

# mangal – making complex ecological network analysis simpler

T. Poisot, D. Gravel

Jan. 2014

1 The study of ecological networks is severely limited by (i) the difficulty to access  
2 data, (ii) the lack of a standardized way to link meta-data with interactions, and (iii)  
3 the disparity of formats in which ecological networks themselves are represented.  
4 To overcome these limitations, we conceived a data specification for ecological net-  
5 works. We implemented a database respecting this standard, and released a R  
6 package ( `rmangal`) allowing users to programmatically access, curate, and deposit  
7 data on ecological interactions. In this article, we show how these tools, in conjunc-  
8 tions with other frameworks for the programmatic manipulation of open ecological  
9 data, will streamline the analysis process, and improve reproducibility in studies  
10 of ecological networks.

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

## 12 Introduction

13 Ecological networks enable ecologists to accommodate the complexity of natural communi-  
14 ties, and to discover mechanisms contributing to their persistence, stability, resilience, and  
15 functioning. Most of the “early” studies of ecological networks were focused on understand-  
16 ing how the structure of interactions within one location affected the ecological properties of  
17 this local community. This led to classical results, such as the buffering impact of modularity

1 on species loss {ref}, the increase in robustness along with increases in connectance {ref}, and  
2 {missing}. More recently, there have been new studies introducing the idea that different net-  
3 works can be meaningfully compared, either to understand the importance of environmental  
4 gradients on the realisation of ecological interactions {ref}, or to understand the mechanisms  
5 behind variation in the structure of ecological networks {refs}. Yet, meta-analyses of a large  
6 number of ecological networks are still extremely rare, and most of the studies comparing sev-  
7 eral networks do so within the limit of particular systems {refs}. In part, this can be attributed  
8 to the limited methods allowing to compare networks in which no species are in common  
9 {ref}. However, the severe shortage of data in the field also restricts the power of large-scale  
10 analyses. Indeed, most of the studies working on several types of interactions focused on  
11 comparing emerging properties {refs}.

12 An increasing number of approaches are being put forth to *predict* the structure of ecological  
13 networks, either relying on latent variables {ref} or actual traits {ref}. These approaches, so  
14 as to be adequately calibrated, require easily accessible data. Comparing the efficiency of  
15 different methods will also be facilitated if there is an homogeneous way of representing  
16 ecological interactions, and the associated metadata. In this paper, we (i) establish the need  
17 of a data specification serving as a *lingua franca* among network ecologists, (ii) describe this  
18 data specification. Finally, we (iii) describe `mangal`, a R package and `compagnon` database,  
19 relying on this data specification. We provide some use cases showing how this new approach  
20 makes complex analyzes simpler, and allows for the integration of new tools to manipulate  
21 biodiversity resources.

## 22 **Networks need a data specification**

23 Ecological networks are (often) stored as their *adjacency matrix* (or as the quantitative link  
24 matrix), that is a series of 0 and 1 indicating, respectively, the absence and presence of an  
25 interaction. This format is extremely convenient for *use* (as most network analysis packages,  
26 *e.g.* `bipartite`, `betalink`, `foodweb`, require data to be presented this way), but is extremely

1 inefficient at *storing* meta-data. In most cases, an adjacency matrix will inform on the identity  
2 of species (in cases where rows and columns headers are present), and the presence or absence  
3 of interactions. If other data about the environment (*e.g.* where the network was sampled) or  
4 the species (*e.g.* the population size, trait distribution, or other observations) are available,  
5 they are most either given in other files, or as accompanying text. In both cases, making  
6 a programmatic link between interaction data and relevant meta-data is difficult and error-  
7 prone.

8 By contrast, a data specification provides a common language for network ecologists to inter-  
9 act, and ensure that, regardless of their source, data can be used in a shared workflow. Most  
10 importantly, a data specification describes how data are *exchanged*. Each group retains the abil-  
11 ity to store the data in the format that is most convenient for in-house use, and only needs to  
12 provide export options (*e.g.* through an API) respecting the data specification. This approach  
13 ensures that *all* data can be used in meta-analyses, and will in time increase the impact of data  
14 {ref}.

## 15 **Elements of the data specification**

16 {complete}The data specification (Fig. XX) is built around the idea that (ecological) networks  
17 are collections of relationships between ecological objects, each element having particular  
18 meta-data associated. In this section, we detail highlight the way networks are represented  
19 in the mangal specification. An interactive webpage with the elements of the data specifi-  
20 cation can be found online at <http://mangal.uqar.ca./doc/spec/>. The data specification is  
21 available either at the API root (*e.g.* <http://mangal.uqar.ca/api/v1/?format=json>), or can be  
22 viewed using the `whatIs` function from the R package (see *Supp. Mat. 1*). Rather than giving  
23 an exhaustive list of the data specification (which is available online at the aforementioned  
24 URL), this section will propose an overview of each element, and of how they interact.

25 We propose JSON as the most efficient data format for the following reasons. First, it has  
26 emerged as a *de facto* standard for web platform serving data, and accepting data from users.

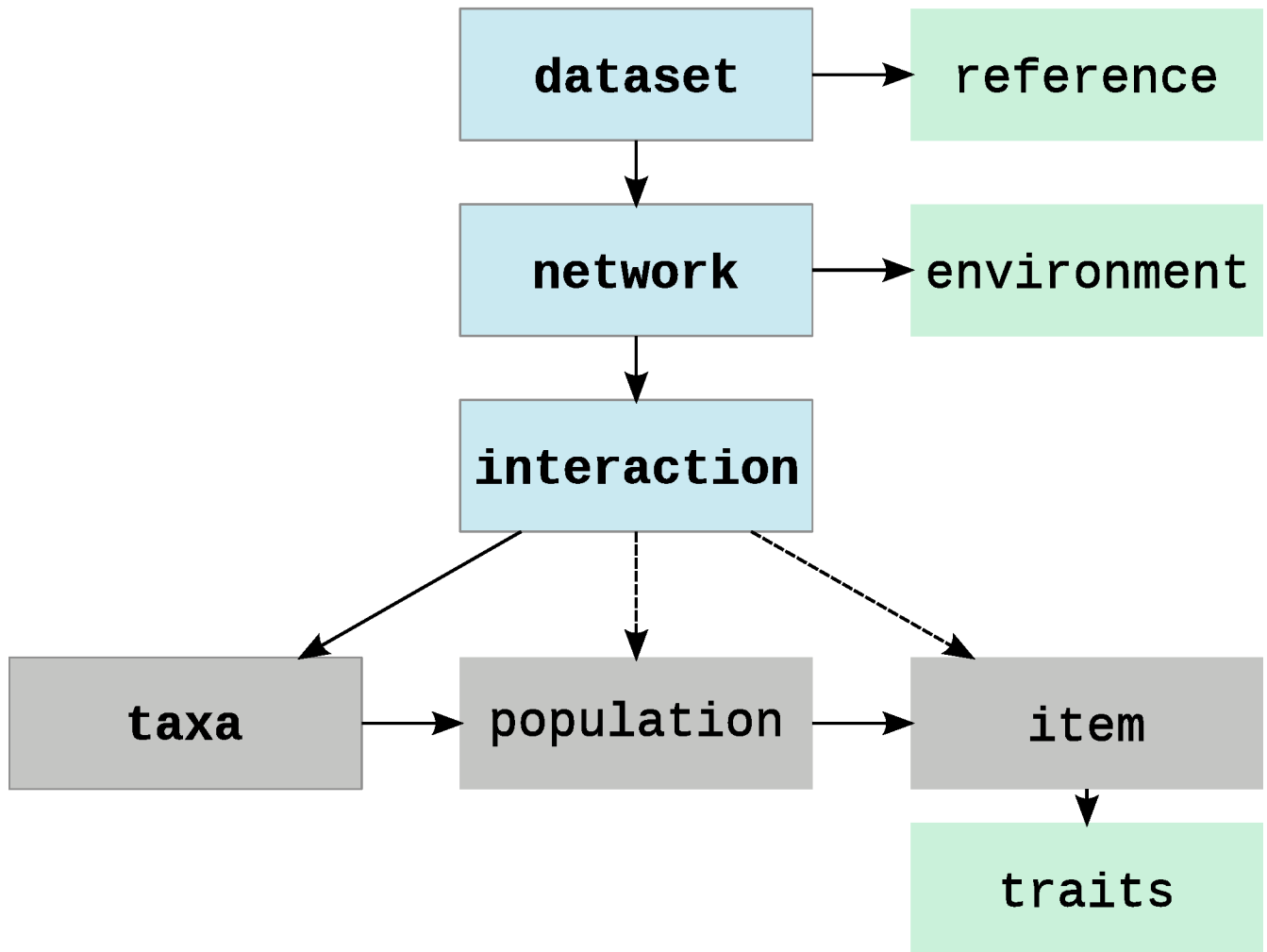


Figure 1: An overview of the data specification, and the hierarchy between objects. Each box correspond to a level of the data specification. Grey boxes are nodes, blue boxes are interactions and networks, and green boxes are metadata. The **bold** boxes (dataset, network, interaction, taxa) are the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

1 Second, it allows *validation* of the data: a JSON file can be matched against a scheme, and one  
2 can verify that it is correctly formatted. Finally, JSON objects are easily and cheaply (memory-  
3 wise) parsed in the most common programming languages, notably R (equivalent to `list`) and  
4 python (equivalent to `dict`). For most users, the format in which data are transmitted will be  
5 entirely transparent, as the interaction will happen within R.

## 6 **Node informations**

### 7 **Taxa**

8 Taxa are a taxonomic entity of any level, identified by their name, vernacular name, and  
9 their identifiers in a variety of taxonomic services. Associating the identifiers of each taxa is  
10 important to leverage the power of the new generation of open data tools, such as `taxize`  
11 `{ref}`. The data specification currently accomodates `ncbi`, `gbif`, `itis`, `eol` and `bold` identifiers.  
12 Correspondances between these and other services can be made through other tools, such as  
13 *e.g.* `taxize`.

### 14 **Population**

15 A population is one observed instance of a taxa object. If your experimental design is repli-  
16 cated through space, then each taxa will have a population object corresponding to each  
17 locality. Populations do not have associated meta-data, but serve as “containers” for item  
18 objects.

### 19 **Item**

20 An item is an instance of a population. Items have a `level` argument, which can be either  
21 `individual` or `population`; this allows to represent both individual-level networks (*i.e.* there  
22 are as many items attached to a population than there were individuals of this population  
23 sampled), and population-level networks. When `item` represents a population, it is possible

1 to give a measure of the size of this population. The notion of `item` is particularly useful  
2 for time-replicated designs: each observation of a population at a time-point is an `item` with  
3 associated trait values, and possibly population size.

## 4 **Network informations**

### 5 **Interaction**

### 6 **Network**

### 7 **Dataset**

## 8 **Meta-data**

### 9 **Trait value**

### 10 **Environmental condition**

### 11 **User**

12 `paternity {ref}`

## 13 **References**

## 14 **Use cases**

15 In this section, we present use cases using the `rmangal` package for R, to interact with a  
16 database implementing this data specification, and serving data through a RESTful API (<http://mangal.ucl.ac.uk>).  
17 It is possible for users to deposit data into this database, through the R package. Data are made  
18 available under a *CC-0 Waiver*. Detailed informations about how to upload data are given in  
19 the vignettes and manual of the `rmangal` package. So as to save room in the manuscript,

1 we source each example. The complete r files to reproduce the examples of this section are  
2 attached as *Suppl. Mat.*.

3 The data we use for this example come from {ref}. They were previously available on the  
4 *InteractionWeb DataBase* as a single xls file. We uploaded them in the mangal database at  
5 <http://mangal.uqar.ca/api/v1/dataset/{todo}>.

## 6 **Link-species relationships**

7 In the first example, we visualize the relationship between the number of species and the  
8 number of interactions, which @martinez\_constant\_1992 propose to be linear (in food webs).

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

9 Producing this figure requires less than 10 lines of code. The only information needed is the  
10 identifier of the network or dataset, which we suggest should be reported in publications as:  
11 “These data were deposited in the mangal format at <URL>/api/v1/dataset/<ID>”. This will  
12 encourage re-use of the data.

## 13 **Network beta-diversity**

14 In the second example, we use the framework of network  $\beta$ -diversity (Poisot *et al.* 2012) to  
15 measure the extent to which networks that are far apart in space have different interactions.  
16 Each network in the dataset has a latitude and longitude, meaning that it is possible to measure  
17 the geographic distance between two networks.

18 For each pair of network, we measure the geographic distance (in km.), the species dissimilar-  
19 ity ( $\beta_S$ ), the network dissimilarity when all species are present ( $\beta_{WN}$ ), and finally, the network  
20 dissimilarity when only shared species are considered ( $\beta_{OS}$ ).

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

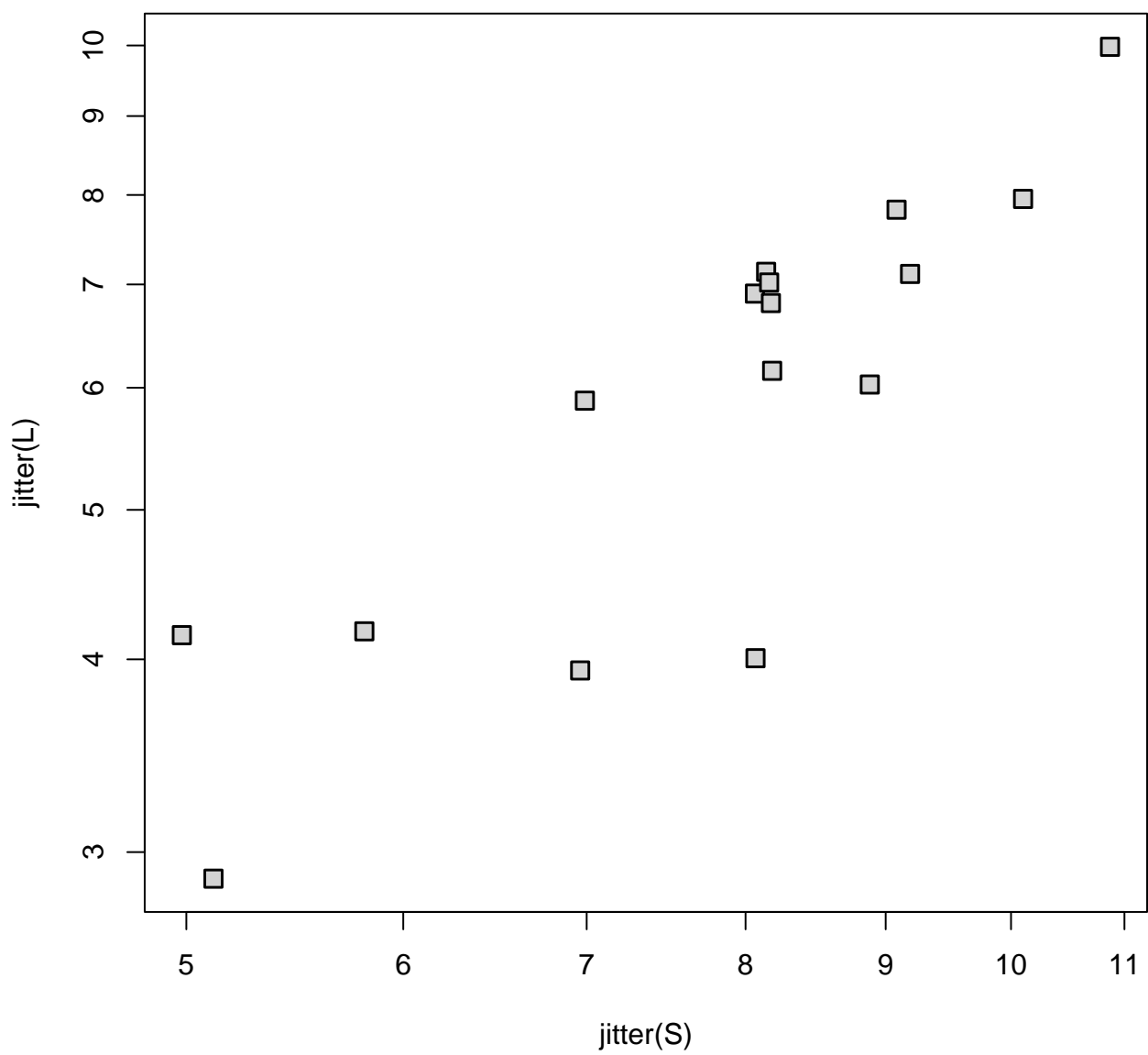


Figure 2: Relationship between the number of species and number of interactions in the anemonefish-fish dataset.



```

1  ## Installing github repo betalink/master from tpoisot
2  ## Downloading betalink.zip from https://github.com/tpoisot/betalink/archive/master.zip
3  ## Installing package from /tmp/Rtmp91U4kX/betalink.zip
4  ## arguments 'minimized' and 'invisible' are for Windows only
5  ## Installing betalink
6  ## '/usr/lib/R/bin/R' --vanilla CMD INSTALL \
7  ##   '/tmp/Rtmp91U4kX/devtools50a3776d264/betalink-master' \
8  ##   --library='/home/tpoisot/R/i686-pc-linux-gnu-library/3.0' \
9  ##   --install-tests

```

10 As shown in *Fig. XX*, while species dissimilarity and overall network dissimilarity increase  
11 when two networks are far apart, this is not the case for the way common species interact.  
12 This suggests that in this system, network dissimilarity over space is primarily driven by  
13 species turnover. The ease to gather both raw interaction data and associated meta-data make  
14 producing this analysis extremely straightforward. We foresee that with an increase in the  
15 number of deposited datasets, new properties of ecological networks will be uncovered.

## 16 **Spatial visualisation of networks**

17 Bascompte (2009) proposes an interesting visualisation for spatialized networks, in which each  
18 species is laid out on a map at the center of mass of its area of occurrence; interactions are then  
19 drawn between species, to show how species distribution determines biotic interactions. In  
20 this final use case, we propose to reproduce a similar figure, using the *RgoogleMaps* package.

## 21 **References**

- 22 Bascompte, J. (2009). Disentangling the web of life. *Science (New York, N.Y.)*, **325**, 416–9.
- 23 Poisot, T., Canard, E., Mouillot, D., Mouquet, N. & Gravel, D. (2012). The dissimilarity of  
24 species interaction networks. *Ecology Letters*, **15**, 1353–1361.

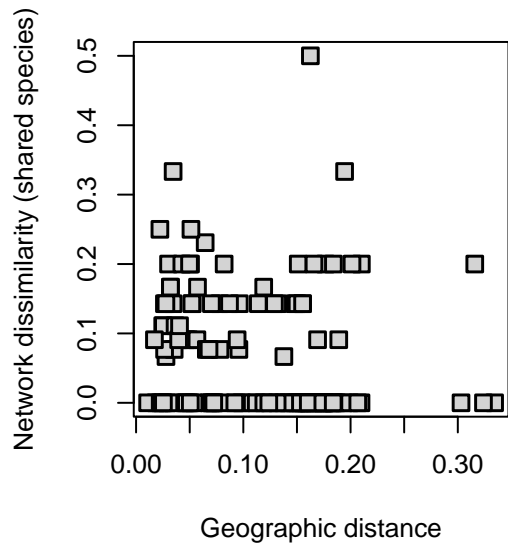
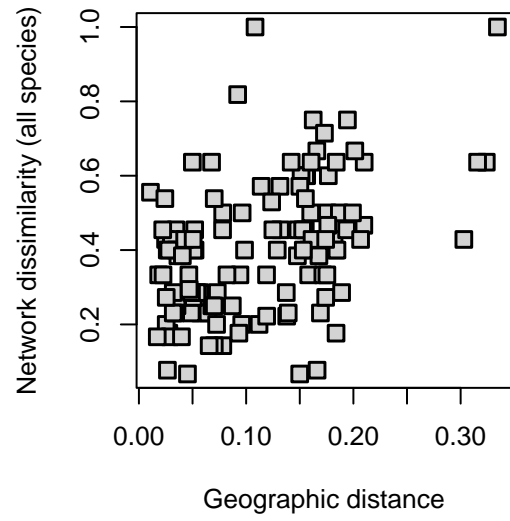
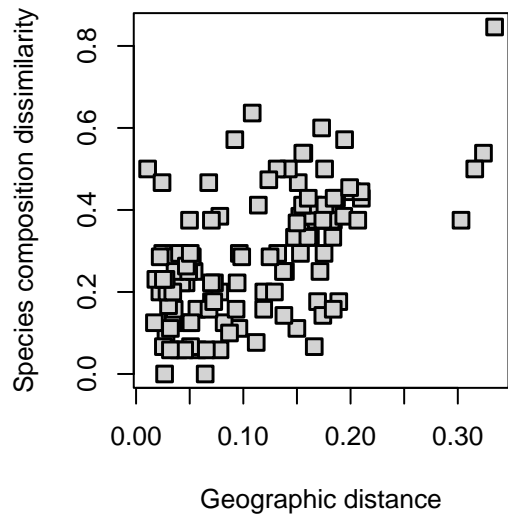


Figure 3: Relationships between the geographic distance between two sites, and the species dissimilarity, network dissimilarity with all, and only shared, species.