`mangal` – making complex ecological network analysis simpler

Timothée Poisot et al. *(see complete author list below)*

**Author list: **Timothée Poisot (1,2,)*, Benjamin Baiser (3), Jennifer A. Dunne (4,5), Sonia Kéfi (6), François Massol (7,8), Nicolas Mouquet (6), Tamara N. Romanuk (9), Daniel B. Stouffer (10), Spencer A. Wood (11,12), Dominique Gravel (1,2)

1. Université du Québec à Rimouski, Département de Biologie, 300 Allées des Ursulines, Rimouski (QC) G5L 3A1, Canada

2. Québec Centre for Biodiversity Sciences, Montréal (QC), Canada

3. Department of Wildlife Ecology and Conservation, University of Florida, Gainesville

4. Sante Fe Institute, 1399 Hyde Park Road, Santa Fe NM 87501

5. Pacific Ecoinformatics and Computational Ecology Lab, 1604 McGee Ave., Berkeley, CA 94703

6. Institut des Sciences de l'Évolution, UMR CRNS 5554, Université Montpellier 2, 3405 Montpellier, France

7. Laboratoire Génétique et Evolution des Populations Végétales, CNRS UMR 8198, Université Lille 1, Bâtiment SN2, F-59655 Villeneuve d'Ascq cedex, France

8. UMR 5175 CEFE – Centre d'Ecologie Fonctionnelle et Evolutive (CNRS), 1919 Route de Mende, F-34293 Montpellier cedex 05, France

9. Department of Biology, Dalhousie University

10. University of Canterbury, School of Biological Sciences, Christchurch, New Zealand

11. Natural Capital Project, School of Environmental and Forest Sciences, University of Washington, Seattle, WA 98195, USA

12. Department of Biological Sciences, Idaho State University, Pocatello, ID 83209, USA

Author for correspondence: t.poisot@gmail.com.

Short title: Automated retrieval of ecological networks data

Keywords: R, API, database, open data, ecological networks, species interactions

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

Introduction

Ecological networks enable ecologists to accommodate the complexity of natural communities, and to 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. Such analyses revealed the contribution of 'average' network properties, such as the buffering impact of modularity on species loss (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). More recently, new studies introduced the idea that networks can vary from one realization to another. They can be meaningfully compared, either to understand the importance of environmental gradients on the realization of ecological interactions [@tylianakis_habitat_2007], or to understand the mechanisms behind variation in the structure of ecological networks (Poisot *et al.* 2012). Yet, meta-analyses of a large number of 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). The severe shortage of data in the field also restricts the scope of large-scale analyses.

An increasing number of approaches are being put forth to *predict* the structure of ecological networks, either relying on latent variables (Rohr *et al.* 2010) or actual traits (Gravel *et al.* 2013). Such approaches, so as to be adequately calibrated, require easily accessible data. Comparing the efficiency of different methods is also facilitated if there is an homogeneous way of representing ecological interactions, and the associated metadata. In this paper, we (i) establish the need of a data specification serving as a *lingua franca* among network ecologists, (ii) describe this data specification, and (iii) describe `rmangal`, a `R` package and companion database relying on this data specification. The `rmangal` package allows to easily retrieve, but also deposit, ecological interaction networks data from a database. We provide some use cases showing how this new approach makes complex analyzes 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 `0` and `1` indicating, respectively, the absence and presence of an interaction. This format is extremely convenient for *use* (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 informs on the identity of species (in 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 most 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 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 ensure 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 webserver, 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 *et al.* 2007; Piwowar & Vision 2013).

Elements of the data specification

The data specification (Fig. 1) is built around the idea that (ecological) networks are collections of relationships between ecological objects, each element having particular meta-data associated. 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.uqar.ca./doc/spec/`. The data specification is available either at the API root (*e.g.* `http://mangal.uqar.ca/api/v1/?format=json`), or can be viewed using the `whatIs` function from the `R` package (see *Supp. Mat. 1*). 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 format equivalent to `XML`, as an efficient way to uniformise 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 *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 common 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. The `rmangal` package takes care of converting the data into the correct `JSON` format to upload them in the database.

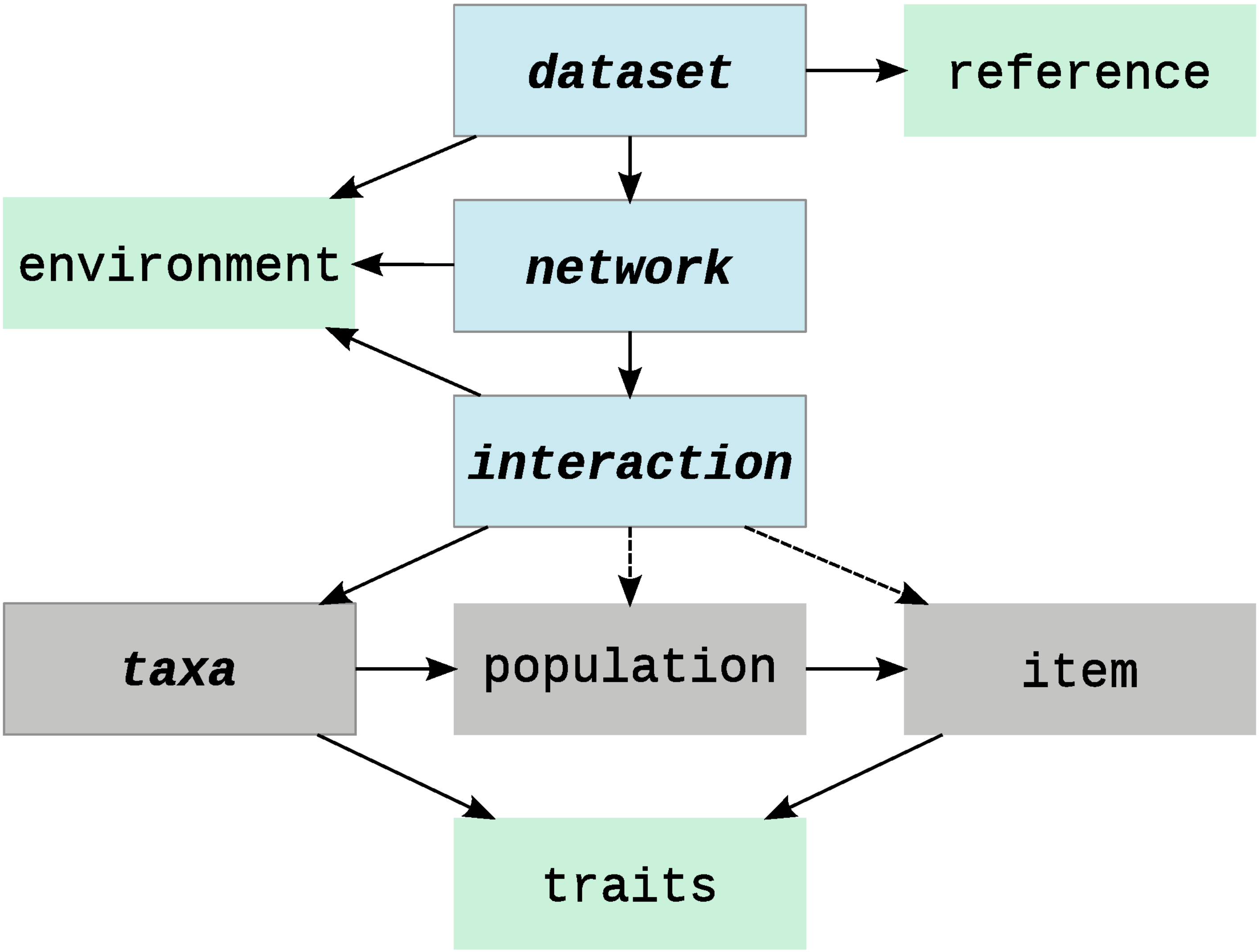

Fig. 1: An overview of the data specification, and the hierarchy between objects. Each box correspond to a level of the data specification. Grey boxes are nodes, blue boxes are interactions and networks, and green boxes are meta-data. The **bold** boxes (`dataset`, `network`, `interaction`, `taxa`) are the minimal elements needed to represent a network.

Node information

Taxa

Taxa are a taxonomic entity of any level, identified by their name, vernacular name, and their identifiers in a variety of taxonomic services. Associating the identifiers of each taxa is important to leverage the power of the new generation of open data tools, such as `taxize` [@chamberlain_taxize_2013]. The data specification currently has fields for `ncbi`, `gbif`, `itis`, `eol` and `bold` identifiers. We also provide the taxonomic status, *i.e.* whether a taxa is a true taxonomic entity, a "trophic species", or a morphospecies.

Population

A `population` is one observed instance of a `taxa` object. If your experimental design is replicated through space, then each taxa have a `population` object corresponding to each locality. Populations do not have associated meta-data, but serve as "containers" for `item` objects.

Item

An `item` is an instance of a population. Items have a `level` argument, which can be either `individual` or `population`; this allows to represent both individual-level networks (*i.e.* there are as many `items` attached to a `population` than there were individuals of this `population` 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

Interaction

An `interaction` links, *a minima*, two `taxa` objects (but can also link pairs of `populations` or `items`). The most important attributes of `interactions` are the type of interaction (of which we provide a list of possible values, see *Supp. Mat. 1*), and its `nature`, *i.e.* how it was observed. This field help differentiate direct observations, text mining, and inference. Note that the `nature` field can also take `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` object, along with (i) informations on its spatial position (provided at the latitude and longitude), (ii) the date of sampling, and (iii) 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 describing, respectively, the source of the data, and the papers in which these data have been significantly used. Datasets are the preferred entry point in the resources.

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.

Environmental condition

Environmental conditions are associated to 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.

References

References are associated to datasets. They accommodate the DOI, JSON or PubMed identifiers, or a URL. When possible, the DOI should be preferred as it offers more potential to interact with other on-line 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 serving data through an API (`http://mangal.uqar.ca/api/v1/`). It is possible for users to deposit data into this database, through the `R` package. Data are made available under a *CC-0 Waiver* (Poisot *et al.* 2013). Detailed informations about how to upload data are given in the vignettes and manual of the `rmangal` package. So as to save room in the manuscript, we source each example; the complete `r` files to reproduce the examples of this section are attached as *Suppl. Mat.*. 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 were previously available on the *Interaction-Web DataBase* as a single `xls` file. We uploaded them in the `mangal` database at `http://mangal.uqar.ca/api/v1/dataset/1`.

Link-species relationships

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

```r
source("usecases/1_ls.r")
```

Producing this figure 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>`", possibly in the acknowledgements. So as to encourage data sharing, we encourage users of the database to cite the original dataset or publication.

Network beta-diversity

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

```r
source("usecases/2_beta.r")
```

As shown in *Fig. XX*, 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 producing this analysis extremely straightforward.

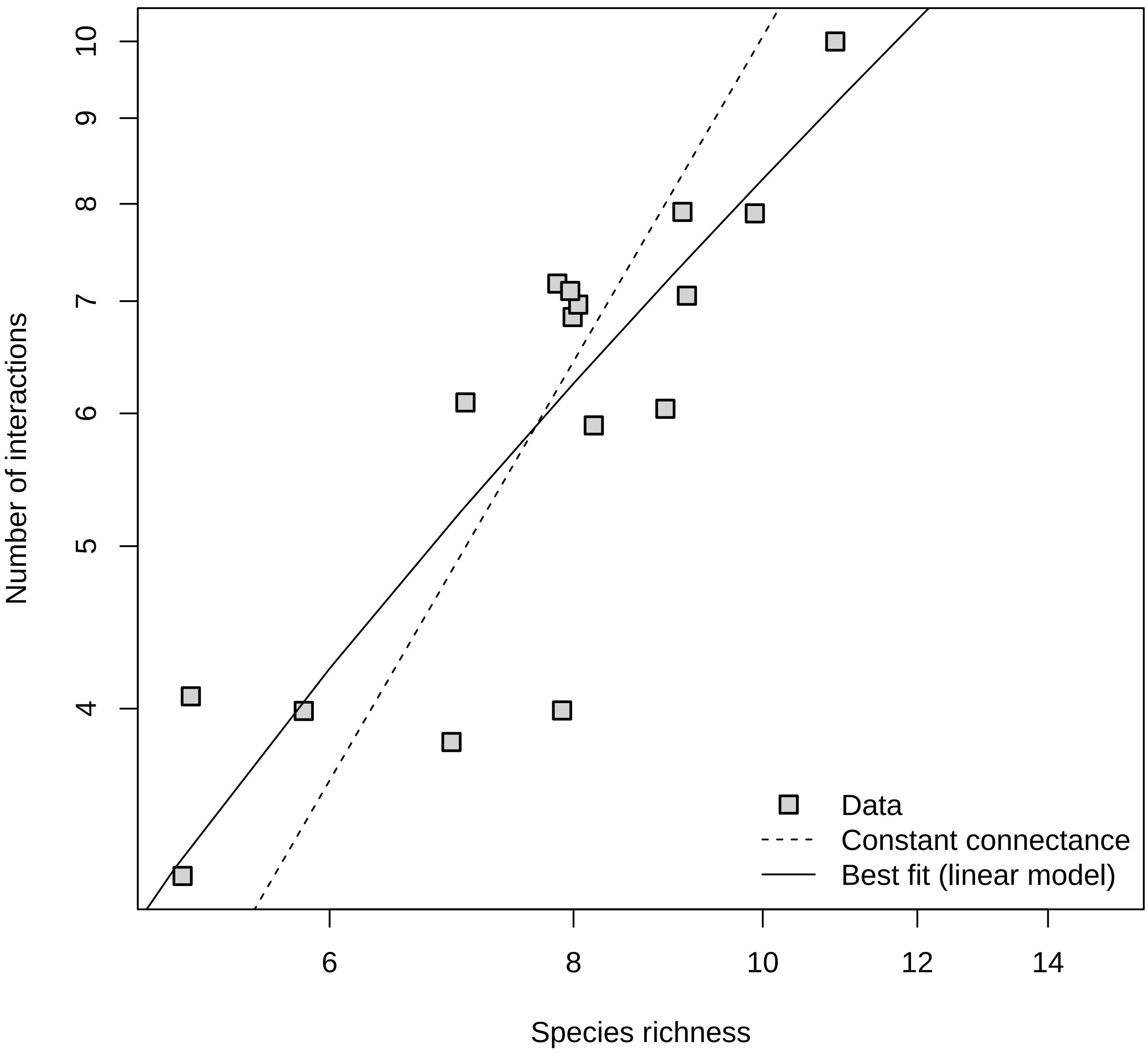

Fig. 2: Relationship between the number of species and number of interactions in the anemonefish-fish dataset.

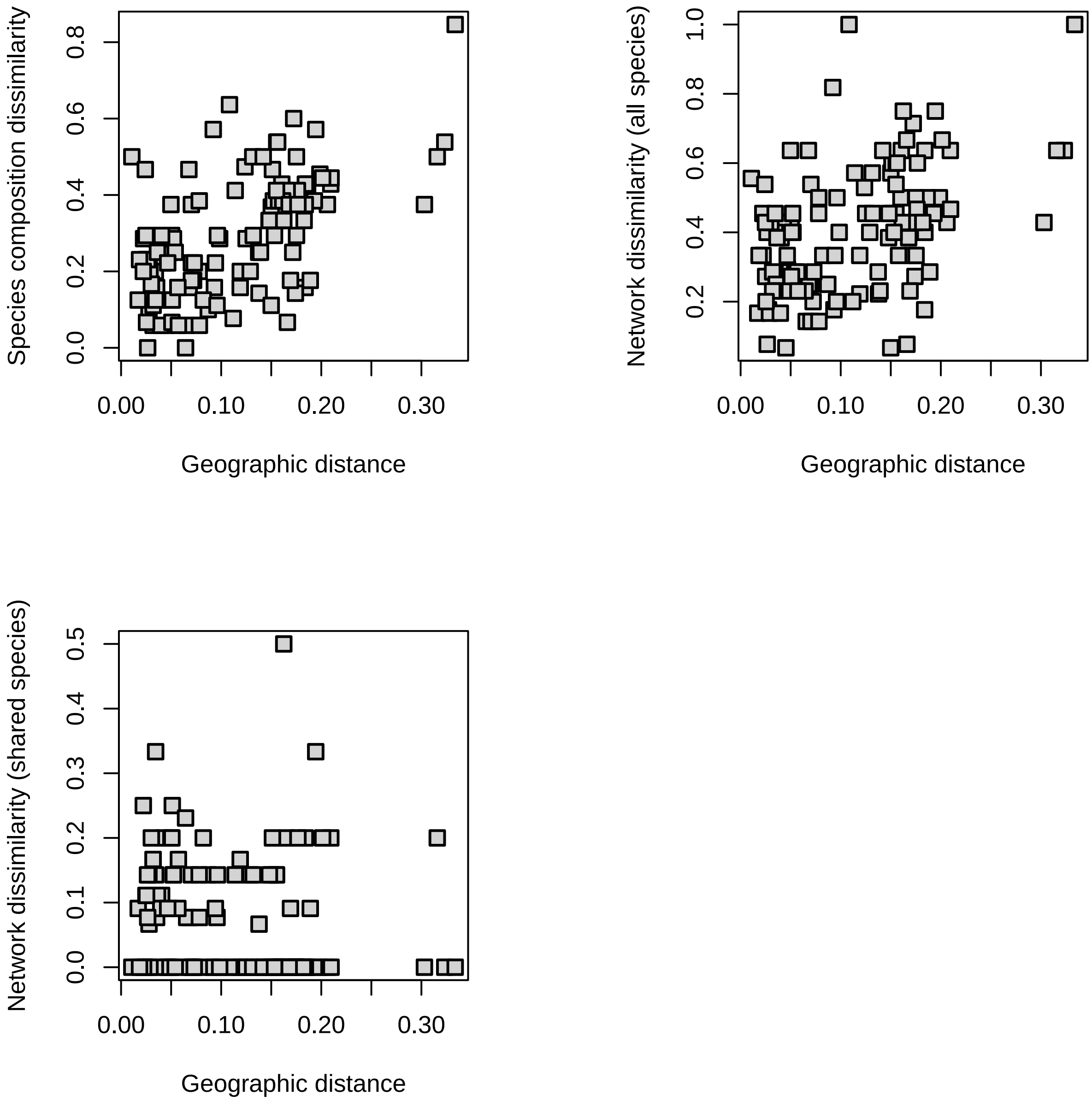

Fig. 3: Relationships between the geographic distance between two sites, and the species dissimilarity, network dissimilarity with all, and only shared, species.

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.

```r
source("usecases/3_spatial.r")
```

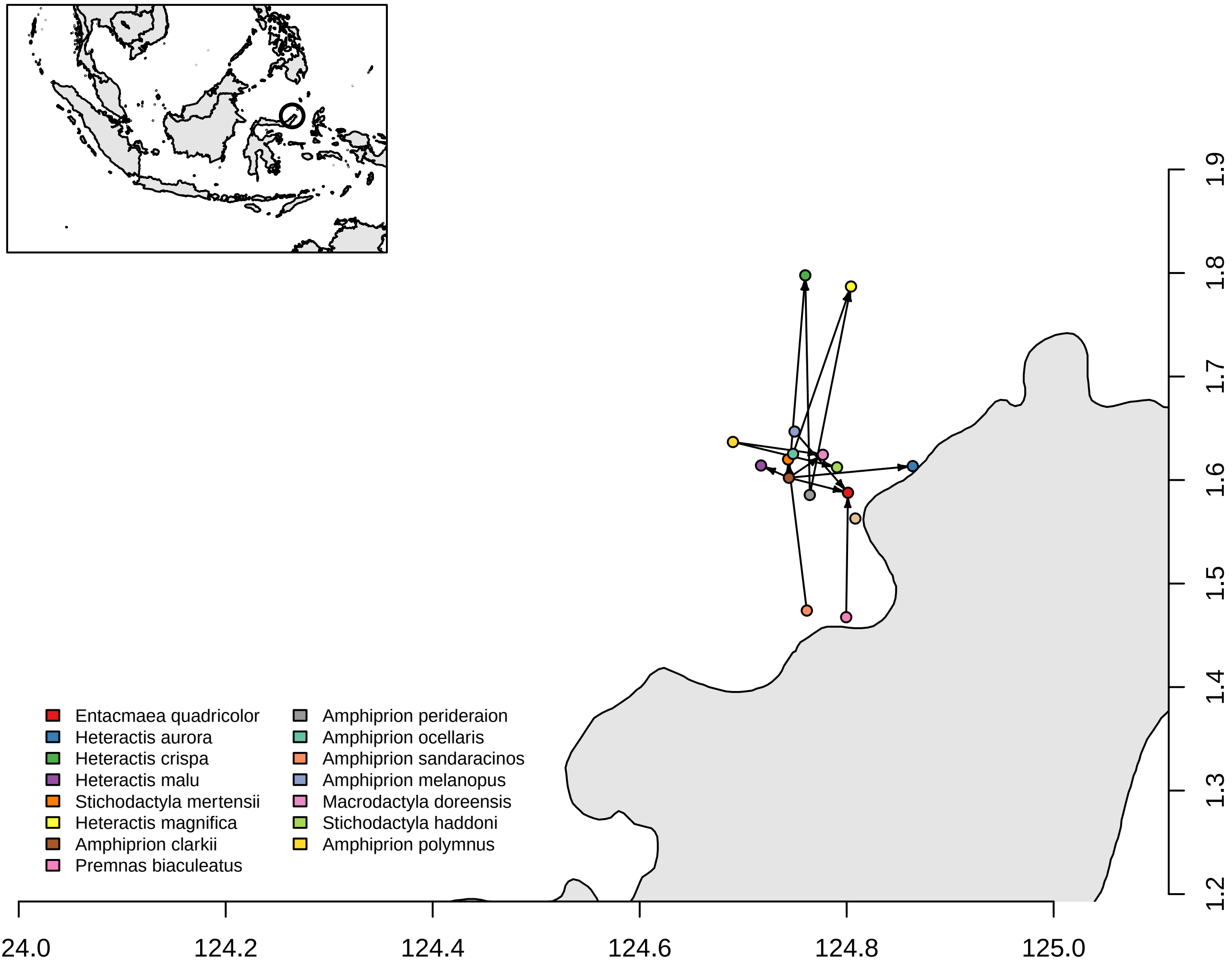

Fig. 4: Spatial plot of a network, using the `maps` and `rmangal` packages. The circle in the inset map show the location of the sites. Each dot in the main map represents a species, with interactions drawn between them.

Conclusions

In this contribution, we presented mangal, a data format for the exchange of ecological networks and associated metadata. We deployed an online database with an associated API, relying on this data specification. Finally, we introduced rmangal, a R package designed to interact with APIs using the mangal format. We expect that the data specification will evolve based on the needs of the community. At the moment, users are welcome to propose such changes on 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/.

References

Bascompte, J. (2009). Disentangling the web of life. *Science (New York, N.Y.)*, **325**, 416–9.

Bastolla, U., Fortuna, M.A., Pascual-García, A., Ferrera, A., Luque, B. & Bascompte, J. (2009). The architecture of mutualistic networks minimizes competition and increases biodiversity. *Nature*, **458**, 1018–1020.

Dalsgaard, B., Trøjelsgaard, K., González, A.M.M., Nogués-Bravo, D., Ollerton, J., Petanidou, T., Sandel, B., Schleuning, M., Wang, Z., Rahbek, C., Sutherland, W.J., Svenning, J.-C. & Olesen, J.M. (2013). Historical climate-change influences modularity and nestedness of pollination networks. *Ecography*, no–no. Retrieved May 14, 2013,

Dunne, J.A., Williams, R.J. & Martinez, N.D. (2002). Network structure and biodiversity loss in food webs: robustness increases with connectance. *Ecology Letters*, **5**, 558–567.

Gravel, D., Poisot, T., Albouy, C., Velez, L. & Mouillot, D. (2013). Inferring food web structure from predator-prey body size relationships. *Methods in Ecology and Evolution*.

Martinez, N.D. (1992). Constant connectance in community food webs. *The American Naturalist*, **139**, 1208–1218.

Pimm, S.L., Lawton, J.H. & Cohen, J.E. (1991). Food web patterns and their consequences. *Nature*, **350**, 669–674.

Piwowar, H.A. & Vision, T.J. (2013). Data reuse and the open data citation advantage. *PeerJ*, **1**. Retrieved October 05, 2013,

Piwowar, H.A., Day, R.S. & Fridsma, D.B. (2007). Sharing detailed research data is associated with increased citation rate. (J. Ioannidis, Ed.). *PloS one*, **2**, e308.

Poisot, T., Canard, E., Mouillot, D., Mouquet, N. & Gravel, D. (2012). The dissimilarity of species interaction networks. *Ecology Letters*, **15**, 1353–1361.

Poisot, T., Mounce, R. & Gravel, D. (2013). Moving toward a sustainable ecological science: don’t let data go to waste!

Ricciardi, F., Boyer, M. & Ollerton, J. (2010). Assemblage and interaction structure of the anemonefish-anemone mutualism across the Manado region of Sulawesi, Indonesia. *Environmental Biology of Fishes*, **87**, 333–347. Retrieved January 10, 2014,

Rohr, R.P., Scherer, H., Kehrli, P., Mazza, C. & Bersier, L.-F. (2010). Modeling food webs: exploring unexplained structure using latent traits. *The American naturalist*, **176**, 170–7.

Schleuning, M., Blüthgen, N., Flörchinger, M., Braun, J., Schaefer, H.M. & Böhning-Gaese, K. (2011). Specialization and interaction strength in a tropical plant-frugivore network differ among forest strata. *Ecology*, **92**, 26–36.

Yodzis, P. (1981). The stability of real ecosystems. *Nature*, **289**, 674–676.