Software notes

Ecography 38: 001–007, 2015
doi: 10.1111/ecog.00976

Subject Editor: Michael Borregaard. Editor-in-Chief: Miguel Araújo. Accepted 7 April 2015

mangal – making ecological network analysis simple

Timothée Poisot, Benjamin Baiser, Jennifer A. Dunne, Sonia Kéfi, François Massol, Nicolas Mouquet, Tamara N. Romanuk, Daniel B. Stouffer, Spencer A. Wood and Dominique Gravel

T. Poisot (tim@poisotlab.io.) and D. B. Stouffer, Univ. of Canterbury, School of Biological Sciences, Christchurch, New Zealand. TP also at: Dépt de sciences biologiques, Univ. de Montréal, Pavillon Marie-Victorin, C.P. 6128, succ. Centre-ville, Montréal, QC H3C 3J7, Canada. – B. Baiser, Dept of Wildlife Ecology and Conservation, Univ. of Florida, Gainesville, USA. – J. A. Dunne, Sante Fe Inst., 1399 Hyde Park Road, Santa Fe, NM 87501, USA. – S. Kéfi and N. Mouquet, Inst. des Sciences de l'Évolution, Univ. de Montpellier, CNRS, IRD, EPHE, CC065, Place Eugène Bataillon, FR-34095 Montpellier Cedex 05, France. – F. Massol, Laboratoire Génétique et Evolution des Populations Végétales, CNRS UMR 8198, Univ. Lille 1, Bâtiment SN2, FR-59655 Villeneuve d'Ascq cedex, France, and UMR 5175 CEFE – Centre d'Ecologie Fonctionnelle et Evolutive (CNRS), 1919 Route de Mende, FR-34293 Montpellier Cedex 05, France. – T. N. Romanuk, Dept of Biology, Dalhousie Univ., Canada. – S. A. Wood, Natural Capital Project, School of Environmental and Forest Sciences, Univ. of Washington, Seattle, WA 98195, USA, and Dept of Biological Sciences, Idaho State Univ., Pocatello, ID 83209, USA. – D. Gravel, Univ. du Québec à Rimouski, Dépt de Biologie, 300 Allées des Ursulines, Rimouski, QC G5L 3A1, Canada. DG and TP also at: Québec Centre for Biodiversity Sciences, Montréal, QC, Canada.

The study of ecological networks is severely limited by 1) the difficulty to access data, 2) the lack of a standardized way to link meta-data with interactions, and 3) the disparity of formats in which ecological networks themselves are stored and represented. To overcome these limitations, we have designed a data specification for ecological networks. We implemented a database respecting this standard, and released an R package (rmangal) allowing users to programmatically access, curate, and deposit data on ecological interactions. In this article, we show how these tools, in conjunction with other frameworks for the programmatic manipulation of open ecological data, streamlines the analysis process and improves replicability and reproducibility of ecological network studies.

Ecological networks are efficient representations of the complexity of natural communities, and help discover mechanisms contributing to their persistence, stability, resilience, and functioning. Most of the early studies of ecological networks were focused on understanding how the structure of interactions within one location affected the ecological properties of this local community. They revealed the contribution of average network properties, such as the buffering impact of modularity on species loss (Yodzis 1981, Pimm et al. 1991), the increase in robustness to extinctions along with increases in connectance (Dunne et al. 2002), and the fact that organization of interactions maximizes biodiversity (Bastolla et al. 2009). New studies introduced the idea that networks can vary from one locality to another. They can be meaningfully compared, either to understand the importance of environmental gradients on the presence of ecological interactions (Tylianakis et al. 2007), or to understand the mechanisms behind variation itself (Poisot et al. 2012, 2014). Yet, meta-analyses of numerous ecological networks are still extremely rare, and most of the studies comparing several networks do so within the limit of particular systems (Schleuning et al. 2011, Dalsgaard et al. 2013, Poisot et al. 2013b, Chamberlain et al. 2014, Olito and Fox 2015). The severe shortage of publicly shared data in the field also restricts the scope of large-scale analyses.

It is possible to predict the structure of ecological networks, either using latent variables (Rohr et al. 2010, Eklöf et al. 2013) or actual trait values (Gravel et al. 2013). The calibration of these approaches require accessible data, not only about the interactions, but about the traits of the species involved. Comparing the efficiency of different methods is also facilitated if there is a homogeneous way of representing ecological interactions, and the associated metadata. In this paper, we 1) establish the need for a data specification serving as a common language among network ecologists, 2) describe this data specification, and 3) describe rmangal, a R package and companion database relying on this data specification. The rmangal package allows one to easily deposit and retrieve data about ecological interactions and networks in a publicly accessible database. We provide use-cases showing how this new approach makes complex analyses simpler, and allows for the integration of new tools to manipulate biodiversity resources.

Networks need a data specification

Ecological networks are (often) stored as an adjacency matrix (or as the quantitative link matrix), that is a series of 0s and 1s indicating, respectively, the absence or presence of an interaction. This format is extremely convenient (as most network analysis packages, e.g. bipartite, betalink, foodweb, require data to be presented this way), but is extremely inefficient at storing meta-data. In most cases, an adjacency matrix provides information about the identity of species (in the cases where rows and columns headers are present) and the presence or absence of interactions. If other data about the environment (e.g. where the network was sampled) or the species (e.g. the population size, trait distribution, or other observations) are available, they are often either given in other files or as accompanying text. In both cases, making a programmatic link between interaction data and relevant meta-data is difficult and, more importantly, error-prone.

By contrast, a data specification (i.e. a set of precise instructions detailing how each object should be represented) provides a common language for network ecologists to interact, and ensures that, regardless of their source, data can be used in a shared workflow. Most importantly, a data specification describes how data are exchanged. Each group retains the ability to store the data in the format that is most convenient for in-house use, and only needs to provide export options (e.g. through an API, i.e. a programmatic interface running on a web server, returning data in response to queries in a pre-determined language) respecting the data specification. This approach ensures that all data can be used in meta-analyses, and increases the impact of data (Piwowar and Vision 2013). Data archival also offers additional advantages for ecology. The aggregation of local observations can reveal large-scale phenomena (Reichman et al. 2011), which would be unattainable in the absence of a collaborative effort. Data archival in databases also prevents data rot and data loss (Vines et al. 2014), thus ensuring that data on interaction networks – which are typically hard and costly to produce – continue to be available and usable.

Elements of the data specification

The data specification introduced here (Fig. 1) is built around the idea that (ecological) networks are collections of relationships between ecological objects, and each element has particular meta-data associated with it. In this section, we detail the way networks are represented in the mangal specification. An interactive webpage with the elements of the data specification can be found online at <http://mangal.io/doc/spec/>. The data specification is available either at the API root (e.g. <http://mangal.io/api/v1/?format=json>), or can be viewed using the whatIs function from the rmangal package. Rather than giving an exhaustive list of the data specification (which is available online at the aforementioned URL), this section serves as an overview of each element, and how they interact.

We propose JSON, a user-friendly format equivalent to XML, as an efficient way to standardise data representation for two main reasons. First, it has emerged as a de facto standard for web platform serving data, and accepting data from users. Second, it allows strict validation of the data: a JSON file can be matched against a scheme, and one can verify that it is correctly formatted (this includes the possibility that not all fields are filled, as will depend on available data). Finally, JSON objects are easily and cheaply (memory-wise) parsed in the most commonly-used programming languages, notably R (equivalent to list) and python (equivalent to dict). For most users, the format in which data are transmitted is unimportant, as the interaction happens within R – as such, knowing how JSON objects are organized is only useful for those who want to interact with the API directly. As such, the rmangal package takes care of converting the data into the correct JSON format to upload them in the database.

Functions in the rmangal package are names after elements of the data specification, in the following way: verb + Element. verb can be one of list, get, or patch; for example, the function to get a particular network is getNetwork. The function to modify (patch) a taxon is patchTaxa. All of these functions return a list object, which means that chaining them together using, e.g. the plyr package, is time-efficient. There are examples of this in the use-cases.

Node information

Taxa

Taxa are a taxonomic entity of any level, identified by their name, vernacular name, and their identifiers in a variety of

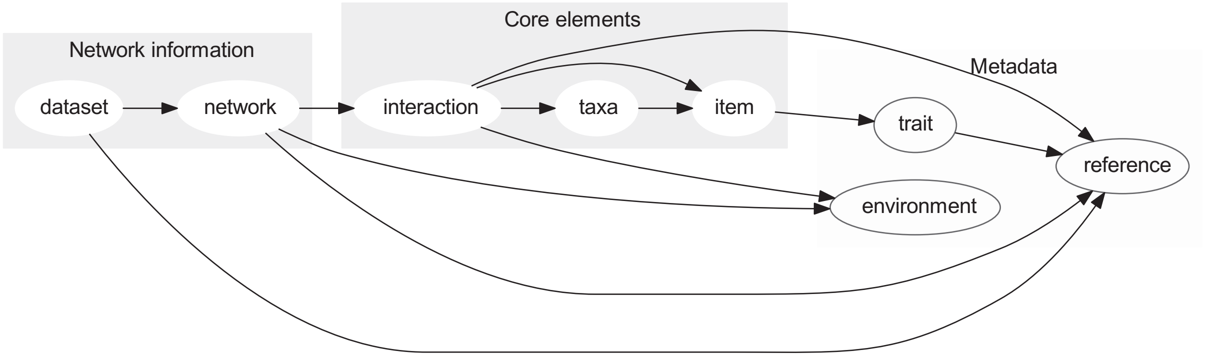

Figure 1. An overview of the data specification, and the hierarchy between objects. Every box corresponds to a level of the data specification. Grey boxes are nodes, blue boxes are interactions and networks, and green boxes are metadata. The bold boxes (dataset, network, interaction, taxa) are the minimal elements needed to represent a network.

taxonomic services. Associating the identifiers of each taxa allows using the new generation of open data tools, such as taxize (Chamberlain and Szöcs 2013), in addition to protecting the database against taxonomic revisions. The data specification currently has fields for NCBI (National Center for Biotechnology Information), GBIF (Global Biodiversity Information Facility), TSN (Taxonomic Serial Number, used by the Integrated Taxonomic Information System), EOL (Encyclopedia of Life) and BOLD (Barcode of Life) identifiers. We also provide the taxonomic status, i.e. whether the taxon is a true taxonomic entity, a 'trophic species', or a morphospecies. Taxonomic identifiers can either be added by the contributors, or will be automatically retrieved during the automated curation routine.

Item

An item is any measured instance of a taxon. Items have a level argument, which can be either individual or population; this allows representing both individual-level networks (i.e. there are as many items of a given taxa as there were individuals of this taxon sampled), and population-level networks. When item represents a population, it is possible to give a measure of the size of this population. The notion of item is particularly useful for time-replicated designs: each observation of a population at a time-point is an item with associated trait values, and possibly population size.

Network information

All objects described in this sub-section can have a spatial position, information on the date of sampling, and references to both papers and datasets.

Interaction

An interaction links two taxa objects (but can also link pairs of items). The most important attributes of interactions are the type of interaction (of which we provide a list of possible values), and its obs_type, i.e. how it was observed. This field helps differentiate direct observations, text mining, and inference. Note that the obs_type field can also take confirmed absence as a value; this is useful for, e.g. 'cafeteria' experiments in which there is high confidence that the interaction did not happen.

Network

A network is a series of interaction objects, along with 1) information on its spatial position (provided at the latitude and longitude), 2) the date of sampling, and 3) references to measures of environmental conditions.

Dataset

A dataset is a collection of one or several network(s). Datasets also have a field for data and papers, both of which are references to bibliographic or web resources that describe, respectively, the source of the data and the papers in which these data have been studied. Datasets or networks are the preferred entry point into the resources, although in some cases it can be meaningful to get a list of interactions only.

Meta-data

Trait value

Objects of type item can have associated trait values. These consist in the description of the trait being measured, the value, and the units in which the measure was taken. As the measurment was taken at a different time and/or location that the interaction was, they have fields for time, latitude and longitude, and references to original publication and original datasets.

Environmental condition

Environmental conditions are associated with datasets, networks, and interactions objects, to allow for both macro and micro environmental conditions. These are defined by the environmental property measured, its value, and the units. As traits, they have fields for time, latitude and longitude, and references to original publication and original datasets.

References

References are associated with datasets. They accommodate the DOI, JSON or PubMed identifiers, or a URL. When possible, the DOI is preferred as it offers more potential to interact with other online tools, such as the CrossRef API.

Use cases

In this section, we present use-cases using the rmangal package for R, to interact with a database implementing this data specification, and to serve data through an API (<http://mangal.io/api/v1/>). It is possible for users to deposit data into this database through the R package. Note that data are made available under a CC-0 Waiver (Poisot et al. 2013a). Detailed information about how to upload data are given in the vignettes and manual of the rmangal package. In addition, the rmangal package comes with vignettes explaining how users can upload their data into the database through R.

The data we use for this example come from Ricciardi et al. (2010). These data were previously available on the InteractionWeb DataBase as a single xls file. We uploaded them in the mangal database at <http://mangal.io/api/v1/dataset/2>. The rmangal package can be installed this way:

```
# Prepare the environment
library(devtools)
# This line is needed on some linux distributions
if(getOption('unzip')=='') options ('unzip' = 'unzip')
# This installs the rmangal package
install_github('mangal-wg/rmangal')
library(rmangal)
```

Once rmangal is installed and loaded, users can establish a connection to the database this way:

```
mangal_url <-'http://mangal.io/'
api <-mangalapi(mangal_url)
```

Create taxa and add an interaction

In the first use-case, we will create an interaction between two taxa. We ask of readers not to execute this code as it is, but rather to use it as a template for their own analyses. A complete, step-by-step guide to upload is given in the vignettes of the rmangal package. Uploading anything requires an username and API key, which can be obtained at the following URL: <http://mangal.io/dashboard/login>. Your API key be generated automatically after registration. You can use it to connect to the database securely:

```
api_secure <- mangalapi("http://mangal.io", usr="MyUserName",
key="AbcDefIjkL1234")
```

The first step is to create two taxa objects, with the species that we observed interacting:

```
seal <- list(
    name = "Hydrurgaleptonix",
    vernacular = "Leopardseal",
    eol = 328637
)
cod <- list(
    name = "Gadusmorhua",
    vernacular = "Atlanticcod"
)
```

Now, we will send these two objects to the remote database:

```
seal <- addTaxa(api_secure, seal)
cod <- addTaxa(api_secure, cod)
```

Note that it is suggested to overwrite the local copy of the object, because the database will always send back the remote copy. This makes the syntax of further addition considerably easier, as we show below. Once the two objects are created, we can create an interaction between them:

```
seal_eats_cod <- list(
    taxa_from = seal,
    taxa_to = cod,
    int_type = "predation",
    obs_type = "observed"
)
```

Then using the same approach, we can send this information in the remote database:

```
seal_eats_cod <- addInteraction(api_secure, seal_eats_cod)
```

To create networks, datasets, etc, one needs to follow the same procedure, as is explained in the online guide for data contributors, available at <http://mangal.io/doc/upload/>.

Link–species relationships

In the first example, we visualize the relationship between the number of species and the number of interactions, which Martinez (1992) proposed to be linear (in food webs).

```
library(plyr)
library(igraph)
# Retrieve the dataset of interest
dataset <-getDataset(api, 2)
# Get each network in the dataset as a graph object
graphs <- alply(dataset$networks, 1, function(x) toIgraph(api, x))
# Make a data.frame with the number of links and species
ls <-ldply(graphs, function(x) c(S = length(V(x)), L = length(E(x))))
ls$X1 <-aaply(as.numeric(as.vector(dataset$networks)), 1,
               function(x) getNetwork(api, x)$name)
colnames(ls)[1] <- 'Network'
# Now plot this dataset
source("suppmat/usecase_ls.r")
```

Getting the data to produce Fig. 2 requires less than 10 lines of code. The only information needed is the identifier of the network or dataset, which we suggest should be reported in publications as: 'these data were deposited in the mangal format at <URL>/api/v1/dataset/<ID>' (where <URL> and <ID> are replaced by the corresponding values), preferably in the methods, possibly in the acknowledgements. To encourage data sharing and its recognition, we encourage users of the database to always cite the original datasets or publications.

Network beta-diversity

In the second example, we use the framework of network β-diversity (Poisot et al. 2012) to measure the extent to

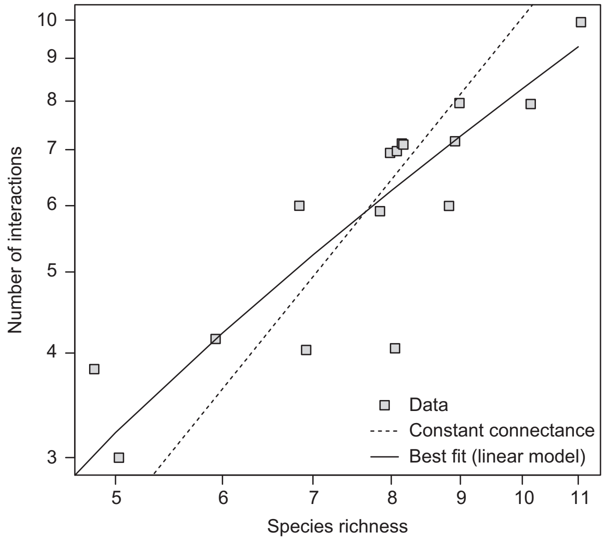

Figure 2. Relationship between the number of species and number of interactions in the anemonefish-fish dataset. Constant connectance refers to the hypothesis that there is a quadratic relationship between these two quantities.

which networks that are far apart in space have different interactions. Each network in the dataset has a latitude and longitude, meaning that it is possible to measure the geographic distance between two networks. For each pair of networks, we measure the geographic distance (in km), the species dissimilarity (β_s), the network dissimilarity when all species are present (β_{WN}), and finally, the network dissimilarity when only shared species are considered (β_{os}).

```
# We need the betalink package to measure network beta-diversity
install_github('PoisotLab/betalink')
library(betalink)

library(plyr)
library(igraph)
library(sp)
# We first retrieve all information about the networks
Networks <- alply(dataset$networks, 1, function(x) getNetwork(api,x))
# Extract the lat/lon data
LatLon <- ldply(Networks, function(x) c(name = x$name, lat =
x$latitude, lon = x$longitude))
rownames (LatLon) <- LatLon$name
LatLon$lat <- as.numeric(LatLon$lat)
LatLon$lon <- as.numeric(LatLon$lon)
LatLon <- LatLon[,c('lat', 'lon')]
# Then we measure the distances between all pairs of sites
GeoDist <- spDists(as.matrix(LatLon, latlon=TRUE))
colnames(GeoDist) <- rownames(GeoDist) <- rownames(LatLon)
GeoDist <- as.dist(GeoDist)
# Now, we measure the beta-diversity of the networks
names(graphs) <- aaply(names(graphs), 1, function(x)
Networks[[x]]$name)
# Finally, we measure the beta-diversity
BetaDiv <- network_betadiversity(graphs)
# We add the geographic distance
BetaDiv$GEO <- GeoDist
# Plotting
source("suppmat/usecase_beta.r")
```

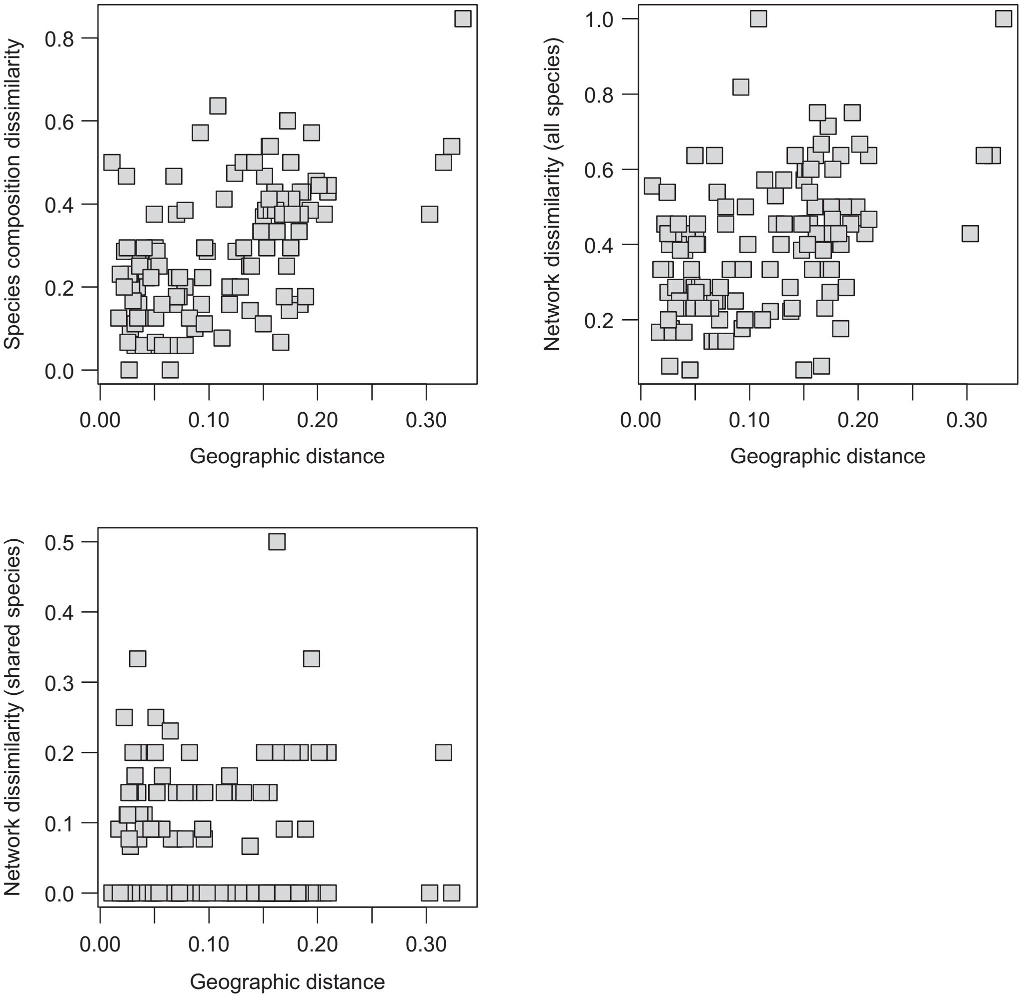

Figure 3. Relationships between the geographic distance between two sites, and the species dissimilarity, network dissimilarity with all species, and network dissimilarity with only shared species.

As shown in Fig. 3, while species dissimilarity and overall network dissimilarity increase when two networks are far apart, this is not the case for the way common species interact. This suggests that in this system, network dissimilarity over space is primarily driven by species turnover. The ease to gather both raw interaction data and associated meta-data make conducting this analysis extremely straightforward.

Spatial visualization of networks

Bascompte (2009) uses an interesting visualization for spatial networks, in which each species is laid out on a map at the center of mass of its distribution; interactions are then drawn between species to show how species distribution determines biotic interactions. In this final use case, we propose to reproduce a similar figure (Fig. 4).

```r
library(maps)
library(mapdata)
library(RColorBrewer)
library(sp)
library(plyr)
library(igraph)
# We fill a community data matrix
sp_by_site <- llply(graphs, function(x) unlist(V(x)$name))
sp_list <- unique(unlist(sp_by_site))
M <- matrix(0, ncol = length(sp_list), nrow = length(sp_by_site))
colnames(M) <- sp_list
rownames(M) <- names(sp_by_site)
for (site in c(1:length(sp_by_site))) M[names(sp_by_site)[site],
sp_by_site[[site]]] = 1
# Next, we get the center position for each species
# (i.e. the mean position of the sites it occurs at)
sp_center <- adply(M, 2, function(x) colMeans(LatLon[names(x)[x>0],
]))
rownames(sp_center) <- sp_center[, 1]
sp_center <-sp_center[, -1]
# We now create a regional network using betalink::metaweb
Mw <- metaweb(graphs)
# Plotting
source("suppmat/usecase_map.r")
```

Conclusions

The mangal data format will allow researchers to put together datasets with species interactions and rich meta-data that are needed to address emerging questions about the structure of ecological networks. We deployed an online database with an associated API relying on this data specification. Finally, we introduced rmangal, an R package designed to interact with APIs using the mangal format. We expect that the data specification will evolve based on the needs and feedback of the community. At the moment, users are welcome to propose such changes on

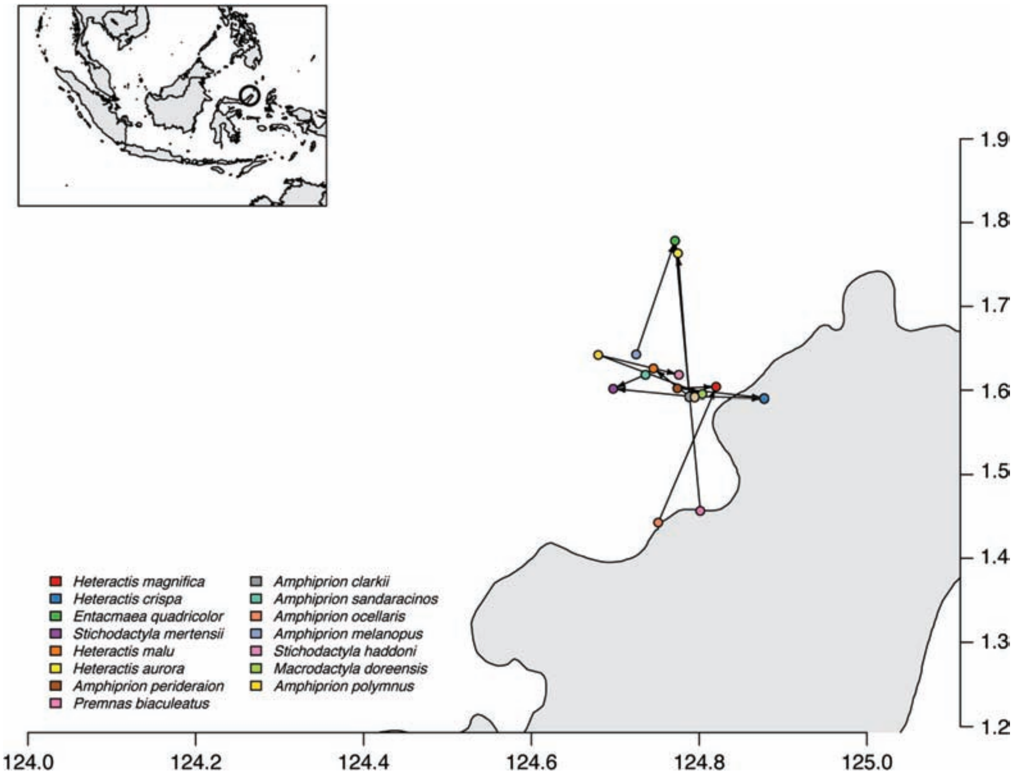

Figure 4. Spatial plot of a network, using the maps and rmangal packages. The circles in the inset map show the location of the sites. Each dot in the main map represents a species, with symbiotic mutualisms drawn between them. The land is in grey.

the project issue page: <https://github.com/mangal-wg/mangal-schemes/issues>. A python wrapper for the API is also available at <http://github.com/mangal-wg/pymangal/>. Additionally, there are plans to integrate this database with GLOBI, so that data can be accessed from multiple sources (Poelen et al. 2014).

To cite the mangal.io, or acknowledge its use, please cite this software note. To cite the rmangal R package, or acknowledge its use, please cite:

Poisot, T. et al. rmangal (ver. 1.0.1). – doi: 105281/zenodo.16998

Acknowledgements – This paper was developed during a workshop hosted at the Santa Fe Institute. TP, DBS, and DG acknowledge funding from the Canadian Inst. of Ecology and Evolution. We thank Scott Chamberlain and one anonymous reviewer for comments on the manuscript. TP is funded by a start-up grant from the Univ. de Montréal. We thank the rOpenSci team and developers for inspiration.

References

Bascompte, J. 2009. Disentangling the Web of Life. – Science 325: 416–419.

Bastolla, U. et al. 2009. The architecture of mutualistic networks minimizes competition and increases biodiversity. – Nature 458: 1018–1020.

Chamberlain, S. A. and Szöcs, E. 2013. taxize: taxonomic search and retrieval in R. – F1000Research in press.

Chamberlain, S. A. et al. 2014. Traits and phylogenetic history contribute to network structure across Canadian plant–pollinator communities. – Oecologia: 1–12.

Dalsgaard, B. et al. 2013. Historical climate-change influences modularity and nestedness of pollination networks. – Ecography 36: 1331–1340.

Dunne, J. A. et al. 2002. Network structure and biodiversity loss in food webs: robustness increases with connectance. – Ecol. Lett. 5: 558–567.

Eklöf, A. et al. 2013. The dimensionality of ecological networks. – Ecol. Lett. 16: 577–583.

Gravel, D. et al. 2013. Inferring food web structure from predator–prey body size relationships. – Methods Ecol. Evol. 4: 1083–1090.

Martinez, N. D. 1992. Constant connectance in community food webs. – Am. Nat. 139: 1208–1218.

Olito, C. and Fox, J. W. 2015. Species traits and abundances predict metrics of plant–pollinator network structure, but not pairwise interactions. – Oikos 124: 428–436.

Pimm, S. L. et al. 1991. Food web patterns and their consequences. – Nature 350: 669–674.

Piwowar, H. A. and Vision, T. J. 2013. Data reuse and the open data citation advantage. – PeerJ 1: e175.

Poelen, J. H. et al. 2014. Global biotic interactions: an open infrastructure to share and analyze species-interaction datasets. – Ecol. Inform. in press.

Poisot, T. et al. 2012. The dissimilarity of species interaction networks. – Ecol. Lett. 15: 1353–1361.

Poisot, T. et al. 2013a. Moving toward a sustainable ecological science: don't let data go to waste! – Ideas Ecol. Evol. 6: e4632.

Poisot, T. et al. 2013b. Facultative and obligate parasite communities exhibit different network properties. – Parasitology 140: 1340–1345.

Poisot, T. et al. 2014. Beyond species: why ecological interaction networks vary through space and time. – Oikos 124: 243–251.

Reichman, O. J. et al. 2011. Challenges and opportunities of open data in ecology. – Science 331: 703–705.

Ricciardi, F. et al. 2010. Assemblage and interaction structure of the anemonefish – anemone mutualism across the Manado region of Sulawesi, Indonesia. – Environ. Biol. Fish. 87: 333–347.

Rohr, R. P. et al. 2010. Modeling food webs: exploring unexplained structure using latent traits. – Am. Nat. 176: 170–177.

Schleuning, M. et al. 2011. Specialization and interaction strength in a tropical plant–frugivore network differ among forest strata. – Ecology 92: 26–36.

Tylianakis, J. M. et al. 2007. Habitat modification alters the structure of tropical host–parasitoid food webs. – Nature 445: 202–205.

Vines, T. H. et al. 2014. The availability of research data declines rapidly with article age. – Curr. Biol. 24: 94–97.

Yodzis, P. 1981. The stability of real ecosystems. – Nature 289: 674–676.