'$	ρ_8	ρ_9	ρ_{10}	ρ_{11}	$\rho_{12}{}'$
Minimum	0.127	0.060	0.185	0.030	0.512	0.190	0.067	0.017	0.056	0.117	0.097	0.233
Quantile 1	0.219	0.198	0.271	0.078	0.719	0.345	0.172	0.027	0.143	0.192	0.154	0.320
Quantile 2	0.260	0.303	0.340	0.104	0.767	0.401	0.263	0.030	0.251	0.225	0.209	0.428
Quantile 3	0.285	0.382	0.446	0.196	0.799	0.426	0.339	0.037	0.380	0.297	0.252	0.532
Maximum	0.466	0.600	0.619	0.310	0.853	0.564	0.486	0.083	0.688	0.451	0.353	0.714

Table A-5 GHuST values for retweet graphs.

	ρ_1'	ρ_2	ρ_3'	ρ_4'	ρ_5'	ρ_6	ρ_7'	ρ_8	ρ_9	ρ_{10}	ρ_{11}	ρ_{12}'
Minimum	0.014	0.609	0.132	0.001	0.244	0.293	0.000	0.000	0.000	0.001	0.057	0.004
Quantile 1	0.065	0.730	0.209	0.002	0.408	0.439	0.050	0.000	0.463	0.015	0.161	0.019
Quantile 2	0.127	0.810	0.327	0.003	0.506	0.483	0.084	0.000	0.538	0.027	0.237	0.041
Quantile 3	0.182	0.902	0.575	0.004	0.621	0.618	0.118	0.001	0.678	0.048	0.476	0.069
Maximum	0.616	0.968	0.926	0.086	0.838	0.795	0.194	0.026	0.959	0.202	0.722	0.347

Table A-6. GHuST values for road graphs.

	ρ_1'	ρ_2	ρ_3'	ρ_4'	ρ_5'	ρ_6	ρ_7'	ρ_8	ρ_9	ρ_{10}	ρ_{11}	ρ_{12}'
Minimum	0.042	0.019	0.119	0.018	0.723	0.098	0.186	0.000	0.000	0.002	0.006	0.242
Quantile 1	0.065	0.035	0.182	0.048	0.766	0.121	0.224	0.001	0.026	0.003	0.018	0.311
Quantile 2	0.185	0.058	0.207	0.066	0.801	0.570	0.405	0.009	0.041	0.055	0.021	0.378
Quantile 3	0.284	0.196	0.246	0.102	0.842	0.816	0.770	0.021	0.102	0.162	0.044	0.464
Maximum	0.294	0.217	0.371	0.218	0.846	0.889	0.882	0.021	0.121	0.165	0.096	0.683

Table A-7. GHuST values for web graphs.

	ρ_1'	ρ_2	ρ_3'	ρ_4'	ρ_5'	ρ_6	ρ_7'	ρ_8	ρ_9	ρ_{10}	ρ_{11}	ρ_{12}'
Minimum	0.369	0.497	0.012	0.000	0.555	0.029	0.000	0.000	0.689	0.218	0.069	0.001
Quantile 1	0.532	0.608	0.022	0.000	0.672	0.075	0.032	0.017	0.916	0.380	0.172	0.003
Quantile 2	0.701	0.687	0.146	0.001	0.735	0.185	0.059	0.060	0.984	0.591	0.229	0.045
Quantile 3	0.872	0.802	0.247	0.007	0.800	0.243	0.116	0.275	0.995	0.739	0.554	0.066
Maximum	0.989	0.910	0.481	0.023	0.938	0.352	0.413	0.999	1.000	0.989	0.984	0.216

Exhibit B

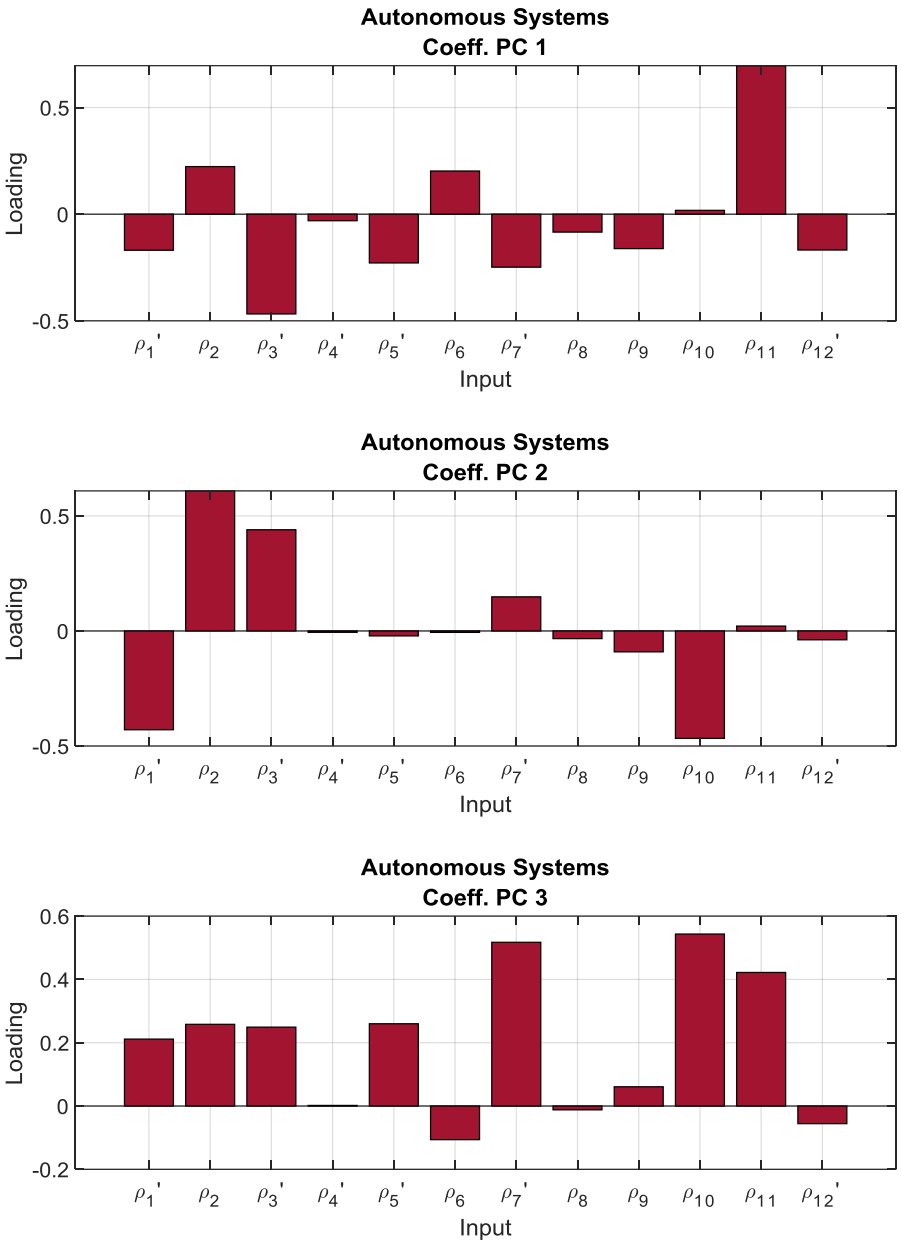


Figure B-1. Contribution of each dimension of the GHuST framework to the first 3 principal components obtained for the autonomous-system set of networks analyzed.

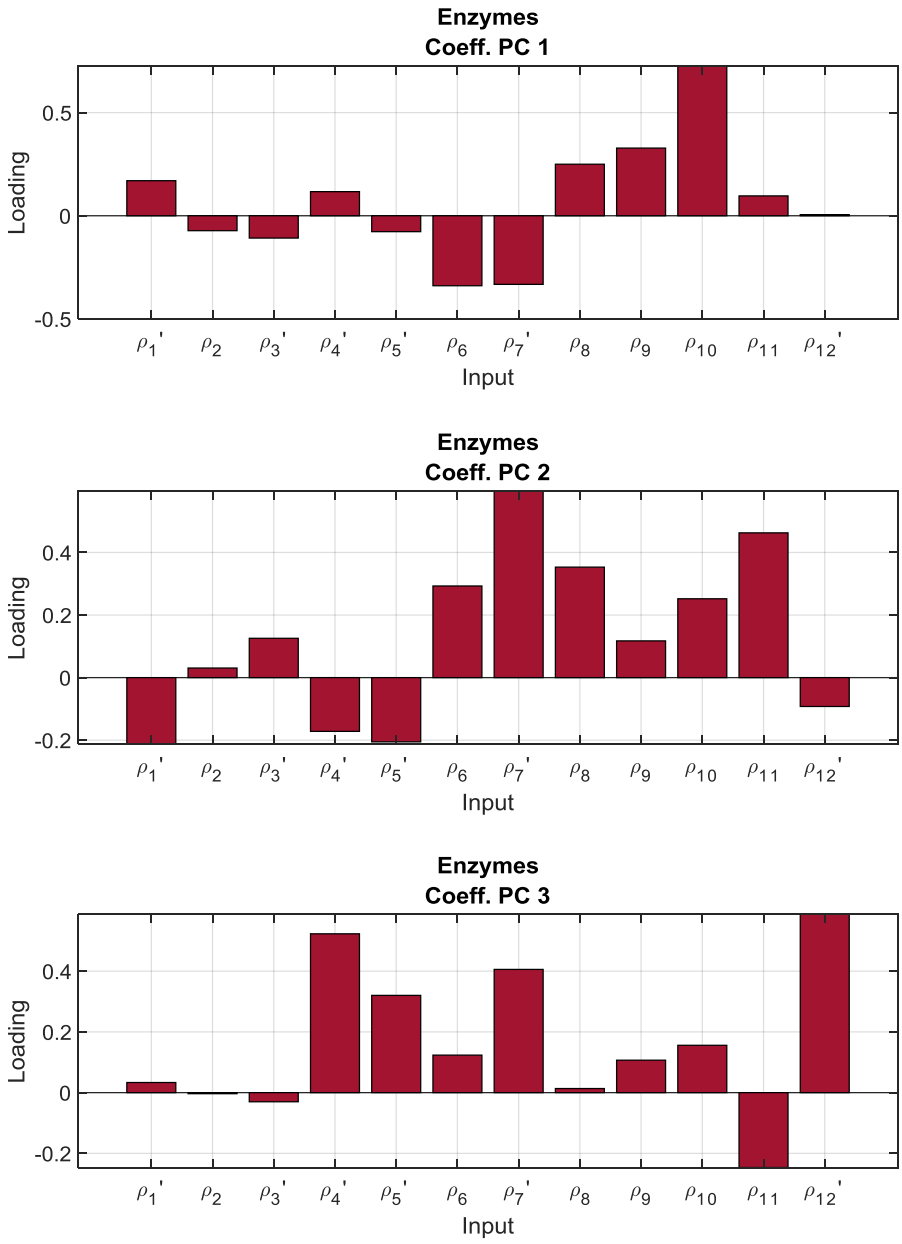


Figure B-2. Contribution of each dimension of the GHuST framework to the first 3 principal components obtained for the enzyme set of networks analyzed.

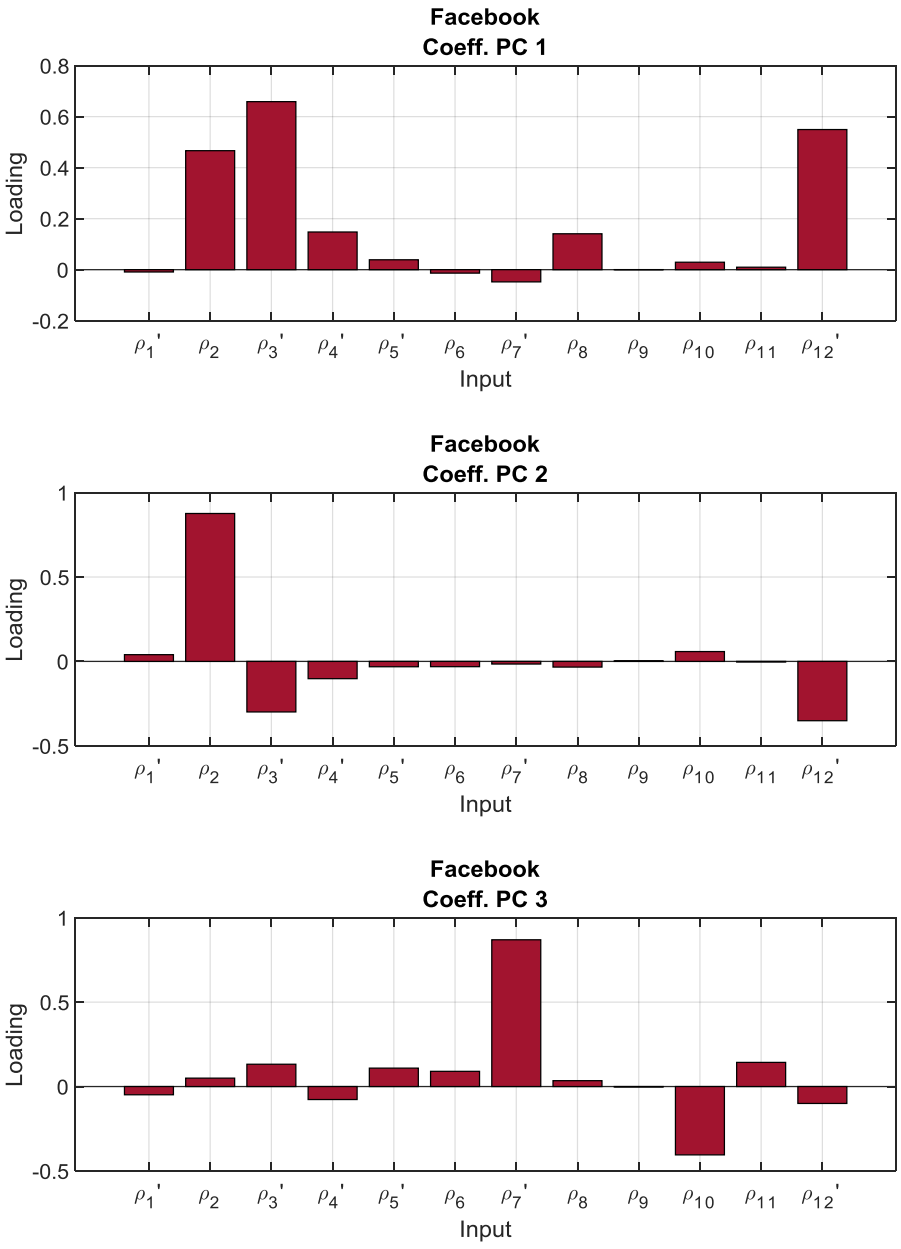


Figure B-3. Contribution of each dimension of the GHuST framework to the first 3 principal components obtained for the Facebook set of networks analyzed.

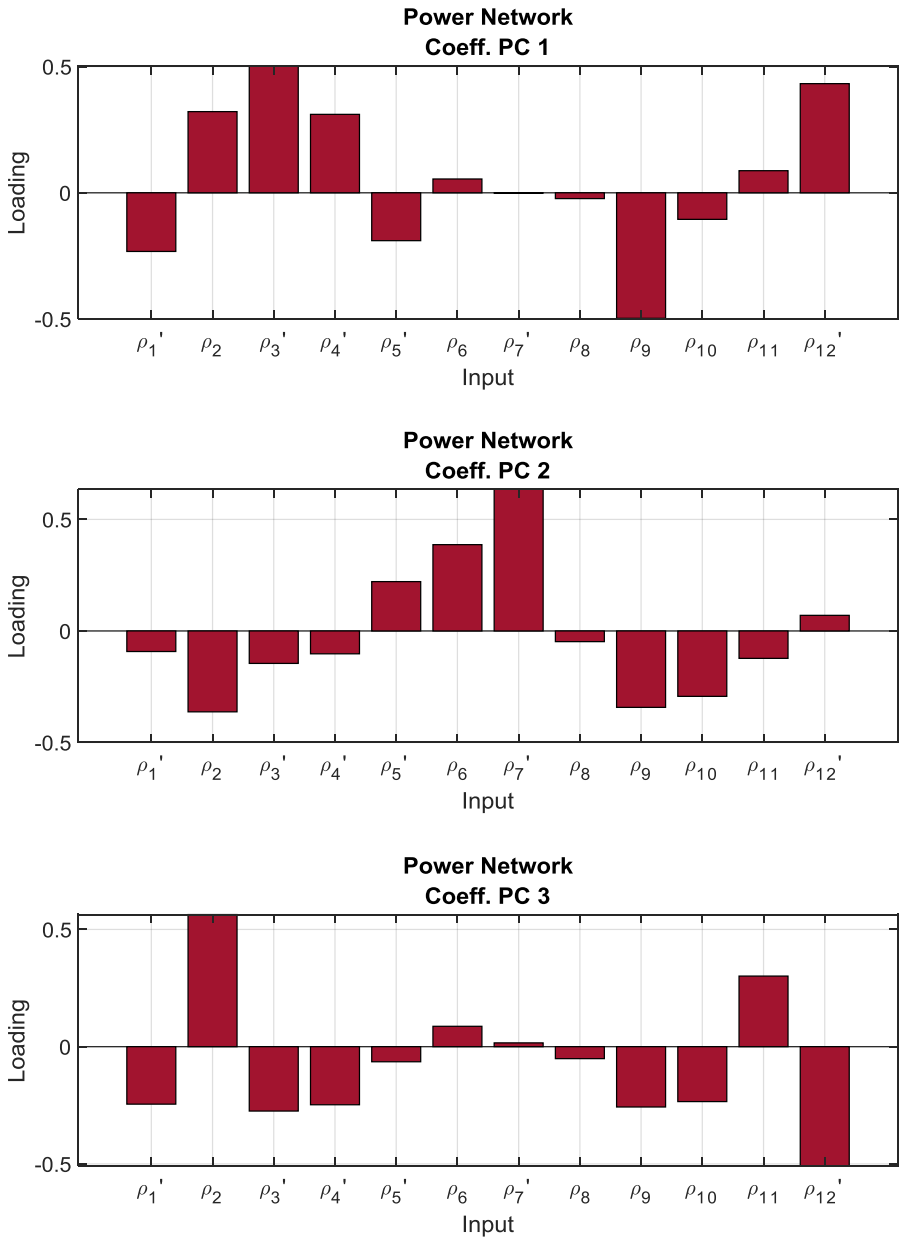


Figure B-4 Contribution of each dimension of the GHuST framework to the first 3 principal components obtained for the power-network set of networks analyzed.

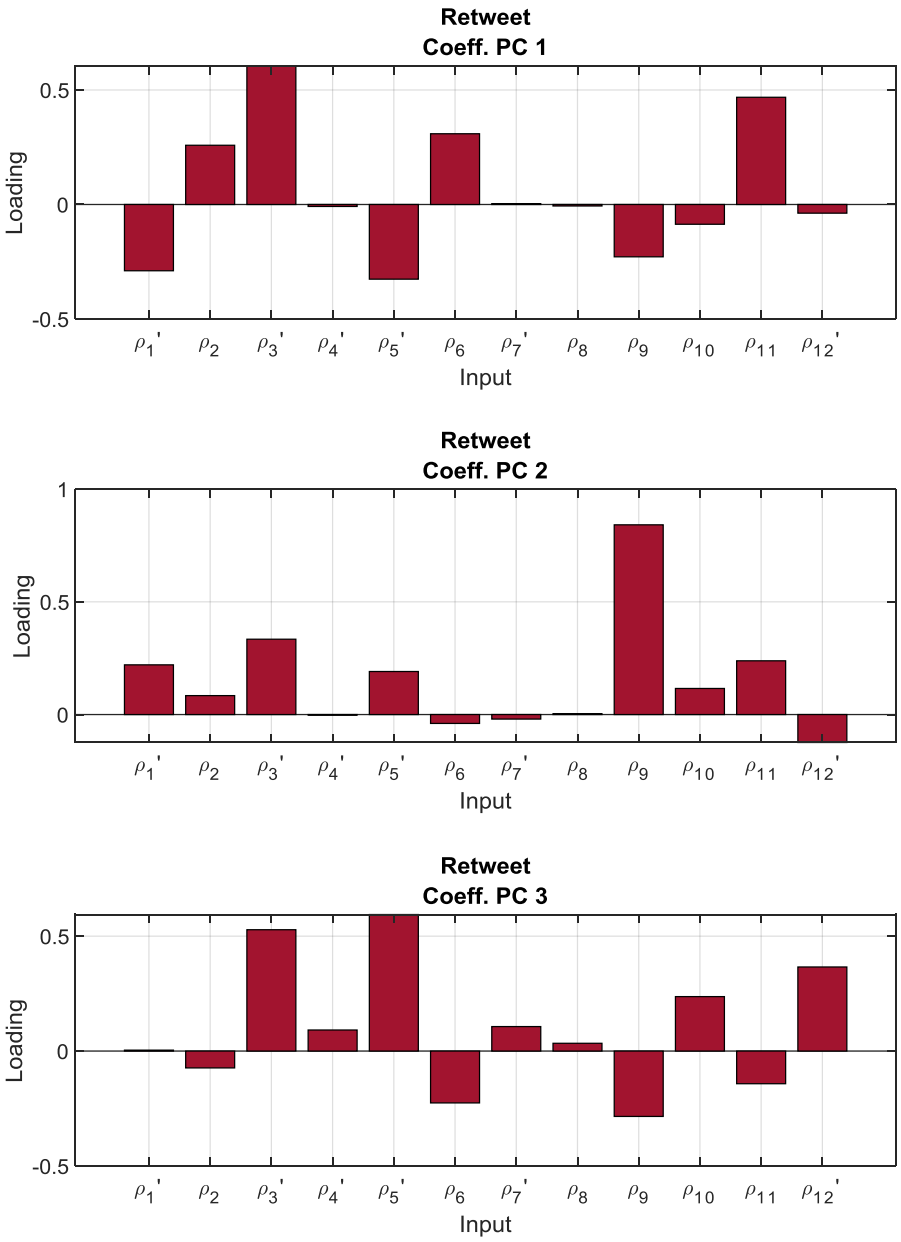


Figure B-5. Contribution of each dimension of the GHuST framework to the first 3 principal components obtained for the retweet set of networks analyzed.

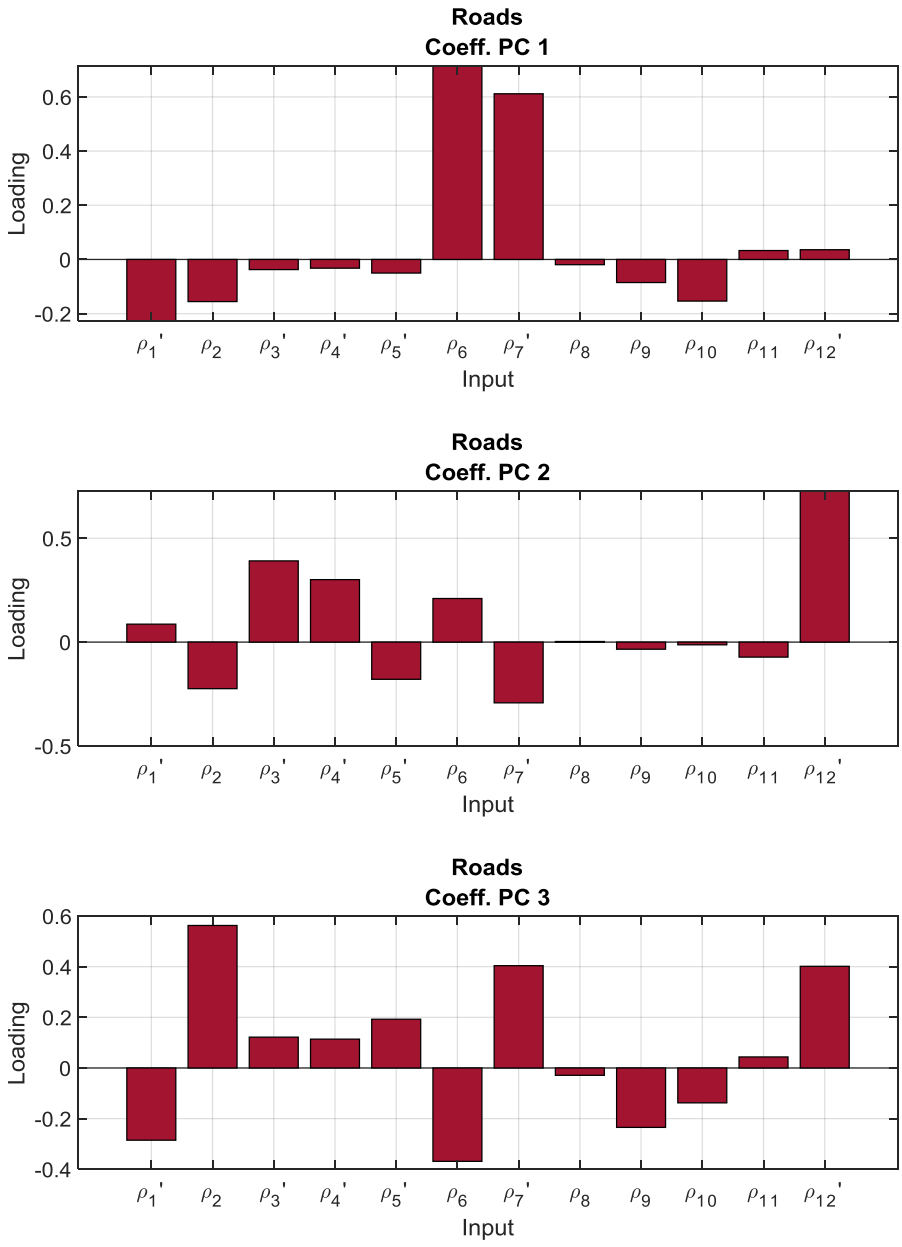


Figure B-6. Contribution of each dimension of the GHuST framework to the first 3 principal components obtained for the road set of networks analyzed.

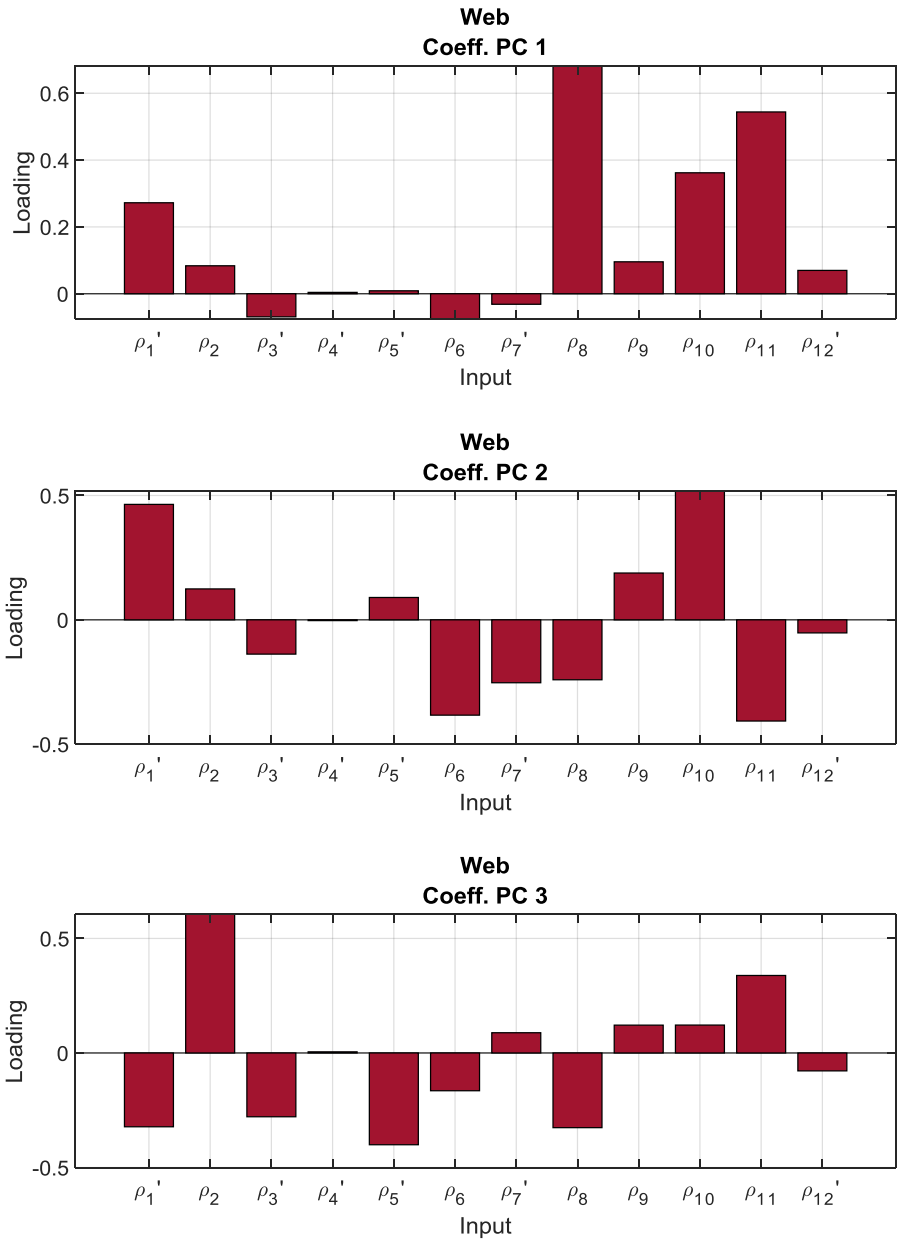


Figure B-7. Contribution of each dimension of the GHuST framework to the first 3 principal components obtained for the web set of networks analyzed.

Exhibit C

A. Global efficiency:

$$E = \frac{1}{N(N-1)} \sum_{i \neq j} \frac{1}{d_{ij}} \tag{C-1}$$

N is the total number of nodes in the network and d_{ij} is the shortest path between nodes i and j

B. Damage

$$D = \frac{E(G_0) - E(G_f)}{E(G_0)} \tag{C-2}$$

$E(G_0)$ is the value of a vulnerability index before a failure and $E(G_f)$ is the value of the vulnerability index after a failure.

C. Modified Global Efficiency

$$E = \frac{1}{N_D N_G} \sum_{i \in G} \sum_{j \in D} \frac{1}{d_{ij}} \tag{C-3}$$

N_D is the number of demand nodes in the network, N_G is the number of generators in the network, d_{ij} is the shortest path between nodes i and j

D. Loss of load

$$LOL = \frac{1}{D} \sum_{s} \Delta L_i \tag{C-4}$$

$$\Delta L_i = \begin{cases} D_i - C_i, & D_i > C_i \\ 0, & D_i < C_i \end{cases} \tag{C-5}$$

D is the total demand of the network before component failure, s is the number of islands in the system after a component failure, D_i total power demanded in an island, and C_i total generation capacity on an island.

E. Connectivity loss

$$CL = 1 - \langle \frac{N_g^i}{N_g} \rangle_i \quad \text{(C-6)}$$

N_g^i is the number of generators connected to node i and N_g the number of generators in the network.

F. Electrical degree centrality

$$D_i = \frac{\sum F_{i,j}}{N-1} \quad \text{(C-7)}$$

N is the number of nodes in the network and $F_{i,j}$ is the power flow through the lines that are connected to node i.

G. Electrical betweenness centrality

$$B_i = \sum \frac{P_{st,i}}{P_{st}} \quad \text{(C-8)}$$

P_{st} is the maximum amount of power that can flow through line st, and $P_{st,i}$ is the power that is injected in node i when the power through line st is equal to the transmission line capacity.

H. Structural vulnerability index

$$SVI = \frac{1}{N_D N_G} \sum_{i \in G} \sum_{j \in D} \frac{P_{Gi}}{P_{Dj}\, e^{z_{ij}}} \quad \text{(C-9)}$$

N_D is the number of demand nodes, N_G the number of generators nodes, P_{Gi} maximum generation capacity of generator i, and P_{Dj} is the maximum power demanded by node j

I. Directed global efficiency

$$Directed\ Global\ Efficiency = \frac{1}{N_D N_G} \sum_{i \in G} \sum_{j \in D} \frac{1}{z_{ij}} \quad \text{(C-10)}$$

N_D is the number of demand nodes, N_G the number of generators nodes, and z_{ij} is the electrical distance between nodes i and j.

J. Net ability

$$Net\ Ability = \frac{1}{N_D N_G} \sum_{i \in G} \sum_{j \in D} \frac{c_{ij}}{z_{ij}} \quad \text{(C-11)}$$

N_D is the number of demand nodes, N_G is the number of generation nodes, c_{ij} is the maximum power that can be injected in node i to be withdrawn in node j.

K. Effective graph resistance

$$Graph\ resistance = \sum_{i} \sum_{j \neq i} R_{i,j} \quad \text{(C-12)}$$

$R_{i,j}$ is the effective resistance between nodes i and.

L. Electrical centrality

$$c_i = \frac{1}{\bar{e}_a} \quad \text{(C-13)}$$

$$\bar{e}_a = \sum_{\substack{b=1 \\ b \neq a}} \frac{e_{ab}}{n-1} \quad \text{(C-14)}$$

e_{ab} is the electrical distance between nodes a and b and n is the number of nodes in the network.

M. Centrality index

$$CI_{uv} = \sum_{i} \sum_{j} f_{ij}^{uv} \quad \text{(C-15)}$$

f_{ij}^{uv} is the maximum power that can be injected in node i to be withdrawn in node j.

N. Extended betweenness

$$Extended\ betweenness\ (l) = \max\left(T^P(l), T^P(l)\right) \quad \text{(C-16)}$$

$$T^P(l) = \sum_{i \in G} \sum_{j \in D} c_{ij}\, f_l^{ij} \text{ if } f_l^{ij} > 0 \quad \text{(C-17)}$$

$$T^P(l) = \sum_{i \in G} \sum_{j \in D} c_{ij}\, f_l^{ij} \text{ if } f_l^{ij} < 0 \quad \text{(C-18)}$$

c_{ij} is the maximum power that can be injected in node i to be withdrawn in node j, and f_l^{ij} is the change of power through line l when there is an injection of power in node i and it is withdrawn in node j.