Please note that this is an author-produced PDF of an article accepted for publication following peer review. The definitive publisher-authenticated version is available on the publisher Web site

Molecular Ecology Resources
January 2015, Volume 15, Issue 1, Pages 117-122
http://dx.doi.org/10.1111/1755-0998.12290
© 2014 John Wiley & Sons Ltd

Archimer
http://archimer.ifremer.fr

EDENetworks: A user-friendly software to build and analyse networks in biogeography, ecology and population genetics

Mikko Kivelä[1,2], Sophie Arnaud-Haond[3,*] and Jari Saramäki[2]

[1] Oxford Centre for Industrial and Applied Mathematics, Mathematical Institute, University of Oxford, Oxford, UK
[2] Department of Biomedical Engineering and Computational Science, School of Science, Aalto University, Helsinki, Finland
[3] Ifremer, UMR «Ecosystèmes Marins Exploités», Sète Cedex, France

*: Corresponding author : Sophie Arnaud-Haond, fax: (+33)(0)4 99 57 32 95 ; email address : sarnaud@ifremer.fr

Abstract:

The recent application of graph-based network theory analysis to biogeography, community ecology and population genetics has created a need for user-friendly software, which would allow a wider accessibility to and adaptation of these methods. EDENetworks aims to fill this void by providing an easy-to-use interface for the whole analysis pipeline of ecological and evolutionary networks starting from matrices of species distributions, genotypes, bacterial OTUs or populations characterized genetically. The user can choose between several different ecological distance metrics, such as Bray-Curtis or Sorensen distance, or population genetic metrics such as F_{ST} or Goldstein distances, to turn the raw data into a distance/dissimilarity matrix. This matrix is then transformed into a network by manual or automatic thresholding based on percolation theory or by building the minimum spanning tree. The networks can be visualized along with auxiliary data and analysed with various metrics such as *degree*, *clustering coefficient*, *assortativity* and *betweenness centrality*. The statistical significance of the results can be estimated either by resampling the original biological data or by null models based on permutations of the data.

Keywords: biogeography ; biological communities ; graph theory ; microbial ecology ; network analysis ; population genetics

1. Introduction

Network analysis based on graph theory has turned out to be an invaluable tool for exploring the structure of many complex systems in diverse range of fields from sociology to economy and from physics to cell biology (Newman 2010; Ueda *et al.* 2004; Alon 2003). In this approach, it is assumed that most of the complexity of the system can be captured by the topology of a network formed by a set of nodes – or agents – that are connected to each other by links (see Table 1 for a glossary of terms). In addition to domain-specific characteristics, most networks display universal features, such as clustered and modular structures (Lancichinetti *et al.* 2010), the small-world property (Watts& Strogatz 1998), and broad connectivity distributions (Barabasi & Albert 1999). The topology of a network can be crucial for the function and robustness of the system (Albert *et al.* 2000) and dynamics taking place on top it (Barrat *et al.* 2009). Network topology can also suggest possible mechanisms by which the network has evolved to its current state (Dorogovtsev & Mendes 2003).

In a distinct way, graph-based network theory provides a promising approach for studies in ecology and evolution (Bascompte *et al.* 2003; Hernández-García *et al.* 2007; Proulx *et al.* 2005). Population genetics data (Multi Locus Genotypes) can be turned into networks among individuals (Becheler *et al.* 2010; Hernandez-Garcia *et al.* 2006; Rozenfeld *et al.* 2007) or among some pre-determined populations (Fortuna *et al.* 2009; Rozenfeld *et al.* 2008) by considering each pair of individuals or populations connected if they are genetically similar enough. At both of these levels, network theory has proven to be a useful method to analyse population genetics data and to help unravel ecological and evolutionary processes acting at local and regional scales. It has recently been proposed that the structure of ecological networks may illustrate relationships among communities based on ecological dissimilarities of their taxonomic composition (species composition or presence absence), and that the analysis of the topology of such networks may be used to define biogeographic provinces and to reconstruct their history of divergence (Dos Santos *et al*., 2008; Moalic *et al.* 2012). In addition, the network approach was recently successfully tested to illustrate and analyse the relatedness and clustering of both eukaryotic MOTUs (Molecular Operational Taxonomic Units) and microbial OTUs forming communities characterized by new generation sequencing technologies (data reanalysed from Aires et al., 2013).

Network-based methods of data exploration are free of many of the '*a priori*' assumptions, such as geographic clustering (genetic similarity spatially close populations), that usually underlie the population-genetic interpretation of molecular data, as well as some ecological data analysis. In addition, the tools and indices developed in the framework of network theory allow unravelling unique properties such as the importance of each agent (individual, population or community) in populations, metapopulation or biogeographic systems (Becheler *et al.* 2010; Moalic *et al.* 2012; Rozenfeld *et al.* 2008). Finally, networks offer a natural way to graphically present inherently multidimensional data such as genetic relationships. This is an advantage over the classical methods based on phylograms, or trees, in cases where

some of their assumptions, such as binary branching or absence of loops, are known to be violated due the reticulate nature of relationships. To this extent, the rationale behind this kind of network analysis converges with the objectives of sequence, or haplotypes networks (Posada & Crandall, 2001; Huson & Bryant, 2006) proposed to illustrate genes evolution taking into account the uncertainties in mutational pathways or possibilities of sporadic reticulate events (such as recombination, lateral transfer or hybridization). Contrastingly, the methods proposed here are adapted from network analysis developed in the framework of graph theory and aim at unravelling the history and dynamics of naturally reticulated systems of interconnected communities, populations or individuals through the analysis of species distribution or population genetic data.

Until now, most network analysis for biogeography and population genetics has been performed using ad-hoc scripts and separate network visualization tools such as Pajek (Batajelj & Mrvar, 2002). Here we introduce EDENetworks, a user-friendly software package that makes network analysis and visualization accessible to a wide range of researchers. EDENetworks has been developed for constructing and analysing ecological and evolutionary networks starting from genetic or ecological data. It implements a straightforward pipeline that standardizes network construction and analysis in this context (Figure 1). This should facilitate a more widespread use of network methods in the community of ecologists and population geneticists, as well as provide tools for constructing ecological and evolutionary networks for network scientists. Further, for assessing the statistical significance of findings, EDENetworks provides a way for constructing randomized reference ensembles of networks that are based on biologically-motivated null models. In the null models, randomization takes place at the level of source data, either by randomly shuffling alleles among individuals or samples among populations. This is in contrast to the purely structural null models commonly used by network scientists (e.g., the configuration model for randomly rewiring networks; Newman *et al.* 2001) that do not correspond to any biologically motivated null hypothesis, since the randomization takes place only after the source data has been processed into a network representation. Finally, although EDENetworks already includes a wide variety of analysis methods, the user can choose to export the networks and to analyse them with some general-purpose network analysis tools such as Gephi (Bastian *et al.* 2009), Cytoscape (Smoot *et al.* 2011), or igraph (Hartvigsen 2011).

Data input and export formats

EDENetworks can handle a wide range of data types in simple delimited text file formats as described in the manual. Ecological distance networks between communities can be constructed from data matrices of presence/absence or abundance of species, eukaryotic MOTUs or microbial OTUs in the characterized communities. Genetic distance networks between individuals or populations can be constructed from data matrices of genotypes of individuals. In addition, EDENetworks can read pre-computed distance matrices (e.g. when the user wants to experiment with a distance metric that is not available in EDENetworks) or files that directly contain network structures (e.g. in Graph Markup Language). The user can also provide an input file containing auxiliary data for the nodes, which can contain, for

example, individual or community labels, geographic locations or custom color codes which can be used for network visualization.

All results of network analysis can be exported as image files or text files that can be read with any standard spreadsheet or text processing software. The networks themselves can be saved in standard file formats or visualized and saved as images in vector or raster formats. For further visualization with external software packages such as Gephi, the layout coordinates used in network visualization can be saved in a text file.

Analysis

The analysis pipeline in EDENetworks is shown in Figure 1 for various data types and is described in more detail below:

(1) Data input and distance matrix construction. The user provides an input file and chooses the type of data it contains. For some data types it is possible to automatically infer the exact format of the data (e.g. if the distance matrix is upper or lower triangular). The distance/dissimilarity metric is chosen from a list appropriate distances for the input data. An auxiliary node data file can also be given if desired.

(2) Analyze distance data and derive networks. The distance matrix constructed from input data can be thresholded manually, or automatically at the identified percolation threshold to produce a network. Alternatively, the distance matrix can be used to construct a minimum spanning tree. There is a possibility to randomize the genetic data by resampling or through sample/allele shuffling to produce any number of reference networks. This procedure allows testing for the significance of various networks statistics.

(3) Network analysis. Some summary statistics of the network such as number of nodes, edges and components, the average degree (a node degree is the number of connections a node has), and the average clustering coefficient (Watts& Strogatz 1998) are produced automatically. A number of network and node properties can be extracted from the network, including the degree distribution, edge weight/distance distribution, clustering coefficient as a function of degree, and average neighbor degree as a function of a degree. If the last function is increasing, nodes of high degree tend to connect to other nodes of high degree and the network is assortative (see, for example, Newman, 2010 for an introduction to the basic topological properties of networks).

(4) Network visualization. The software automatically generates a visualization of the network, optimized for clarity. The resulting network visualization can be customized using an interactive user interface. Node properties such as betweenness centrality (Freeman 1977) or any user given auxiliary attributes can be used to color the nodes, to label them, or to change their size. The network visualization can be saved as an image file (Figure 2).

Some examples of the use of such methods include the definition of biogeographic provinces on the basis of biodiversity inventories in hydrothermal vents (Moalic et al., 2012), the test of hypothesis of ancestral polymorphism versus present day hybridization to explain shared genetic polymorphism between two closely related species (Moalic et al., 2011), or the geographic pattern of genetic differentiation and connectivity among populations (Becheler et al., 2010), including the identification of putative source and pathways areas (Cowart et al., 2013; Rozenfeld et al., 2008). Some of the hypotheses that can be tested with those methods can be generically detailed with the example of the genetic network analysis of *Posidonia oceanica* meadows (Figure 2) in the Mediterranean, based on microsatellite polymorphism (Rozenfeld et al., 2008). The matrix of microsatellites genotype processed through step (1) together with a set of *n* randomizations delivered populations pairwise differences used to (2) build the network (Figure 2) and compare its properties at the percolation threshold to their distributions obtained by randomization. The occurrence of two clusters of populations in Eastern and Western Mediterranean, supported by the departure of the high *clustering* value compared to the range obtained through randomization, allows rejecting the hypothesis of a lack of hierarchical differentiation at the scale of the Mediterranean and supports the existence of at least two clusters of populations. The high and significant *betweeness centrality* (Figure 2) of meadows located in the Siculo-Tunisian Straight permit rejecting the hypothesis of an equivalent role of populations in the gene flow across the system, showing populations located in the Straight contribute more importantly in facilitating or allowing connectivity across the Mediterranean.

Comparison to other software packages

Whereas some functions of EDENetworks have also been implemented in other software packages, it is at the moment the only software containing the entire pipeline from computation of distance matrices to network analysis and visualization and to statistical significance testing. For pure network visualization, the most widely used programs are Pajek, Gephi, and Cytoscape; these also allow for computation of some network characteristics either directly or via plugins. For network analysis by command-line scripting (e.g. in R or Python), there is a number of options such as Networkx and iGraph that however require considerable programming expertise of their users. Additionally, interesting exploration may be envisaged by using network analysis in conjunction with methods based on circuit theory to predict gene flow, such as Circuitscape (McRae 2006).

As mentioned above, a major difference between EDENetworks and existing packages is that it implements the whole analysis pipeline, eliminating the need to use multiple software packages and to transfer files between them. Additionally, instead of attempting to be a general-purpose tool for any network analysis and visualization, EDENetwork focuses on the functions required for analyzing ecological networks, while allowing exporting of network data for further analysis e.g. in iGraph. Note that because of the GUI of EDENetworks, no scripting or programming is required. Further, the analysis pipeline of EDENetworks has elements that are not covered by any existing software package. First, EDENetworks computes distance matrices and networks directly from raw molecular (genotypes or SNPs-presence absence) and ecological (abundance or presence absence) data, using appropriate metrics computed internally by the program itself. Second, because of this, EDENetworks can use random permutations and jackknifing of raw data for null hypothesis testing and inference of the statistical significance of network parameters (clustering, betweeness centrality). Third, EDENetworks has built-in methods for thresholding distance matrices to

networks and computing spanning trees that do not require network-theory expertise from the user. As discussed in the next section, and in detail in the manual, all computations done by EDENetworks for typical data sets are reasonably fast. Further, benchmarking presented in the manual shows that the computation times for most important procedures scale optimally with the data size.

Example data sets

A number of example data sets are distributed with the program. Their earlier analysis and interpretation is detailed in the references listed here. The distance-thresholding method has been used to address biogeography of communities, species hybridization (Moalic et al. 2011) and gene flow among populations (Rozenfeld et al. 2008), and also to study individual relatedness networks (Becheler *et al.* 2010; Moalic *et al.* 2011; Rozenfeld *et al.* 2007). All this research has been performed with similar algorithms and methods as those implemented in EDENetworks, which has later been successfully used to repeat all the relevant analysis in these articles. In addition, one data sets on microbial diversity containing about 40 samples encompassing about 30000 OTUs was successfully analyzed (from Aires et al., 2013). A detailed tutorial of the implemented genetic distances, the flow of analysis guidelines with examples, and warnings of the interpretation of results are included in the EDENetworks manual.

Finally, detailed benchmarking is available in the manual, showing that the computation times for a typical data set are unnoticeable (milliseconds) for most operations, and reasonable (a few seconds) for more demanding procedures such as distance matrix generation and permutations. As an example, one run of the population level pipeline similar to the one presented in Rozenfeld et al. 2008 takes only 50 ms).

Requirements

EDENetworks is freely available at http://www.becs.hut.fi/edenetworks/ with binaries for Windows and Linux systems together with full documentation and example input files. The Windows version comes with an installer and the Linux version is distributed as .dep package. Source code is available at https://github.com/bolozna/EDENetworks. EDENetworks is an open-source (licensed under GPL2) program written entirely in Python, and as such it can be installed to many other systems as long as Python and the third party libraries (Numpy, Matplotlib and Himmeli) it uses are available.

Acknowledgements

We wish to thank all the participants of the EDEN project for great discussion and interesting suggestions. We would like to thank Frédérique Viard for her help on the beta version of the software. MK acknowledges that his contribution was mainly done when he was working at Aalto University.

Funding

This work was supported by EDEN [043251] project funded by European Commission through the NEST-PATHFINDER Call on "Tackling Complexity in Science" of the Sixth Framework Program, and by the ANR project Clonix.

References

Aires T, Serrao EA, Kendrick G, Duarte CM, Arnaud-Haond S (2013) Invasion is a community affair: Clandestine followers in the bacterial community associated to green algae, *Caulerpa racemosa*, track the invasion source. Plos One 8, e68429-e68429.

Albert R, Jeong H, Barabási A-L (2000) Error and attack tolerance of complex networks. *Nature* **406**, 378-382.

Alon U (2003) Biological networks: The tinkerer as an engineer. *Science* **301**, 1866-1867.

Barabási A-L, Albert R (1999) Emergence of scaling in random networks. *Science* **286,** (5439): 509–512.

Barrat A, Barthelemy M, Vespignani A (2009) Dynamical Processes on Complex Networks. *Cambridge University Press.*

Bascompte J, Jordano P, Melian CJ, Olesen JM (2003) The nested assembly of plant-animal mutualistic networks. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 9383-9387.

Bastian M, Heymann S, Jacomy M (2009) Gephi: an open source software for exploring and manipulating networks. *International AAAI Conference on Weblogs and Social Media.*

Batagelj V, Mrvar A (2002) Pajek—analysis and visualization of large networks. *Graph Drawing* **2265**, 477–478.

Becheler R, Diekmann O, Hily C, Moalic Y, Arnaud-Haond S (2010) The concept of population in clonal organisms: mosaics of temporally colonized patches are forming highly diverse meadows of Zostera marina in Brittany. *Molecular Ecology* **19**, 2394-2407.

Dorogovtsev SN, Mendes JFF (2003) Evolution of Networks: From Biological Nets to the Internet and WWW. *Oxford University Press.*

Dos Santos DA, Fernandez HR, Cuezzo MG, Dominguez E (2008) Sympatry inference and network analysis in biogeography. *Systematic Biology* **57**,432–448.

Ferres L, Parush A, Li ZH, Oppacher Y, Lindgaard G (2006) Representing and querying line graphs in natural language: The iGraph system. Smart Graphics, Proceedings 4073, 248-253.

Fortuna MA, Albaladejo RG, Fernandez L, Aparicio A, Bascompte J (2009) Networks of spatial genetic variation across species. *Proceedings of the National Academy of Sciences* **106**, 19044-19049.

Freeman, Linton (1977) A set of measures of centrality based on betweenness. *Sociometry* **40** 35–41.

Hartvigsen G (2011) Using R to Build and Assess Network Models in Biology. *Mathematical Modelling of Natural Phenomena* **6**, 61-75.

Hernández-García E, Herrada EA, Rozenfeld AF*, et al.* (2007) Evolutionary and Ecological Trees and Networks. Nonequilibrium Statistical Mechanics and Nonlinear Physics In: *AIP Conference Proceedings*, pp. pp. 78-83. American Institute of Physics, New York, 2007

Hernandez-Garcia E, Rozenfeld AF, Eguiluz VM, Arnaud-Haond S, Duarte CM (2006) Clone size distributions in networks of genetic similarity. *Physica D* **224**, 166-173

Huson DH, Bryant D (2006) Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol* **23**, 254-267.

Lancichinetti A, Kivelä M, Saramäki J, Fortunato S (2010) Characterizing the community structure of complex networks. PloS One 5(8), e11976

McRae, B.H. (2006) Isolation by resistance. *Evolution* **60** 1551-1561.

Moalic Y, Arnaud-Haond S, Perrin C, Pearson GA, Serrao EA (2011) Travelling in time with networks: Revealing present day hybridization versus ancestral polymorphism between two species of brown algae, Fucus vesiculosus and F. spiralis. *BMC Evolutionary Biology* **11**.

Moalic Y, Desbruyeres D, Duarte CM*, et al.* (2012) Biogeography Revisited with Network Theory: Retracing the History of Hydrothermal Vent Communities. *Systematic Biology* **61**, 127-137.

Newman MEJ (2010) *Networks – An Introduction*. Oxford University Press, Oxford, UK.

Newman MEJ, Strogatz SH, Watts DJ (2001) Random graphs with arbitrary degree distributions and their applications. *Phys. Rev. E* **64**, 026118.

Posada D, Crandall KA (2001) Intraspecific gene genealogies: trees grafting into networks. *Trends in Ecology & Evolution* 16, 37-45.

Proulx SR, Promislow DEL, Phillips PC (2005) Network thinking in ecology and evolution. *Trends in Ecology & Evolution* **20**, 345-353.

Rozenfeld AF, Arnaud-Haond S, Hernández-García E*, et al.* (2007) Spectrum of genetic diversity and networks of clonal populations. *Journal of the Royal Society Interface* **4**, 1093-1102.

Rozenfeld AF, Arnaud-Haond S, Hernandez-Garcia E*, et al.* (2008) Network analysis identifies weak and strong links in a metapopulation system. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 18824-18829.

Smoot M, Ono K, Ruscheinski J, Wang P-L, Ideker T (2011) Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* 27(3): 431-432.

Ueda HR, Hayashi S, Matsuyama S*, et al.* (2004) Universality and flexibility in gene expression from bacteria to human. *Proceedings of the National Academy of Science, USA* **101**, 3765-3769.

Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world' networks. *Nature* **393**, 440-442.

Figures

Figure 1: Schematic overview of the workflow.

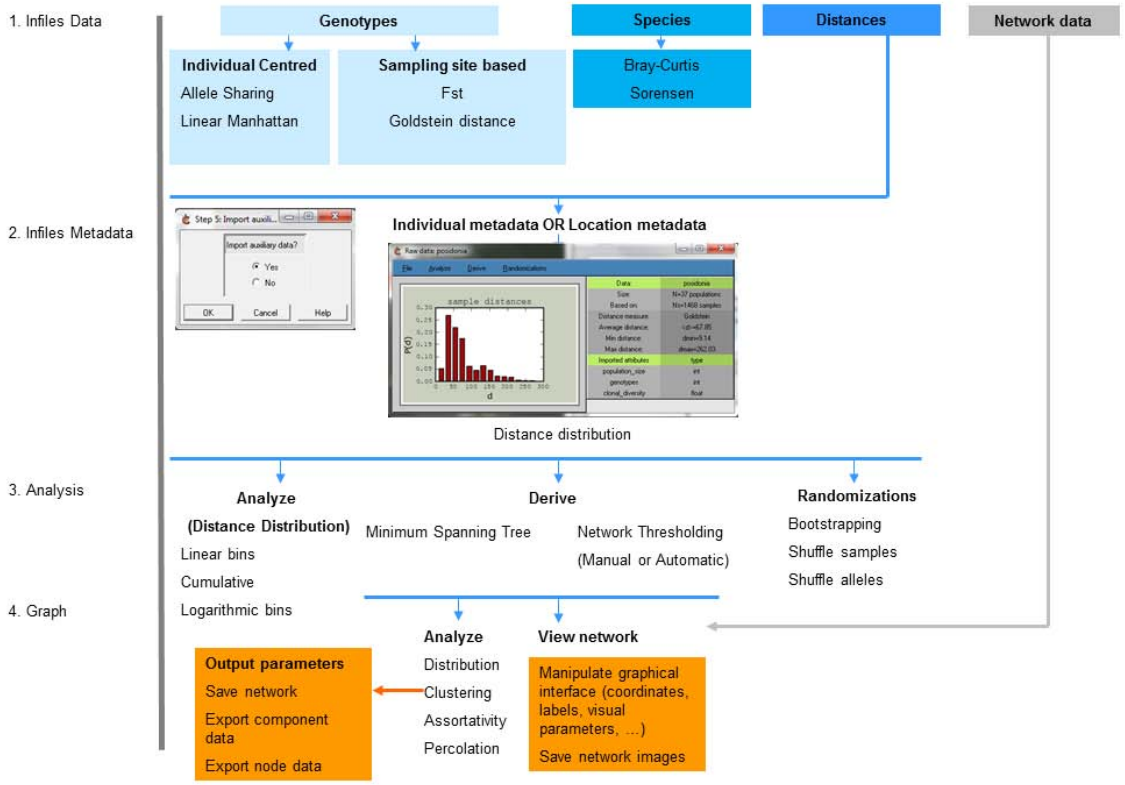
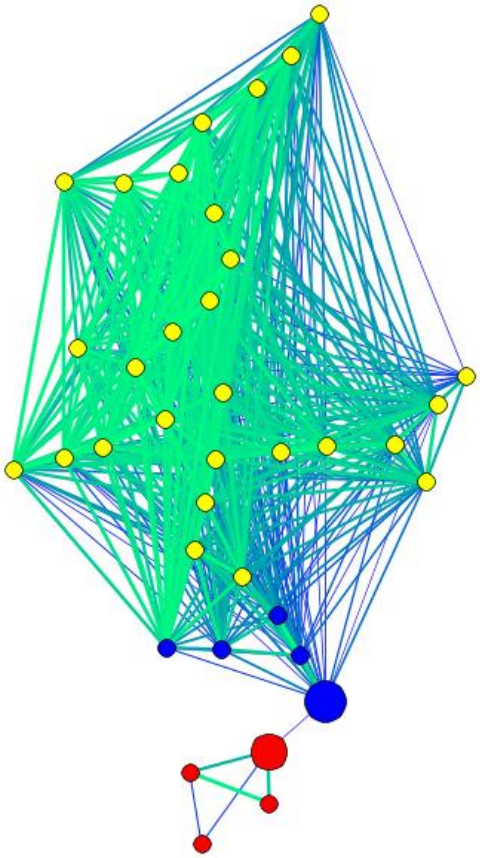


Figure 2: Examples of genetic networks of seagrass (*Posidonia oceanica*). Network nodes represent populations as defined by sampling sites, and edges represent genetic distances. The figure is produced by EDENetworks from a genotype matrix by applying an automatic thresholding algorithm, using data analyzed by Rozenfeld et al. (2008). Node colors (Yellow for Western, red for Central and Blue for Eastern Mediterranean) and sizes represent geographical divisions and *betweenness centrality* values, respectively.



2. Tables

Table 1: Glossary of terms used to describe network topology

Network	set of nodes (or vertices) connected by links (or edges).
Weighted network	a network where a weight is associated with each edge. The weights can e.g. represent genetic similarity.
Neighbour of a node	a node connected to the focal node.
Degree	the number of edges connected to a node, i.e. the number of neighbours.
Path	a sequence of adjacent links.
Shortest path	the path between two nodes that requires traversing the smallest number of links.
Component	a set of nodes where paths exist between all nodes.
Clustering coefficient	the ratio between existing and possible links between a node's neighbours, $c_i=2e_i/[k_i(k_i-1)]$, where e_i = the number of links between neighbours of node i and k_i = degree of i.
Assortativity	the tendency of high-degree nodes to connect to other high-degree nodes. Can be measured e.g. by calculating the Pearson correlation coefficient between degrees of connected nodes.
Betweenness centrality	a measure of the importance of a node (or link) in connecting other nodes through shortest paths. Formally, the fraction of all shortest paths going through a node (or link).
Thresholding	removing links with weights below a given threshold from a weighted network, so that only the most important links are retained.
Percolation threshold	the critical fraction of links that needs to be removed in order to break the network into disconnected components. Often, the composition of these disconnected components is informative.