mangal – making complex ecological network analysis simpler

T. Poisot, D. Gravel

Jan. 2014

The study of ecological networks is severaly 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, will streamline the analysis process, and improve reproducibility in studies of ecological networks.

```
## Error: impossible de trouver la fonction "as"
```

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. This led to classical results, such as the buffering impact of modularity on species loss {ref}, the increase in robustness along with increases in connectance {ref}, and {missing}. More recently, there have been new studies introducing the idea that different networks can be meaningfully compared, either to understand the importance of environmental gradients on the realisation of ecological interactions {ref}, or to understand the mechanisms behind variation in the structure of ecological networks {refs}. 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 {refs}. In part, this can be attributed to the limited methods allowing to compare networks in which no species are in common {ref}. However, the severe shortage of data in the field also restricts the power of large-scale analyses. Indeed, most of the studies working on several types of interactions focused on comparing emerging properties {refs}.

An increasing number of approachs are being put forth to *predict* the structure of ecological networks, either relying on latent variables {ref} or actual traits {ref}. These approaches, so as to be adequately calibrated, require easily accessible data. Comparing the efficiency of different methods will also be 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. Finally, we (iii) describe `mangal`, a `R` package and compagnon database, relying on this data specification. 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 their *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 will inform 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 wassampled) 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 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) respecting the data specification. This approach ensures that *all* data can be used in meta-analyses, and will in time increase the impact of data.

Elements of the data specification

The data specification (Fig. **XX**) 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 highlight 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 aforementionned URL), this section will propose an overview of each element, and of how they interact.

We propose `JSON` as the most efficient data format for the following 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

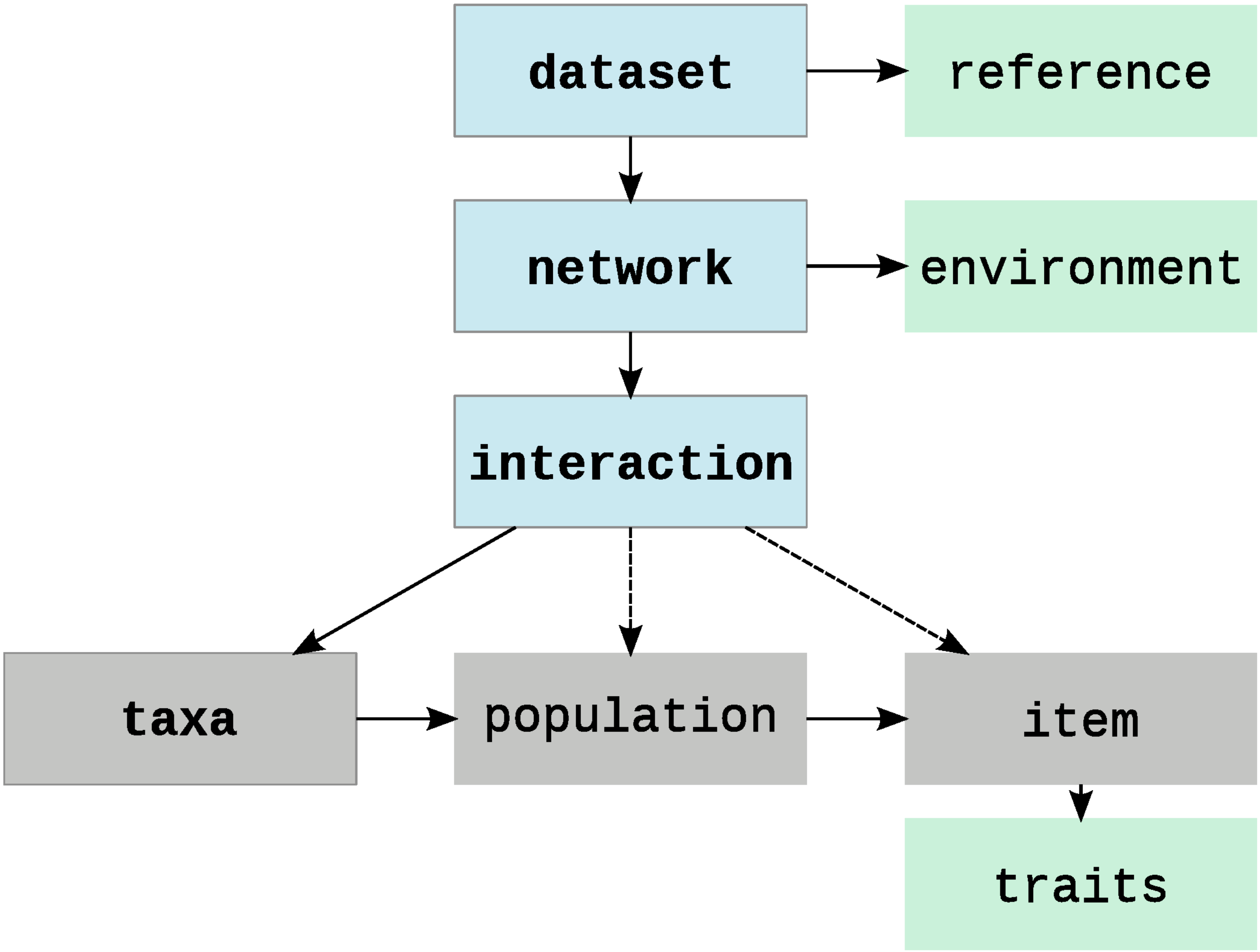

Figure 1: An overview of the data specification, and the hierarchy between objects. Each box correspond to a level of the data specification. Grey boxes are nodes, blue boxes are interactions and networks, and green boxes are metadata. The **bold** boxes (`dataset`, `network`, `interaction`, `taxa`)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

can verify that it is correctly formatted. 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 will be entirely transparent, as the interaction will happen within R – as such, knowing how JSON objects are organised is only useful for those who want to interact with the API directly.

Node informations

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 accomodates ncbi, gbif, itis, eol and bold identifiers.

Population

A population is one observed instance of a taxa object. If your experimental design is replicated through space, then each taxa will 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 informations

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 will help differentiate from direct observations, text mining, and inference. Note that the `nature` field can also take `absence` as a value; this will be 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 prefered 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 network. These are defined by the environmental property measured, its value, and the units.

References

References are associated to datasets. They accomodate 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 a `REST`ful API (`http://mangal.uc` It is possible for users to deposit data into this database, through the `R` package. Data are made available under a *CC-0 Waiver*. 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.*.

The data we use for this example come from Ricciardi et al. (2010). These were previously available on the *InteractionWeb DataBase* as a single `xls` file. We uploaded them in the `mangal` database at `http://mangal.uqar.ca/api/v1/dataset/{todo}`.

Link-species relationships

In the first example, we visualize the relationship between the number of species and the number of interactions, which @martinez_constant_1992 propose to be linear (in food webs).

```
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>`". This will encourage re-use of the data.

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}).

```
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 straigthforward.

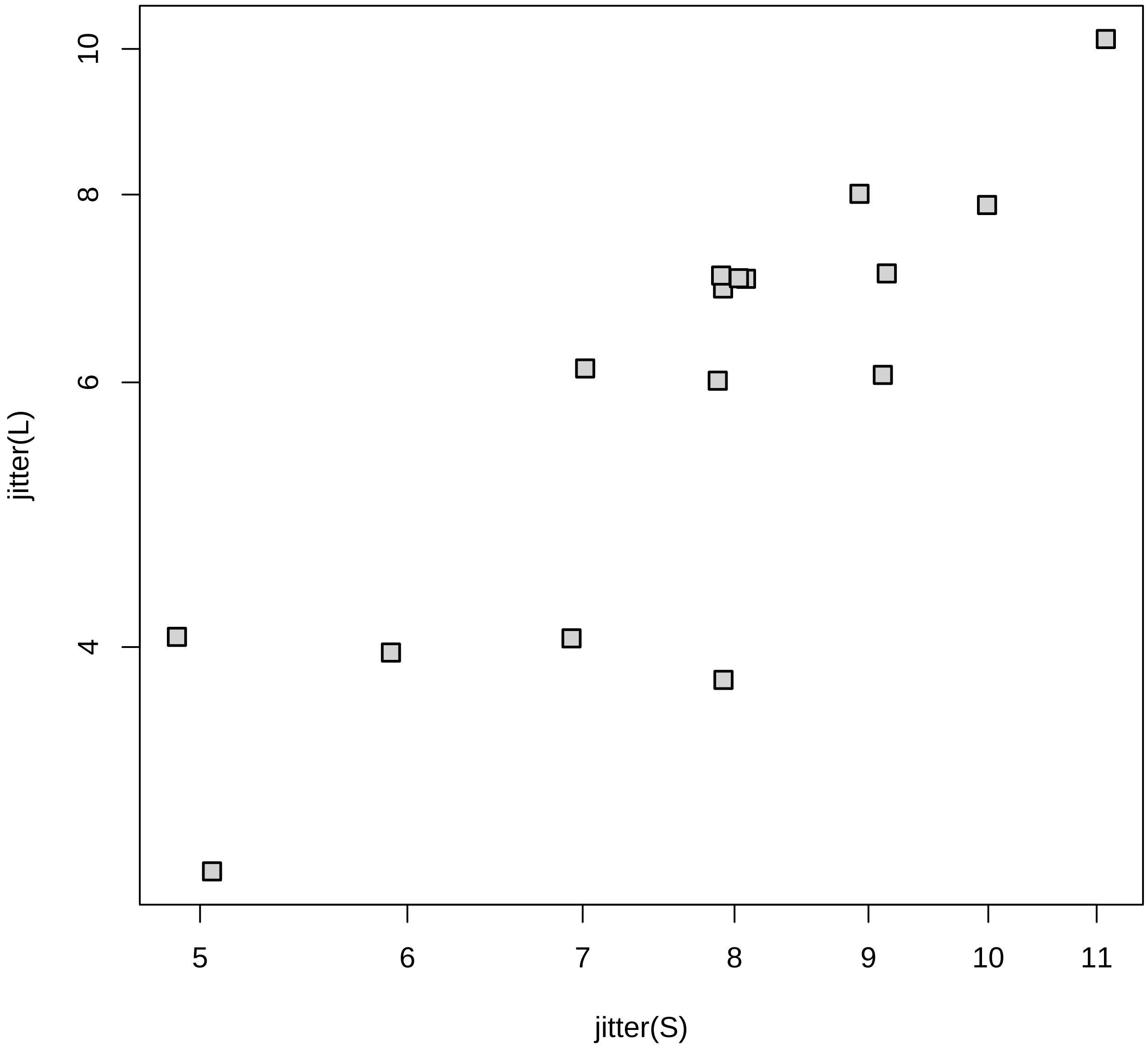

Figure 2: Relationship between the number of species and number of interactions in the anemonefish-fish dataset.

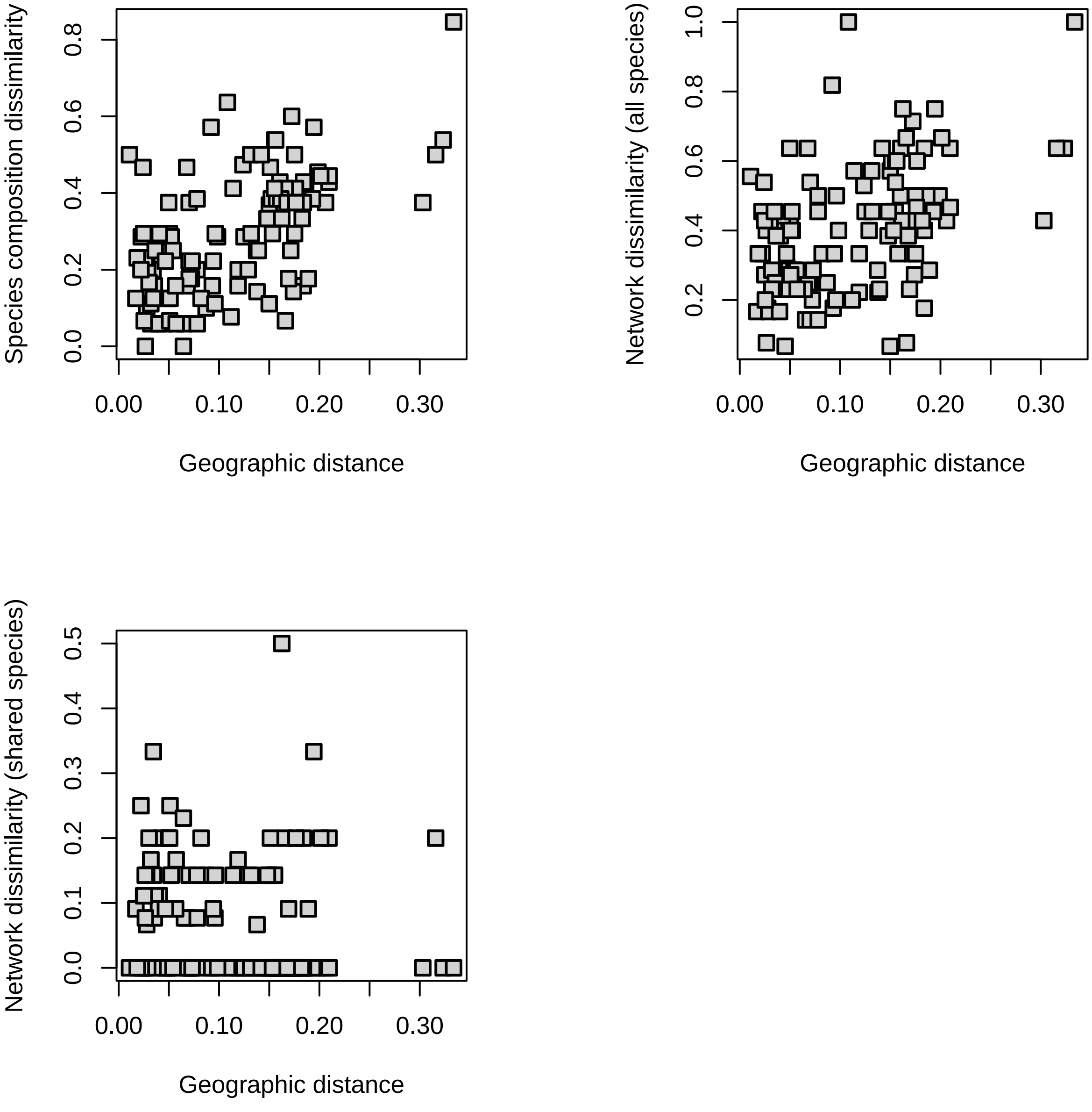

Figure 3: Relationships between the geographic distance between two sites, and the species dissimiliarity, network dissimilarity with all, and only shared, species.

Spatial visualisation of networks

Bascompte (2009) proposes an interesting visualisation for spatialized networks, in which each species is laid out on a map at the center of mass of its area of occurence; 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, using the `RgoogleMaps` package.

Conclusions

In this contribution, we presented `mangal`, a data format for the exchange of ecological networks and associated meta-data. 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.

References

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

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

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,