Modeling and Extending Ecological Networks Using Land Similarity

Gianni Fenu, Pier Luigi Pau and Danilo Dessì

Abstract Complex network analysis is being applied on topological models of ecological networks, to extrapolate their advanced properties and as part of the activity of land management. Commonly employed methods tend to focus on single target species. This is satisfactory for cognitive analysis, but the limited view provided by these models results in a lack of general information needed for land planning. Similarity scores computed for pairs of nature protection areas are proposed as a building block of a general model to address this shortcoming.

1 Introduction

Nature protection areas are established to protect endangered habitats and species from possible destruction due to the effects of increasing urbanization. Over the decades, policies have shifted toward the creation of ecological networks with a focus on the preservation of biodiversity. In the European Union, the establishment of a wide ecological network is the main goal of the Natura 2000 project.

Current methods to build graph models for ecological networks keep the focus on a species of interest. The resulting graphs are useful to perform quantitative analysis with respect to the target species, but the analysis of a large number of graphs is necessary to assess general properties of the network. In this paper, similarity scores between nature protection areas are proposed as a building block for graph models with a higher degree of generality, and different approaches are evaluated according to their aptness to the process of proposing network modifications.

The paper is organized as follows: in Section 2, basic information is provided concerning ecological networks, their graph representations, and goals of analysis. In Section 3, the aptness of available data on Natura 2000 sites to this study is

Gianni Fenu (e-mail: fenu@unica.it)✉ · Pier Luigi Pau (e-mail: pierluigipau@unica.it)✉ · Danilo Dessì (e-mail: danilo_dessi@unica.it)
Department of Mathematics and Computer Science, University of Cagliari, Via Ospedale 72, 09124 Cagliari, Italy

H. Cherifi et al. (eds.), *Complex Networks & Their Applications V*,
Studies in Computational Intelligence 693,
DOI 10.1007/978-3-319-50901-3_56

discussed. In Section 4, the sites located in Sardinia are presented as a case study, and similarity-based graph models are introduced; three approaches to their construction are provided. In Section 5, correlations are sought between graph models, in order to determine which is most useful for land management and planning. Lastly, in Section 6, conclusions are drawn and possibilities for future work are discussed.

2 Ecological Networks and Graph Models

The expansion of human activities in every sector has caused radical modifications in land use, with a destruction of portions of habitats, and a fragmentation of those still in place. To protect habitats and species at risk of extinction, nature protection areas have been created. As the effectiveness of these areas is strongly reduced if habitat patches are too small or too distant from similar ones, policies have converged toward the creation of ecological networks, with each area contributing to large-scale preservation goals, and more endeavors to preserve the possibility of migration of species, in order to protect biodiversity [12]. Where necessary, migration can be encouraged by the establishment of man-made 'habitat corridors', either contiguous or in the form of 'stepping stones', i.e. sets of disconnected patches.

In the European Union, an ecological network is maintained as part of the project denominated "Natura 2000". Its elements are sites designated as Special Protection Areas (SPA), as defined in the EU Birds Directive (2009/147/EC), and Special Areas of Conservation (SAC), as defined in the EU Habitats Directive (92/43/EEC); the latter are preliminarily designated as Sites of Community Interest (SCI). The boundaries of a SPA can overlap with those of SACs or SCIs, and vice versa; sites of the same category can be adjacent to one another.

The maintenance of ecological networks is becoming an essential aspect of land management and planning: local administrations are directly involved when Natura 2000 sites are in their jurisdiction, and can be affected by the presence of neighboring sites as well, due to their involvement in the possible creation of habitat corridors. Administrations are involved with the identification of threats and the proposal of a course of action to address standing issues with proper land management planning, which requires the consideration of several technical, regulatory, and political aspects. Tools to perform quantitative analysis on models representing an ecological network could make for an important contribution to the solution of these problems.

In analogy with many other kinds of networks and complex systems, a mathematical model for ecological networks is generally based on a graph, consisting of a set of nodes and a set of edges. A node may represent a site or habitat patch, depending on the desired scale, while edges represent connections. Graph models are built to represent functional connectivity with respect to a target species [11], while structural connectivity is analyzed with Geographic Information System tools. Quantitative analysis can uncover advanced properties of a network, which are not easily devised from its geographical map. Moreover, it enables comparison of graph models built for different target species in the same area, for a single target species in different areas, or representing different proposals for network modification.

Complex network analysis involves the study of statistical properties of graphs, related to node degree, shortest path length, and other features. Among the most commonly used indices are the clustering coefficient, related to the degree of redundancy of links; and the betweenness centrality index, often used to rank nodes by importance, according to their occurrence in shortest paths. The meaning of indices ought to be investigated according to each kind of real-world network being represented [3]; interpretations of several complex network indices have been proposed for ecological networks [4]. Global indices can be used as a measure of 'health' of the network, and local indices may assist in identifying vulnerabilities in topological networks [6], often associated to resiliency to node removal [5]. In general, the comparison of indices of a given network with those of modified versions is useful to predict the effect of modifications.

3 Similarity of Natura 2000 Sites

In order to collect data on the habitats and species found within the area, and to evaluate the impact of changes over time, reports are filed periodically for each Natura 2000 site. Information is gathered on-site and written to a data base conforming to a Standard Data Form, released with Commission Implementing Decision 2011/484/EU. Each Natura 2000 site is made up of patches of different habitat types, and each patch may host a different set of species. However, habitats and species found within a Natura 2000 site are reported to be present in the site, but no explicit relationship is established between each species and the habitat patch where it is found. This is sensible for the purposes of the Natura 2000 project, but it has a drawback in the fact that the knowledge of which habitat type is ideal for each species is not stored; rather, it is assumed to be part of expert knowledge or found in external documents. As a consequence, it is not straightforward to represent constraints that apply when proposing modifications.

To address these problems at least partially, it is possible to represent each site as a vector and compute similarity scores of these vectors, thus estimating a similarity score for pairs of sites. A minimum score between a pair of sites can be a prerequisite for the proposal to add an edge to the network. Adopting a similarity score taking values from 0 to 1 (where 1 is associated with pairs of identical vectors), such as the Jaccard coefficient or cosine distance, makes it easy to choose a threshold value.

The reported presence of species and that of habitats are two viable choices to build vectors representing Natura 2000 sites, using only data collected for the Natura 2000 project. A third viable approach is given by computing the intersection of sites with land use data from the CORINE program (Coordination of Information on the Environment). For this study, this was done using the open source QGIS [8] software suite. It is notable that CORINE land use data is available for areas outside of Natura 2000 sites; this is important to be able to combine graph-theoretic approaches and GIS functions [7], to determine whether it is possible to establish contiguous corridors. Land use types are categorized in a hierarchical manner with five levels of increasing detail. Only the first three levels of detail were used, as the fourth and

fifth were not available consistently; thus, a vector for each site was built by counting land patches intersecting the site, corresponding to each 3-digit code.

4 Case Study

In this work, the subset of Natura 2000 sites found in Sardinia is presented as a case study. At the time of writing, there is a total of 124 sites counting those designated as SPA and SCI; however, seven sites where excluded from this study because of unavailable land use data. If the boundaries of a SPA and a SCI overlap, two nodes are created, but they are considered to be at zero distance from each other. In all graph instances, an edge is not drawn between a pair of nodes if their approximate distance (calculated between borders on a map projection, using SQLite with the Spatialite extension) is greater than a set threshold (30 Km). The resulting network has 117 nodes, each corresponding to a Natura 2000 site. When all pairs of nodes where the geographical distance is up to 30 Km are linked, there is a total of 850 edges in the graph model. This shall be referred to as the *raw-distance graph* (Figure 1a).

A *single-species graph* is a model built to represent the state of the network with respect to a single species. In order to build a single-species graph, node pairs are linked with an edge if their distance is below the threshold and the presence of the species has been reported in both sites. Single-species graphs were built for all species listed in Annex II to Directive 92/43/EEC, plus others for which a species code consistent across site reports was given; an example is in Figure 1b. The open source Cytoscape suite (version 3.4.0) was used for graph visualization [9] and analysis, through the native NetworkAnalyzer plugin [2].

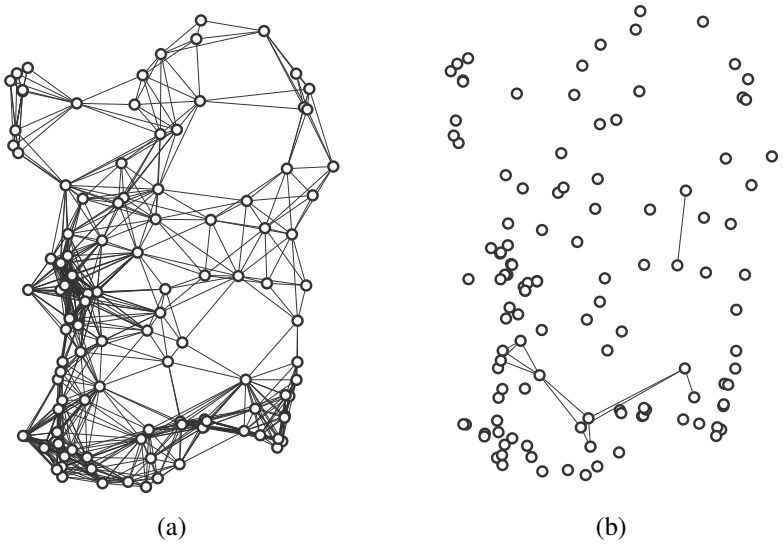
To represent the state of the network from a more general point of view, it is possible to build a graph instance based on site similarity, which shall be generally referred to as a *similarity-based graph*. This corresponds to a modification of the raw-distance graph, with the removal of edges that link node pairs with a similarity score below a set threshold. Clearly, different ways to compute similarity scores result in different graphs. In this study, three graphs are built for analysis, each based on Jaccard coefficients calculated on different vector representation of sites: the set of species reported to be in a site (*species-set graph*), the set of habitats found in the site according to Natura 2000 project data (*habitat graph*), and the set of level 3 land use codes according to the CORINE program (*land-use graph*). A similarity score of 0.5 shall be used as a threshold for all similarity-based graphs. In fact, recalling that the raw-distance network has 850 edges, similarity scores of 0.6 and above turn out to be strong requirements, removing over 85% of edges in all cases (Table 1).

5 Analysis of Edge Hit Rates and Complex Network Indices

The analysis of a single-species graph, with the extraction of its indices, is meant to give insight on the state of the network for the purpose of conservation of that species. As land management proposals may be reflected by modifications on the graph model, the improvement of indices according to set goals can act as a criterion

Table 1: Number of edges in similarity-based graphs of Natura 2000 sites in Sardinia

Minimum similarity	Land use-based	Habitat-based	Species-based
0.0 (raw-distance)	850	850	850
0.4	360	295	198
0.5	232	205	117
0.6	104	120	53



(a) (b)

Fig. 1: Graph models of Natura 2000 sites in Sardinia. (a) 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. (b) Single-species graph for *Cervus elaphus corsicanus* (species code 1367). For comparison purposes, all nodes from the raw-distance graph are represented; technically, only linked nodes are part of this graph.

to identify favorable modifications. In many complex network applications, the set of nodes is to be kept unmodified; assuming the network is initially connected, proposed modifications can fall into one of three categories [1]:

- Addition of edges (also referred to as 'updating'; proposed edges are referred to as 'virtual edges'). Assuming that adding links in the real-world network has a cost, this problem is related to that of finding a set of new links which brings as great a benefit as possible, while respecting budget constraints.
- Removal of edges ('downdating'). Assuming that the network has some degree of redundancy, this problem is related to that of finding a set of edges that can

be removed to decrease maintenance costs, while affecting the efficiency of the network as little as possible. The network as a whole should not be disconnected.
- Rewiring, i.e. removing and subsequently adding one or more edges. This problem is related to that of improving the efficiency of a network, while avoiding an increase in maintenance costs.

In land management for ecological networks, an interesting problem may be to find a site to relocate part of the population of a species, among those where it has not been reported, while preserving or enhancing the emerging network effect; this can be done to improve network indices or to merge components which are not initially connected. In the graph model, this is reflected as an addition of nodes; this poses a few problems, most notably the identification of suitable candidate sites for node addition.

As previously mentioned, the Natura 2000 dataset was not designed with this problem in mind, hence it is not straightforward to suggest good candidate nodes to extend any given single-species graph. Newly connected sites should be within a set geographical distance from an already connected node, and should host the preferred habitat for the target species, or a suitable set of habitats for a temporary settlement of the species, if the node is to act as a 'bridge'.

One of the methods to calculate site similarity scores may qualify as a way to formalize this criterion when data on habitat suitability is missing or incomplete. Then, similarity-based graphs become a useful tool to express this notion. In formal terms, a good candidate has the property that, in a similarity-based graph, it is adjacent to a node that is part of a connected component in the single-species graph. In symbols, let V be the full set of nodes representing Natura 2000 sites in the region of interest, and let $G_s = (V, E_s)$ be a similarity-based graph built on node set V with a suitable geographical distance threshold. Let $G' = (V', E')$ be a connected component in the single-species graph built on V for the target species ($V' \subseteq V$), with the same geographical distance threshold used for G_s. Then, if

$$i \in V', \quad j \in V, \quad j \notin V', \quad (i, j) \in E_s, \tag{1}$$

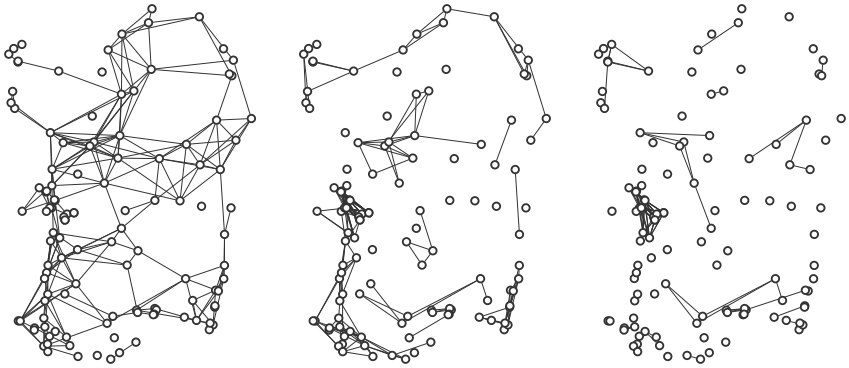
then $j \in V$ is a good candidate node, and (i, j) is a candidate edge to link j to G'.

Since there are more ways to build G_s, an interesting question is which similarity-based graph is best for the purpose of determining good node candidates. Intuitively, if edges in single-species graphs often appear as edges in G_s, then G_s should provide better candidates for graph updating. To measure the aptness of the similarity-based graphs built on Jaccard coefficients of species sets, habitats and land use codes, the three graphs were compared with 351 single-species graphs, using the same 30 Km distance threshold. Results for a few sample species are reported in Table 2, together with average rates. The three similarity-based graphs rank about the same way at a 34% average hit rate, with a slight disadvantage for the species-set graph at 31%.

These low hit rates do not show a clear winner among the three criteria under consideration for building similarity-based graph; not only that, but they suggest that there may be remarkable differences among the three graphs, which is confirmed by comparing them visually (Figure 2).

Table 2: Excerpt of the table of hit rates. For each species, the number of edges in the single-species graph is reported. Then, for each similarity-based graph, it is shown how many of those edges are present in the similarity-based graph (hits), and the corresponding rate. The last row reports the average of all hit rates for each similarity-based graph.

		Land use-based		Habitat-based		Species-based	
Species code	Edges	Hits	Rate	Hits	Rate	Hits	Rate
...							
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
...							
Average hit rate			0.33650		0.34341		0.31308



(a) (b) (c)

Fig. 2: Similarity-based graph models of Sardinian Natura 2000 sites, with a 0.5 similarity score threshold and a 30 Km distance threshold. (a) Based on CORINE land use codes. (b) Based on Natura 2000 habitat codes. (c) Based on species sets.

To establish whether these differences are significant, it is interesting to assess whether any pair of similarity-based graphs behave similarly with respect to hit rates; essentially, if the hit rate for a similarity graph G_s were high for the same species as that of another similarity graph G_t, it could be argued that G_s and G_t express a similar concept. To do so, Spearman correlation indices are calculated between pairs of columns reporting hit rates in Table 2. It is notable that, while no correlation is detected between the land use graph and the others, the species-set and habitat graphs appear to have a strong correlation, above 0.8 (Table 3). This is not only consistent with the fact that land use data originates from a different project; it confirms that

nearby sites with similar habitat sets also host similar sets of species, thus reinforcing the notion that the classification of habitats within the Natura 2000 project is more suitable to describe sites than land use codes are.

Table 3: Spearman correlation between sets of hit rates.

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

To corroborate this notion, it is possible to extend the comparison to complex network indices calculated for nodes on the three similarity-based graph instances. Indices are calculated for nodes representing Natura 2000 sites on the three graph instances (see an excerpt in Table 4). The question is, for each index, whether a higher value calculated on a graph corresponds to a higher value calculated on another. Considered indices are node degree, closeness and betweenness centrality indices, clustering coefficient, and topological coefficient [10].

Table 4: Excerpt of the table of normalized betweenness centrality indices calculated for each node (site) on each similarity-based graph.

	Betweenness centrality index		
Site	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
...			

Then, correlation are sought between sets of values for the same index on the three possible pairs of graph instances, once again by calculating their Spearman correlation coefficients (see Table 5 and a visual representation in Figure 3). Contrary to hit rates, there is no value suggesting a strong correlation; however, a moderate degree of correlation can be identified between the species-set and the habitats graph for three measures (degree, topological coefficient and clustering coefficient), thus reinforcing the previous observations that these two graphs are more similar to one another, than the land-use graph is to either.

Table 5: Spearman correlation of various complex network indices, between pairs of similarity-based graphs.

Index	Land-use/Species	Land-use/Habitats	Species/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

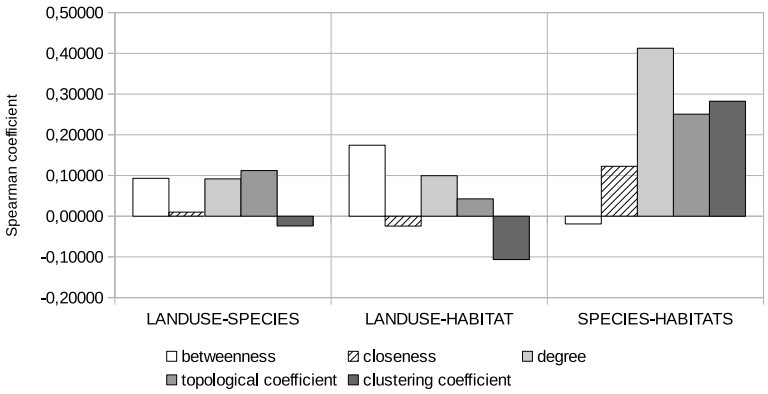


Fig. 3: Histogram representation of Spearman correlation of various complex network indices, between pairs of similarity-based graphs.

6 Conclusions and Future Work

Current methods to apply complex network analysis on ecological networks tend to focus on single species of interest, making it hard to evaluate and represent high-level properties of the network. The analysis of single-species graphs proves to be useful to assess the state of the network, but in the context of the Natura 2000 project in the European Union, methods for data collection and storage were not designed to assist researchers in proposing network modifications for its improvement. In this paper, the construction of graph models based on site similarity is proposed as a way to address this shortcoming. Multiple ways to build similarity-based graphs are discussed and compared; results suggest that land use data expresses different concepts than the species sets and habitat sets associated to each site. This represents a challenge for land managers seeking to detect or establish habitat corridors, since only land use data is available for land outside of Natura 2000 sites.

Future work will focus on an extension and application of the network updating problem on single-species models. The linking of nodes that are not initially con-

nected will be considered, subject to constraints based on site similarity and on the degree of contiguity of land use outside of Natura 2000 sites.

Acknowledgements This essay is written within the Research Program "Natura 2000: Assessment of management plans and definition of ecological corridors as a complex network", funded by the Autonomous Region of Sardinia (Legge Regionale 7/2007) for the period 2015-2018, under the provisions of the Call for the presentation of "Projects related to fundamental or basic research" in year 2013, implemented at the Department of Mathematics and Computer Science of the University of Cagliari, Italy. 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.).

References

[1] Arrigo, F., Benzi, M.: Updating and Downdating Techniques for Optimizing Network Communicability. SIAM Journal on Scientific Computing **38**(1), B25–B49 (2016)
[2] Assenov, Y., Ramírez, F., Schelhorn, S.E., Lengauer, T., Albrecht, M.: Computing topological parameters of biological networks. Bioinformatics (Oxford, England) **24**(2), 282–284 (2008). DOI 10.1093/bioinformatics/btm554
[3] Borgatti, S.: Centrality and network flow. Social Networks **27**(1) (2005)
[4] Estrada, E., Bodin, Ö.: Using Network Centrality Measures to Manage Landscape Connectivity. Ecological Applications **18**(7), 1810–1825 (2008)
[5] Iyer, S., Killingback, T., Sundaram, B., Wang, Z.: Attack Robustness and Centrality of Complex Networks. PLOS ONE **8**(4), e59,613 (2013)
[6] Mishkovski, I., Biey, M., Kocarev, L.: Vulnerability of complex networks. Communications in Nonlinear Science and Numerical Simulation **16**(1), 341–349 (2011)
[7] Pinto, N., Keitt, T.H.: Beyond the least-cost path: evaluating corridor redundancy using a graph-theoretic approach. Landscape Ecology **24**(2), 253–266 (2008)
[8] QGIS Development Team: QGIS Geographic Information System. Open Source Geospatial Foundation (2009). URL `http://qgis.osgeo.org`
[9] Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., Ideker, T.: Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Research **13**(11), 2498–2504 (2003). DOI 10.1101/gr.1239303
[10] Stelzl, U., Worm, U., Lalowski, M., Haenig, C., Brembeck, F.H., Goehler, H., Stroedicke, M., Zenkner, M., Schoenherr, A., Koeppen, S., Timm, J., Mintzlaff, S., Abraham, C., Bock, N., Kietzmann, S., Goedde, A., Toksöz, E., Droege, A., Krobitsch, S., Korn, B., Birchmeier, W., Lehrach, H., Wanker, E.E.: A human protein-protein interaction network: a resource for annotating the proteome. Cell **122**(6), 957–968 (2005). DOI 10.1016/j.cell.2005.08.029
[11] Urban, D.L., Minor, E.S., Treml, E.A., Schick, R.S.: Graph models of habitat mosaics. Ecology Letters **12**(3), 260–273 (2009)
[12] Vimal, R., Mathevet, R., Thompson, J.D.: The changing landscape of ecological networks. Journal for Nature Conservation **20**(1), 49–55 (2012)