UNIONE EUROPEA
Fondo sociale europeo

REGIONE AUTONOMA DELLA SARDEGNA

FSE 2007-2013
POR
SARDEGNA
obiettivo competitività regionale e occupazione

UNIVERSITY OF CAGLIARI

DEPARTMENT OF MATHEMATICS AND COMPUTER SCIENCE
PH.D. COURSE IN COMPUTER SCIENCE
CYCLE XXVIII

PH.D. THESIS

Criteria for Modification of Complex Infrastructure Networks

S.S.D. INF/01

Candidate:
Pier Luigi PAU

Ph.D. Coordinator:
Prof. Giovanni Michele PINNA

Supervisor:
Prof. Gianni FENU

Final examination academic year 2015/2016

UNIVERSITY OF CAGLIARI

Abstract

Faculty of Science
Department of Mathematics and Computer Science

Doctor of Philosophy

Criteria for Modification of Complex Infrastructure Networks

by Pier Luigi PAU

Complex network theory enables the analysis and comparison of graphs with a very large number of nodes, or with non-trivial topological properties. Graph models exist for many kinds of networks, ranging from computer networks to representation of protein-protein interactions, and analysis techniques are often shared between fields of application.

Infrastructure networks are an active field of application of complex network analysis, which is frequently aimed at finding ways to improve on the structure of a network, while respecting budget constraints. In this activity, complex network analysis is often cross-referenced with simulations or operational research.

Power grids stand out among the most prominent examples of infrastructure network analyzed with techniques derived from complex network theory, due to their importance as a service, their properties of quick response to events, and the desired transition to a smart grid paradigm. With the growing interest for the protection of endangered species and habitats, the modeling and analysis of green infrastructure has also received increasing attention from scholars.

These classes of infrastructure provide case studies for the exemplification of a common process for the analysis of various kinds of infrastructure networks, which involves the identification of vulnerabilities, the exploration of a search space for possible modifications, and the definition of a comparable measure of health of the network.

Acknowledgements

Pier Luigi Pau gratefully acknowledges Sardinia Regional Government for the financial support of his PhD scholarship (P.O.R. Sardegna F.S.E. Operational Programme of the Autonomous Region of Sardinia, European Social Fund 2007-2013 - Axis IV Human Resources, Objective l.3, Line of Activity l.3.1.).

La presente tesi è stata prodotta durante la frequenza del corso di dottorato in Informatica dell'Università degli Studi di Cagliari, a.a. 2015/2016 – XXVIII ciclo, con il supporto di una borsa di studio finanziata con le risorse del P.O.R. SARDEGNA F.S.E. 2007-2013 – Obiettivo competitività regionale e occupazione, Asse IV Capitale umano, Linea di Attività l.3.1 "Finanziamento di corsi di dottorato finalizzati alla formazione di capitale umano altamente specializzato, in particolare per i settori dell'ICT, delle nanotecnologie e delle biotecnologie, dell'energia e dello sviluppo sostenibile, dell'agroalimentare e dei materiali tradizionali".

UNIONE EUROPEA
Fondo sociale europeo

REPVBBLICA ITALIANA

REGIONE AUTONOMA DELLA SARDEGNA

FSE 2007-2013
POR
SARDEGNA
obiettivo competitività regionale e occupazione

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Graphs are a powerful mathematical tool, used to represent networks of many kinds, whether physical (computer networks, power grids, etc.) or conceptual (protein-protein interactions, networks of knowledge, etc.). The most discussed theoretical problems from graph theory generally correspond to one or more practical problems in applications. One such example is the problem known as ‘graph coloring’ (more specifically, vertex coloring), which consists in mapping the nodes in a graph to a minimal set of labels (colors), while respecting the constraint that no two adjacent nodes can be mapped to the same label. This is clearly analogous to the problem of coloring countries in a world map; in fact, it is this latter problem that fueled initial research into the theoretical aspects of graph coloring, beginning with the restricted case of planar graphs.

Other practical problems exist, which can be solved with an instance of the vertex coloring problem, although the relation is not as obvious. One such problem is that of job scheduling, in which a number of jobs is to be assigned to time slots, while keeping in mind that some resources that may be needed for multiple jobs can only be assigned to one job at any given time. The resources may be workers themselves, in which case the problem can also be applied to the definition of class schedules (a teacher may give lectures to multiple classes, but may only give one in any time slot). Another related problem is found in the field of computer science: given a function written in C (or another language compiled to machine code), compilers may optimize performance by assigning CPU registers to store local variables, but registers are available in an extremely limited number compared to storage in RAM, and a register can only be allocated to one variable at any given time. This leads to the problem of determining whether the scope of variables in a certain function is such that the number of variables that should be stored in registers at the same time is within the available number of CPU registers [10].

The concept of finding a theoretical problem, to which practical problems are reduced, is recurrent in many branches of the vast field that is graph theory, and is one of the reasons for its popularity as a research field.

One of these branches is complex network theory. Since the beginning of the 21st century, this area of research has received an increasing amount of attention from scholars with different backgrounds, including physics, mathematics, computer science, biology, and engineering. Complex network theory deals with the extraction and analysis of statistical properties of graphs, and aims at finding a comparable set of measures for graphs with non-trivial topological properties, as well as extremely large graphs, with a number of nodes in the millions, or even at higher orders of magnitude. Complex network theory is considered by some to be a separate field of research, as opposed to a branch of graph theory, due to its specific collocation encompassing graph theory and statistics, together with the large number of its applications and their heavy influence on research questions.

From theoretical results, techniques for complex network analysis are derived, which can be applied to conceptual and physical networks alike, and it is not uncommon for ideas from a field of application to be adapted for use in a different field. The same is true for software tools, with software initially designed for analysis of specific classes of networks (e.g. social networks, protein-protein interaction networks) being adapted for general-purpose use in subsequent versions.

Complex network theory is often used to provide the foundations and an established theoretical background for the construction of models for infrastructure networks. This enables the application of methods to assess network properties and predict the effect of changes, whether intentional or due to external factors, to assist in devising development plans, as well as guidelines for response to random failures or attacks from malicious users; if the process involves determining whether the network – or a significant portion thereof – is susceptible to collapsing when certain elements or links (i.e. nodes or edges in the graph model) are removed or damaged, this is referred to as a vulnerability assessment. The success of these methods of analysis is heavily dependent on the construction of a faithful model of the infrastructure being analyzed.

The present work takes two kinds of network infrastructure that have become popular applications of complex network analysis in the past two decades: ecological landscape networks and power grids.

The former are systems of "green infrastructure", made up of nature protection areas and habitat corridors, and are being established in various parts of the world, as a way to overcome the limitations of traditional natural reserves: specifically, their inadequacy at preserving biodiversity, and their limited effectiveness in conditions of insufficient size or excessive distance from other suitable habitat patches for the endangered species, which they are supposed to host in a protected environment. Ecological landscape networks have been an active field of discussion by scholars with a background in ecology since at least the 1970s, although the first mentions of habitat corridors, without a definite view of the "big picture" of green infrastructure, can be dated to as far back as the 1920s.

The establishment and maintenance of a continent-wide ecological network is a task that requires considerable investments and the coordination of efforts at local, national and international level. In the European Union, this is done within the project denominated "Natura 2000", aimed at the preservation of biodiversity with the creation of a continent-wide ecological network. A common framework is defined in European directives to determine large-scale conservation goals, with a list of protected habitats and species and a set of common rules for the activities across the continent, while local administrations are involved in the management of single nature protection areas. In this context, complex network analysis tools are a valuable complement for geographic information systems (GIS) used in land management and planning, enabling a better understanding of network behavior in habitat patches at a local level, as well as the evaluation of large-scale consequences of local changes.

Power grids represent another class of infrastructure, which has received a large amount of attention from scholars as a field of application for complex network analysis. This can be attributed to the properties for which power grids stand out among commodity distribution networks, such as their quick response to changes and the necessity to adjust the output of power plants to meet demand in real time; moreover, the special interest for power grids is fueled by current plans for a smooth transition toward a "smart grid" paradigm. This aims at increasing the reliability, flexibility and efficiency of the grid, by complementing the power infrastructure with an information infrastructure, in order to enable the grid to self-regulate, and provide the possibility for end users to contribute to power generation and grid operation.

In spite of a difference in goals, these two classes of application share the motives for the creation of complex network models: the understanding of global and local properties of the network, with the identification of measures of health of the infrastructure, and the evaluation of the impact of proposed modifications.

1.1 Dissertation Structure

The following is an outline of the organization of the present work:

- Chapter 2 provides a brief review of theoretical aspects that are common to both case studies.

- Chapter 3 introduces the concept of ecological landscape networks in general, together with the methods in use for building graph models of green infrastructure. The Natura 2000 network is taken as a frame of reference; following an analysis of the form in which public data is made available within the project, details are given on which ways it is possible to create complex network models from available data.

- In Chapter 4, after a brief introduction concerning the work of land management and planning, a case study is presented, which is based on Natura 2000 sites located in Sardinia. A method to build a model of an ecological network based on site similarity is discussed, and models created from multiple data sources are compared to determine which approach is most useful for land management.
- Chapter 5 presents a second case study, consisting of a model of the Sardinian power grid. An optimization problem is defined on a graph model of the grid, in order to determine a set of power flows toward which average operation should converge. A measure of collateral damage from the failure of a node or set of nodes is proposed, and multiple complex network models are analyzed, seeking a correlation between network indices and collateral damage.
- General conclusions are drawn in Chapter 6. Final comments concerning the aptness of each of the proposed network models are given, and possibilities of future work are discussed.

1.2 Software tools

The open source Cytoscape suite [54] was used for complex network visualization and analysis. For some applications, custom modifications were applied to the software. The IBM CPLEX Optimization Studio was used to solve linear optimization problems. QGIS [49] and SQLite with the Spatialite extension were used for computation of spatial data.

Chapter 2

Complex Networks

2.1 Graph Models of Complex Systems

Graphs are a powerful and versatile mathematical tool, and have been used to model a wide range of complex systems of many kinds. By their definition, they are suitable as a mathematical model for networks, including connected sets of physical entities, as well as virtual and conceptual networks built from datasets of different origins.

Properties of a graph model will reflect those of the real-world network they represent, and the effect of modifications applied to a graph model may be used to predict the effect of changes on the corresponding real network. A trivial example of this is given by computer networks: at the smallest possible scale, most wired LANs are made up of a number of hosts connected to a switch; in a graph model corresponding to this network, the switch and each host are represented by nodes, and an edge represents each physical connection. As the only connections are between the switch and a host, the resulting graph will be an instance of a special case of graph, denominated 'star graph', which consists of a number of nodes that are only adjacent to a special central node, which in turn is adjacent to all the others. In the real-world network, the failure of a host does not affect the connectivity of the other hosts, whereas the failure of the switch causes a loss of connectivity for every station connected to it; likewise, in the graph model, the removal of peripheral nodes simply results in another instance of the star graph with fewer nodes, while the removal of the central node results in a set of isolated nodes, with a complete lack of edges, i.e. no connections.

Clearly, most networks in the real world are not so simple as to be represented by a star graph. However, the ability to predict the effect of changes by examining the modifications of a graph model is desirable for many kinds of physical, virtual and conceptual networks. In practice, this can only be done intuitively for small networks; it may be possible to do so for some medium-sized networks, so long as their topology is regular, or simple enough to be represented visually. An analytical approach is required for larger networks, as well as those with a non-trivial topology.

This approach is made possible by applications of complex network theory, a branch of graph theory that deals with the study of statistical properties of graphs, thus providing tools to compare and categorize networks at every scale and every degree of complexity. Fields of application of complex network theory include, but are not limited to:

- computer networks at every scale (LAN, WLAN, MAN, WAN);
- social networks (centralized, e.g. Facebook, Twitter, LinkedIn; or decentralized, e.g. e-mail, in-person interaction);
- transportation networks (roads, railways, airplane connections);
- commodity distribution networks, including power grids;
- protein-protein and gene-protein interactions;
- metabolic networks;
- financial and trade networks;
- networks of knowledge, e.g. those used in the semantic web.

Some fields of application exist, which encompass multiple kinds of networks at the same time. For instance, law enforcement may benefit from a combined study of social networks and distribution networks to identify drug dealers acting as 'hubs', thus prioritizing actions that may disrupt the distribution of illegal substances in the quickest and most effective way.

The aim of complex network theory is to provide a robust framework for the comparison of networks and their categorization, which may work for very large network and in the presence of non-trivial topological structures. Complex network analysis techniques often involve the comparison of numerical indices and statistical distributions.

2.2 Graphs and Their Properties

This section revises the most important definitions and some basic concepts of graph theory. There exists a vast literature on graph theory; one of the most comprehensive works is [16].

2.2.1 Basic Definitions

A graph is a pair $G = (V, E)$, where V is a set of vertices (or nodes) and $E \subseteq V^2$ is a set of edges (or lines), which represent the connection between a pair of nodes. An edge $(i, j) \in E$ is said to be incident to nodes i and j, and these nodes are said to be adjacent to one another. The order of a graph is the number of its nodes $|V|$, while the number of edges $|E|$ is referred to by size of a graph.

Node pairs are usually assumed to be unordered. In cases where they are ordered, i.e. (i, j) and (j, i) are to be treated as two separate edges, then the graph and its edges are said to be directed. A directed edge (i, j) is also referred to as an arc, with head j and tail i; consistently with this definition, a directed graph is often denoted as $G = (V, A)$. A graph in which undirected edges and arcs are allowed to exist at the same time is said to be a mixed graph, denoted as $G = (V, E, A)$.

It is normally assumed that only one edge may exist joining any two nodes in an undirected graph, or two arcs of opposite direction in a directed graph. To avoid confusion, a graph where multiple edges are allowed to share a node pair is called a multigraph, and one where this is disallowed is called a simple graph. An oriented graph is a directed graph with no pair of opposite arcs between any two nodes.

A loop is an edge (i, i) joining a node with itself. In most applications, loops are disregarded or forbidden; thus, unless stated otherwise in the context of study, it is usually assumed that no loops are present.

A graph is said to be weighted if each edge has a numerical attribute associated to it; this attribute may represent a distance (or cost) for traversing the link, or a strength of the link (for example, the bandwidth of a link between two routers).

2.2.2 Graph Concepts

Subgraphs

A subgraph of a graph G is a graph G' with all of the following properties:

$$\begin{aligned} &V(G') \subseteq V(G), \\ &E(G') \subseteq E(G), \\ &(i, j) \in E(G') \Rightarrow i, j \in V(G'). \end{aligned} \tag{2.1}$$

It is also said that G is a supergraph of G'. A subgraph such that $V(G') = V(G)$ is said to be a spanning subgraph of G.

If a subgraph G' includes every edge in G that satisfies the conditions to be a valid edge in G', then it is said to be the subgraph of G induced (or implied) by its node set S, and is denoted by $G[S]$. If $G' = G[S]$:

$$\begin{aligned} &S = V(G') \subseteq V(G) \quad \text{and} \\ &E(G') = \{(i, j) \,|\, i, j \in S,\ (i, j) \in E(G)\}. \end{aligned} \tag{2.2}$$

Walks, Paths and Cycles

A walk on an undirected graph consists in an ordered list of nodes, such that an edge exists between any two consecutive nodes. A walk is closed if the first and last node are the same node, otherwise it is said to be open. A walk is simple if there is no repetition of nodes, with the sole exception of the repetition of the starting node as the final node in a simple closed walk. A simple open walk is also called a path; a simple closed walk is also called a cycle.

On directed graphs, a walk is made up of an ordered list of nodes, such that an arc exists between consecutive nodes; if an arc oriented according to the order in which the pair is found in the walk exists for all pairs of consecutive nodes in the walk, then the walk is directed; otherwise, it is undirected. Depending on the application, it may be required that only directed walks are considered. Analogous definitions are given for directed and undirected paths and cycles.

Connected Graphs

A graph is said to be connected if a path exists between any pair of nodes chosen arbitrarily. A connected graph without cycles is said to be a tree.

If a graph G is not connected, its components (or connected components) are the subgraphs of G with the following properties:

- the subgraph is induced from the supergraph by its node set;
- the subgraph is a connected graph;
- no path exists in the supergraph between any node in the subgraph and any node that is not part of the subgraph.

If an edge exists between every pair of nodes, the graph is said to be complete. The number of edges in a complete undirected graph is

$$|E| = \frac{n(n-1)}{2}, \tag{2.3}$$

where $n = |V|$.

A clique is a complete subgraph of a given graph. A maximal clique of G is a clique of G that is not a subgraph of another clique of G.

2.2.3 Basic Properties

Node Degree

The degree of a node n, denoted as $deg(n)$ or k_n, is the number of edges incident to it. If loops are allowed, they are to be counted twice, i.e. both ends of the loop contribute to the node degree. In a simple undirected graph, the degree of a node is equal to the number of its neighbors.

If every node in a graph has the same degree, the graph is said to be regular.

In directed and mixed graphs, node degree is usually evaluated separately for arcs that are incident to a node as their head, or as their tail. The in-degree of a node n is the number of arcs for which n is the head, and the out-degree of n is the number of edges for which n is the tail.

Path Length

Let P be the ordered set of edges that make up the path from i to j in an unweighted graph. The length of the path $l(P)$ is given by:

$$l(P) = |P|. \tag{2.4}$$

On a weighted graph, in which the weight represents a distance or a cost associated with the action of traversing an edge, the length of a path is defined as the sum of the weights of the edges in the path:

$$l(P) = \sum_{(i,j)\in P} w_{ij}. \tag{2.5}$$

Conversely, if the weight of a link represents its strength (for example, a link bandwidth in a computer network), the length of a path can be defined as the sum of weight reciprocals:

$$l(P) = \sum_{(i,j)\in P} \frac{1}{w_{ij}}. \tag{2.6}$$

In this case, it is required that all $w_{ij} \neq 0$ for all edges in a path. Depending on the physical meaning of edge weights, however, such edges might not exist, or be forbidden from being traversed.

Shortest Paths

The shortest path between two nodes is the path of minimum length connecting the two nodes. Its length is the distance between the two nodes, and may be denoted as $d_G(i, j)$, $d(i, j)$, or d_{ij}. To extend this definition to include cases where i and j are the same node, and the case where there is no path connecting the two nodes, $\mathcal{P}_{ij}$ is defined as the set of all paths from

i and j; then:

$$d_G(i,j) = \begin{cases} 0 & \text{if } i = j, \\ \infty & \text{if } \mathcal{P}_{ij} = \emptyset, \\ \min_{P \in \mathcal{P}_{ij}} l(P) & \text{otherwise.} \end{cases} \tag{2.7}$$

The full set of shortest paths from a given starting node is generally found by performing a breadth-first search on unweighted graphs, or with Dijkstra's algorithm on weighted graphs with non-negative weights representing distances. For directed graphs, it is intended that only directed paths are to be considered.

2.3 Complex Network Properties

It is possible to identify and extract complex network measures of different kinds, according to whether they are global or local properties, and whether they are simple or compound values (such as vectors, matrices, and statistical distributions).

Global properties are calculated as a single value for an entire network. Such are, for example, the diameter of a network or its characteristic path length. Conversely, local properties are evaluated for each node, by analyzing only its immediate neighborhood. An example thereof is the local clustering coefficient. Some properties are found in the middle of this spectrum, like most centrality indices, which are calculated for nodes or edges, but their value depends on the topology of the network in its entirety. A survey of complex network measurements is found in [12].

2.3.1 Global Measures

Diameter and Average Path Length

The diameter of a network $D(G)$ is the maximum length of its shortest paths, considering all pairs of nodes:

$$D(G) = \max_{i,j \in V} d(i,j). \tag{2.8}$$

This measure is defined for connected graphs or connected components of a graph. If a graph is not connected, the maximum diameter among those of its connected components is sometimes considered to be the diameter of the whole graph.

The average path length of a network $L(G)$ is the average of all shortest path lengths in the network, considering all pairs of nodes:

$$L(G) = \frac{1}{n(n-1)} \sum_{i \neq j \in V} d(i,j). \tag{2.9}$$

This measure is defined for connected graphs. In some applications, it is calculated for disconnected graph by setting $d(i,j) = 0$ if no path exists between i and j. The average path length is sometimes referred to as the characteristic path length.

Eccentricity and Network Radius

The eccentricity $\epsilon(i)$ of a node i is the maximum length of a shortest path from i to another node in the same connected component, or 0 for isolated nodes.

$$\epsilon(i) = \begin{cases} 0 & \text{if } k_i = 0, \\ \max_{j \in C_i} d(i,j) & \text{otherwise,} \end{cases} \tag{2.10}$$

where C_i is the connected component in which node i is found.

Network radius is the minimum non-zero value for eccentricity, considering every node in the network:

$$r(G) = \min_{i \in V, \epsilon(i) > 0} \epsilon(i). \tag{2.11}$$

Network Density

The density of a network is the ratio of the number of its edges to the number of its possible edges. For simple directed unweighted networks, this is given by

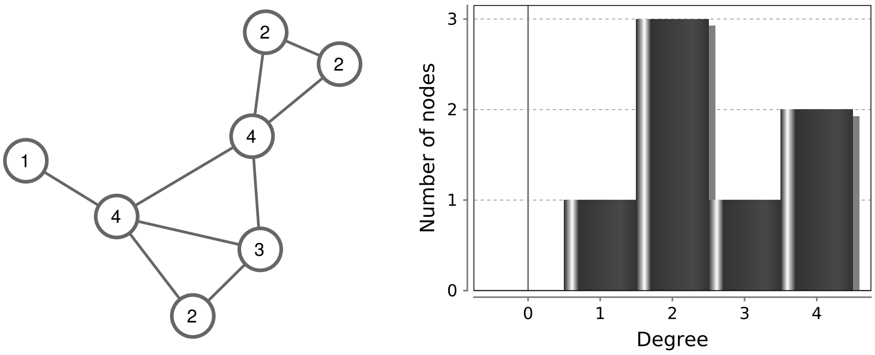
$$\textit{Density(G)} = \frac{|E(G)|}{n(n-1)}, \tag{2.12}$$

where $n = |V(G)|$. For undirected networks, the numerator has to be doubled, to compensate for the fact that $n(n-1)$ counts each node pair twice:

$$\textit{Density(G)} = \frac{2\,|E(G)|}{n(n-1)}. \tag{2.13}$$

The following alternative definition may also be used [17]:

$$\textit{Density(G)} = \frac{\sum_i \sum_{j \neq i} a_{ij}}{n(n-1)} = \frac{S_1(\mathbf{k})}{n(n-1)} = \frac{\overline{k_i}}{n-1}, \tag{2.14}$$



(a) (b)

FIGURE 2.1: Degree distribution. (a) Sample graph, labeled with node degrees. (b) A visualization of the node degree distribution, as a histogram chart with the number of nodes on the y-axis.

where a_{ij} are the elements of the adjacency matrix, $\mathbf{k}$ is a vector where each element is given by the degree of the corresponding node, and $\overline{k_i}$ is the average node degree. The S_p function of a vector $\mathbf{v}$ is defined as:

$$S_p(\mathbf{v}) = \sum_i \mathbf{v}_i^p \, \mathbf{1}^T. \tag{2.15}$$

Network density takes values from 0 to 1, where 0 is the density of a set of isolated nodes, and 1 is the density of a complete graph. Thus, density expresses how close a network is to one of the ends of this spectrum.

Degree Distribution

One of the most straightforward methods to compare complex networks is to study their node degree distribution. This consists in counting the number of nodes for each value of node degree found in the network (for a simple example, see Figure 2.1). To make comparison possible, this is reduced to a probability distribution by normalizing to the total number of nodes:

$$P(k) = \frac{n_k}{n}, \tag{2.16}$$

where $n = |V|$ and n_k is the number of nodes with degree k. In this form, the degree distribution can be compared with known probability distribution laws, or with other network instances, regardless of their order.

2.3.2 Local Measures

Clustering Coefficient

The clustering coefficient of a node expresses a measure of how much its neighborhood is made up of fully connected clusters.

The definition of this index is based on those of triplets and triangles. In unweighted undirected graphs, a triplet is a set of three nodes, made up of a central node and two of its adjacent nodes. If the non-central nodes are also adjacent to one another, the set is a closed triplet; otherwise, it is an open triplet. A triangle is a clique of three nodes; each triangle can be thought of as three closed triplets, where each triplet has a different central node.

The **local clustering coefficient** is defined for each node i as

$$C_i = \frac{N_\triangle(i)}{N_3(i)}, \tag{2.17}$$

where $N_\triangle(i)$ is the number of triangles that contain i, and $N_3(i)$ is the number of triplets where i is the central node.

An equivalent definition of $N_\triangle(i)$ is the number of edges between neighbors of i. Moreover, a strict relation exists between the number of triplets of a node and its degree:

$$N_3(i) = \frac{k_i(k_i - 1)}{2}. \tag{2.18}$$

That means that $N_3(i) = 0$ if $k_i < 2$. In that case, $N_\triangle(i)$ also takes the value of 0, and (2.17) takes an indeterminate form. By convention, the clustering coefficient of any such node is considered to be 0. Thus, an alternate definition of C_i is:

$$C_i = \begin{cases} 0 & \text{if } k_i < 2, \\ \dfrac{2l_i}{k_i(k_i - 1)} & \text{if } k_i \geq 2, \end{cases} \tag{2.19}$$

where $l_i = N_\triangle(i)$. It is simple to verify that $0 \leq C_i \leq 1$.

The **network average clustering coefficient** is the average value of the local clustering coefficient calculated for every node in a graph:

$$\overline{C} = \frac{1}{|V|} \sum_{i \in V} C_i. \tag{2.20}$$

This value is sometimes used as a global clustering coefficient. Another definition of such a measure exists, which is analogous to (2.17):

$$C = \frac{3N_{\triangle}}{N_3}, \tag{2.21}$$

where $N_{\triangle}$ is the total number of triangles in the graph, thus $3N_{\triangle}$ is the total number of closed triplets; and N_3 is the total number of triplets, inclusive of open and closed triplets. It is possible to extend the convention that $C = 0$ if $N_3 = 0$ to this version of the clustering coefficient; with this in effect, this measure also takes values from 0 to 1, where 1 is the value of this coefficient in a complete graph. In general, the value of C differs from that of $\overline{C}$; essentially, (2.20) gives the same weight to each node, while (2.21) gives the same weight to each triangle. C is sometimes referred to as **global clustering coefficient**, but its formulation should be made explicit to avoid confusion with the network average clustering coefficient.

If A is the adjacency matrix of an unweighted undirected graph, $N_{\triangle}$ and N_3 can be calculated from its entries a_{ij} as follows:

$$N_{\triangle} = \sum_{k>j>i} a_{ij}a_{ik}a_{jk}, \tag{2.22}$$

$$N_3 = \sum_{k>j>i} a_{ij}a_{ik} + a_{ji}a_{jk} + a_{ki}a_{kj}, \tag{2.23}$$

where the sum is taken over all triples of distinct nodes i, j, k only once.

Centrality Indices

A centrality index is a measure of how central a node is in the network; multiple indices exist, in accordance with different definitions of centrality, and each may have a different degree of importance depending on the field of application and the goal of the analysis. Some centrality indices are defined for edges, as well.

In general, the definitions of centrality are such that the indices are local properties, but their computation requires evaluations at a global level. For this reason, there have been many efforts to find fast algorithms to compute centrality indices.

Three centrality indices are commonly used across most fields of application:

- closeness centrality;
- stress centrality;
- betweenness centrality.

Closeness centrality expresses the degree to which a node is close to all others, in the sense that the node has a low distance to all the other nodes. This can be thought of as being central in a topological sense: intuitively, the distance between two peripheral nodes at opposite sides of a complex structure will be higher than the distance from a central node to any peripheral node.

The closeness centrality index of a node v is defined as

$$C_C(v) = \frac{1}{\sum_{i \in V} d_G(v,i)}, \tag{2.24}$$

where $d(v,i)$ is the geodesic distance from v to i. This definition holds if V is connected; if this condition is not satisfied, the simplest possible strategy is to compute centrality indices separately for each connected component.

The definitions of stress centrality and betweenness centrality involve the concept of shortest path. Stress centrality is defined as the number of shortest paths passing through a node; betweenness centrality is a measure of how frequently a node appears in shortest paths.

In formulas, the stress centrality of a node v is:

$$C_S(v) = \sum_{s \neq v \neq t \in V} \sigma_{st}(v), \tag{2.25}$$

where $\sigma_{st}(v)$ is the number of shortest paths from s to t that include v. The betweenness centrality of v is given by

$$C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}, \tag{2.26}$$

where $\sigma_{st}(v)$ is defined the same way as above, and σ_{st} is the total number of shortest paths from s to t.

The betweenness centrality index is often normalized to the number of node pairs excluding v; for simple unweighted graphs, this corresponds to

$$\frac{(N-1)(N-2)}{2}, \tag{2.27}$$

where $N = |V|$. This normalizes the index to a scale from 0 to 1, making it simpler to identify the most central nodes.

2.4 Reference Models

In the analysis of a network with unknown properties, it is often useful to compare its indices with those of a reference model. This ought to be a network with known properties with some common ground with the network under analysis, such as the same number of nodes and edges.

The most common reference models are:

- Erdős-Rényi model (or random network model);
- Watts-Strogatz model (or small-world model);
- Barabási-Albert model (or scale-free network model).

These are defined as parametrized constructing algorithms.

2.4.1 Erdős-Rényi Model

The Erdős-Rényi model is held as a reference for random networks. Two variants of the model exist: one is aimed at the construction of a random graph with a set number of nodes and edges, while the other sets a fixed probability for any pair of nodes to be joined by an edge.

The first variant is also referred to as the $G(n, m)$ model, where n is the number of nodes and m is the number of edges. It was introduced by Erdős and Rényi [20] and is formalized as follows:

1. Consider a set V of n nodes;
2. Build the set $\hat{E}$ of $\binom{n}{2}$ possible edges on V;
3. Build E by choosing m edges from $\hat{E}$, uniformly at random, without replacement (i.e. avoiding multiple edges between any pair of nodes).

The resulting graph $G = (V, E)$ is not guaranteed to be connected and may have some isolated nodes.

The alternate variant was introduced by Gilbert [36], but is commonly referred to as the $G(n, p)$ version of the Erdős-Rényi model:

1. Consider a set V of n nodes;
2. For each pair of nodes $i, j \in V, i \neq j$, add an edge (i, j) to E with probability p.

The resulting graph $G = (V, E)$ is guaranteed to be a set of isolated nodes if $p = 0$, or a complete graph if $p = 1$. Clearly, an increase in p means a greater likelihood for the resulting graph to include a larger number of edges.

2.4.2 Watts-Strogatz Model

The reference model proposed by Watts and Strogatz [59] is meant to provide a method to build a random network with two specific properties: a high clustering and a short average path length. These are referred to as 'small-world' properties.

The model takes three parameters: a number of nodes (n), a number of initially connected neighbors for each node (k), and a probability of rewiring (p). It is required for k to be an even integer, with $k \ll n$. The constructing algorithm is as follows:

1. Build a regular ring network of n nodes, where each node is connected to its k nearest neighbors (half on each side);

2. For each edge in the starting network, rewire the edge with probability p, where rewiring is done by detaching one of the incident nodes and attaching the edge to another node, chosen with uniform probability, avoiding loops and duplicate edges.

This model was built in an attempt to replicate properties observed in some real-world network, most notably social networks, in which common acquaintances between strangers act as 'bridges' that make it possible to reach any node with a small number of steps. Another goal of this model is to provide a method to build a network that falls somewhere between random networks built with the Erdős-Rényi model and regular networks, in the sense that the former may have low average path lengths, but tend to low clustering coefficients, while the latter have a high clustering coefficient, but with a high average path length.

2.4.3 Barabási-Albert Model

For many applications, the Watts-Strogatz model has a shortcoming in the fact that its generated instances tend to lack hubs, i.e. nodes with a significantly higher degree than the network average. Hubs are observed in many real-world networks, particularly those describing self-organizing systems. The degree distribution in such networks obeys a power law, in the form

$$P(k) \sim k^{-\gamma}, \tag{2.28}$$

for some fixed parameter γ, which is independent of the size of the network. This is referred to as the scale-free property of a network.

The Barabási-Albert model [3] was designed as a method to build networks exhibiting this property; its building algorithm is an iterative growth process based on the concept of preferential attachment. The starting point is a small initial set of m_0 nodes. At every step, a new node is added to the network, and linked to existing nodes by $m < m_0$ new edges. The probability for the new node i to be joined with an existing node j is to be proportional to the degree of j:

$$P(i \rightarrow j) = \frac{k_j}{\sum_{v \in V} k_v} \tag{2.29}$$

That means that new nodes tend to connect to "popular" nodes. After a sufficient number of steps, a network built with this method exhibits the scale-free property. More specifically, for large numbers of nodes, the degree distribution tends to a power law (2.28) with $\gamma = 3$; this was confirmed analytically [8]. Scale-free networks with a different value of γ in the power-law degree distribution can result from adjustments in the growth and preferential attachment process [1].

Chapter 3

Ecological Networks

3.1 A Paradigm Shift in Nature Preservation

Over the past few centuries, increasing portions of human population have moved away from rural areas toward urban agglomerates, to the point that a majority of human activities are nowadays focused on cities. This has resulted in the urbanization of increasing amounts of land, including on the coastline. Moreover, a connection of cities with contiguous infrastructure has become a requirement for the provision of essential services, such as the transportation of people and goods via roads and railways, communication (telephone and Internet backbones, etc.), and the distribution of electrical power. A direct consequence of this process of urbanization is the elimination of portions of habitat patches from both land and sea. Human infrastructure has become a distributional barrier for many species. In extreme cases, these barriers are impassable, not unlike the oceans for terrestrial animals, or non-forest habitats for forest species; in other cases, while not preventing migration, the presence of human infrastructure acts as a factor to reduce the population of certain animals. An example of this is given by the phenomenon of roadkills.

Concern for the protection of the environment, particularly species and habitats at risk of extinction, has given rise to the creation of nature reserves. Historically, these have existed in the form of isolated regions, mainly for two reasons: first, reserves were tailored for the preservation of the most endangered species, and secondly, the contiguity of infrastructure for human activities was held as a priority. Eventually, this led to a heavy fragmentation of habitat patches. By the second half of the 20th century, the limited effectiveness of this approach had become evident [15]. If the plan to establish a reserve is intended precisely to host a target species with a given estimated population, the area assigned to it might be sized accordingly, resulting in a small nature reserve. An insufficient land size acts as a limiting factor in more than one way: not only will a small reserve be able to host a smaller population of the target species, but it will be possible to host fewer different species in the reserve overall, as well. Moreover, a similar phenomenon is observed in relation to the distance between reserves: if said distance is excessive, the number of species that it is possible to protect is also reduced.

As a way to address these shortcomings, there has been an increasing push toward the creation of ecological networks [6] [58]; that is, rather than being thought of as a self-contained entity, each nature reserve is to be designed to contribute to large-scale conservation goals. Migration and dispersal of animals and plants are the essential elements of a network behavior that should be observed in a set of nature reserves. The presence of migration paths for animals can increase their chances of survival in exceptional cases, such as the occurrence of local natural disasters; however, the importance of migration paths lies in their effects in normal conditions. In literature, a set of populations of a species, found in different areas, is often referred to as a 'metapopulation' [43]. In the extreme case where no migration can happen between sites, each population makes up a separate metapopulation, and in time, the genetic material in each site can diverge, while becoming poorer due to excessive inbreeding. Migration of animals and dispersal of plants act as factors in merging the genetic pools of the populations of different areas, reducing the frequency of inbreeding and increasing the degree of biodiversity, which in turn also reduces the risk of extinction of a species.

Network behavior emerges spontaneously when conditions for its occurrence are met. In some cases, it is sufficient to make sure that the distance between core areas of nature reserves is not excessive; most notably, this is the case for most birds, but it may be true for other animals as well, depending on the quality of the surrounding matrix [53]. In other cases, the connectivity between suitable habitat patches for a target species may be enhanced by the creation of 'habitat corridors' (or 'green corridors'), defined as "linear strips of habitat connecting two or more larger patches of habitat, surrounded by a dissimilar matrix" [5]. It must be noted that, while corridors are meant as passageways for most animals, they represent a permanent settlement for plants and some smaller animals, such as insects.

Several studies have confirmed habitat corridors to be an effective way to improve biodiversity in a region, at least in the short term [37]. In designing a corridor, it should be kept in mind that their efficacy depends on several factors, including their quality relative to that of the surrounding matrix [53]. In general, a corridor ought to be designed to meet the needs of the species intended to use it [33]; in some cases, it is more sensible to arrange a set of disconnected patches to form a line, rather than a contiguous strip of habitat: this kind of corridor is referred to as a set of 'stepping stones'. Lastly, to enhance the degree of separation between nature reserves and urban areas, a transitional area (referred to as a "buffer zone") may be established, as a way to keep reserves at a distance from heavily urbanized areas. Figure 3.1 illustrates all of these elements in a sample configuration.

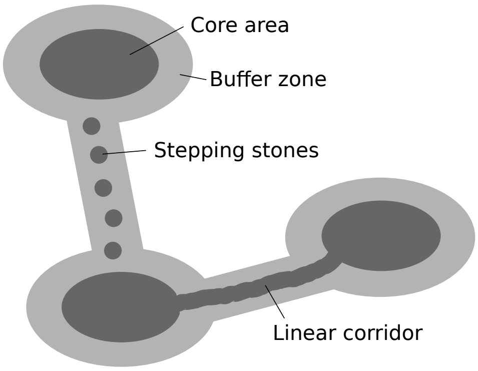


FIGURE 3.1: Sample configuration of three nature reserves and two habitat corridors: a set of stepping stones (upper-left) and a contiguous corridor (lower-right).

A comprehensive review of ecological network concepts, inclusive of details on these elements, is found in [6].

The costs to design, implement and maintain an artificial habitat corridor represent a serious issue, particularly when intersections with infrastructure such as roads and railways are involved, as these introduce the necessity for bridges, overpasses or underpasses. It stands to reason that the problem of maximizing a benefit/cost ratio since the planning stages can be considered extremely important.

The first mention of habitat corridors (then called "greenways") as a proposed solution to the shortcomings of the traditional approach to nature preservation can be dated back to the 1920s. However, the first implementations were many decades later. In Europe, nature conservation became a topic of great political importance only after World War II, under the pressure of a disastrous state of natural resources and a necessity of recovering from it [41]. Initially, policies for nature conservation were formulated in many countries, at a national level; then, as the European Union was formed, member states eventually coordinated their efforts toward a EU-wide ecological network, under the project denominated "Natura 2000".

3.2 Graph Models for Ecological Networks

Most complex systems consisting in a network can be represented mathematically by a graph, with the purpose of determining their properties or comparing them with other instances of the same kind of system. Ecological networks make no exception, and their graph models have been an object of study since the beginning of the 21st century, when it was proposed that a graph-theoretic approach to the analysis of ecological networks brings the advantage of being applicable to larger-scale case studies than previous approaches allowed [56], and moreover, it is useful in building a bridge between the field of biology and ecology and that of environmental and land engineering, so long as methods based on graph theory and complex network theory are seen as complement to other analysis techniques, rather than a replacement thereof.

The latter point is particularly important, given the nature of graphs as a mathematical object. The properties of a graph, including those derived from complex network theory, are invariant to its presentational features, which include the position taken by any node in its visualization. Fields of application exist, such as the study of protein-protein interactions, where the displacement of nodes and their presentational layout does not indeed play any role, other than to provide the end user of complex network analysis software with an easy way to read results. In contrast, ecological networks fall under the category of spatial networks, i.e. those networks in which nodes are located in a space equipped with a metric [4]. Spatial features and constraints can not be ignored, whatever method is considered for building a graph model of an ecological network.

Keeping this in mind, it seems sensible to build graph models by determining edges and weights in a such a way, that it represents structural connectivity as it is found in the ecological network; essentially, in an attempt to describe what the network is like. However, this approach has proven to be unsuccessful, the main reason being that it is extremely difficult to determine edge weights to represent structural connectivity in a meaningful way [57]. Theoretically, these should be made to reflect the presence of physical features found between areas represented by nodes, which may have a positive or a negative impact on the migration of species: for example, the presence of mountains between any two nodes may be a cause for the lack of edges, or for edge weights representing very low chances for flows between nodes. In practice, the variety of helping elements and possible obstacles to migration makes it so that representing them by the presence of an edge and by its weight turns out to be an oversimplification.

Instead, graph models of ecological networks are commonly built to represent functional connectivity, by placing edges and adjusting weights according to amounts of migrations of species, whether actual, potential, or estimated. This restricts the scope of a graph model to a single target species, introducing the requirement to build multiple graphs to represent the state of the network as a whole; nonetheless, the functional approach has proven to be more effective and has been adopted in a wide range of studies. Rather than using graph-theoretical approaches and complex network analysis software, structural connectivity is generally analyzed using Geographical Information System (GIS) tools; together with spatial databases, these can also act as valuable tools to build functional models.

3.2.1 GIS Data Types

Spatial data used for habitat conservation studies generally falls into one of these categories [56]:

- collections of spatial points in a landscape;
- samples of measurements, referred to points in a landscape;
- subdivisions of a landscape in a lattice, where each region is assigned some value or associated with a measurement.

GIS tools can work with different representations of spatial data, and can present different views thereof. Most importantly, any subdivision of a landscape into regions can be represented as raster data or vector data. This distinction is not unlike the one found in applications of computer graphics: raster data (also referred to as a 'bitmap') consists of collection of pixels, which may be associated with their color value; in a similar way, a portion of landscape can be subdivided into a grid of elements with a fixed size, and each element can be associated with a value. Vector data, on the other hand, consists of a collection of paths, which may make up polygons to represent boundaries of regions. Vector data allows a higher degree of precision, as regions represented as vectors may have boundaries with arbitrary directions, but the computational costs associated with its use can be significantly higher than those of raster data.

Another important aspect to be considered is the system of coordinates. Spatial data is represented on a map projection, i.e. a two-dimensional representation of the surface of the Earth. Map projections fall into one of three classes, according to the basis of their method of creation: there exist cylindrical projections, conical projections, and planar projections, depending on whether the globe is projected onto a cylinder, a cone or a flat surface. Each projection type introduces distortions, with a different effect on map features. Depending on different methods of projection, it is possible to preserve angular conformity (orthomorphic projection or conformal projection), have a constant scale for distances (equidistant projection), or preserve the proportion of surface areas (equal area projection), but not all three at the same time. As preserving one property generally causes a larger degree of distortion in the other two, a compromise solution of minimizing distortion in all three features, while preserving none, is sometimes sought (compromise projection). Systems of coordinates are often based on latitude, longitude and height on the globe, but the use of a map projection may be associated with a reference system placed on the projection, with an alternative system of coordinates associated to it.

The choice of a map projection and system of coordinates should depend on which of the three properties is most important to preserve for the

TABLE 3.1: Design choices for graph models of ecological networks.

Choice	Examples
Species	An endangered species; an umbrella species
Scale	A municipal area; a region; a continent
Granularity	A node for each habitat patch or each conservation area
Corridors	As nodes or as edges

goals of a study. In practice, however, the choice is sometimes dictated by national or local standards, which mandate that public spatial databases use a specific system, usually justified by the fact that a specific map projection has been optimized to minimize distortion in an area of interest. Depending on the properties of the base system used in data sources, it may be necessary to convert spatial data to a different system prior to performing calculations, or accept a degree of approximation in results.

3.2.2 From Landscape Data To Graphs

The creation of a graph model for an ecological network, using raw data and computation results from GIS tools, is heavily influenced by a number of design choices, to be made consistently with the fundamental goals of analysis. First, the motive behind the creation of a graph model has to be considered. Representing the state of the network is generally not enough; the success of a graph model depends on its versatility and, most importantly, its ability to predict the effect of changes on the network. The comparison of results of analysis performed on a graph representing the current state of the network with those obtained on a modified instance of the graph, where nodes, edges, or edge weights are adjusted according to changes that may happen naturally or as a consequence of human intervention, is supposed to provide insight on the effects of these changes.

The method to build a graph generally consists of determining a set of nodes, a set of edges and, if applicable, edge weights and other attributes. It is important to have considered issues involving both nodes and edges before any final decision is made. The most important and recurring design choices to be made are summarized in Table 3.1 and described in the remainder of this section.

Species

As already pointed out, a model representing functional connectivity keeps the focus on a target species, and the characterizing features of that species have an influence on most design choices. It stands to reason for the most endangered species to be common choices for target species, but research goals may involve a number of species at the same time, not all of them being at risk of extinction. When multiple species are considered as a target, it is possible to build multiple graphs to represent the state of the network with respect to each species; this may not be a viable option if the number of species is considerable.

A common approach in this case is that of finding an "umbrella species". This is defined as a "species with large area requirements, which if given sufficient protected habitat area, will bring many other species under protection" [9]. In other words, a species is to be chosen, such that the end result of the implementation of beneficial measures to that species brings an improvement for a large number of other species. This concept has been applied in the management and planning of conservation areas, and has proven to be successful in many situations, especially under strict time constraints for analysis and planning [51]. The same approach can be considered for large-scale goals, keeping its limitations in mind.

Scale

The size of the area under consideration may vary from that of a municipal area to that of a region, or even a whole continent.

Scale can be determined by the goals at hand, such as what research questions and what land management plans are the most immediate concern. Examples range from continent-wide studies for the prediction of the effects that local changes in climate and land use [42] to analyses of the resilience of small-scale green infrastructure in peri-urban settings [14]. In general, biologists and ecologists are likely to be concerned with studies at different scales, while scholars and practitioners in the field of land and environmental engineering may be most interested in small-scale studies; however, the very purpose of ecological networks is that of enhancing the ability of local reserves to contribute to large-scale goals, and as such, large-scale studies should be considered by land managers as well as ecologists.

Granularity

Granularity refers to whether a node should represent each single habitat patch (fine granularity), a whole conservation area (coarse granularity) or some intermediate degree. This choice is heavily influenced by that of scale, as a fine granularity paired with a large scale may result in a very large graph, with a corresponding increase in computational costs.

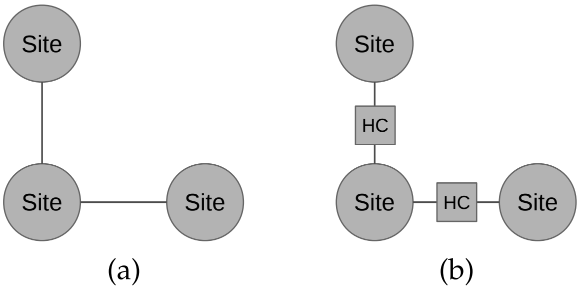


FIGURE 3.2: Graph model representations of the sample configuration of Figure 3.1. (a) Habitat corridors modeled as edges. (b) Habitat corridors modeled as nodes.

Representation of Corridors

Habitat corridors may be represented as edges, treating them as connecting elements, or as nodes, treating them as additional habitat patches (see Figure 3.2).

The choice should be consistent with that of granularity and with the species of interest: as already observed, habitat corridors are permanent settlements for plants and some smaller animals, and in that case, it may be most sensible to represent them as nodes, adapting the choice of granularity if necessary.

3.2.3 Basic Analysis

Once a graph model of an ecological network has been established, its analysis can uncover advanced properties, which may not be simply derived from the corresponding geographical map.

A preliminary observation, which is simple but proves most important, is to check whether the graph is connected or made up of several connected components. This is needed because some indices are defined for a connected component, and it has a meaning in itself in the fact that, assuming the graph has been instantiated for a single species and edges reflect the actual possibilities of migration, each connected component is associated with the existence of a separate metapopulation of the species [43], and merging them can be identified as a goal of environmental planning.

Ranking network elements by their centrality is another common kind of analysis. This is useful in determining criteria to produce modified version of a graph model, for comparison with the instance representing the current state of the network. In general, the removal of nodes with a high betweenness centrality index affects a larger number of shortest paths, with an impact on network connectivity [7]. This mostly affects the network as a whole at large scale, while local agglomerates of patches with a strong organization in cliques may be unaffected for the most part [21].

3.3 Natura 2000

3.3.1 Basic Elements of Natura 2000

In the European Union, policies for the protection of nature and biodiversity have been extended to include the creation of an ecological network, denominated "Natura 2000". The inception of Natura 2000 can be dated back to May 1992, when Council Directive 92/43/EEC ("Habitat Directive") was approved. This Directive complements the previous Council Directive 79/409/EEC ("Birds Directive"), which was later replaced by Council Directive 2009/147/EC, in giving the definitions for a network to cover the cases for conservation of birds and other species.

The main elements of Natura 2000 are nature protection areas, categorized into two distinct sets:

- Special Protection Areas (SPA), designated by member states according to the EU Birds Directive;
- Special Areas of Conservation (SAC), designated by member states according to the EU Habitats Directive.

While SPAs are simply designated by member states, the process to designate a Special Area of Conservation generally consists of two phases. First, a site is proposed by a member state of the Union to become a Site of Community Importance (SCI); once the site has been approved as a SCI by the EU, the member state can designate it as a SAC.

If it is considered important for the protection of birds as well as other forms of wildlife, a site can be recognized both as a SPA and as a SAC (or SCI), at the same time. Moreover, the boundaries of a SPA may be intersected with those of SACs and SCIs, and vice versa; that way, any area may be part of a SPA and a SAC at the same time, even though the relevant sites are designated independently. Sites designated as SPA may be adjacent to one another, without overlapping; the same applies to SCIs and SACs.

Privately owned land and areas dedicated to human activities may also become part of Natura 2000 sites. If that is the case, the site is to be considered an area where the EU seeks a sustainable management of natural resources in the appropriate context.

TABLE 3.2: The first few habitats listed in Annex I, with their Natura 2000 codes.

Code	Priority	Name
1110		Sandbanks which are slightly covered by sea water all the time
1120	yes	Posidonia beds (Posidonion oceanicae)
1130		Estuaries
1140		Mudflats and sandflats not covered by seawater at low tide
1150	yes	Coastal lagoons
...	...	...

3.3.2 Categorization of Habitats and Species

Within the Natura 2000 project, a categorization of habitats and species is provided. Habitats are listed in Annex I of the Habitats Directive; species are listed in the Birds Directive and in Annex II of the Habitats Directive.

Habitats are assigned a four-digit code, where each digit represents a taxonomic rank; for example, habitat codes beginning with a ‘1’ are used for “coastal and halophytic habitats”, and at the second level, ‘11’ denotes “Open sea and tidal areas”, ‘12’ is used for “Sea cliffs and shingle or stony beaches”, etc. (see Table 3.2). Where necessary to preserve this hierarchical categorization, a letter was used in some habitat codes, instead of a digit. Specific habitats are flagged as a ‘priority’ for conservation purposes. An official Interpretation Manual of European Union Habitats contains a thorough description of each habitat, including the corresponding categorization under different projects.

Lists of species of interest, for which it is required to setup conservation areas, are given in the relevant Directives. Each species is assigned a unique code for identification. Unlike those in use for habitats, these codes do not reflect any form of taxonomy, with the sole exception of the use of the letter ’A’ in the first position to identify birds (see Table 3.3).

3.3.3 Collection of Reports

As part of the activities that make up the Natura 2000 project, sites are periodically visited in order to collect data on their current state, concerning the recognized set of species of interest and, at the discretion of local experts, any other species considered relevant. Reports are written following a Standard Data Form found in Commission Implementing Decision

TABLE 3.3: Excerpts from a list of species with their Natura 2000 codes.

Code	Category	Name
...	...	...
1190	Amphibian	Discoglossus sardus
6207	Amphibian	Speleomantes flavus
...	...	...
A400	Bird	Accipiter gentilis arrigonii
A293	Bird	Acrocephalus melanopogon
...	...	...
1095	Fish	Petromyzon marinus
6135	Fish	Salmo trutta macrostigma
...	...	...

2011/484/EU, replacing a previous version of the form, released with EU Commission Decision 97/266/EC.

The current Standard Data Form consists of seven sections:

1. Site Information: this includes an identification code, the classification (as SPA, SCI, SAC), dates of first compilation and latest update;

2. Site Location: coordinates on Earth of a site centroid, the extension of the site and which percentage is made up of marine area, which administrative region the site is part of, a subdivision into biogeographical regions (e.g. Alpine, Atlantic, etc.);

3. Ecological Information: detailed data on each habitat found within the site, including an evaluation of quality (i.e. its state of conservation, etc.); data on each recognized species of interest found within the site, with an evaluation (e.g. size of population, whether a settlement is permanent, etc.); optionally, data on any other species deemed relevant by local reporters.

4. Site Description: includes the percentage of extension cover for each habitat, notes on quality and importance, a classification of threats and activities, and optionally data on the ownership of the site;

5. Site Protection Status (optional): references to local designations, relations to other sites;

6. Site Management: administrative references to the entities in charge of site management, and documents detailing the management plans;

7. Map of the Site (optional): a graphical map, which may be attached in PDF format.

Collected data is eventually stored in a public data base, and a web interface exists for easy access by the general public. From the point of view of data analysis, when the whole dataset is retrieved, two potential roadblocks emerge.

The first issue is that part of the data is collected in the form of free text, i.e. it has to be stored as unstructured data; this includes the evaluation of threats and description of activities. Moreover, while most of the species data is stored in structured form, it is not available in a consistent manner, due to difficulties in their collection at the source. Namely, the size of population can be reported as a number of specimen, a number of couples, or an estimation in thousands or other units; however, in many cases, it is impossible to produce an exact or even an approximate number, particularly for smaller animals. In fact, reporters are asked to include an evaluation of the "quality of data" they managed to collect, i.e. whether they were able to gather exact or approximate data, rough estimates, or insufficient data.

The second issue is the fragmentation of data that results from a number of factors: individual reporters may choose to include data on any species they consider relevant for the site (within the data base, this is referred to as "other species"); however, in doing so, it is not necessary to report unique species codes or names, and in fact, a large number of entries do not contain any species code, or use alternative names for species or subspecies. Moreover, some data instances are found with typographic errors, possibly persisted from the original forms, or introduced by manual data entry.

3.3.4 Data Model

The entities, on which data is collected, are the following:

- Main entities
  - Site
  - Habitat
  - Species ("other species" are stored in a separate table)
- Administrative references
  - Region
  - Site Relation (references to other projects)
  - Management Body
  - Management Plan
  - Contact Information

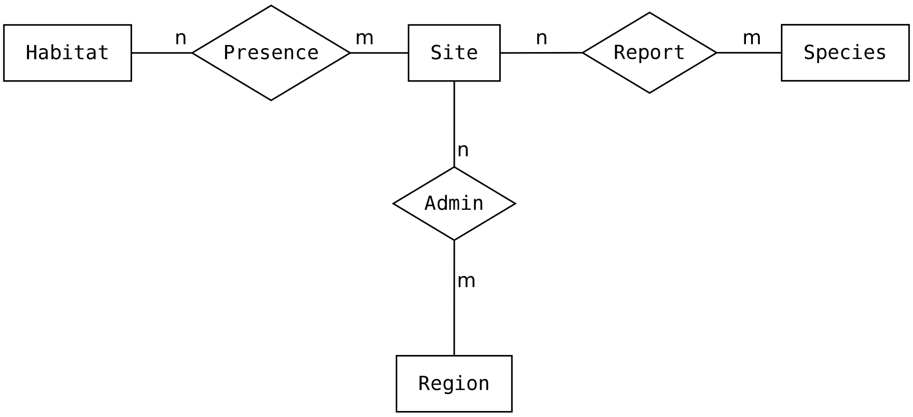


FIGURE 3.3: A simplified version of the actual Entity-Relationship Diagram of the Natura 2000 Data Base.

    - Document (generic)

- Categorizations
    - Impact type
    - Bio-geographical type
    - Habitat class

Among the main entities, the site acts as a central entity, to which all the others are linked (Figure 3.3); this reflects the original organization of reports. Each Natura 2000 site is linked to the different habitat types and the set of species that were reported in it, but no explicit relationship is established between species and habitats.

Considering that a site is made up of a mosaic of habitat patches, this means that the available data is not sufficiently fine-grained to determine which set of plants makes up each patch, or which habitat type is preferred by a species of animals; this is assumed to be part of expert knowledge, or to be available from external documents. While the lack of fine-grained information poses no particular problems for the original purposes of the Natura 2000 project, a direct consequence is that there is no simple (and error-resilient) way to determine whether any species can (or should not) be relocated to another site, by ensuring that a proper habitat or set of habitats is already found in the target site and that a balanced ecosystem can be preserved.

The solution to this problem would be made easier by adopting an extended model, where the habitat patch is represented as an entity of its own (Figure 3.4) [29]. The species could then be associated with the habitat patch

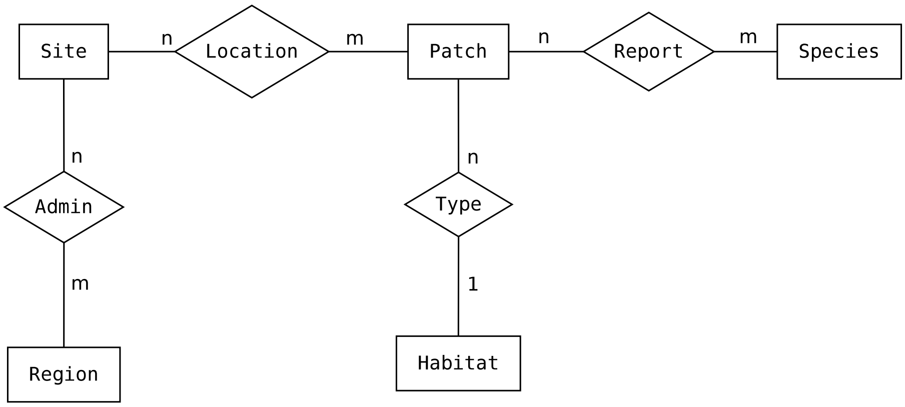


FIGURE 3.4: Alternate Entity-Relationship Diagram of the Natura 2000 Data Base, with the habitat patch as a central entity.

(or patches) where they are prevalently found, and it would be possible to infer an association with habitat types with an automatic process.

3.3.5 CORINE Land Cover Data

Another way to address the limitations of the Natura 2000 dataset is to complement it with external data sources. A viable candidate is the database from the CORINE program (Coordination of Information on the Environment), initiated in 1985 in the European Union and currently managed within the European Environment Agency. While the ultimate goal of both projects is to contribute to the preservation of biodiversity, the CORINE program is exclusively aimed at the collection and storage of information, with a focus on consistency across member states and compatibility of data. The program is intended to provide scholars and land managers with a common framework, on which different approaches may be based, thus reaching a higher degree of coordination between different fields of activity, and among practitioners in different areas. Clearly, the goals differ from those of Natura 2000, which include a proactive approach to the conservation of biodiversity.

The database provides geographical information on land use data of member states, referred to as CORINE Land Cover (CLC). The mapping results from the interpretation of satellite images and is available at 1:100000 scale, with a minimum mapping unit of 25 hectares, and a minimum width of 100 meters for linear elements. Land is categorized into classes according to a taxonomic model of land use, with five levels expressing increasing degree of detail. Codes are made up of sequences of digits (written simply by concatenating digits in the form `ij` at Level 2 or `ijk` at Level 3; alternatively, dots are used to separate digits: `i.j` at Level 2, etc.). The length of a code corresponds to the level of detail it expresses, and codes are assigned following a strict prefix rule, so that truncating a code to a certain number of digits returns the correspondent class at the broader level. For example, class `3.1.2` corresponds to coniferous forests, which are a subcategory of the Level 2 class `3.1`, associated with forests of all kinds. As an example, Level 3 codes for the subclasses of Level 1 class `3` are reported in Table 3.5.

TABLE 3.4: CORINE Land Cover (CLC) Level 1 and Level 2 nomenclature

	Level 1	Level 2
1	Artificial surfaces	11 Urban fabric
		12 Industrial, commercial and transport units
		13 Mine, dump and construction sites
		14 Artificial, non-agricultural vegetated areas
2	Agricultural areas	21 Arable land
		22 Permanent crops
		23 Pastures
		24 Heterogeneous agricultural areas
3	Forest and seminatural areas	31 Forests
		32 Scrub and/or herbaceous vegetation association
		33 Open spaces with little or no vegetation
4	Wetlands	41 Inland wetlands
		42 Maritime wetlands
5	Water bodies	51 Inland waters
		52 Marine waters

TABLE 3.5: CLC Level 3 nomenclature: codes beginning with '3'

Level 2 code	Level 3	
31	311	Broad-leaved forest
	312	Coniferous forest
	313	Mixed forest
32	321	Natural grasslands
	322	Moors and heathland
	323	Sclerophyllous vegetation
	324	Transitional woodland-shrub
33	331	Beaches, dunes, sands
	332	Bare rocks
	333	Sparsely vegetated areas
	334	Burnt areas
	335	Glaciers and perpetual snow

Unfortunately, a regular time frame for database updates is not set within the CORINE program. The first release of CORINE Land Cover is referred to as CLC1990; data collection for this version happened over a very long period, from 1985 (at the initiation of the CORINE program) to 1998. This version was followed by CLC2000, CLC2006, and lastly CLC2012. In these three version, the number is an indication of the year of data collection, with a margin of error of one year.

Chapter 4

Land Management and Complex Networks

4.1 Urban Planning and Natura 2000

A trait shared by Natura 2000 and other initiatives of the European Union is the division of competences between EU and national bodies. Ultimately, since environmental planning makes up part of the activity of urban planning and management, administrative bodies down to the local level (such as municipalities) are involved with Natura 2000.

In fact, urban planning is not exclusively concerned with the management of urbanized areas; many local administrations are responsible for a territory that includes a city, as well as the rural or natural areas in its proximity. Urban planning is an interdisciplinary activity, encompassing civil engineering, architecture, economy, health, sociology, and ecology, among other fields.

As part of the activity of urban planning, strategies for the management of a territory are to be outlined in official documents, referred to as management plans. These documents report on the current situation and goals over several periods of time (short-term, mid-term, long-term). A management plan for an area including a Natura 2000 site should take into account its features from multiple points of view (boundaries, extension, climate, presence of pollution, habitats, landscape, social and economic aspects, subdivision into zones), provide a list of detected or possible threats to the survival of protected species and the conservation of habitats, and lastly, propose actions to address these threats and improve the conservation of species and habitats. The priority of each action should also be given.

There are technical, regulatory and political aspects involved in devising management plans. In the past decades, software tools have proven to be able to provide valuable assistance in the evaluation of technical aspects. GIS tools continue to be the most important class of software in use for environmental urban planning, although complex network analysis has also risen in popularity, and in fact, the ability to extract graph models from landscape features is being incorporated into GIS suites.

4.2 Building Graph Models

As previously mentioned (Section 3.2), among topological models, the state of an ecological network is best represented by a graph model expressing functional connectivity, with respect to a single target species. This kind of model shall be referred to as *single-species graph*, to avoid confusion with different models to be introduced later. A topological model can be built from data available from the Natura 2000 project, although design choices are heavily influenced by the availability of data. The examples that will be provided are based on the subset of Natura 2000 sites found in Sardinia, considered as a case study. The dataset in use was collected at the end of year 2014.

Keeping the framework introduced in Section 3.2.2 as reference, the limitations on design choices that derive from the exclusive use of the Natura 2000 dataset are the following:

- **Species**: all species of interest, listed in the Birds Directive and in Annex II of the Habitats Directive, can be considered as target species, since data on these is consistently available for the whole set of Natura 2000 sites. The same is not true for other species, although a subset thereof can be identified, for which a species code is consistently given; these species can optionally be considered in building a graph model, but extreme care should be taken to distinguish the absence of data from the absence of the species in a site; external data sources may have to be used.

- **Scale**: regional and national scale are possible with hardly any limitation; studies at continental scale may have to integrate datasets from non-EU countries, whereas studies at sub-regional scale may hit limitations due to the granularity of data. The examples provided in this work will be made at a regional scale.

- **Granularity**: in general, a node is to represent an entire Natura 2000 site, since the available data is hardly suitable for a finer granularity (see Section 3.3.4). Finer granularity may be made possible by the use of map data, if available, and by the integration of expert knowledge.

- **Representation of Corridors**: no data is made available on the presence or implementation of habitat corridors. Therefore, barring the availability of external data sources, it will be necessary to make assumptions on possible migration paths, to be represented by edges, based on some criteria.

To build graphs for the case study, a node list is built from the list of Natura 2000 sites in the administrative region of Sardinia, which includes the island of Sardinia (second-largest in the Mediterranean Sea) and the smaller islands which surround it.

At the time of this study, no site has been designated as a SAC within this administrative region; 87 sites are designated as a SCI, 31 as a SPA, and 6 sites have both designations, with the same boundaries. Thus, the total number of Natura 2000 sites in Sardinia is 124. Among these, 107 are located on the main island. In the graph models, each site shall be represented by a node, regardless of its designation: if a site is designated as a SPA and a SCI (or SAC), it is considered one node; if a SPA and a SCI (or SAC) overlap, a node is created for each site, but their geographical distance is considered to be zero.

The criteria to draw edges between pairs of nodes can be based on:

- geographical distance: this can be calculated between boundaries, or between centroids, and a pair of nodes is not to be linked if their distance is larger than a definite threshold;
- presence of a given species in both sites in a pair.

To estimate geographical distances between sites, perimeter and centroid data was imported in a spatial database, using the open source SQLite engine with the Spatialite extension. In these examples, distance is computed between boundaries, keeping in mind that they are to be treated as an approximation, since they are calculated on a cylindrical conformal map projection (EPSG:32632, WGS 84 / UTM zone 32N). At a regional scale, the degree of distortion has been deemed acceptable.

The simplest method to build a single-species graph is as follows:

1. Consider the full set of nodes for the area being analyzed (full graph) or, alternatively, the set of graph where the target species is reported to be present (local graph).
2. Add an edge between any two nodes in which the target species is reported to be present, so long as their geographical distance is within a set threshold.

Figure 4.1 shows an example for *Cervus elaphus corsicanus*.

4.3 Graph Modifications

The extraction of complex network indices from a single-species graph is meant to contribute to the understanding of the aptness of the ecological network for the conservation of the target species. The analysis can provide

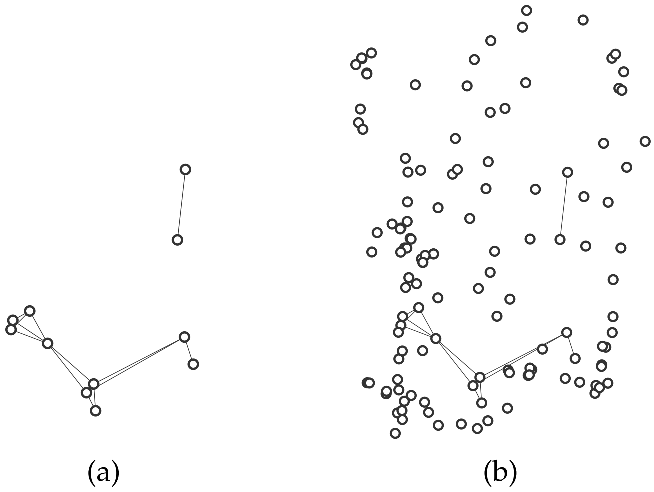


FIGURE 4.1: Single-species graph models of Natura 2000 sites in Sardinia, built for *Cervus elaphus corsicanus* (species code 1367). The position of each node roughly corresponds to the coordinates of the site centroid. (a) Local graph, in which only sites with reported presence of the species are represented by nodes. (b) Full graph, in which every Natura 2000 site in the Region of Sardinia is represented.

the basis for a large-scale evaluation of land management proposals, by way of comparing the current graph model with those that reflect proposed local modifications, in order to identify favorable modifications as those that improve the indices that correspond to set goals.

The real-world problem generally consists in finding a set of modifications that can be applied while respecting a set budget, maximizing the gain, which is represented in the graph model in the form of better network indices. In many applications, it is assumed that the set of nodes is not to be modified, and any proposal for modification is to involve only the set of edges. This leaves three subclasses for this problem, depending on which modifications are allowed [2]:

- 'updating': only the addition of edges is allowed (considered edges, which are not found in the initial graph, are referred to as 'virtual edges');
- 'downdating': only the removal of edges is allowed;
- 'rewiring': both the addition and the removal of edges are allowed; the usual approach is to remove a set number of edges, and subsequently add the same number of virtual edges.

It is generally intended for these problems to be solved on a connected graph; if the real-world network is not connected, the problem of connecting its component should be addressed beforehand, or each connected component should be thought of as a separate scope of application for these classes of problems.

The updating problem corresponds to the real-world problem of finding a set of new links, such that its addition can be performed while respecting budget constraints, resulting in as great a benefit as possible. In the graph representation, the budget can be represented as a maximum number of edges, or if necessary, as a maximum value for a cost function, to which each virtual edge contributes in different measures.

The downdating problem, on the other hand, applies to networks with some degree of redundancy and a maintenance cost associated with each link. It consists in finding a set of edges that can be removed from the network in order to decrease maintenance costs, keeping the impact on the efficiency of the network as small as possible. Any solution that causes the network as a whole to be disconnected should be disregarded.

The rewiring problem is closely related to the downdating problem, in that it considers link maintenance costs carefully; the goal, rather than reducing these costs, is to improve the efficiency of a network, while avoiding a hike in maintenance costs.

A problem that may occur in environmental planning is the *relocation problem*, which involves finding a suitable site for the relocation of part of the population of a species. If the site has to be found among those where the species has not been reported yet, this translates to a problem on the graph model; however, if the network is modeled into a single-species graph as previously described, it consists in the addition of nodes, and as such, it does not fall into any of the classes described so far. In some instances, the goal may be that of merging components in an initially disconnected graph model.

In order to add nodes to the single-species graph, suitable candidate sites have to be identified. In most cases, these should host habitat patches compatible with it, and be located within a maximum geographical distance from an already connected node. This distance should correspond to the dispersal distance of the target species. Due to the previously discussed shortcomings of the Natura 2000 dataset, the selection of candidate sites is not a straightforward activity to perform. The next section is focused on proposing a method to address the problem.

4.4 Similarity of Natura 2000 Sites

As previously mentioned (Section 3.3.4), the Natura 2000 dataset does not include records of any explicit relationship between each species and the single habitat patches where it is found, or any indication of the preferred habitat type for each species. While this is understandable, considering the original purposes that were considered before the collection of data, the drawback is that it is not straightforward to obtain a machine-readable representation of constraints that apply when proposing modifications.

In order to provide at least a partial solution to this problem, it is possible to consider similarity between sites as a criterion to determine whether a site is suitable to host a certain target species. Similarity scores are commonly used for the comparison of text documents and their classification according to their subject; the basis for these activities is the creation of vectors representing the occurrence of keywords in each document. The same principle could be applied to vectors built to represent Natura 2000 sites.

In the choice of a similarity measure for this application, it is desirable to adopt a measure with a definite lower and upper bound (e.g. 0 to 1), as this makes it simple to choose a threshold, i.e. a minimum similarity score between a pair of sites to be a prerequisite for a proposal to be considered, such as the relocation of a species population or the implementation of a habitat corridor. Among the measures that hold this property are the Jaccard coefficient for binary vectors, and cosine similarity for non-binary vectors.

The Jaccard coefficient of a pair of binary vectors is defined as follows. Let f_{11} be the number of attributes that are *true* (1) in both vectors; f_{10} the number of attributes that are *true* in the first vector and *false* (0) in the second; and in an analogous manner, f_{01} the number of attributes that are *true* in the second vector and *false* in the first. Then, the Jaccard coefficient J is given by:

$$J = \frac{f_{11}}{f_{01} + f_{10} + f_{11}}. \tag{4.1}$$

Cosine similarity is defined as the cosine of the angle between two vectors with non-zero magnitude in a multi-dimensional space. This measure is particularly useful for the comparison and categorization of text documents, due to its independence from document length. A simple formula to compute cosine similarity is:

$$cos(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{||\mathbf{x}||\,||\mathbf{y}||}, \tag{4.2}$$

where $\mathbf{x}$ and $\mathbf{y}$ are vectors, $\mathbf{x} \cdot \mathbf{y}$ is their scalar product, and $||\mathbf{v}||$ denotes the magnitude of a vector $\mathbf{v}$.

There are three viable choices to build vectors to represent Natura 2000 sites:

- species sets: the reported presence of species;

- habitats: the reported composition as habitat types;
- land use: the intersection of sites with patches from the CORINE dataset of land use codes, or a compatible dataset.

The first two sets of vectors can be built using data from the Natura 2000 project, while the last set involves cross-referencing sites with an external dataset. Associating a Natura 2000 site to the land use codes in its extension require that the intersections of CORINE Land Cover patches with the site are computed, which can be done with GIS software, such as the QGIS software suite [49]. For this case study, a public dataset made available by the Region of Sardinia was used, updated in year 2008 and based on the land use codes defined within the CORINE project. Upon computing the intersections, it became apparent that land use data was unavailable for 7 sites in the Region of Sardinia, among which 3 are located on the main island. More specifically, these sites have no intersection with any land patch in the CLC dataset in use. This leaves a set of 117 sites (104 on the main island), for which data is consistently available, making it possible to evaluate the usefulness of each data source for vector creation.

4.4.1 Similarity-based Graphs

The same kind of visualization used for single-species graphs can be used to provide a meaningful and straightforward way to visualize pairs of similar sites within a set geographical distance. In this section, a number of reference graphs are built on the set of 117 nodes, representing sites for which vectors can be built, with a maximum distance between sites of 30 Km. The number of edges in a graph with a link between every pair of sites within the set geographical distance threshold is 850 (Figure 4.2). This graph can be referred to as the *raw-distance graph* of this portion of the Natura 2000 network.

A graph instance based on site similarity, or *similarity-based graph*, can be built from the raw-distance graph by removing edges between node pairs with a similarity score below a set threshold. Different graphs result from a change of similarity measure, threshold values, or the vector set in use (created from species sets, habitats, or land use). Since graphs will be built for comparison using the same measure and threshold, while varying the origins of vectors, the denominations *species-set graphs*, *habitat graphs*, and *land-use graphs* shall be used as a way to distinguish data sources.

The Jaccard coefficient will be considered in this case study, as it is the simplest choice for a similarity measure in this context, being intended for binary vectors, which are trivially built from available data. Elements of a *species set* vector for a site will take a 0 for a species that was not reported, and a 1 for a species that was reported in the site, and likewise for *habitat*

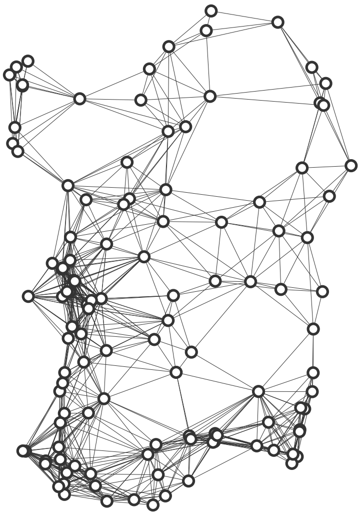


FIGURE 4.2: Raw-distance graph. Edges link pairs of nodes with a geographical distance up to 30 Km between boundaries. The position of each node roughly corresponds to the coordinates of the site centroid.

vectors. For *land use* vectors, the third level classes of CORINE Land Cover were chosen as labels of vector elements, because data on the fourth and fifth level were not consistently available across the territory, while the first two levels offer an insufficient degree of detail. Thus, the binary land use vector for a Natura 2000 site has an element for each Level 3 CLC code, set to a value of 1 if the site is intersected with at least one patch with that CLC code, and to 0 otherwise.

Taking the graph of Figure 4.2 as reference, and increasing a minimum similarity threshold from 0 to 1 in steps of $1/10$, it is possible to observe how the number of edges decreases for the three types of similarity-based graph [25]. Results are reported in Table 4.1 and visualized in Figure 4.3. For similarity of species sets, it can be observed that the number of edges decreases sharply at low threshold values: over a half of the edges are removed with the application of a threshold of 0.3. The decrease in the number of links is not as steep when similarity of land use is considered, with over half of the original edges removed at 0.4; however, a drop to zero edges is observed somewhere over the 0.8 mark. The most peculiar case is that of habitat similarity, which shows a very steep decrease at very low thresholds, followed by the highest number of kept edges at higher thresholds.

Further observations are possible on the visualized form of similarity-based graphs. Figure 4.4 shows graphs built with a 0.5 similarity threshold; while the land-use graph is made up of a large connected component and a

TABLE 4.1: Number of edges after applying a minimum Jaccard coefficient threshold

	Edges		
Threshold	**Land use-based**	**Habitat-based**	**Species-based**
0.0	850	850	850
0.1	777	655	739
0.2	661	533	573
0.3	509	415	350
0.4	360	295	198
0.5	232	205	134
0.6	104	120	53
0.7	34	58	24
0.8	5	36	16
0.9	0	14	14

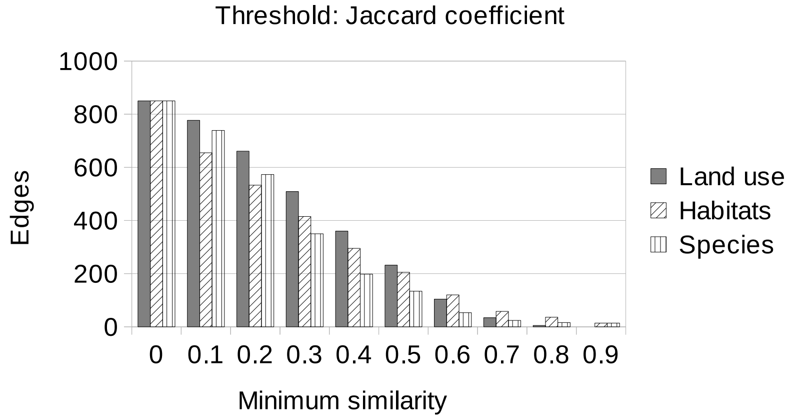


FIGURE 4.3: Number of edges in the graph model of the Natura 2000 network in Sardinia, when a threshold based on Jaccard coefficients is applied, calculated on land use, habitat configuration, and species presence data.

small number of isolated nodes and smaller components, both the species-set and habitat graph are made up of several disconnected components at this threshold level. More instances of the land-use graph with stricter similarity threshold values are shown in Figure 4.5. A striking difference between the land-use graphs at 0.5 and 0.6 threshold is that the latter has two large connected components, which roughly correspond to a subdivision into Northern and Southern Sardinia, in addition to a larger number of isolated nodes [28]. At a 0.7 similarity threshold, the number of edges reduced to about one third of those in the graph at 0.6, and only a few connected components of multiple nodes are left. This suggests that a threshold of 0.7 can be considered a very strong requirement for the linking of nodes, and a threshold of 0.6 may be considered as a reasonable requirement in many cases.

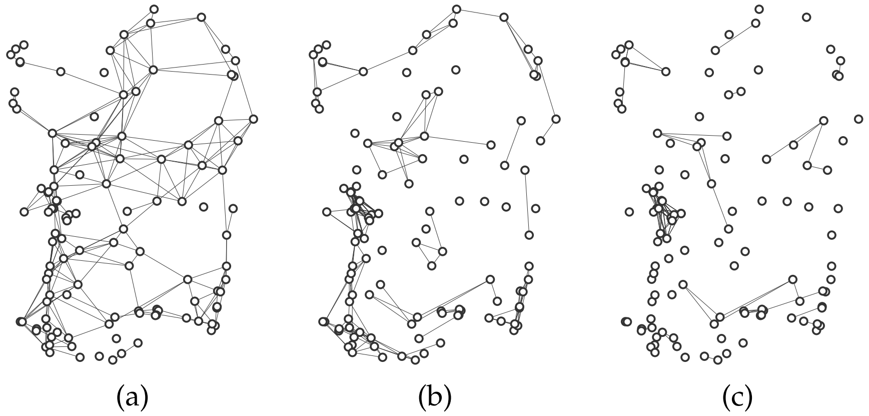


FIGURE 4.4: Similarity-based graph models of Natura 2000 sites in Sardinia, built with a 30 Km distance threshold, with edges drawn for a 0.5 similarity score or above. (a) Based on CORINE land use codes. (b) Based on Natura 2000 habitat codes. (c) Based on species sets.

4.4.2 Analysis of Hit Rates and Complex Network Indices

One of the measures for site similarity may qualify as a substitute for data on habitat suitability, when this is missing or incomplete; paired with geographical distance, this may be a criterion to identify suitable nodes for the relocation problem introduced in Section 4.3. In this case, a similarity-based graph becomes a tool in formalizing the method and visualizing its possible solutions. Suppose that a similarity-based graph is built from the same node set as the single-species graph and the same geographical distance threshold; then, a candidate node for relocation has the following property: in the similarity-based graph, it is adjacent to a node corresponding to one that is part of a connected component in the single-species graph.

Formally, let V be the full set of nodes that represent Natura 2000 sites in the region of interest. Let $G = (V, E)$ be the full single-species graph for which modifications are to be proposed, built with a geographical distance threshold found to be suitable for the target species. Then, let $G_s = (V, E_s)$ be a similarity-based graph built on node set V, with the same geographical distance threshold as G and a similarity threshold deemed sufficient. If

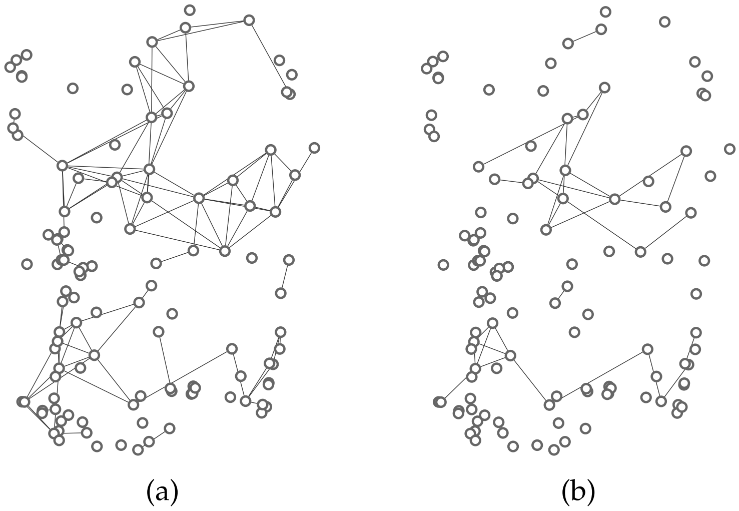


FIGURE 4.5: Land-use graph models of Natura 2000 sites in Sardinia, built with a 30 Km distance threshold. (a) Edges drawn for a 0.6 similarity score or above. (b) 0.7 similarity score or above.

$G' = (V', E')$ is a connected component in G ($V' \subseteq V$), and the following conditions are met:

$$i \in V', \quad j \in V, \quad j \notin V', \quad (i, j) \in E_s, \tag{4.3}$$

then $j \in V$ is a viable candidate node, and (i, j) is a candidate edge to link j to G'.

This leaves the question of which dataset, among those in use to build a similarity-based graph, is best for finding suitable candidate nodes. To answer this question intuitively, it can be argued that, if edges in a large number of single-species graphs are often found as edges in G_s, then G_s is likely to provide better candidates for relocation.

To compare the three datasets being considered, a graph should be built from each dataset, while keeping the same thresholds; this was done for a 30 Km distance threshold and a 0.5 similarity threshold. A comparison with 351 single-species graphs was performed and the number of hits and the hit rate was analyzed [32]. Results are reported for a few species in Table 4.2 as an example: in each row corresponding to a species, n denotes the number of edges in the single-species graph, l in each row denotes the number of edges that the land-use graph has in common with the single-species graph (and likewise for h, s counting the edges in the habitat graph and species-set graph, respectively). Hit rates were calculated as l/n, etc., and considered to be 0 for species with $n = 0$.

To compare hit rates as a measure of aptness for each similarity-based graph, it is possible to average results over every species. Considering the unweighted average over all rows, habitats and species sets have about a 42% average hit rate, while land use data can be ranked slightly worse, standing at 38%; if averages are weighted according to the number of edges (i.e. the ratio is calculated on the summation of hits vs. considered edges), the land-use graph and the habitat graph stand at approximately a 34% average hit rate, while the species-set graph is slightly behind, at 31%.

TABLE 4.2: Excerpt of the table of hit rates. The number of edges in each single-species graph in the set is reported. For each similarity-based graph, the number of hits (edges in the similarity-based graph that are found in the single-species graph), and the corresponding rate are reported. Weighted and unweighted average hit rates are also given.

		Land use-based		Habitat-based		Species-based	
Species code	Edges n	Hits l	Rate l/n	Hits h	Rate h/n	Hits s	Rate s/n
...							
6137	186	93	0.5	55	0.29570	24	0.12903
1367	15	9	0.6	8	0.53333	6	0.4
1373	8	6	0.75	2	0.25	2	0.25
...							
Sum	32538	10949	0.33650	11304	0.34741	10187	0.31308
Average hit rate			0.38063		0.41714		0.42587

Since all average hit rates can be considered quite low, and the differences among the three can be argued to be hardly significant, none of the three types of graph emerges as an undisputed best option to consider for the node linking process. However, strong differences between columns confirm that the three graphs are far from being analogous, as can be noticed visually (Figure 4.4). As a way to determine to what degree these differences are significant, Spearman correlation indices are calculated between pairs of columns reporting hit rates in Table 4.2, to assess whether any pair of graphs behave in a similar way with respect to hit rates. That is, if the hit rate for a pair of similarity-based graphs G_s, G_t were high for the same subset of species, it could be inferred that the criteria behind the definition of similarity scores used in G_s and G_t express a similar concept, or two closely related concepts.

Results of this study are summarized in Table 4.3. It is striking that the hit rates for the species-set graph and the habitat graph appear to be strongly correlated, with an index above 0.8, while neither has a significant

TABLE 4.3: Spearman correlation between sets of hit rates.

	Habitat-based	Species-based
Land use-based	0.08446	0.03489
Habitat-based		0.80397

degree of correlation with the land use graph. The latter observation may come as no surprise, seeing as that land use data originates from a different project and has not been updated in the same time frame; this test ought to be repeated when more recent data is made available. Concerning the higher correlation index for the species-set graph and the habitat graph, this can be interpreted as a confirmation that nearby sites characterized by similar sets of habitats can be inferred to host similar sets of species; compared to the set of land use codes taken as reference, the classification of habitats made within the Natura 2000 project appears to provide a better indication of suitability of a site for a species.

Correlations can be sought on complex network indices calculated on the three graph instances, as well (an excerpt is in Table 4.4). Node degree, closeness and betweenness centrality indices, clustering coefficient, and topological coefficient [55] were selected for this study. For each index, a high degree of correlation would imply that a higher value calculated on a graph corresponds to a higher value calculated on another, thus reinforcing the notion that the considered pair of graphs may express a similar concept.

Once again, their Spearman correlation coefficients are calculated (Table 4.5; a visual representation is in Figure 4.6). This time, no value suggests a strong correlation between the sets of indices. All coefficients comparing the land-use graph with the other two are within the ± 0.2 range, which was quite expected. A moderate degree of correlation can be identified for three indices (degree, topological coefficient and clustering coefficient) between the species-set graph and the habitat graph; Spearman coefficients are within an upper bound of only $+0.42$, but overall results are consistent with the observation that these two graphs have more analogies to one another, than the land-use graph has with either.

4.4.3 Further Comparison of Average Hit Rates

Extending the study beyond the restricted case of non-binary vectors may bring more meaningful results, on the condition that vector representations reflect the intended features accurately, and subject to the availability of data. The considerations in this section are based on a second case study, intended to be more focused on land animals: sites are considered only if

TABLE 4.4: Excerpt of the table of normalized node betweenness centrality indices calculated on each similarity-based graph.

	Betweenness centrality index		
Site ID	Land-use	Habitats	Species-set
...			
ITB030034	0.01671	0.11557	0.04915
ITB030035	0.00014	0.04341	0.09402
ITB030036	0.01046	0	0.00641
...			

TABLE 4.5: Spearman correlation of complex network indices between pairs of similarity graphs.

Correlation of Index	Land-use with Species	Land-use with Habitats	Species with Habitats
Betweenness centrality	$+0.09309$	$+0.17446$	-0.01905
Closeness centrality	$+0.01001$	-0.02426	$+0.12268$
Degree	$+0.09172$	$+0.09961$	$+0.41257$
Topological coefficient	$+0.11214$	$+0.04271$	$+0.25071$
Clustering coefficient	-0.02396	-0.10644	$+0.28248$

they are located on the main island on Sardinia and land use data is available, resulting in graphs with 104 nodes. Moreover, only the species of interest listed in official Natura 2000 Directives shall be accounted for in computing similarity based on species sets.

Non-binary vectors can be built by assigning attribute values based on the number of occurrences of the attribute in the data base, or by summing the values of a meaningful feature associated to the attribute. In both cases, the resulting vectors can be used in the computation of cosine similarity. For habitat sets and land use codes, a summation of the surface area of each land patch can be used. Concerning species sets, a meaningful way to assign attribute values would be the summation of population sizes of a species in each site; unfortunately, the availability of this attribute is very sparse in the dataset: out of 4461 associations of 131 species with Sardinian sites, only 495 have unambiguous data on population size, and most of those are in the form of ranged estimates. Therefore, non-binary vectors were avoided for species sets, and only binary vectors and Jaccard coefficients are used. Vectors based on the number of occurrences were disregarded, because while multiple records can occur in the association of a species with a site, these are merely an indication of additional details in a

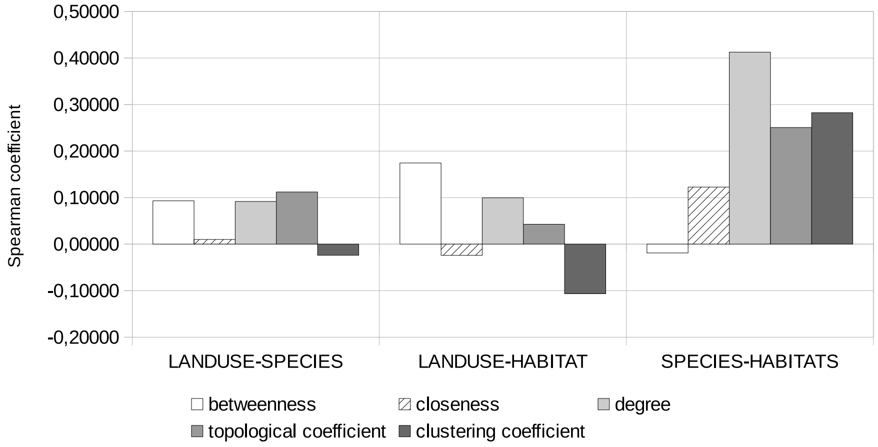


FIGURE 4.6: Histogram representation of Spearman correlation of complex network indices between pairs of similarity graphs.

report (e.g. two records may be added to report on a seasonal presence and the occurrence of reproduction in the site).

The raw-distance graph in this 104-site case study has 706 edges. The number of edges in different similarity-based graphs can be analyzed in the same way as the 117-site case; results for Jaccard coefficients of binary vectors are reported in Table 4.6, and results for cosine similarity of non-binary vectors are in Table 4.7, with a distinction of vectors built from the number of occurrences or the summation of surface areas.

Hit rates were calculated for this case study, with the same method described in the previous section, taking a similarity score threshold of 0.5 as reference. Table 4.8 reports on a few sample species, for similarity-based graphs based on Jaccard coefficients of pairs of binary vectors. The study was extended to cosine similarity of non-binary vectors, built on occurrences and surface areas from the records of Natura 2000 habitats and land use codes [31]. The indication of surface area was missing in 17 out of 1225 associations of habitats with Natura 2000 sites; in order to avoid having to exclude the affected sites, a surface area of 1 hectare was assumed. Cosine similarity is unaffected by the magnitude of vectors, so this does not introduce any error for sites with only one habitat associated to them; the error introduced for the other sites can be considered within acceptable limits.

Table 4.9 reports unweighted average hit rates and a normalized version of the same measure. Let n be the number of edges in the raw-distance graph (706 in this instance); the number of edges in a similarity-based graph, denoted by $|E|$, can be used to define a 'relative density' as the ratio $|E|/n$.

TABLE 4.6: Number of edges in similarity-based graphs of Natura 2000 sites in Sardinia, using Jaccard coefficients: 104-site case study

	Edges		
Threshold	**Land use-based**	**Habitat-based**	**Species-based**
0.0 (raw-distance)	706	706	706
0.1	658	533	554
0.2	575	436	404
0.3	453	339	276
0.4	323	248	179
0.5	210	174	122
0.6	93	103	55
0.7	33	53	18
0.8	4	35	16
0.9	0	13	12

The expected hit rate of a random graph would be higher for graphs with a larger number of edges. To avoid or reduce a bias against graphs of higher size, hit rates can be normalized to the number of edges in the similarity-based graph. For ease of comparison, for this study they are normalized to the relative density:

$$\text{Normalized Hit Rate} = \frac{R}{|E|/n} = \frac{R \cdot n}{|E|}. \tag{4.4}$$

Results show that the species-set graph has the best hit rate, as well as the highest normalized hit rate. This outcome could have been expected, as the same data is used to build this graph and every single-species graph. It is interesting, however, that habitat graphs consistently rank higher than land-use graphs according to their normalized hit rates, consistently with the observations in the previous section. Another interesting result is the better performance of area-based similarity of habitats compared to similarity of vectors based on the number of occurrences and, to a lesser degree, also to Jaccard coefficients of binary vectors.

4.4.4 Limits of Classical Indices

In proposing modifications to the current model of the network, it is important to evaluate local modifications without losing a global perspective on the matter. For example, Figure 4.7 shows a portion of a single-species network built for *Chalcides ocellatus* with a 20 Km distance threshold [25],

TABLE 4.7: Number of edges in similarity-based graphs of Natura 2000 sites in Sardinia, using cosine similarity: 104-site case study

	Edges			
	Occurrence-based		**Area-based**	
Threshold	Land use	Habitat	Land use	Habitat
0.0 (raw-distance)	706	706	706	706
0.1	648	587	507	371
0.2	577	508	412	326
0.3	471	445	354	268
0.4	361	374	289	230
0.5	277	278	242	185
0.6	201	164	184	155
0.7	126	86	132	123
0.8	79	44	97	97
0.9	35	14	56	55

TABLE 4.8: Excerpt of the table of hit rates for the 104-site case study, using Jaccard similarity.

		Land use-based		Habitat-based		Species-based	
Species code	Edges n	Hits l	Rate l/n	Hits h	Rate h/n	Hits s	Rate s/n
...							
6137	160	80	0.5	47	0.29375	27	0.16875
1367	15	9	0.6	8	0.53333	4	0.26667
1373	8	6	0.75	2	0.25	3	0.375
...							
Sum	7819	3056	0.39084	3214	0.41105	3287	0.42039
Average hit rate			0.41607		0.44616		0.47376

and the same portion of the network with a proposed modification. Alphanumeric codes correspond to administrative identifiers of sites within the Natura 2000 project. The proposed modification is a bridge from node `ITB023050` to node `ITB021156` through the node `ITB011102`, which is not originally part of the network. Based on the Jaccard coefficient of land use vectors, this node has a high similarity score with both `ITB023050` (above 0.6) and `ITB021156` (above 0.7).

This modification affects the indices of nodes in the network in various ways; changes in the betweenness centrality index of selected nodes are summarized in Table 4.10. It can be argued that the addition of alternative paths does not decrease the importance of nodes that made up previously

TABLE 4.9: Hit rates in similarity-based graphs. Since each graph has a different number of edges, a ratio of hit rates to relative density is also provided, where relative density is the ratio of edges in the similarity-based graph to edges in the raw-distance graph (n).

Graph	Avg. Hit Rate R	Edges $\|E\|$	Relative Density $\|E\|/n$	Norm. Hit Rate $(R \cdot n)/\|E\|$
Land Use (Jaccard)	0.41607	210	0.29745	1.39878
Habitats (Jaccard)	0.44616	174	0.24646	1.81027
Species Set (Jaccard)	0.47376	122	0.17280	2.74159
Land Use (occurrences)	0.55524	277	0.39235	1.41517
Habitats (occurrences)	0.58591	278	0.39377	1.48795
Land Use (areas)	0.40881	242	0.34278	1.19263
Habitats (areas)	0.50503	185	0.26204	1.92730

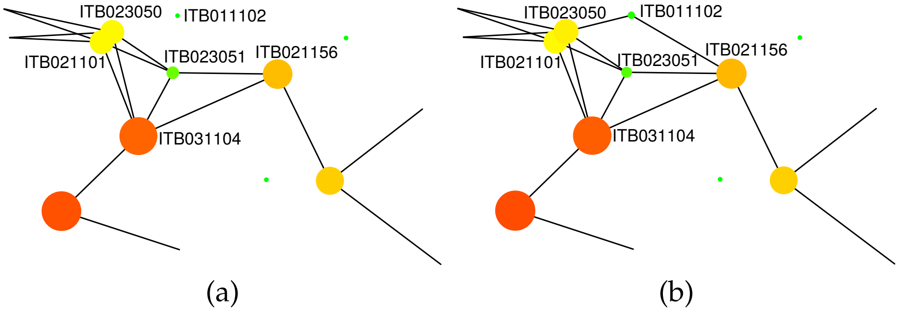
TABLE 4.10: Effect of the proposed modification on network indices

Node	Original Betweenness	Modified Betweenness
ITB011102	N/A	0.00522
ITB021101	0.05105	0.04694
ITB021156	0.19820	0.21408
ITB023050	0.05105	0.07467
ITB023051	0.01502	0.01162
ITB031104	0.42042	0.43409

existing connections, since the new indices are calculated taking into account the paths to and from the newly linked node. The only node with a decreased betweenness centrality index (ITB023051) is one that was not linked to the new node.

In determining whether a proposed modification brings improvements, a larger number of variables and more sophisticated indices should be considered. More refined evaluations can be obtained by considering the actual surface area that is being connected and made available to an endangered target species [42], and by going beyond the classic definition of shortest paths. In fact, while animals moving through a habitat matrix tend to perform better than random walkers [18], they can not be expected to move exclusively through shortest paths.

Working with a raster data representation of a habitat matrix, stochastic models are often taken in consideration in order to represent animal movement. The habitat matrix is often associated with a set of resistance values, and least-cost paths are considered in substitution of shortest paths; an amount of randomness can be added, for example, by considering paths through a random point of passage, or including a random function in the selection of movement direction [48]. An accurate model should also not assume complete knowledge on the part of animals, or even a prior decision on a final destination [46].



(a) (b)

FIGURE 4.7: Example of modification of a single-species network. Administrative codes of some relevant sites are reported. Sites with a higher betweenness centrality index are represented by larger nodes. (a) Single-species network for *Chalcides ocellatus* according to available data. (b) A proposed modification with alternative paths. This lowers the betweenness centrality index of the most central node.

4.5 Conceptual Graph Models

Having devised a method of analysis to be applied on topological graph models of ecological networks, the choice of target species, for which functional models should be built, may still be considered a problem worth addressing.

4.5.1 Shared-report Graphs

A different perspective on the Natura 2000 data base (Figure 3.3) can lead to a conceptual graph model meant to represent relations between species, from which to extract knowledge and establish criteria to select species of interest. In this kind of model, nodes are used to represent species, and edges correspond to interactions or affinities between species.

The simplest way to build this kind of model consists in linking pairs of species that have been reported to be found at the same time in at least a given number of Natura 2000 sites. A graph built with this rule can be referred to as a *shared-report graph*. To provide a frame of reference for their interpretation, basic examples of graphs resulting from minimal site configurations can be seen in Figure 4.8. An extreme case is an area consisting of a single site with n species, or a number of sites all hosting the whole set of n species; the shared-report graph for this situation is a complete graph of order n (Figure 4.8a shows the case for $n = 4$). From this observation, it follows that if the minimum number of sites is set to 1, each site in an area under analysis generates a subgraph, corresponding to the species found in that site, with the property of being complete, i.e. a clique. An example thereof is the ABC triangle in Figure 4.8b.

At the other end of the spectrum, a situation with m sites, each hosting only one species, results in a graph of isolated nodes. It can be argued that this is only possible in theory, as ecosystems require the coexistence of several species, by definition; however, this extreme case could occur in case studies, if data were only available for a very limited set of species present in an area (down to only one). The occurrence of this result could lead researchers to question the quality of data or the choice of species of interest.

Usually, certain subsets of species will occur multiple times. An alternative approach is to treat the resulting shared-report graph as a weighted graph, by assigning weights according to the number of occurrence of a pair, to represent the strength of a link (see Figure 4.8c).

4.5.2 Case Study

To apply this method to the case study, shared-report graphs are built from data on the 93 sites designated as SCI in Sardinia [30]. The official list of species of interest in the Natura 2000 was considered (Birds Directive and Annex II of the Habitats Directive), with the addition of species from the "other species" set, for which a species code was consistently used. Several instances are built by adjusting the minimum number of sites required to link pairs of species, and results are reported in Table 4.11. Two additional instances are built for comparison: one is built including the entire set of species in Sardinia, and the other is built using data for the 2314 sites designated as a SCI or a SAC in Italy, and the same criteria to select a subset of species. In every experiment, the resulting graph was made up of one large connected component and a number of isolated nodes. The table reports the number of isolated nodes and three global complex network indices: density, diameter and characteristic path length.

It is notable that the diameter of the connected component remains low, even as the number of edges is reduced by applying increasing threshold values. The isolated nodes correspond to a number of species that is found

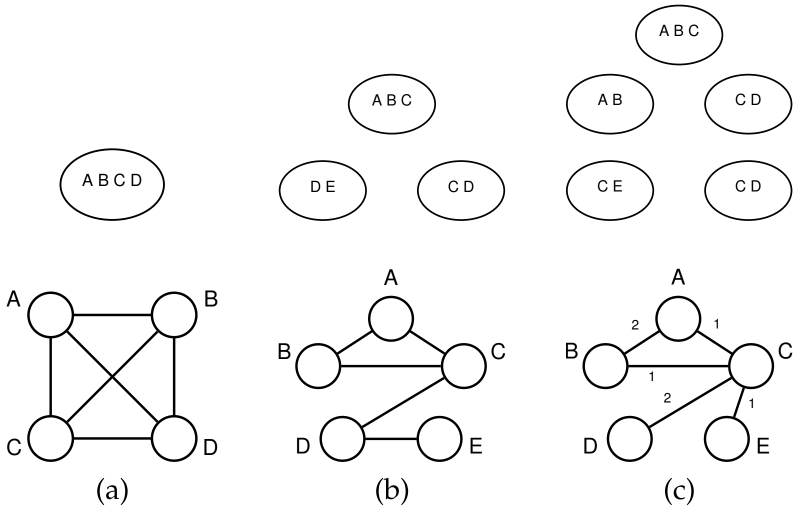


FIGURE 4.8: In the top row, sample configurations of sites, represented as sets of species. Below, the corresponding shared-report graphs. (a) A number of species in a single site generates a complete graph if the minimum number of shared reports is 1. (b) An example with 5 sites. (c) Weighted graph resulting from the multiple occurrence of certain sets of species.

in a very limited number of sites, while it can be interpreted that a 'main' set of species is commonly found together in many sites, which can be a sign of a good degree of homogeneity in common factors among habitats, such as climate. Since diameter is low even in the graph built for the entire Italian territory, it could be inferred that a large set of species may be able to adapt to several different Italian regions.

As far as the approach of studying network properties of the weighted shared-report graph is concerned, Table 4.12 reports the top 10 species by betweenness centrality index, normalized to a scale of 0 to 1, calculated for a weighted graph instance, where the reciprocal of the number of sites was used as a distance. A higher value of this index for a node suggests that the corresponding species is included in a larger proportion of shortest paths between different species. Given the low unweighted diameter, this means that the species often shares reports with both elements in a pair of species, which do not share a report with one another, or do so quite rarely. This can be a sign of a higher degree of adaptability, or a greater tendency to migrate, depending on known features of the species.

TABLE 4.11: Results of complex network analysis on shared-report graphs. Only areas designated as SCI or SAC are considered. The area under analysis is Sardinia, and only for a subset of species with a consistent species code, except in rows marked as follows: (AS) study extended to all species for which data is available, (IT) study extended to sites in Italy.

Min. sites	**Nodes**	**Edges**	**Isolated nodes**	**Density**	**Diam.**	**Charact. Path Length**
1	351	44 967	1	0.732	2	1.263
2	351	34 482	40	0.561	2	1.285
3	351	28 902	61	0.471	2	1.310
4	351	24 711	79	0.402	2	1.330
5	351	21 779	87	0.355	2	1.373
6	351	19 330	95	0.315	3	1.408
7	351	17 020	102	0.277	3	1.450
8	351	15 114	117	0.246	3	1.447
9	351	13 534	128	0.220	3	1.463
10	351	12 106	140	0.197	4	1.463
1 (AS)	860	129 953	23	0.352	3	1.629
1 (IT)	853	169 338	8	0.466	3	1.542

TABLE 4.12: Ranking of species reported in SCIs in Sardinia, by normalized betweenness centrality, calculated on the weighted graph.

Name	**Type**	**Sites**	**Betweenness**
Hyla sarda	Amphibian	71	0.12483
Turdus merula	Bird	71	0.07566
Upupa epops	Bird	60	0.05980
Falco tinnunculus	Bird	64	0.04169
Larus cachinnans	Bird	59	0.03935
Sylvia melanocephala	Bird	70	0.03734
Bufo viridis	Amphibian	61	0.03206
Carduelis chloris	Bird	66	0.02921
Carduelis carduelis	Bird	67	0.02921
Circus aeruginosus	Bird	52	0.02920

Chapter 5

Analysis of a Power Grid

5.1 The Transition to Smart Grids

A power grid is the network of substations, buses and power lines that is responsible for the distribution of electrical power over long distances. Among distribution networks, power grids stand out as a prominent field of application for complex network analysis, due to their inherent network structure, properties such as a quick response to changes, and the importance that power distribution has gained; it comes as no surprise that a large amount of literature exists on the topic of representing power grids using graph models. As icing on the cake, the interest in transitions from traditional power grids to 'smart grids' makes the field particularly appealing.

A traditional power distribution network is essentially unidirectional, in the sense that generation happens exclusively at power plants and specific sites equipped with generators, and power is transmitted and distributed to end users, which simply act as consumer nodes. A modern power grid is made up of two codependent systems: the Energy Management System (EMS) and the Distribution Management System (DMS). The former includes power generators and all the facilities involved with large-scale transmission to substations, while the DMS deals with the distribution of power from substations to end users. Substations are complex facilities, dealing with functions such as voltage transformation and separation of circuits at the occurrence of overloading or to interrupt short circuits.

Multiple definitions of a smart grid exist in literature, reflecting several different points of view on the topic [47]. The smart grid paradigm aims at an improvement of power grids, specifically increasing their reliability, flexibility and efficiency [23, 24]. Some fundamental capabilities of a smart grid are the ability to self-heal and self-regulate, and the possibility for end users to contribute to the operation of the grid.

Concerning the ability to self-heal, the grid ought to be able to detect faults and initiate actions to restore grid components automatically, reducing the duration of blackouts that may happen. Self-regulation refers to being able to adjust grid behavior in response to changes in supply and demand that happen under normal conditions of operation, particularly concerning the intermittent availability of certain renewable sources, such as wind turbines, and the contribution of end users. These may provide energy back to the grid, should they have a surplus from local generation (e.g. solar panels), which is to be treated as an additional intermittent supply. Another form of contribution by end users is the deployment of smart appliances, which may be able to adjust their demand under certain conditions: for instance, low-priority appliances such as water heaters may be configured to switch off temporarily if a signal is received from control centers, concerning the detection of a high risk of blackout; and resume operation at a later time, or when another signal communicates that the grid has been able to adjust to satisfy power demands without risk.

To make all of this possible, the electricity infrastructure should be complemented with information and communication capabilities [38], allowing power as well as information and control signals to flow in every direction over the network, and not exclusively from generation sites to consumer nodes. The information network should be designed to detect the state of every node, and to deliver this information to control centers, in order to alter the behavior of agents in ways that improve the efficiency and reliability of the network as a whole. Privacy and security concerns are to be considered since the planning stages of the implementation of a smart grid, as all of these goals should be achieved while respecting the privacy of customers, preserving the confidentiality of data, and preventing abuse by malicious users.

A full implementation of a smart grid requires changes at every level, from power plants to single home and office appliances. In order to do so, a smart grid ought to be designed and built from scratch, but this turns out to be impractical; thus, smart grid features are to be gradually introduced by incremental changes in the current power grids [39]. This requires a long-term strategy, as well as the capability of identifying short-term decisions with a higher degree of priority. Understanding the topology of the current grid and the properties of power flows is the first step in this direction.

5.2 Graph Models of a Power Grid

In building a graph model, both the topology of a power grid and the electrical properties that govern power flows within it are to be considered and represented; together with analysis goals, these properties determine fundamental choices in the building of graph models.

5.2.1 Approaches to Building a Graph Model

A power grid is made up of buses, each connected to generators, loads and transmission lines; a small-scale, fine-granularity models can reflect this organization, representing each bus with a node, and transmission lines with edges. Since transmission lines are inherently bidirectional, all edges can be undirected. Loads and generators can be thought of as being part of the node entity, which should be assigned a generation-load balance as a numerical attribute. Edges can also be assigned attributes to reflect their capacity; if multiple edges are found between the same pair of nodes, they can be collapsed into one if a simple graph is preferred as a model.

Historically, unweighted models have been considered in a majority of studies until 2006 [45]; their inability to predict the effect of changes on the network in a consistent manner eventually determined a decline in their popularity, and several approaches at weighted models were proposed, aimed at taking into account the laws governing the transmission and distribution of power. Among proposed ways to assign weights are line reactance [19], line impedance [11], and other concepts of "electrical distance" based on resistance, discussed in [13]. It is notable that, in assigning edge weights, electrical properties are deemed more important than strictly topological properties (e.g. geographical distance), as the latter are shown to bring little to no contribution to the prediction of grid behavior, including blackouts and cascading failures. Since electrical properties are not generally static, dynamic network models may have be favored depending on the scope of analysis.

Building a model based on dynamic networks, however, does not necessarily mean that edge weights (and only edge weights) are to be adjusted over time. Changes in production and consumption, including the intermittent availability of renewable sources and power contributions on the part of end users, can be represented by modeling nodes as oscillators, making it possible to determine whether a power grid can achieve stable operation over a period of time, or the way it responds to the introduction of decentralized production [52].

5.2.2 Vulnerability Assessment

As the transition from a traditional power grid to a smart grid is planned, it is common to perform a vulnerability assessment, as way to determine which portion of the network requires improvements most urgently. In a vulnerability assessment, one or more typical load scenarios are chosen (e.g. to reflect average loads, or power demands at peak time, etc.), and the behavior of the network in these scenarios is simulated, first at a full efficiency condition, and then after the removal of a number of nodes and edges, to represent a case where these elements have failed, thus assessing the degree to which the network is able to withstand their removal. This can be done by comparing indices of health of the network, or by measuring the areas where blackouts or brownouts may occur according to the simulation.

A vulnerability assessment can be performed to predict the effects of random failures or targeted attacks, by changing the criteria to mark failed elements – either chosen at random, or depending on some network index, or a combination thereof.

Strategies differ according to the attacker model, and assumptions on security measures at different nodes. If an attacker is assumed to have limited knowledge of network topology, it is sensible to adopt a simple strategy, consisting in the removal of nodes with the highest degree (hubs); at the same time, if it can be taken for granted that stricter security policies are in effect in hubs, it can be considered more interesting to study the effect of attacks on low-degree nodes. Frameworks have been proposed to simulate attacks with various degrees of randomness, with a parameter to express an amount of bias toward larger-degree or smaller-degree nodes [34].

Other refined strategies, which can be considered if the attacker has a complete knowledge of network topology, involve the selection of nodes with the highest values of complex network indices, such as their betweenness centrality index. Studies have shown that no network index stands out as a single final criterion of node vulnerability, as different strategies for node removal result in different effects, such as an increase in characteristic path length, a loss of connectivity, and a different size of areas affected by blackouts [40].

Vulnerability assessments can also be performed on sets of interdependent networks. The scope of this kind of study is not limited to the case of interconnected power grids with a sufficient degree of self-sufficiency; infrastructure networks of different kinds may have the property that a failure in one network, including a local failure, may have an impact on the other one. Power failures, especially prolonged ones, can indeed hamper the operation of several different types of infrastructure, including but not limited to communication and transportation. The opposite is true, as well, as power plants rely on the availability of fuel, which has to be collected from remote locations, and the gradual introduction of smart grid features is going to increase the dependency of power grids from the continued operation of communication networks. Interdependency relations across infrastructure networks of different kinds have been an object of study for over a decade [50], and models for 'network of networks' have been proposed to provide a background to perform robustness analysis of interdependent networks [35].

5.3 Case Study: a Regional Power Grid

At the time of this study, the Sardinian power grid is still mostly based on the traditional paradigm. Wind turbines are installed in various parts of the island, and smart meters have been introduced and installed for most customers, with support for a quite limited set of features. This study builds up on a previous analysis, performed on a large-scale model of this grid with coarse granularity, where each node represents an area, for which data on average load or power generation was available [23, 22]. The model was refined in order to give more insight into the importance of power generation from renewable sources, and in an attempt to consider more realistic sets of power flows, as opposed to a theoretical optimal operation of the grid.

In this model, three node classes are considered: power plant nodes, urban area nodes and industrial area nodes. Each power plant node may represent a thermoelectric or a hydroelectric power plant, or a site dedicated to power generation through a set of wind turbines. Urban area nodes represent a city, a conglomerate or a district, including one or more substations. An urban area node may be adjacent to one or more power plant nodes, or it may have no adjacent power plant node; in the latter case, power has to be received from remote power plants after traversing several urban areas. Lastly, industrial area nodes are used to model sink nodes that aggregate the loads of activities in an industrial area. These nodes are always adjacent to one or more urban area nodes, and receive power through them.

The network is modeled as a directed graph $G = (V, A)$, in which power plant nodes have exclusively outbound arcs, and nodes representing industrial areas have only inbound arcs; the connections between urban areas are modeled as opposite pairs of arcs. The complete model of the network has a total of 133 nodes and 269 arcs (Figure 5.1) and shall be referred to as the model of the network in its *healthy state*.

5.3.1 Power Grid Operation as an Optimization Problem

Since a real power flow simulator can not be used in absence of a model at a finer granularity, a multiple-source, multiple-sink minimum cost flow problem is setup on the model of the network, and its solution shall represent an optimal set of power flows, to which operation should attempt to converge as much as possible.

In order to provide at least an approximate representation of electrical properties, special constraints shall be included in the definition of this problem, which will be applied to the model representing the healthy state of the network and to models of network states resulting from one or more failures, thus performing a vulnerability assessment on the original network.

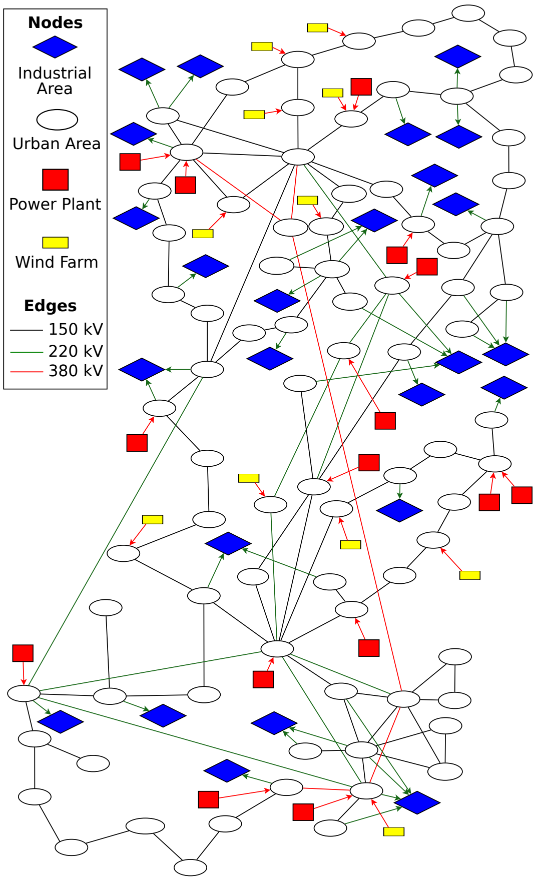


FIGURE 5.1: A visualization of the graph model of the Sardinian power grid. Urban areas are represented with an ellipse, industrial areas with a rhombus, power plants with a tall rectangle, and wind farms with a small rectangles. Pairs of arcs with opposite directions are visualized as an undirected edge.

In a minimum cost flow problem, each node $v \in V$ is labeled with a parameter $b(v) \in \mathbb{R}$, which represents its aggregate supply or demand of a commodity; generally, positive numbers represent supply and negative numbers represent demand. Therefore, nodes with $b(v) < 0$ are sink nodes, while nodes with $b(v) > 0$ are source nodes. If $b(v) = 0$, demand and supply are balanced and v acts as a transit node. Each arc $(u, v) \in A$ is associated with the flow of a commodity, and has a maximum capacity $c(u, v)$ and optionally a lower bound $l(u, v)$ for its flow. Each arc is associated with a cost denoted by $a(u, v)$, i.e. a unitary cost for the flow on that arc. The decision variables are arc flows $f(u, v)$, and the objective function to minimize is:

$$z = \sum_{(u,v)\in A} a(u, v) \cdot f(u, v). \tag{5.1}$$

In this instance of the problem, lower bounds are not in use, i.e. $l(u, v)$ is set to zero for every arc. The capacity is made to correspond to the maximum amount of power that can be sent through the power lines represented by the arc, calculated from data on the voltage and amperage of existing power lines. The value of $b(v)$ for each node was determined according to historical and statistical data provided by the Italian energy distribution company Terna. Source nodes (power plants and wind farms) were assigned values according to their maximum output in a time unit, while the value for sink nodes (urban and industrial areas) was made to correspond to an estimated average consumption in the same time unit, with a negative sign, to match the conventions for the definition of the problem. Each node determines a constraint in the optimization problem, due to its balance value:

$$\forall v \in V, \sum_{(v,u)\in A} f(v, u) - \sum_{(u,v)\in A} f(u, v) = b(v). \tag{5.2}$$

This class of problems is solved by linear optimizers, which apply algorithms working under the assumption that there is a balance between supply and demand, i.e.

$$\sum_{v\in V} b(v) = 0. \tag{5.3}$$

Thus, to solve problems for which this assumption does not hold, an artificial node t is added and linked to all the other nodes, with a value for $b(t)$ such that (5.3) is satisfied; all costs for artificial arcs incident to t ought to be orders of magnitude higher than costs for real arcs, so that flows from source nodes to the artificial node can be treated as a sign that there is a surplus of the commodity on the network, while the presence of flows from the artificial node to real nodes shows that no feasible solutions for the problem.

Two issues with the transmission of electrical power are not represented in a typical commodity flow problem: overloading and power loss.

In the solution of a typical minimum cost flow problem, it is a common occurrence that the most profitable arcs are used at full capacity, while the flow on several arcs is set to zero. This does not reflect the goals of operation of a power grid: the use of a power line at full capacity is to be avoided whenever possible, because it is closer to overloading and causing malfunctions. The satisfaction of demand from multiple lines used at a fraction of their capacity is generally preferred, and may be necessary depending on network topology and physical laws governing power flows.

To represent this, additional constraints are added, approximately mimicking the laws of physics and electrical properties that regulate power flows in sets of substations [44]. Considering all simple cycles of three and more nodes found on the graph, disregarding edge orientation, a 'cycle constraint' is formulated for each cycle, as follows:

$$\sum_{(u,v)\in C_i} d(u,v) \cdot f(u,v) = 0, \tag{5.4}$$

where C_i is the set of arcs which connect nodes making up a cycle (identified by an index i) and, once an orientation on the cycle is chosen (e.g. counter-clockwise), $d(u,v)$ is set to $+1$ for arcs with that orientation, or -1 for arcs with the opposite orientation.

Concerning the issue of power losses, which occur in power transmission due to resistive heating in the wires (Joule heating), the effect of this electrical property is not representable at transmission time in linear optimization problems; in this study, an estimation of power loss was added to the figure for demand at consumption nodes, in order to compensate for this shortcoming.

The amount of power lost in a time unit is given by:

$$P(i,j)_{loss} = I^2 R, \tag{5.5}$$

where I is the current intensity and R is the electrical resistance. Since

$$I = \frac{P(i,j)_{sent}}{V}, \;\; R = \frac{\rho L}{A}, \tag{5.6}$$

where V is voltage, ρ is the resistivity of the material, L is line length and A is the cross sectional area of the cable, an explicit formula for power loss is the following:

$$P(i,j)_{loss} = \left(\frac{P(i,j)_{sent}}{V}\right)^2 \frac{\rho L}{A}. \tag{5.7}$$

Since the length of a power line is bound by geographical constraints, the other variables have to be adjusted in order to minimize power loss. A decrease in ρ is sought by giving preference to materials with a low resistivity. Increasing voltage strongly reduces power losses, but because of the costs associated with this practice, it is common to employ higher levels of voltage exclusively for backbones, as the advantages may not cover costs when shorter distances are involved.

Since the goal of this problem is to determine a set of power flows that operation should attempt to replicate, a measure of power loss is a sensible choice for arc costs, so that the solution acts as a suggestion of a way to limit the phenomenon. However, it is not straightforward to define the problem this way, as $P(i, j)_{loss}$ is calculated from $P(i, j)_{sent}$, and the latter is supposed to be estimated from the optimization process itself.

To break this circular dependency, a first estimation of power loss is calculated based on an assumption that each inbound arc provides an equal power to each node. This estimation is used as a seed in an iterative process (Algorithm 1) [26], in which estimated power flows are obtained from consecutive runs of the optimizer, and used to compute power loss on each arc, using Formula (5.7). This process converges, on the condition that the direction used in pairs of opposite arcs is locked on the first run of the optimizer (line 13 in the algorithm); otherwise, the optimizer would alternate directions of opposite pairs of arcs in consecutive runs, since only one arc in the pair is used at a time, and the cost for the unused arc would drop to zero for the next iteration in the process.

Algorithm 1 Iterated power loss calculation

```
     $i \leftarrow 0$;
     $Flows[0] \leftarrow Seed$;                          ▷ Initial set of assumed flows
     $Costs[0] \leftarrow$ calculatePowerLoss($Seed$);
     repeat
 5:      if $i ==$ MAX_ITERATIONS then
             alertUser                        ▷ Iterative process has not converged
             break
         end if
         $i++$;
10:      if $i == 1$ then
             $Flows[i] \leftarrow$ runOptimizer($Costs[i-1]$);
         else                 ▷ Preserve arc flow directions from previous iteration
             $Costs_{artificial} \leftarrow$ lockDirections($Costs[i-1]$, $Flows[i-1]$);
             $Flows[i] \leftarrow$ runOptimizer($Costs_{artificial}$);
15:      end if
         $Costs[i] \leftarrow$ calculatePowerLoss($Flows[i]$);
     until ($Flows[i] == Flows[i-1]$)
```

5.3.2 Normal Operation of the Power Grid

At the time of data collection, the total supply capacity for the power grid under analysis is higher than the sum of average demand at its consumption nodes. This suggests that during normal operation, the grid should be self-sufficient, and external power sources should be needed only at peak times or under special circumstances. However, this may be true on condition that the arc set has no bottleneck preventing power from reaching load nodes. This is verified by the fact that the optimizer provides a feasible solution for the healthy state of the network.

It must be observed that almost one-fifth of the production capacity of the network under examination is due to wind farms; since their actual output is subject to variations in winds, the same conditions should be verified in a model where wind farms are removed, or the figure for their output is reduced. In fact, the maximum output within the grid is still higher than the total average demand, even when the output of every wind farm is subtracted, but the optimizer does not provide a feasible solution if every wind farm is removed, with unsatisfied demand in one consumption node. This is a sign that wind farm output is crucial to meet local demand in some areas, since some bottleneck is preventing distant power plants from providing the required power [27].

The results of subsequent experiments, in which wind farm output is only partly removed, are reported in Table 5.1. In these, either a single wind farm is removed from the network, or its maximum output is cut by one half. The first column of the table reports an administrative identification of the wind farm by an ID number, the second column specifies whether its maximum output capacity was halved or cut off to zero, and the third column reports the number of nodes with unsatisfied demand resulting from the experiment. An additional experiment was performed with five wind farms being removed at the same time, a case that was considered of interest due to the close location of these five wind farms, the output of which is likely to drop at the same time. In this experiment, the optimizer was able to find a feasible solution.

From these experiment, it seems clear that for the most part, the power grid under analysis can withstand problems deriving from a drop in wind-generated power, with the sole exception of two wind farms that appear to be crucial for the self-sufficiency of specific areas.

5.4 Definition of Collateral Damage

To assess damage on the network deriving from the failure of one or mode nodes, it is not sufficient to simply compare the value of the objective function in the instances of the optimization problem, as doing so does not take

TABLE 5.1: Results of experiments involving removal of wind farms. The rightmost column gives the number of nodes with unsatisfied demand (deficit), following the reduction of maximum output from a specified wind farm to one half of its theoretical maximum ('Half' reduction), or removal of a wind farm from the network ('Full' reduction).

Wind farm ID	Reduction	Nodes with deficit
none	N/A	0
116	Half	0
116	Full	1
134	Half	0
134	Full	1
117	Full	0
118	Full	0
119	Full	0
121	Full	0
126	Full	0
127	Full	0
128	Full	0
130	Full	0
131	Full	0

into account the reduced demand and the effects of a change in the topology of the network, which invalidate some of the cycle constraints defined by (5.4). The problem of normalizing the objective function to a formulation that is comparable across versions of the network is discussed in this section.

Let $G = (V, A)$ be the healthy state model of the network. Let $w \in V$ be a node to be marked as failed. Detaching w from G, together with the arcs incident to it, may create a disconnected component $G_d(w) = (V_d, A_d)$ in the network. If this is the case, and the component $G_d(w)$ has no power plant nodes in itself, then the whole component is to be removed from the network when defining an instance to represent the failure of w; otherwise, only node w is to be removed.

The subset of V determined in this way is denoted as $F(w) = \{w\} \cup V_d$, where V_d may be an empty set. Let $D(w) \subset A$ be the set of arcs that are incident to at least one node in $F(w)$. For brevity, $F(w)$ and $D(w)$ may be denoted simply as F, D when it is clear from the context which set is intended.

Let $G'(w) = (V', A')$ be the graph instance representing the failure of G, i.e. a copy of G with the removal of nodes in F and arcs in D:

$$V' = V \setminus F, \tag{5.8}$$

$$A' = A \setminus D. \tag{5.9}$$

An instance of the optimization problem described in Section 5.3.1 has to be created for $G'(w)$. This is defined by assigning $a'(u,v)$, $b'(u)$, $c'(u,v)$ equal values to $a(u,v)$, $b(u)$, $c(u,v)$ defined for the problem on G, exclusively for nodes in V' and arcs in A'.

Another instance of the optimization problem is to be created on $G''(w) = (V, A)$, i.e. a copy of G, with a different set of node balances and arc capacities that depend on features of $G'(w)$. Values $a''(u,v)$, $c''(u,v)$ shall be equal to $a(u,v)$, $c(u,v)$ from the problem on G, while $b''(u)$ is assigned like this:

$$\forall u \in V,\, b''(u) = \begin{cases} 0 & \text{if } u \in F, \\ b(u) & \text{otherwise.} \end{cases} \tag{5.10}$$

This essentially removes the demand of failed nodes from the problem, converting them to transit nodes. Additionally, every cycle constraint that was removed from the problem for G', due to the detachment of nodes, is also not considered in the problem for G''.

Let z be the value of the objective function in the optimal solution of the minimum cost flow problem on G:

$$z = \sum_{(u,w) \in A} a(u,w) \cdot f(u,w), \tag{5.11}$$

and likewise, let $z'(w)$, $z''(w)$ be the optimal values for the problems on $G'(w)$, $G''(w)$ respectively. Since these problems have different total demands, these values are not directly comparable.

Considering the problem on G, given the total demand of a network:

$$\hat{b} = \sum_{u \in V, b(u) < 0} |b(u)|\,, \tag{5.12}$$

and the sum of artificial costs in the optimal solution:

$$z^* = \sum_{(u,t) \in A} a(u,t) \cdot f(u,t) + \sum_{(t,u) \in A} a(t,u) \cdot f(t,u), \tag{5.13}$$

where t is the artificial sink node, the total cost with the artificial costs removed can be considered and normalized to the demand on the network like this:

$$y = \frac{z - z^*}{\hat{b}}. \tag{5.14}$$

Repeat the process for the problems on G' and G'', obtaining $\hat{b}'(w)$, $y'(w)$, $\hat{b}''(w)$, $y''(w)$, while checking whether any $f(t, u), u \in V$ is above 0, i.e. the demand of any node is not satisfied, fully or in part. The number of nodes with a deficit shall be referred to as $d(w)$.

$G'(w)$ represents the network G with a failure on node w, and $G''(w)$ represents an ideal situation where G has not failed, but its demand has been removed (together with that of all nodes in $F(w)$, if applicable), and the cycle constraints defined for w have been ignored; essentially, there have been the same modifications to convert the optimization problem on G to the two instances on G' and G'', and consequently, $y'(w)$ and $y''(w)$ are comparable. Then, a measure of collateral damage from the failure of node w is given by their difference:

$$Collateral(w) = y'(w) - y''(w). \tag{5.15}$$

This figure is meant to represent how much the cost of providing service to connected areas is affected by the unavailability of paths. It does not capture the possible existence of deficit nodes in the graph, as the cost of flows on artificial arcs has been removed; for this reason, it has to be considered together with $d(w)$ as defined above.

The same process can be initiated from a set of failed nodes, i.e. choosing a $F \subset V$ such that no disconnected components without power plant nodes are introduced, with the corresponding definitions of $G'(F)$, $G''(F)$, etc. It follows from the definitions that $Collateral(\emptyset) = 0$.

5.5 Network Analysis and Collateral

The measure of collateral damage introduced in Section 5.4 can be used to identify portions of the network where it is most pressing to bring improvements, and which would benefit the most from investments on renewable energy or an increase of redundancy of network connections. However, the calculation of this measure requires multiple runs of an optimization problem on the graph model, which can take a long time for very large networks. Moreover, this limits the scope of application of this measure to offline analysis and planning.

In several studies, correlations between measures of damage and network indices are sought, with the purpose of legitimating the use of network index values as an estimation of vulnerability.

5.5.1 Graph Models for Complex Network Analysis

The graph model in use for the optimization problem can provide the basis for multiple instances to be analyzed with complex network analysis techniques.

The first option is to analyze the same graphs, on which an optimization problem is solved. In these, arc weights represent costs, which can be considered analogous to distances, and the definition of indices for weighted networks can be applied directly. A different approach is to take the solutions, i.e. the set of flows on each arc, and define arc weights accordingly: in this case, weights are meant to represent a strength of the link. In order to apply the same definitions, and most notably to compute betweenness centrality indices, the reciprocal of the estimated power flow is used to represent a 'distance' associated to each arc. Since some arcs with no associated power flow are guaranteed to exist, these arcs are artificially assigned a minimal power flow, i.e. an artificial distance orders of magnitude greater than all others, only in the graph instance on which complex network analysis is applied, to ensure that the selection of these arcs in shortest paths is avoided whenever possible.

Three graph models are considered for analysis, for each instance of the optimization problem, with distances based on:

- seed costs ($Costs[0]$ in Algorithm 1): from these, the *seed cost-based betweenness centrality* (SC-BC) is computed;
- final costs (final values of $Costs[i]$ in Algorithm 1): from these, the *converged cost-based betweenness centrality* (CC-BC) is computed;
- power flows (final values of $Flows[i]$ in Algorithm 1): from these, the *flow-based betweenness centrality* (F-BC) is computed.

In the $Costs[i]$ set, in each pair of opposite arcs, one arc has has a real cost, while the other is assigned an artificial cost. This is discarded in analysis, and the real cost assigned to one of the arcs is used for both arcs in the pair.

The betweenness centrality index calculated for each of the three models built for the healthy state of the network is visualized in Figure 5.2. A few 'hub' nodes have a high value for this index in all of the three instances, but considerable differences are observed for the rest of the nodes.

5.5.2 Correlation with Betweenness Centrality

The calculation of collateral damage described in Section 5.4 is performed in multiple experiments, each time marking one node v, representing an urban area, as a failed node; when $d(v) > 0$, additional experiments are

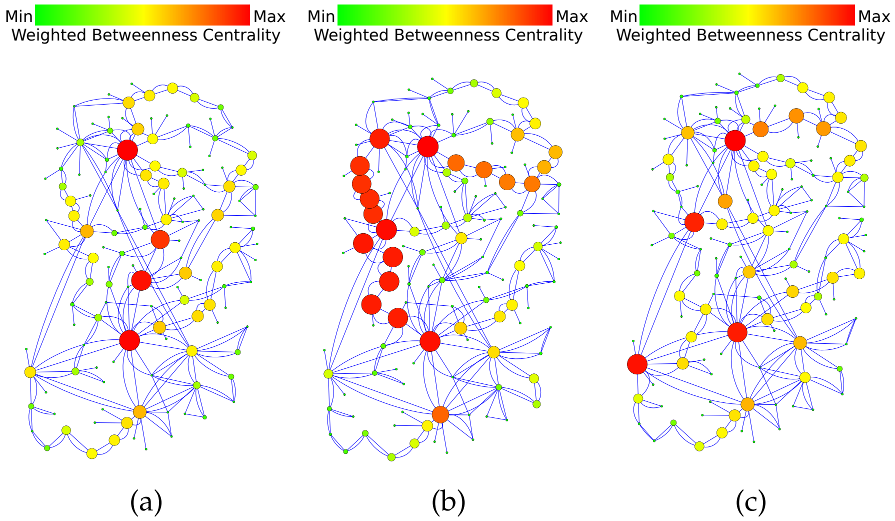

Min
Max
Weighted Betweenness Centrality
Min
Max
Weighted Betweenness Centrality
Min
Max
Weighted Betweenness Centrality
(a)
(b)
(c)

FIGURE 5.2: Visualized analysis results for betweenness centrality. Larger node size corresponds to a higher value of the betweenness centrality index. (a) Seed cost-based. (b) Converged cost-based. (c) Flow-based.

run, building a set of failed nodes (F'), made up of the nodes in $F(v)$ and those that have deficits in the corresponding experiment, and iterating until $d(F') = 0$ if necessary. The final value of $Collateral(F')$ and the number of removed nodes is recorded for each experiment and associated to its starting node. A sample of notable results is in Table 5.2. Among these experiments, some involve a hub as a starting node (e.g. Codrongianos with ID 70 and Villasor with ID 40). These nodes have high values for the betweenness centrality index, and among the highest values for the collateral damages associated with their failure. With few exceptions, nodes with a very small betweenness centrality index have a low value for collateral damage as well.

Table 5.3 reports the Pearson correlation coefficient is calculated between the set of collateral damage values and each of the three sets of betweenness centrality indices of the corresponding starting nodes. Recall that Pearson correlation coefficients are bound within -1 to 1, where values close to -1 represent an inverse linear correlation, and values close to 1 denote a direct linear correlation. An absolute value over 0.7 is considered a sign of a strong linear correlation.

No strong correlation is detected in this study, but moderate correlations are found with the betweenness centrality indices calculated on the model based on seed costs (about 0.65) and network flows (about 0.55). However,

TABLE 5.2: Results of computation of $Collateral(F)$ for some F, for which $d(F) = 0$. The first column represents F. Where multiple nodes are included, the first node in the list is the one for which betweenness centralities are reported, e.g. where $F = \{21, 96\}$ the betweenness centrality of the node with ID 21 is reported.

Removed node IDs	y'	y''	Collateral	F-BC	CC-BC	SC-BC
none	40378.58499	40378.58499	0	N/A	N/A	N/A
6	40480.58231	40480.58231	0	0.00000	0.00000	0.00000
10	40270.57417	40270.57417	0	0.00619	0.00231	0.00000
12	40401.62739	40401.62739	0	0.00000	0.00324	0.00000
...	...	...	...	...	...	...
38	40378.58499	40378.58499	$< 10^{-5}$	0.00000	0.00000	0.00000
63	40352.82864	40352.82864	$< 10^{-5}$	0.01810	0.29389	0.01827
60	40393.82264	40393.73924	0.08340	0.00613	0.02429	0.01024
{21, 96}	40547.61212	40547.16976	0.44236	0.00630	0.29858	0.03620
20	40379.23947	40378.60582	0.63364	0.00000	0.02429	0.09195
78	40416.19985	40415.01453	1.18532	0.04551	0.14394	0.03609
...	...	...	...	...	...	...
80	40461.69260	40381.78177	79.91082	0.04170	0.05748	0.01220
83	40327.80678	40246.23992	81.56686	0.04048	0.00347	0.00000
81	40474.87219	40389.95289	84.91929	0.03007	0.04638	0.02417
65	40432.29062	40341.52756	90.76306	0.00162	0.20657	0.00000
76	40439.03108	40348.16189	90.86919	0.01822	0.03533	0.03591
37	40520.49025	40429.18368	91.30656	0.00069	0.00451	0.00000
53	40341.83277	40244.87584	96.95693	0.07368	0.02232	0.02429
...	...	...	...	...	...	...
{5, 6}	40844.31316	40348.08541	496.22775	0.11051	0.23282	0.13607
64	40659.30957	40135.69493	523.61464	0.02718	0.02643	0.05748
1	40627.79022	40093.67589	534.11433	0.00000	0.00000	0.00619
70	40694.38290	39773.10907	921.27383	0.30893	0.04632	0.40802
{31, 106}	41763.85277	40840.63726	923.21551	0.09507	0.31419	0.01833
40	41211.74186	39633.23393	1578.50793	0.27938	0.33085	0.40204

the former represents only a theoretical construct, while the latter is conceptually analogous on power flows in a real operated network, and converged costs are analogous to power losses. Thus, a moderate degree of correlation with the flow-based indices suggests that collateral damage may be correlated to power flows in a real network, and prompts for further research in this direction.

TABLE 5.3: Pearson correlation coefficients between $Collateral(F')$, for which $d(F') = 0$, built on each starting node representing an urban area, and betweenness centrality indices.

Betweenness centrality index	Pearson correlation with Collateral
Seed Cost-based	0.650460064
Converged Cost-based	0.231452099
Flow-based	0.548274258

Chapter 6

Conclusions

The continued operation of a number of services, such as power generation and distribution, communication, and transportation of goods, has become an essential aspect of society. Whether for their inherent nature or because technological advancements have favored large-scale solutions over local ones, these services are impossible to provide without the active maintenance of complex systems in a wide area. With the realization that regulations concerning land use are necessary to avoid the destruction of habitats at the hand of an uncontrolled expansion of human activities, nature protection areas have been added to the long list of provisions for which maintenance is required.

There are considerable costs in maintaining and repairing most infrastructure systems, as well as in extending them when necessary. Clearly, any proposal for a change on an existing complex system is to be considered carefully to make sure that it brings improvements, before any attempt to implement changes on the real network, since that may result in a loss of resources.

The complex systems that are made up of various kinds of infrastructure for the provision of services, including ecological networks, have become fields of application for complex network modeling and analysis. While models are built to reflect the different nature of each kind of system, it is common for analysis techniques to be shared among different kinds of infrastructure, and even between models of infrastructure networks and conceptual networks. Thus, software tools for complex network analysis and visualization can become useful to assist decision makers in evaluating the potential impact of changes, so long as a proper definition is given for the following parts of the process, consistently with the features of the real system:

- a 'base' model representing normal operation of the system;
- a set of modified models that reflect changes of the system (failures or improvements);

- one or more measures, calculated globally or locally, which can be compared between models to quantify the effect of modifications.

The parameter space represented by the variables that can be adjusted in the operation of the system can be so large as to make it impractical to consider the entire range of possible changes, particularly for large systems and in all cases where the possibility of adding elements to the system is a viable choice. This leads to a further necessity, consisting in the definition of a set of criteria to reduce the range of possibilities. In the present work, models for a portion of an ecological network and a regional power grid have provided examples of systems for which this problem was tackled. This process can be seen as part of a common methodology with a potential for application in multiple contexts.

In the study of ecological networks, the approach of building models to represent functional connectivity proves useful to describe the state of the network with respect to a target species; to some degree, it also provides a way to predict the effect of changes, but it has a shortcoming in the requirement of a large number of graph instances in order to represent the general state of the network. The problem of reducing the search space for proposed changes is made more evident by the necessity to test proposals against every graph that may be affected.

In this work, measures of land similarity are proposed as a way to eliminate unsound proposals. Different datasets, from which similarity measures can be calculated, have been considered and compared: sets of habitats, sets of species, and land use, with the integration of land use data, based on CORINE Land Cover codes. An analysis of how similarity-based graphs compare to single-species graphs suggest that the use of data collected within the Natura 2000 project should be favored, but land use data is potentially most useful for land management, due to its coverage of areas outside of Natura 2000 sites. Therefore, its adoption by practitioners may be necessary in some cases, in which the exclusive use of the Natura 2000 dataset may be perceived as a significant limitation.

The next step in the workflow would involve the choice of a measure of 'health' of the network and a comparison of its values in the base instance and modified instances of the network model. An example of this step was shown in the study of a regional power grid: a large-scale and coarse-granularity model of the regional power grid in Sardinia was used to determine an optimal set of power flows, supporting the notion that the region is self-sufficient in its average operation, but there is room for improvement in the redundancy of links. Then, a measure of damage from failure of transit nodes was proposed. Complex network indices from multiple models of the network were extracted, looking for a correlation of the betweenness centrality index with the collateral damage from the failure of the corresponding node.

Opportunities for future work exist in the completion and refinement of the process in both fields of applications, and its adaptation to more classes of infrastructure. A particularly interesting aspect in the field of ecological networks is the fact that shortest paths – and even least-cost paths – may not be as significant as in other fields, since the natural behavior of animals and plants clearly does not include planning their movements in the same way as it is usually intended, or computing shortest paths with complete information. The choice of a measure of health intended to reflect the possibilities of migration is therefore not straightforward. Other than stochastic models of animal behavior and plant dispersal, it is worth investigating measures that take into account shortest paths as well as suboptimal paths.

The application of this process to power grids, on the other hand, is strongly dependent on the accuracy of data and the availability of software tools for a reliable simulation of the grid at the scale of choice. Part of the challenge of devising a faithful model of the infrastructure network lies in the necessity of some degree of simplification: the collapsing of multiple power lines into single edges is an example thereof, which is understandable as it allows the direct application of analysis techniques for simple graphs. Conversely, other simplifications might be deemed unnecessary and be limiting the effectiveness of current methods. The extension to a hierarchical model of the network would be made possible by the integration of a more accurate representation of the network at a local level; another possibility is that of considering a subdivision of the regional grid into micro-grids, providing the basis for a multi-agent model of the grid in its entirety.

Bibliography

[1] R. Albert and A.-L. Barabási. "Statistical mechanics of complex networks". *Reviews of Modern Physics* 74.1 (2002), pp. 47–97. DOI: 10.1103/RevModPhys.74.47.

[2] F. Arrigo and M. Benzi. "Updating and Downdating Techniques for Optimizing Network Communicability". *SIAM Journal on Scientific Computing* 38.1 (2016), B25–B49. ISSN: 1064-8275. DOI: 10.1137/140991923.

[3] A.-L. Barabási and R. Albert. "Emergence of Scaling in Random Networks". *Science* 286.5439 (Oct. 1999), pp. 509–512. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.286.5439.509.

[4] M. Barthélemy. "Spatial networks". *Physics Reports* 499.1–3 (Feb. 2011), pp. 1–101. ISSN: 0370-1573. DOI: 10.1016/j.physrep.2010.11.002.

[5] P. Beier and R. F. Noss. "Do Habitat Corridors Provide Connectivity?" *Conservation Biology* 12.6 (1998), pp. 1241–1252. ISSN: 1523-1739. DOI: 10.1111/j.1523-1739.1998.98036.x.

[6] G. Bennett, K. J. Mulongoy, and Secretariat of the Convention on Biological Diversity. *Review of experience with ecological networks, corridors and buffer zones*. 2014.

[7] Ö. Bodin and J. Norberg. "A Network Approach for Analyzing Spatially Structured Populations in Fragmented Landscape". *Landscape Ecology* 22.1 (Aug. 2006), pp. 31–44. ISSN: 0921-2973, 1572-9761. DOI: 10.1007/s10980-006-9015-0.

[8] B. Bollobás et al. "The degree sequence of a scale-free random graph process". *Random Structures & Algorithms* 18.3 (May 2001), pp. 279–290. ISSN: 1098-2418. DOI: 10.1002/rsa.1009.

[9] T. M. Caro. "Umbrella species: critique and lessons from East Africa". *Animal Conservation* 6.2 (2003), pp. 171–181. ISSN: 1469-1795. DOI: 10.1017/S1367943003003214.

[10] G. J. Chaitin. "Register Allocation & Spilling via Graph Coloring". *Proceedings of the 1982 SIGPLAN Symposium on Compiler Construction*. SIGPLAN '82. New York, NY, USA: ACM, 1982, pp. 98–105. ISBN: 978-0-89791-074-3. DOI: 10.1145/800230.806984.

[11] M. Cheng, M. Crow, and R. F. Erbacher. "Vulnerability analysis of a smart grid with monitoring and control system". *Proceedings of the Eighth Annual Cyber Security and Information Intelligence Research Workshop*. CSIIRW '13. New York, NY, USA: ACM, 2013, 59:1–59:4. ISBN: 978-1-4503-1687-3. DOI: 10.1145/2459976.2460042.

[12] L. d. F. Costa et al. "Characterization of complex networks: A survey of measurements". *Advances in Physics* 56.1 (2007), pp. 167–242. ISSN: 0001-8732. DOI: 10.1080/00018730601170527.

[13] E. Cotilla-Sanchez et al. "Comparing the Topological and Electrical Structure of the North American Electric Power Infrastructure". *IEEE Systems Journal* 6.4 (2012), pp. 616–626. ISSN: 1932-8184. DOI: 10.1109/JSYST.2012.2183033.

[14] A. De Montis et al. "Urban–rural ecological networks for landscape planning". *Land Use Policy* 50 (2016), pp. 312–327. ISSN: 0264-8377. DOI: 10.1016/j.landusepol.2015.10.004.

[15] J. M. Diamond. "The island dilemma: Lessons of modern biogeographic studies for the design of natural reserves". *Biological Conservation* 7.2 (Feb. 1975), pp. 129–146. ISSN: 0006-3207. DOI: 10.1016/0006-3207(75)90052-X.

[16] R. Diestel. *Graph Theory*. 4th ed. 2010. Corr. 3rd printing 2012 edition. Heidelberg; New York: Springer, Oct. 2010. ISBN: 978-3-642-14278-9.

[17] J. Dong and S. Horvath. "Understanding network concepts in modules". *BMC Systems Biology* 1 (2007), p. 24. ISSN: 1752-0509. DOI: 10.1186/1752-0509-1-24.

[18] K. Driezen et al. "Evaluating least-cost model predictions with empirical dispersal data: A case-study using radiotracking data of hedgehogs (Erinaceus europaeus)". *Ecological Modelling* 209.2–4 (2007), pp. 314–322. ISSN: 0304-3800. DOI: 10.1016/j.ecolmodel.2007.07.002.

[19] A. Dwivedi, X. Yu, and P. Sokolowski. "Identifying vulnerable lines in a power network using complex network theory". *2009 IEEE International Symposium on Industrial Electronics*. July 2009, pp. 18–23. DOI: 10.1109/ISIE.2009.5214082.

[20] P. Erdős and A. Rényi. "On Random Graphs I." *Publicationes Mathematicae (Debrecen)* 6 (1959), pp. 290–297.

[21] E. Estrada and Ö. Bodin. "Using Network Centrality Measures to Manage Landscape Connectivity". *Ecological Applications* 18.7 (2008), pp. 1810–1825. ISSN: 1939-5582. DOI: 10.1890/07-1419.1.

[22] G. Fenu and M. Nitti. "An emerging complex network approach for grid analysis". *2013 IEEE International Symposium on Industrial Electronics (ISIE)*. 2013, pp. 1–6. DOI: 10.1109/ISIE.2013.6563705.

[23] G. Fenu, M. Nitti, and P. L. Pau. "A complex network approach for a regional power grid analysis". *2012 Second International Conference on Digital Information Processing and Communications (ICDIPC)*. July 2012, pp. 45–50. DOI: `10.1109/ICDIPC.2012.6257271`.

[24] G. Fenu, M. Nitti, and P. L. Pau. "Performance Analysis and Grid Computing on Wide Area". *International Journal of New Computer Architectures and their Applications (IJNCAA)* 2.4 (2012), pp. 500–510.

[25] G. Fenu and P. L. Pau. "A land similarity approach to modeling complex ecological networks". *2016 International Multidisciplinary Conference on Computer and Energy Science (SpliTech)*. July 2016, pp. 1–6. DOI: `10.1109/SpliTech.2016.7555921`.

[26] G. Fenu and P. L. Pau. "A Model of Assessment of Collateral Damage on Power Grids based on Complex Network Theory". *Procedia Computer Science*. The 5th International Conference on Ambient Systems, Networks and Technologies (ANT-2014), the 4th International Conference on Sustainable Energy Information Technology (SEIT-2014) 32 (2014), pp. 437–444. ISSN: 1877-0509. DOI: `10.1016/j.procs.2014.05.445`.

[27] G. Fenu and P. L. Pau. "Evaluating complex network indices for vulnerability analysis of a territorial power grid". *Journal of Ambient Intelligence and Humanized Computing* (Mar. 2015), pp. 1–10. ISSN: 1868-5137, 1868-5145. DOI: `10.1007/s12652-015-0264-0`.

[28] G. Fenu and P. L. Pau. "Graph Models of Network Behavior in Environmental Planning". *Procedia Computer Science*. Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 20th International Conference KES-2016 96 (2016), pp. 73–80. ISSN: 1877-0509. DOI: `10.1016/j.procs.2016.08.097`.

[29] G. Fenu and P. L. Pau. "Modelli funzionali delle reti ecologiche: dal particolare al generale". *Conferenza ASITA 2016 - Federazione delle Associazioni Scientifiche per le Informazioni Territoriali e Ambientali*. Nov. 2016, pp. 397–404. ISBN: 978-88-941232-6-5.

[30] G. Fenu and P. L. Pau. "Topological and Conceptual Complex Network Models for Environmental Planning". *Procedia Computer Science*. The 7th International Conference on Ambient Systems, Networks and Technologies (ANT 2016) / The 6th International Conference on Sustainable Energy Information Technology (SEIT-2016) / Affiliated Workshops 83 (2016), pp. 123–130. ISSN: 1877-0509. DOI: `10.1016/j.procs.2016.04.107`.

[31] G. Fenu, P. L. Pau, and D. Dessì. "Functional Models and Extending Strategies for Ecological Networks" (2017). Submitted to Applied Network Science, Springer International Publishing.

[32] G. Fenu, P. L. Pau, and D. Dessì. "Modeling and Extending Ecological Networks Using Land Similarity". *Complex Networks & Their Applications V* (Nov. 2016). Springer International Publishing, pp. 709–718. DOI: 10.1007/978-3-319-50901-3_56.

[33] A. M. Fleury and R. D. Brown. "A framework for the design of wildlife conservation corridors With specific application to southwestern Ontario". *Landscape and Urban Planning* 37.3–4 (1997), pp. 163–186. ISSN: 0169-2046. DOI: 10.1016/S0169-2046(97)80002-3.

[34] L. K. Gallos et al. "Stability and Topology of Scale-Free Networks under Attack and Defense Strategies". *Physical Review Letters* 94.18 (2005), p. 188701. DOI: 10.1103/PhysRevLett.94.188701.

[35] J. Gao et al. "Robustness of a Network of Networks". *Physical Review Letters* 107.19 (Nov. 2011), p. 195701. DOI: 10.1103/PhysRevLett.107.195701.

[36] E. N. Gilbert. "Random Graphs". *The Annals of Mathematical Statistics* 30.4 (Dec. 1959), pp. 1141–1144. ISSN: 0003-4851, 2168-8990. DOI: 10.1214/aoms/1177706098.

[37] L. Gilbert-Norton et al. "A meta-analytic review of corridor effectiveness". *Conservation Biology: The Journal of the Society for Conservation Biology* 24.3 (June 2010), pp. 660–668. ISSN: 1523-1739. DOI: 10.1111/j.1523-1739.2010.01450.x.

[38] M. Hashmi, S. Hanninen, and K. Maki. "Survey of smart grid concepts, architectures, and technological demonstrations worldwide". *2011 IEEE PES Conference on Innovative Smart Grid Technologies (ISGT Latin America)*. 2011, pp. 1–7. DOI: 10.1109/ISGT-LA.2011.6083192.

[39] R. Hassan and G. Radman. "Survey on Smart Grid". *Proceedings of the IEEE SoutheastCon 2010 (SoutheastCon)*. Mar. 2010, pp. 210–213. DOI: 10.1109/SECON.2010.5453886.

[40] P. Hines, E. Cotilla-Sanchez, and S. Blumsack. "Topological Models and Critical Slowing down: Two Approaches to Power System Blackout Risk Analysis". *2011 44th Hawaii International Conference on System Sciences (HICSS)*. 2011, pp. 1–10. DOI: 10.1109/HICSS.2011.444.

[41] R. H. G. Jongman. "Nature conservation planning in Europe: developing ecological networks". *Landscape and Urban Planning* 32.3 (1995), pp. 169–183. ISSN: 0169-2046. DOI: 10.1016/0169-2046(95)00197-O.

[42] A. D. Mazaris et al. "Evaluating the Connectivity of a Protected Areas' Network under the Prism of Global Change: The Efficiency of the European Natura 2000 Network for Four Birds of Prey". *PLoS ONE* 8.3 (Mar. 2013), e59640. DOI: 10.1371/journal.pone.0059640.

[43] E. S. Minor and D. L. Urban. "A Graph-Theory Framework for Evaluating Landscape Connectivity and Conservation Planning". *Conservation Biology* 22.2 (Apr. 2008), pp. 297–307. ISSN: 1523-1739. DOI: 10.1111/j.1523-1739.2007.00871.x.

[44] T. J. Overbye. "Power system simulation: understanding small- and large-system operations". *IEEE Power and Energy Magazine* 2.1 (2004), pp. 20–30. ISSN: 1540-7977. DOI: 10.1109/MPAE.2004.1263413.

[45] G. A. Pagani and M. Aiello. "The Power Grid as a complex network: A survey". *Physica A: Statistical Mechanics and its Applications* 392.11 (2013), pp. 2688–2700. ISSN: 0378-4371. DOI: 10.1016/j.physa.2013.01.023.

[46] S. C. F. Palmer, A. Coulon, and J. M. J. Travis. "Introducing a 'stochastic movement simulator' for estimating habitat connectivity". *Methods in Ecology and Evolution* 2.3 (2011), pp. 258–268. ISSN: 2041-210X. DOI: 10.1111/j.2041-210X.2010.00073.x.

[47] J. Petinrin and M. Shaaban. "Smart power grid: Technologies and applications". *2012 IEEE International Conference on Power and Energy (PECon)*. 2012, pp. 892–897. DOI: 10.1109/PECon.2012.6450343.

[48] N. Pinto and T. H. Keitt. "Beyond the least-cost path: evaluating corridor redundancy using a graph-theoretic approach". *Landscape Ecology* 24.2 (Nov. 2008), pp. 253–266. ISSN: 0921-2973, 1572-9761. DOI: 10.1007/s10980-008-9303-y.

[49] QGIS Development Team. *QGIS Geographic Information System*. Open Source Geospatial Foundation, 2009.

[50] S. M. Rinaldi, J. P. Peerenboom, and T. K. Kelly. "Identifying, understanding, and analyzing critical infrastructure interdependencies". *IEEE Control Systems* 21.6 (2001), pp. 11–25. ISSN: 1066-033X. DOI: 10.1109/37.969131.

[51] J.-M. Roberge and P. Angelstam. "Usefulness of the Umbrella Species Concept as a Conservation Tool". *Conservation Biology* 18.1 (Feb. 2004), pp. 76–85. ISSN: 1523-1739. DOI: 10.1111/j.1523-1739.2004.00450.x.

[52] M. Rohden et al. "Impact of network topology on synchrony of oscillatory power grids". *Chaos: An Interdisciplinary Journal of Nonlinear Science* 24.1 (Feb. 2014), p. 013123. ISSN: 1054-1500. DOI: 10.1063/1.4865895.

[53] D. K. Rosenberg, B. R. Noon, and E. C. Meslow. "Biological Corridors: Form, Function, and Efficacy". *BioScience* 47.10 (Nov. 1997), pp. 677–687. ISSN: 0006-3568, 1525-3244. DOI: 10.2307/1313208.

[54] P. Shannon et al. "Cytoscape: a software environment for integrated models of biomolecular interaction networks". *Genome Research* 13.11 (Nov. 2003), pp. 2498–2504. ISSN: 1088-9051. DOI: 10.1101/gr.1239303.

[55] U. Stelzl et al. "A Human Protein-Protein Interaction Network: A Resource for Annotating the Proteome". *Cell* 122.6 (Sept. 2005), pp. 957–968. ISSN: 0092-8674, 1097-4172. DOI: 10.1016/j.cell.2005.08.029.

[56] D. Urban and T. Keitt. "Landscape Connectivity: A Graph-Theoretic Perspective". *Ecology* 82.5 (2001), pp. 1205–1218. ISSN: 1939-9170. DOI: 10.1890/0012-9658(2001)082[1205:LCAGTP]2.0.CO;2.

[57] D. L. Urban et al. "Graph models of habitat mosaics". *Ecology Letters* 12.3 (Mar. 2009), pp. 260–273. ISSN: 1461-0248. DOI: 10.1111/j.1461-0248.2008.01271.x.

[58] R. Vimal, R. Mathevet, and J. D. Thompson. "The changing landscape of ecological networks". *Journal for Nature Conservation* 20.1 (2012), pp. 49–55. ISSN: 1617-1381. DOI: 10.1016/j.jnc.2011.08.001.

[59] D. J. Watts and S. H. Strogatz. "Collective dynamics of 'small-world' networks". *Nature* 393.6684 (1998), pp. 440–442. ISSN: 0028-0836. DOI: 10.1038/30918.