diverse: an R Package to Analyze Diversity in Complex Systems

by Miguel R. Guevara, Dominik Hartmann and Marcelo Mendoza

Abstract The package **diverse** provides an easy-to-use interface to calculate and visualize different aspects of diversity in complex systems. In recent years, an increasing number of research projects in social and interdisciplinary sciences, including fields like innovation studies, scientometrics, economics, and network science have emphasized the role of diversification and sophistication of socioeconomic systems. However, so far no dedicated package exists that covers the needs of these emerging fields and interdisciplinary teams. Most packages about diversity tend to be created according to the demands and terminology of particular areas of natural and biological sciences. The package **diverse** uses interdisciplinary concepts of diversity—like variety, disparity and balance— as well as ubiquity and revealed comparative advantages, that are relevant to many fields of science, but are in particular useful for interdisciplinary research on diversity in socioeconomic systems. The package **diverse** provides a toolkit for social scientists, interdisciplinary researcher and beginners in ecology to (i) import data, (ii) calculate different data transformations and normalization like revealed comparative advantages, (iii) calculate different diversity measures, and (iv) connect **diverse** to other specialized R packages on similarity measures, data visualization techniques and statistical significance tests. The comprehensiveness of the package, from matrix import and transformations options, over similarity and diversity measures, to data visualization methods, makes it a useful package to explore different dimensions of diversity in complex systems.

Introduction

While measuring diversity is a natural topic in ecology and biology, the rise of complexity and big data research has also provided new opportunities for researchers in various fields of social sciences to understand the evolution of diversity in social, economic and political systems (Nature, 2014).

Today, a large range of scientific fields make use of diversity measures, including ecologists calculating the diversity of species (Forey et al., 1994), sociologists measuring the structure of communities (Haughton and Mukerjee, 1995), economists studying the diversification of exports or financial assets (Hidalgo and Hausmann, 2009), scientometrists analyzing the diversity and interdisciplinarity of research fields (Rafols, 2014; Chavarro et al., 2014; Wagner et al., 2011), or computer scientists searching for new diversity methods to ensemble algorithms (Kuncheva and Whitaker, 2003).

Consequently, different fields of science have created several specialized R packages on diversity that explore a wide range of fields. This includes packages that allow for the analysis of species and biodiversity, (e.g. **entropart** (Marcon and Hérault, 2015), **vegan** (Oksanen et al., 2016), **biodiversityR** (Kindt and Coe, 2005)), social distances (e.g. **Blaunet** (Wang et al., 2016)), genetics (e.g. **diveRsity** (Keenan et al., 2013)), biological systems (e.g. **divo** (Pietrzak et al., 2016)), functional ecology (e.g. **FD**, (Laliberté and Legendre, 2010)), species complexity (e.g. **hierDiversity** (Marion et al., 2015)), bootstrapping diversity indices (e.g. **simboot** (Scherer and Pallmann, 2014)), disparity of phylogenetic trees (e.g. **treescape** (Jombart et al., 2016)) or phylogenetic patterns (e.g. **SYNCSA** (Debastiani and Pillar, 2012)).

Most packages on diversity are in the fields of ecology, biology and other natural sciences. Each discipline and respective package uses the particular terminology of its scientific field. It is important to translate the existing mathematical diversity formulas into the relevant concepts and language of each community, and thereby also helps to create new specialized measures considering the particular research topics and demands of each community. But the thematic specialization can also make interdisciplinary communication difficult and reduces the chances of the adoption of these new measures, concepts and specialized packages by researchers outside of the particular scientific field.

In recent years, an increasingly large number of research projects in social and interdisciplinary sciences are exploring the role of diversity in complex socioeconomic systems. These new approaches in social and interdisciplinary sciences use existing diversity concepts from biological and natural sciences, but also have their own particular needs and concepts. For instance, recent work in economics, scientometrics and network science has highlighted the importance of diversification processes in complex systems, such as research, financial and energy portfolios, cultural diversity, the diversity of ties in social and economic networks, or the emergence of new or related scientific and economic fields (Hidalgo et al., 2007; Frenken et al., 2007; Rafols et al., 2010; Chavarro et al., 2014; Guevara et al., 2016; Eagle et al., 2010; Farchy and Ranaivoson, 2011).

Here we present the package **diverse** which aims to provide a useful toolkit for social scientists

and interdisciplinary teams to measure and visualize diversity in socioeconomic systems, by providing several of the most used measures of diversity and allowing for versatility with existing R packages on diversity, focusing, for example, on the calculation of similarity and distance measures (**proxy** (Meyer and Buchta, 2015)), bias corrected diversity measures (**entropart** (Marcon and Hérault, 2015)) or the visualization of diversity in matrices, treemaps and networks (**pheatmap** (Kolde, 2015), **treemap** (Kindt and Coe, 2005) and **igraph** (Kindt and Coe, 2005)).

The package applies a diversity taxonomy that includes the variety, balance and disparity of complex systems (Stirling, 2007). The package **diverse** allows researchers to:

1. Read, input and process data from complex systems in a simple manner.
2. Compute some of the most commonly used measures of diversity across sciences—including Shannon-Entropy, Herfindahl-Hirschman Index, Gini-Simpson Index or Berger-Parker Index.
3. Calculate complementary measures that are related to diversity, such as ubiquity, disparity or similarity between categories and entities.
4. Apply advanced diversity measures such as Rao-Stirling diversity and other diversity measures including weighting parameters.
5. Visualize different dimensions of diversity as variety, balance or disparity.

The package **diverse** is available at the CRAN repository. The newest development version is accessible at the branch *development* of the Git repository github.com/mguevara/diverse. In this Git repository interested users are also very welcome to submit issues and requirements.

The remainder of the article is organized as follows. In Section Diversity we describe different dimensions of diversity. In Section Input data we explain which type of data can be read/imported into the package and how it can be normalized, using for instance either binary, absolute or relative values. In Section Measuring diversity we present the measures available in this package discussing how researchers can use them to calculate different dimensions of diversity. In Section Synthetic data and performance tests we explain functions that are included in the package to simulate data and conduct bias, coverage and performance tests. In Section Conclusions we summarize and briefly discuss the limitations and advantages of the package.

Diversity

In this section we explain key properties of diversity with the help of example datasets.

Example data

To illustrate the use of the package **diverse**, we will work with three datasets: Pantheon, Scidat and Geese.

- Pantheon is a sample of 10 countries from MIT's Pantheon project (MediaLab, 2014). This dataset allows for a comparison of the diversity of occupations of the globally famous people from each country. The complete dataset includes 11341 persons classified in 88 distinct occupations and assigned to 195 countries (Yu et al., 2016).
- Scidat is an aggregation of the number of scientific publications assigned to 27 areas of science. This dataset was aggregated over the raw data of SCImago (2007). Scidat includes a sample of 10 countries from the year 2013.
- The third dataset is on the geese population in the Netherlands and was published by the Sovon Dutch Centre for Field Ornithology (Nederland, 2015). This dataset presents observations of 4 species of geese over a period of 11 years.

The three datasets are included in the package **diverse**. The subset of the Pantheon dataset is included as a '`dataframe`' object and both the Scidat and Geese datasets are included as a '`matrix`' object.

The actors and concepts of diversity

We use the term *entity* to describe the systems or agents that host a set of categories. Entities could be, for example, persons, companies, countries, regions, institutions or years.

We also use the term *category* to identify the different types of species that define the diversity of an entity. Categories could include types of animals, species of plants, fields of research, taxonomies of products or technologies. The package assumes that the imported dataset has a previously given classification scheme.

The term *value* or value of abundance is used for the amount of a category in each entity. This could be the quantity of each species in an ecosystem, or the total value of the different types of export goods of an economy.

In Pantheon, entities are countries, categories are different types of occupations, and values are the respective number of globally famous persons a country has in each category. In Scidat, entities are countries, categories are SCimago's areas of science, and values are the total number of citable documents that a country has published in each area in 2013. In Geese, entities are years, categories are species of geese, and values are the number of each species of geese observed in the Netherlands in the respective year.

Pantheon is a good example of data where some entities have missing values in some categories. Scidat is a useful example where most entities have values in each category yet have very large absolute differences between their values. The Geese dataset is a good example of the temporal evolution of natural species.

It must be noted that in most diversity measures (e.g. variety or Shannon entropy) the information about the number and types of categories of a single entity is sufficient to calculate this entity's diversity. However **diverse** is oriented to also work with multiple entities. Therefore it allows for the calculation of different distance and similarity matrices across categories and entities, and uses these distance measures in diversity measures like the Rao-Stirling Index (Stirling, 2007). Moreover, **diverse** allows for the calculation of relative specialization measures like the activity index or revealed comparative advantages (RCAs) that takes the portfolio and size of other entities into account when evaluating their relative specialization or comparative advantages (e.g. Belgium versus USA) (see Section Data transformation and normalization). Subsequently, we will mainly use data examples with multiple entities and categories. Nonetheless, many measures embedded in **diverse** can also be used to track the evolution of the diversity within a single entity.

Regarding the concept of diversity, previous interdisciplinary studies on diversity (Rao, 1982; Stirling, 1998; Mcdonald and Dimmick, 2003; Stirling, 2007) showed that the concept of diversity is related to three main questions:

1. How many categories does an entity (and/or does each entity in a system) have?
2. How much of each category does an entity (and/or each entity in a system) have?
3. How distinct are the categories of an entity (and/or the categories of each entity in a system)?

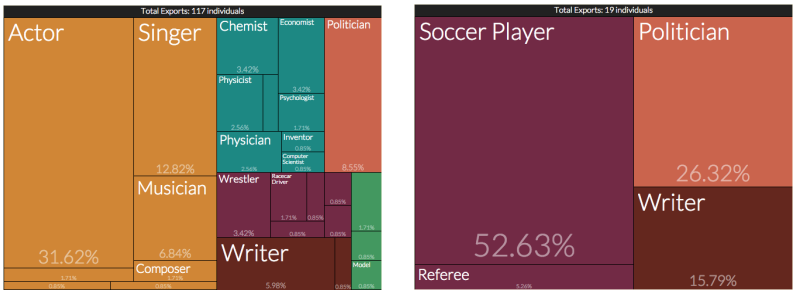
Stirling (1998, 2007) categorized these three properties of diversity as *variety*, *balance* and *disparity*. Most diversity measures combine and emphasize these aspects with varying weights. Comprehensive measures take all three dimensions deliberately into account. Moreover data visualization methods and R packages like for instance **treemap** (Tennekes, 2016) or **igraph** (Csardi and Nepusz, 2006) can help to visualize these three dimensions of diversity. For instance, treemaps—allow for an emphasis on variety and balance (Hausmann et al., 2011, p.105)—or network overlays maps allow for an emphasis on disparity (Hidalgo et al., 2007; Rafols et al., 2010). The disparity dimension is often implied by a previous classification scheme, like a given classification of types of animals, scientific fields or exports, phylogenetic trees and/or can alternatively be calculated based on a similarity or distance matrix (see also Section Matrix of dissimilarities between entities).

As an example, Figure 1 presents treemaps about the diversity of occupations of globally famous individuals from Canada and Uruguay according to MIT's Pantheon. Variety is represented by the number of boxes, balance is indicated by the differences in the size of the boxes (= percentage of the category), and disparity is represented by different colors.

First, regarding the *variety*, it is clear that Canada has a larger number of different occupations (27 boxes in Figure 1a) than Uruguay (4 boxes in Figure 1b). Second, regarding the *balance*, we can observe that Uruguay's concentration in terms of soccer players is very high (52.63%), while Canada's balance is less concentrated on one category, but is spread across more occupations of the Pantheon dataset. The R package **treemap** or other specialized data visualization programs like D3plus allows for the creation of such treemaps with different colors, text sizes and further visualization options.

To illustrate disparity, Figure 2a shows a similarity network between areas of science that was obtained by considering each column as a vector of features (i.e. the number of articles of each country in that category) and then computing the cosine (dis-)similarity between those vectors. The pre-process options embedded in the package allows for the calculation of different types of similarity and distance matrices and considers both the absolute shares as well as relative strengths / specializations / comparative advantages of entities (see Section on Data transformation and normalization).

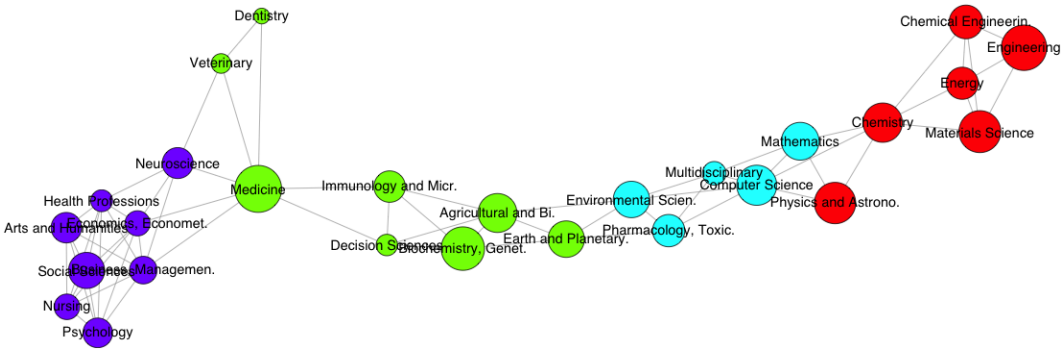
As expected, we can observe that "Social Sciences" are close to "Arts and Humanities", but more distant from "Mathematics" or "Engineering". We can also see the clustering between natural sciences and technological fields, like "Chemistry" and "Material Sciences". The disparity within a country is high if the dissimilarities between the areas/categories, in which the country has values, are also high.



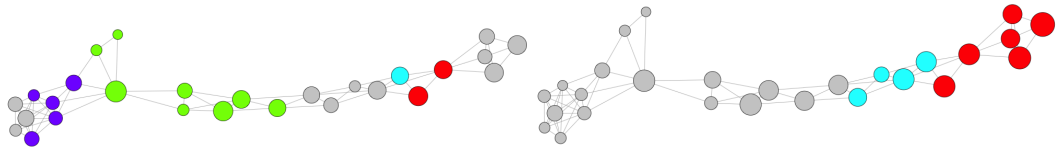
(a) Canada **(b)** Uruguay

Figure 1: Globally famous people according to Pantheon dataset. The size of the boxes is proportional to the number of people assigned to each occupation and born in that country. The color is according to main domains of occupations. Source `http://pantheon.media.mit.edu`

In so-called overlay networks (Rafols et al., 2010), the values of entities (which are countries in the present dataset) are overlaid on the global network structure. Moreover, the variety is represented by the number of colored nodes, and the dimension balance can be represented by the size of the nodes. Figure 2b illustrates that in the Scidat example, Germany has *comparative advantages* (see the section on Data transformation and normalization for details) in many fields of science across the network, while China (Figure 2c) is more specialized (concentrated) in technological areas and engineering. In consequence the disparity, variety and balance in Germany are higher than in China. Such network overlay maps can, for instance, be made with the R package **igraph**.



(a) Network of similarities between areas of science.



(b) Overlay network for Germany **(c)** Overlay network for China

Figure 2: Cosine similarity network of 27 areas of science obtained with Scidat dataset. Links represent the (dis)similarities between areas. Links below the threshold of 0.015 of Cosine similarity are not illustrated. The force-directed algorithm Fruchterman-Reingold was used for the network layout. The size of the nodes in 2a represents the total number of papers authored by the 10 countries included in Scidat; the size of the nodes in 2b and 2c is proportional to the papers authored for each country. Colors are according to communities detected by the algorithm *fastgreedy*. The grey-colored nodes identify areas with *Revealed Comparative Advantages RCA* below 1 (see section on Data transformation and normalization for details).

In the following sections we will detail how to import and transform data, and how to use **diverse** to quantify the described properties of diversity.

Input data

This section details the type of the data object that is required by the package, how to import data from an external data file and how to pre-process or normalize the raw data.

Input formats

Since **diverse** was created to be able to work with multiple categories (N) and multiple entities (M) simultaneously, the data objects used for most of the functions in the package **diverse** can be either a 'dataframe' or a 'matrix' of $M \times N$.

In the case of a 'dataframe'—meaning that the data is shaped as an *edge* list— it has to have three columns in this order: entity, category and value. The first two columns are of the type 'factor' and the third column is of type 'numeric'. The 'pantheon' 'dataframe' is an example of this type of data object.

```
str(pantheon)
'data.frame':        119 obs. of  3 variables:
 $ Country   : Factor w/ 10 levels "Canada","Chile",..: 10 5 4 9 2 8 7 6 3 1 ...
 $ Occupation: Factor w/ 52 levels "Actor","Architect",..: 40 40 40 40 40 40 40 40 40 40
 $ Value     : int  6 2 8 5 10 9 17 36 38 10 ...
```

When the data is in a 'matrix' format, each cell has to contain numeric values and the 'rownames' and 'colnames' must be defined with the names of entities and the names of categories. Non-existent values ('NA') or 0 have to be used to indicate the lack of a category in an entity. The matrix of the 'scidat' dataset is an example of this type of object.

```
str(scidat)
 num [1:10, 1:27] 3507 35351 15603 1346 4158 ...
 - attr(*, "dimnames")=List of 2
  ..$ : chr [1:10] "Argentina" "China" "Germany" "Hungary" ...
  ..$ : chr [1:27] "Agricultural and Biological Sciences" "Arts and Humanities"...
```

If the matrix has categories in the rows and entities in the columns, the parameter 'category_row' must be set to 'TRUE' when using the functions included in **diverse**. The matrix of the 'geese' dataset is an example of this kind of object.

```
str(geese)
 num [1:4, 1:11] 274 10788 4786 39273 247 ...
 - attr(*, "dimnames")=List of 2
  ..$ : chr [1:4] "Little Grebe" "Crested Grebe" "Mute Swan" "Greylag Goose"
  ..$ : chr [1:11] "1996" "1997" "1998" "1999" ...
```

Importing data

To simplify the input of data from external files, **diverse** includes the function `read_data()`. This function reads CSV files and automatically detects whether the data file is a matrix or a edge shape. Moreover, it retrieves a 'dataframe' ready to be passed to the parameter 'data' of all the functions included in **diverse**. The user has to provide the path to the external CSV file by using the argument 'path'.

In addition, to facilitate the import of data from different software formats, the function `read_data()` includes the parameter 'type' that can be used to indicate whether the external data file comes for instance from *Stata* or *SPSS*. This functionality depends on the package **foreign** (R Core Team, 2015).

Data transformation and normalization

Depending on the characteristics of the dataset, researchers often need to normalize, transform or filter the data before measuring diversity. For instance, in some cases the absolute quantity or share of each category in the portfolio of an entity is important. In other cases, the relative specialization and diversification of entities in comparison to a set of other entities (e.g. revealed comparative advantages of countries, or the relative activity in certain research fields) is more important. There are also cases where binary values (e.g. are certain categories present or not present) or discrete steps (e.g. not present, low, middle, high value) are important, depending on the respective research question. For this purpose, **diverse** includes the function `values()` which allows for the filtering and exclusion of data below a certain threshold value, and to binarize or normalize the data.

The normalization process could include proportion values, *Revealed Comparative Advantages* (RCA) (Balassa, 1986) and normalized RCAs (which is also called Activity Index). With the term *proportions* we refer to normalization within an entity (dividing the value of an entity in each category by the sum of values of the entity in all categories). The calculation of RCAs or the Activity Index is a normalization related to the other entities. For instance in economics, RCA computes the ratio

between the proportion of a category within an entity, e.g. a country or region, and the proportion that represents that category in the global system (e.g. the world economy). The purpose of this measure is to understand in which categories an entity is relatively more specialized than others and thus seems to have a comparative advantage (Balassa, 1965). Typically, values of RCA greater than 1 are considered to "reveal" comparative advantages in the respective categories. Values below 1 reveal comparative disadvantages. The same idea can be found in Scientometrics where the RCAs are normalized between -1 and 1, and named 'Activity Index'. Both options, RCA and Activity Index are included in **diverse**.

To use these functionalities, the arguments '`norm`', '`filter`' and '`binary`' should be used. Argument '`norm`' can be set, for instance, to 'p' for proportions, 'rca' for RCAs or 'ai' for Activity Index. The argument '`filter`' allows the user to indicate a threshold, below which all the values are discarded (replaced with NA). The argument '`binary`' has to be set to TRUE if binary values are required. If the three arguments are applied, then the function `values()` first applies the normalization, then the filter and finally creates binary values.

The following matrix visualizations show the importance of the normalization process in datasets like Scidat where most entities produce all categories and the absolute differences (e.g. between the values of a small and a large country) are very large.

```
library(pheatmap)
colfunc <- colorRampPalette(c("deepskyblue4", "deepskyblue", "cyan"))
plot_mat <- function(data)
        pheatmap(data, colfunc(100), cluster_rows = FALSE, cluster_cols = FALSE)

col_l <- names(sort(colSums(values(scidat))))  #order
row_l <- names(sort(rowSums(values(scidat)), decreasing = TRUE))
plot_mat(values(scidat)[row_l,col_l])
plot_mat(values(scidat, norm = 'p')[row_l,col_l])
plot_mat(values(scidat, norm = 'rca')[row_l,col_l])
plot_mat(values(scidat, norm = 'rca', filter = 1)[row_l,col_l])
```

In Figure 3a we see the absolute values of authored papers by country in each area. The large number of papers from the United States in "Medicine" and "Biochemistry", as well as from China in "Engineering" and "Material Sciences" are the outstanding features of this matrix. If we consider proportions instead of absolute values, we can observe that "Medicine" is an important field of science for most countries, while a large proportion of the publication portfolios in Argentina or Mexico are in agricultural and biological sciences (see Figure 3b). Moreover, if we want to compare the relative specialization and comparative advantages of each country within the global system, an RCA based matrix will be more useful. Figure 3c presents the values of all RCAs, while Figure 3d presents the values of an RCA matrix in which values below 1 are represented by empty cells.

Measuring diversity

In this section we explain the measures included in the package **diverse** by illustrating their use with our sample datasets.

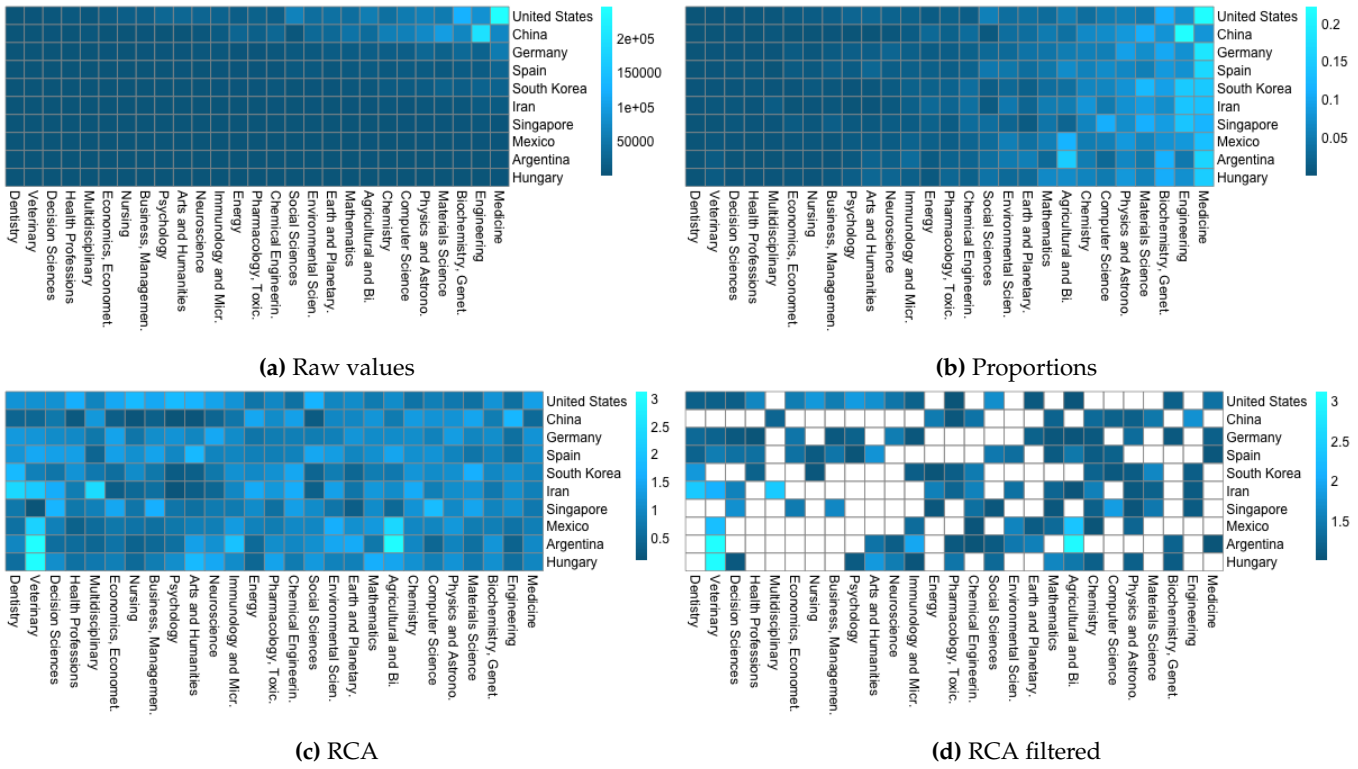
Measures included

The diversity measures included in the package **diverse** allow for different dimensions of diversity—like variety, balance and disparity—to be analyzed separately or jointly.

To compute these measures, the main function `diversity()` must be used. All diversity measures available in the package **diverse** are listed in Table 1. These measures are organized from simple to complex, considering the properties of diversity they take into account.

Regarding the parameters of the function `diversity()`, the dataset to be analyzed should be provided in the '`data`' parameter, and the required diversity measure(s) should be provided in the '`type`' parameter.

The argument '`data`' has to fulfill the characteristics analyzed in the previous section on Input data. The argument '`type`' can be a single '`string`' or a '`vector`' of strings, with either the complete name of the measure or a mnemonic term (see column *ID* in Table 1). In the following sections we will explain the function `diversity()` and the related functions `variety()`, `balance()`, and `disparity()`.



(a) Raw values

(b) Proportions

(c) RCA

(d) RCA filtered

Figure 3: Pheatmaps for matrices. The lighter the color, the higher the value. The white color represents empty cells.

ID	Measure	Formula	Reference
v	Variety	$v = \sum_i (p_i^0)$	
hhi	Herfindahl–Hirschman Index	$HHI = \sum_i (p_i^2)$	Rhoades (1993)
b, gs	Blau Index, Gini-Simpson	$B = 1 - \sum_i (p_i^2) = 1 - HHI$	Blau (1977); Gini (1912)
s	Simpson	$D_S = \sum_i n_i(n_i - 1)/N_t(N_t - 1)$	Simpson (1949)
bp	Berger-Parker	$D_{BP} = \max_i (p_i)$	Berger and Parker (1970)
e	Shannon Entropy	$H = -\sum_i (p_i \log p_i)$	Shannon (1948)
ev	Pielou Evenness	$J = -\sum_i (p_i \log p_i) \,/\, \log v$	Pielou (1970)
re	Rényi-Entropy	${}^qH = (1-q)^{-1} \log \left(\sum_i p_i^q\right)$	Rényi (1961)
hcdt	HCDT Entropy	${}^qH = (q-1)^{-1} \left(1 - \sum_i p_i^q\right)$	Havrda and Charvát (1967); Tsallis (1988)
hn	Hill Numbers	${}^qD_{HN} = \left(\sum_i p_i^q\right)^{1/(1-q)}$	Hill (1973)
d	Disparity	$DIS = \sum_{ij} d_{ij}/N$	
rao	Rao	$D_{RAO} = \sum_{ij} d_{ij} p_i p_j$	Rao (1982)
rs	Rao-Stirling	$\Delta = \sum_{ij} {d_{ij}}^{\alpha} (p_i p_j)^{\beta}$	Stirling (2007)

Table 1: Summary of measures available in the package **diverse**. The first block of measures are associated mainly with the dimensions variety and balance of the diversity, while the second block presents measures that use also the dimension disparity. C is the set of categories present in the entity. $i, j \in C$. $i \neq j$ and $ij \neq ji$; n_i is the value of abundance and p_i the proportion of the category i in the entity. $v = n(C)$ is the number of categories present in the entity—the variety. $N_t = \sum n_i$. log is the logarithm usually natural. $q, \alpha, \beta \geqslant 0$. For HCDT and Rény entropies when $q \to 1$ their result is Shannon entropy. Additionally, for Hill numbers, when $q \to 1$, it results in the exponential of Shannon Entropy.

Variety or richness

Variety measures *how many categories or types an entity has*. Variety is useful as a first approach to the diversity of an entity since the number of categories (e.g. species, scientific fields or export categories) is easy to understand and calculate. Users can compute variety both within the function `diversity()` indicating '`type='v''`, or with the function `variety()`. Both options return a '`dataframe`' with the values of variety. In the case of the function `variety()` values are sorted in an decreasing order. For an increasing order, the argument '`decreasing`' should be set to '`FALSE`'.

For instance, we can compare the variety of the 10 countries included in our sample of Pantheon. Canada and China rank at the top of variety, while Uruguay and Vietnam rank at the bottom. In Scidata, US and Germany (see Figure 2b) have the highest level of variety, while China (see Figure 2c) and Mexico have the lowest variety. It is important to note that we are only considering fields of science in which these countries have Revealed Comparative Advantages (RCAs) equal to or higher than 1.

```
%variety(data = pantheon)
 %            variety
%Canada            27
%China             24
%...
%Uruguay            4
%Vietnam            4

#using function values() to normalize the dataset
 scidat_rca_fil <- values(data = scidat, norm = 'rca', filter = 1)
 variety(scidat_rca_fil)
              variety
United States      17
Germany            16
...
China              10
Mexico              9
```

Being related to the concept of variety, it is helpful in some cases to know the *ubiquity* or rareness of each category by considering its presence in all entities. Ubiquity could also be considered as the variety of entities that each category has (Hidalgo et al., 2007). We include this concept and measure through the function `ubiquity()` that returns the number of entities in which the category is present. A decreasing order is retrieved by default. In our sample of Pantheon "politicians" and "soccer players" are more common (ubiquitous) than "referees" or "wrestlers".

```
ubiquity(data = pantheon)
                  ubiquity
Politician               10
Writer                    8
Soccer.Player             6
...
Referee                   1
Wrestler                  1
```

Diversity measures that emphasize abundance and balance

Balance measures *how much of each category the entity has*. The raw indicators of balance are the values of abundance or the relative values of abundance which are the proportions p_i of each i-th category. The package **diverse** includes the function `balance()` which retrieves the matrix of entities-categories with their correspondent shares, proportions or probabilities.

The word *balance* is used when the values of abundance are more equally distributed across the categories. For a given variety, a more balanced system is considered more diverse. Extreme cases are those where the quantity of elements for each category is exactly the same (i.e. perfect balance) or conversely, where all the elements are concentrated in just one category (i.e. total concentration).

As pointed out by Tuomisto (2012), measuring balance alone is a complicated task because it is difficult to remove the effect of variety on it. The only measure of "balance" in a strict sense that is facilitated in this package is Pielou's evenness (Pielou, 1970), which according to Jost (2010) is also the best measure of balance.

However, **diverse** also allows for the calculation of a series of commonly used "balance" measures like the Herfindahl-Hirschman Index (HHI), Gini-Simpson or Blau-Index that emphasize the evenness or balance of a system (while also being affected by the variety of categories).

Diversity measures related to the property "balance" could be understood as statistical dispersion and are mainly a function of p_i. While some of them measure the *evenness* or *heterogeneity* of the distribution such as the *Blau Index*, others emphasize the *concentration*, such as the *Herfindahl-Hirschman Index (HHI)*.

The Herfindahl-Hirschman Index (HHI), for example, computes the probability that two individuals taken randomly belong to the same category. This probability is calculated with replacement, which means that after taking the first individual into account, it is replaced with an identical one; and thus neither affecting the total number of individuals in that category (n_i) nor the total amount of individuals in the entity (N_t). HHI is used in economics, for instance, to estimate the concentration of markets or wealth (Ceriani and Verme, 2011).

Taking into account that balance is the opposite to concentration, the *Gini-Simpson* Index ($1 - HHI$) subtracts HHI from 1 to estimate balance. The same idea is behind the *Blau Index*. The Blau Index was created to measure the heterogeneity of social communities and its use is very common in sociology and other social sciences.

Similar to HHI, *Simpson* measure D_s has the same probabilistic idea of measuring concentration, but it computes the probability without replacement—meaning that the values of N_t and n_i decreases in 1 after the first probability is calculated (see Table 1). This measure of concentration and its equivalent balance or index of diversity ($1 - D_s$) are widespread in ecology. Moreover, the reciprocal index ($R_S = 1/D_S$) can be calculated.

In the following example, the *Herfindahl-Hirschman Index (HHI)*, the *Gini-Simpson Index* and the *Blau Index* from Pantheon are computed by using the function `diversity()`. We can observe that Uruguay and Vietnam have a higher HHI value and are thus more concentrated and less balanced than Canada and Chile. Note that the opposite occurs with the Gini-Simpson or Blau indexes. Besides the Gini-Simpson index, the concentration `gini.simpson.C` and the reciprocal of the concentration `gini.simpson.R` are also retrieved.

```
round(diversity(data = pantheon, type = c('hhi', 'gs', 'b','ev')), 3)
              HHI gini.simpson gini.simpson.C gini.simpson.R blau.index evenness
Canada      0.372        0.628          0.372          2.689      0.628    0.843
Chile       0.133        0.867          0.133          7.538      0.867    0.959
...
Uruguay     0.235        0.765          0.235          4.263      0.765    0.820
Vietnam     0.139        0.861          0.139
```

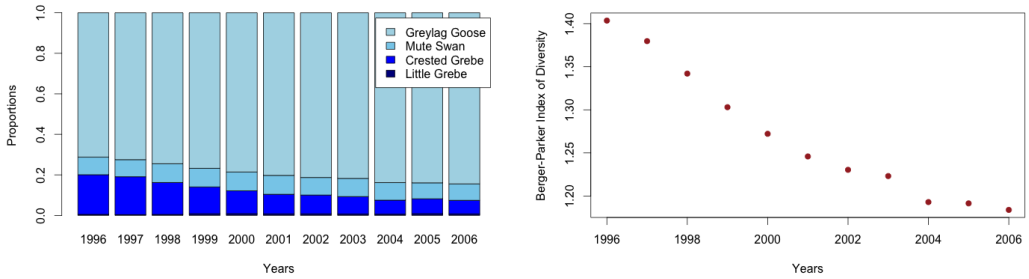
Graphical representations can help to understand the importance of balance and how it is captured by specific diversity measures. Figure 4a illustrates how the share of the dominant species Greylag Goose increases over time and how the share of Crested Grebe declines. The result is an unbalance between the species in this ecosystem.

The decrease in diversity of geese —understood here mainly as the balance of the abundance of different types of geese—can for instance be captured by *Berger-Parker* measures. The Berger-Parker Dominance Index (D_{BP}) is a measure based on the dominant category ($max(p_i)$) and thus captures the dominance of the Greylag Goose. On the other hand, the Berger-Parker Index of Diversity ($I_{BP} = 1/D_{BP}$) captures the balance between the species. Figure 4b shows how the Berger-Parker Index of Diversity decreases over time.

```
bal <- balance(geese, category_row = TRUE) #note the function balance
barplot(t(bal), legend = TRUE, xlab = "Years", ylab = "Proportions",
col=c("darkblue","blue","sky blue", "light blue") )

 bp <- diversity(geese, type = 'bp', category_row = TRUE)
 plot(bp$berger.parker.I~rownames(bp), xlab = "Years",
       ylab = "Berger-Parker Index of Diversity", pch = 19, col = "brown")


diversity(data = geese, type = c('e','ev','s','bp'), category_row = TRUE)
       entropy  evenness simpson.D simpson.I simpson.R berger.parker.D berger.parker.I
1996 0.7993160 0.5765846 0.5534977 0.4465023  1.806692       0.7124871        1.403534
1997 0.7764028 0.5600563 0.5674638 0.4325362  1.762227       0.7247953        1.379700
...
2005 0.5790954 0.4177290 0.7160910 0.2839090  1.396471       0.8392823        1.191494
2006 0.5633026 0.4063369 0.7245616 0.2754384  1.380145       0.8446653        1.183901
```



(a) Proportions of geese species. **(b)** Berger Parker Index of diversity.

Figure 4: Analysis of balance for Geese dataset.

Entropy measures and Hill numbers

Shannon Entropy is a frequently used measure of balance and diversity. Entropy is a measure first created and used in information theory (Shannon, 1948) and has been widely adopted by other disciplines such as computation, ecology, and economy.

Entropy H measures the minimum volume of communication required to code a message. Furthermore, as pointed by Hidalgo (2015, p.17), "entropy is a measure of the multiplicity of states". A high value of multiplicity of states (categories) implies more evenness and less concentration: as a consequence, the higher the variety and the balance, the higher the entropy. In the example above we could observe how the entropy in Geese decreased from 0.78 in 1996 to 0.56 in 2006. This is a consequence of an increase of the population of the dominant species.

A generalization of Shannon Entropy that is also included in **diverse** is Rényi's entropy (see Table 1). Rényi's entropy allows the users to give more or less relative importance to rare categories through the parameter q.

Another parameterized entropy is the HCDT entropy (Havrda and Charvát, 1967; Daróczy, 1970; Tsallis, 1988). It is noteworthy that Variety, Shannon, Blau's and Berger-Parker's indexes are special cases of HCDT (respectively with the parameter $q = 0$, 1, 2 and infinity).

Finally, Hill numbers (Hill, 1973) are a mathematically unified family of diversity indexes that differ only by a parameter q and that take the effective number of categories into account, i.e. the number of equally abundant species that would be needed to give the same value of a diversity measure (Chao et al., 2014b). Hill numbers are of particular interest since entropy is not linear to the number of categories hosted by an entity Jost (2006). Moreover, several widely used diversity indexes, like variety/richness, Shannon entropy, Gini-Simpson Index, Rényi's or HDCT entropy, can be obtained from Hill numbers (Chao et al., 2014a).

By using **diverse** we can observe the similarities between entropy measures and Hill numbers, when q has values of 0, 1 and 2.

When $q = 0$, variety, HCDT entropy and Hill numbers are the same. Rényi entropy is equal to $\log variety$. When $q = 1$, Rényi entropy and HCDT entropy are equal to Shannon entropy (H), while Hill numbers are equal to the exponential of H. When $q = 2$, HCDT entropy is equal to Gini-Simpson, while the Hill numbers index is equal to the reciprocal (`gini.simpson.R`) of the index of concentration of Gini (or Herfindahl-Hirschman Index (`gini.simpson.C`)).

```
diversity(pantheon, type=c("v","hcdt","hn","re"), q=0)[1,]
       variety hcdt.entropy hill.numbers renyi.entropy
Canada      27           27           27      3.295837

diversity(pantheon, type=c("e","re","hcdt", "hn"), q=1)[2,]
       entropy renyi.entropy hcdt.entropy hill.numbers
Chile 1.626709      1.626709     1.626709     5.087107

diversity(pantheon, type=c("hcdt","gs","hn"), q=2)[3,]
      hcdt.entropy gini.simpson gini.simpson.C gini.simpson.R hill.numbers
China    0.8168554    0.8168554      0.1831446       5.460167     5.460167
```

Generally for the three parameterized measures that are included in **diverse** (i.e. Rényi Entropy, HCDT and Hill numbers), the parameter value $q = 0$ calculates variety, $q < 1$ considers rare categories more important for diversity, $q = 1$ considers all categories as equally important, and $q > 1$ mainly shows the impact of dominant categories in diversity.

Disparity

Another significant dimension of diversity is the disparity or dissimilarity between categories or entities (Stirling, 2007; Rafols and Meyer, 2009). Disparity is important for all diversity measures, though, often pre-given in form of classification schemes, like, for example, phylogenetic trees or types of species in ecology, or the type of research fields in scientometrics. Here, we explicitly take the diversity dimension disparity into account.

The dimension of disparity provides a notion of *how different the categories of an entity are*. For example the areas "Mathematics" and "Physics" are arguably more similar than "Mathematics" and "Nursing". Measures of diversity, therefore, are also closely related to distance and similarity measures like Euclidean distances, cosine similarity, Jaccard-Index or expert classifications of different categories.

Matrix of dissimilarities between entities

Beside computing disparity, the dissimilarity matrix between entities is also useful for the visualization of networks, such as those proposed to evaluate economic complexity (Hidalgo et al., 2007; Hartmann et al., 2016) or the research capabilities of scholars (Guevara et al., 2016). Moreover, it helps to analyze the portfolio of entities in so-called network overlay maps and to explore the path of diversification as a function of the disparity in the network (Rafols et al., 2010; Guevara et al., 2016).

Based on the 10 countries included in Scidat, we calculate the dissimilarities between categories and then we create a network of areas of science in the following example. The resulting network is the one presented in Figure 2a in Section Diversity.

```
adj <- dis_categories(data = scidat, method = 'cosine')
adj[adj > 0.015] <- 0  #filter

library(igraph)
g <- graph.adjacency(adjmatrix = adj, mode = 'undirected', weighted = TRUE)
totals <- colSums(values(scidat))
V(g)$size = log(totals[match(V(g)$name, names(totals))], base = 2) -9
fc <- fastgreedy.community(g); colors <- rainbow(max(membership(fc)))
V(g)$color = colors[membership(fc)]
set.seed(67); g$layout <- layout.fruchterman.reingold(g)
plot.igraph(g,  vertex.label.cex = 0.9, vertex.label.font = 0,
        vertex.label.family = 'Helvetica', vertex.label.color='black', asp = FALSE)
```

Calculating dissimilarities between entities

The function `dis_entities()` can be used to calculate a matrix of dissimilarities between entities. The following example computes the matrices of dissimilarities between countries (entities) for the 10 countries included in Scidat. In this example, Argentina is more similar to Mexico (0.04) and less similar to China (0.32). In addition, Germany is more similar to Hungary (0.02) and less similar to Singapore (0.10).

```
round(dis_entities(scidat, method = 'cosine'), 2)
              Argentina China Germany Hungary Iran Mexico Singapore...
Argentina          0.00  0.32    0.09    0.07 0.17   0.04      0.25
China              0.32  0.00    0.20    0.19 0.06   0.18      0.06
Germany            0.09  0.20    0.00    0.02 0.07   0.05      0.10
Hungary            0.07  0.19    0.02    0.00 0.07   0.03      0.11
Iran               0.17  0.06    0.07    0.07 0.00   0.07      0.05
Mexico             0.04  0.18    0.05    0.03 0.07   0.00      0.13
...
```

Average or Sum Disparity

The function `disparity()` computes the average and/or the sum of dissimilarities among categories, either based on a given dissimilarity matrix of the user or through calculating the dissimilarity matrix within the function. The first case is based on a matrix of dissimilarities that the user provides in the argument '`dis`'. The dissimilarity matrix has to include the same names of the categories in the '`rownames`' and in the '`colnames`'.

In the second case, when the argument 'dis' is not provided, **diverse** computes the disparities by using the dissimilarity matrix calculated by using the previously detailed function `dis_categories()`, as in the following example where the argument 'dis' is not defined.

In this example with Scidat, we note that the average dissimilarities of categories in the US are greater than the disparities in Argentina or China.

```
scidat_rca_fil <- values(scidat, norm = 'rca', filter = 1)
disparity(scidat_rca_fil)
              disparity.sum disparity.mean
Argentina         121.12704      0.3450913
China              54.86895      0.1563218
...
Spain             147.35440      0.4198131
United States     190.86552      0.5437764
```

Diversity measures that explicitly take variety, disparity and balance into account

The package includes also "full" measures of diversity that are able to capture variety, balance and disparity at the same time. These measures are *Rao* and *Rao-Stirling*, where the former is widespread in ecology, while the latter is more commonly applied in social sciences and scientometrics (Rafols, 2014; Wang et al., 2015).

Both measures compute the sum of the multiplication of the distances (disparity) and the proportions (balance) between the pairs of two distinct categories i and j (see Table 1). However, Rao-Stirling diversity allows users to assign the weights/parameters α and β according to the importance of the disparity or balance, respectively.

Rao diversity is equivalent to Rao-Stirling diversity with the parameter values $\alpha = \beta = 1$. These values are also the default values in the function `diversity()`. Note that when the argument 'dis' is not provided, the default method 'Euclidean distances' is used for the calculation of the dissimilarity matrix. Users can also provide their own dissimilarity matrix by using the argument 'dis' in the function `diversity()`.

In the following example from Scidat, we calculate Rao diversity as well as the Rao-Stirling diversity with the parameter values α =0.7 and β = 0.3 and cosine dissimilarities between the entities. This example shows that Rao-Stirling diversity provides the possibility to emphasize different aspects of diversity. When we use the Rao Index, then Spain is considered to be more "diverse" than the US, but when we assign more importance to disparity, by increasing the parameter α in the Rao-Stirling index, then the US is more "diverse".

```
scidat_rca_fil <- values(scidat, norm = 'rca', filter = 1)
diversity(data = scidat_rca_fil, type = c('rao', 'rs') ,
   alpha=0.7, beta = 0.3, method = 'cosine')
              rao.stirling       rao
                    rao rao.stirling
Argentina     0.1526983     7.072576
China         0.1346935     4.814975
...
Spain         0.2137356    12.842799
United States 0.1874783    12.864261
```

Thus, the Rao-Stirling index and the package **diverse** allows the user to analyze the impact of different similarity measures as well as different weights of disparity and balance on the resulting diversity values and rankings.

It must be noted that, so far, we focus in "diverse" on the Stirling taxonomy (Stirling, 2007, 1998). In ecology another set of "similarity-based" measures has been developed and can be accessed in the package **entropart** and **treescape**.

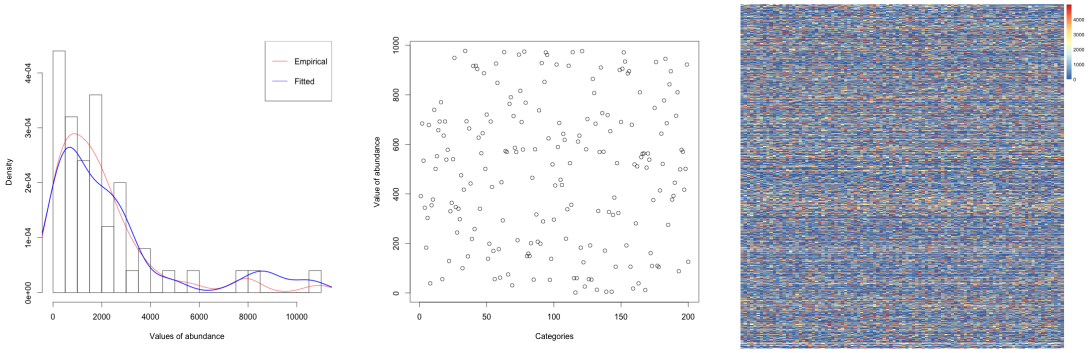
Synthetic data and performance tests

In this Section, we show how to use **diverse** to create synthetic data to the level of *individuals*, entities and datasets. An individual is an independent object that belongs to a category (e.g. a paper in a certain discipline of Scidat or a person in a certain type of occupation in Pantheon). Entities are constituted by a set of categories and their values of abundance. A set of several entities constitutes a synthetic dataset for the type of diversity measures we apply in **diverse**.

The functions included in **diverse** to simulate data are `sim_individuals`, `sim_entities` and `sim_dataset`. In each function the user can define the size (i.e. number of individuals), the required level of variety or richness and the method to define the distribution or values of abundance.

For example, if we want to create individuals of an entity, we can use the function `sim_individuals()` to generate synthetic data with 10000 individuals assigned to 50 different species (categories). Moreover, the values of absolute abundance of species can be, for instance, distributed according to a log normal distribution with $\mu = 0.50$ and $\sigma = 1.183$ (see histogram in Figure 5a and (Beck and Schwanghart, 2010)).

```
set.seed(99)
synt_ind <- sim_individuals(n_categ=50, size=100000,
  category_prefix='ctg', type='log-normal', mean=0.507, sd=1.183)
hist(table(synt_ind), breaks = 30, xlab = "Values of abundance",
 +  probability = TRUE, main = NULL)
lines(density(table(synt_ind)), col="red")
library(fitdistrplus)
f <- fitdist(as.vector(table(synt_ind)), "lnorm")
x = rlnorm(50, mean=f$estimate['meanlog'][[1]], sd = f$estimate['sdlog'][[1]])
lines(density(x), col="blue", lwd=2)
legend("topright",legend = c('Empirical', 'Fitted'), col = c("red", "blue"), lty=1)
head(synt_ind)
[1] "ctg49" "ctg3"  "ctg41" "ctg49" "ctg4"  "ctg25"
```



(a) Histogram of simulated 10K individuals

(b) Simulated entity with 200 categories

(c) Pheatmap of simulated dataset. 15K entities, 100 categories. Colors according to value of abundance

Figure 5: Analysis of simulated data

If the user wants to generate a simulated entity with values of abundance produced by the aggregation of individuals in categories, **diverse** provides the function `sim_entity`. This function allows the user to define a distribution and/or a required number of categories (`n_categ`). See Figure 5b.

```
sim_ent <- sim_entity(n_categ=200, values=sample(1:1000, replace=TRUE))
plot(sim_ent$Value, ylab = "Value of abundance", xlab="Categories")
head(sim_ent)
Category Value
1        1   757
2        2   124
...
```

In a last function, the simulation of a full dataset is also provided with **diverse**. The function `sim_dataset` allows for users to define the number of categories in each entity (variety) as well as the number of required entities. A crucial argument of the function `sim_dataset` is the vector of integers with the desired values of variety for each entity (`n_categ`). In the following example, we create a dataset of 1500 entities and 100 categories with random integer values of abundance between 10 and 5000. The values of variety for each entity are also randomly sampled between 1 and 100. The resulting dataset is retrieved as a dataframe of values of abundance. By using the function `values()` we can plot this dataset as a matrix (see Figure 5c).

```
n_entities <- 1500
v_values <- sample(10:5000, size= n_entities, replace=TRUE)
v_n_categ <- sample(1:100, size = n_entities, replace=TRUE)
data_set <- sim_dataset(n_categ = v_n_categ, values= v_values,
  category_prefix = "C", category_random = TRUE)
pheatmap(values(data_set),cluster_rows = FALSE, cluster_cols = FALSE,
  show_rownames = FALSE, show_colnames = FALSE)
head(data_set)
...
```

Performance

To test the performance of **diverse** we use the previously generated synthetic dataset (1500 entities by 100 categories), then we calculate the time used to perform two measures, namely Shannon entropy and Rao-Stirling diversity. Note that the second one is more time consuming since it involves the computation of a distance matrix. However, the time necessary to compute both measures is reasonable (0.021 and 2.697 seconds respectively). The time used to create the simulated dataset is more time consuming ($\tilde{3}5$ secs.) since the dataset must ensure that the assigned number of categories for each entity accomplishes the requirements. Still it is a reasonable amount of time considering the dimensions and characteristics of the obtained data.

```
system.time(data_set <- sim_dataset(n_categ = v_n_categ, values= v_values,
   +   category_prefix = "C", entity_prefix = "E"))
   user  system elapsed
 29.590   5.245  34.871
system.time(diversity(data_set, type=c("e")))
   user  system elapsed
  0.019   0.001   0.021
system.time(diversity(data_set, type=c("rs")))
   user  system elapsed
  2.478   0.206   2.697
```

Coverage, biases and caveats

The package **diverse** is designed to work with datasets with a known number of categories and a comparatively low level of variety (i.e. scientific fields, occupations or industrial sectors, in comparison to datasets in ecology with millions of species, including many unknown species). For instance, in ecology it has been demonstrated that diversity measures are biased in cases of small samples (e.g. in a very limited spatial area, limited amount of soil, etc.). Accordingly in datasets on biodiversity, it is difficult to sample rare species appropriately (Beck and Schwanghart, 2010). To solve this issue associated to this type of datasets, measures of bias correction, e.g. of Shannon Entropy, have been proposed (Chao and Shen, 2003). These measures, mainly used in the area of ecology and biodiversity, are not yet implemented in **diverse**. Furthermore other advanced measures considering for example phylogenetic diversity or functional diversity, such as the generalization of the Rao's cuadratic entropy (Chao et al., 2014a) are not yet included. But to address these current limitations it must be noted that **diverse** can also be used in combination with several specialized packages as **entropart**, **vegan** or **spadeR**. For instance, **diverse** provides a function (`to_entropart`) that allows the user to transform the datasets from the package **diverse** into values of abundance to be used in **entropart**. Here we present a simple example to compute the richness of the *metacommunity* (see **entropart** manual for details (Marcon and Hérault, 2015)) generated with our synthetic dataset (variable '`data_set`' of previous example).

```
library(entropart)
abundance <- to_entropart(data_set)
mc <- MetaCommunity(abundance)
Richness(mc$Ps);Shannon(mc$Ps)
```

It must also be noted that the ability to statistically test differences in diversity measures across time and across systems could provide important insights for researchers. The following example shows a strong linear correlation between the Shannon Entropy in two different time steps of Scidat (dataframes `scidat` for 2013, and `scidat_2` for 2003). Further statistical test functions need to be implemented in subsequent versions of of **diverse**, in order to also allow for the testing of the differences across systems. **diverse** will continue to learn from other disciplines, with the aim of implementing and

adapting statistical test functions to the particular needs of researchers exploring the diversity in complex socioeconomic systems.

```
d_1 <- diversity(scidat, type="e")
d_2 <- diversity(scidat_2, type="e")
cor.test(d_1[,1], d_2[,1])

        Pearson's product-moment correlation

        data:  d_1[, 1] and d_2[, 1]
t = 3.7171, df = 8, p-value = 0.005896
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.3330683 0.9496172
sample estimates:
     cor
0.795807
```

Conclusions

This paper introduced the package **diverse** which allows users to compute some of the most common measures of diversity from different fields of science. In summary, measuring diversity has become an important topic in many disciplines which analyze complex systems. The R package **diverse** allows for a combination of common measures from several disciplines and recent approaches from interdisciplinary research.

It must be noted that **diverse** has limitations that we aim to address in subsequent versions of the package. Possible future improvements include methods of considering diversity at different levels of aggregation (in hierarchical classification schemes, like mammals and insects, or agricultural or industrial goods, natural or social sciences, and their respective subcategories). Moreover, further emphasis on the role of different similarity measures at different levels of aggregations, as well as analyzing estimation error biases in incomplete samples are important future research areas in the measurement of diversity in socioeconomic systems, where social sciences can significantly learn from ecology and biology. Finally **diverse** can also continue to learn from ecology, biology and other disciplines about how to apply statistical tests on the differences of diversity measures across systems.

In general, **diverse** offers a toolkit to analyze and visualize the diversity of entities, categories and complex systems that is useful in particular for social scientists and interdisciplinary social research, as well as beginners in ecology and natural sciences. The package **diverse** provides different data import and export options and allows for the calculation of the different data transformations and similarity matrices, diversity measures and diversity visualization options.

In order to present the functions provided by the package, we took advantage of an interdisciplinary taxonomy of diversity that defines variety, balance and disparity as three dimensions of diversity (Stirling, 2007). This taxonomy favors the creation of interdisciplinary bridges and helps in understanding how each diversity measure captures different aspects of diversity.

Acknowledgments

We would like to thank Ismael Rafols, Diego Chavarro, Daniele Rotolo and Andy Stirling for the example codes and valuable comments on diversity measures. We are also grateful for the valuable comments made by two anonymous reviewers. MG and MM acknowledge the Program of Incentives to Scientific Initiation (PIIC) from DGIP at Universidad Técnica Federico Santa María. MG thanks internal project ING01-1516 from Universidad de Playa Ancha. DH acknowledges support from the Marie Curie International Outgoing Fellowship No. 328828 within the 7th European Community Framework Programme. MM and MG acknowledges support from project FONDECYT 11121435.

Bibliography

B. Balassa. Trade Liberalisation and "Revealed" Comparative Advantage. *The Manchester School*, 33(2): 99–123, 1965. ISSN 1467-9957. doi: 10.1111/j.1467-9957.1965.tb00050.x. URL http://onlinelibrary.wiley.com/doi/10.1111/j.1467-9957.1965.tb00050.x/abstract. [p6]

B. Balassa. Comparative advantage in manufactured goods: a reappraisal. *The Review of Economics and Statistics*, pages 315–319, 1986. [p5]

J. Beck and W. Schwanghart. Comparing measures of species diversity from incomplete inventories: an update. *Methods in Ecology and Evolution*, 1(1):38–44, Mar. 2010. ISSN 2041-210X. doi: 10.1111/j.2041-210X.2009.00003.x. URL http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2009.00003.x/abstract. [p13, 14]

W. H. Berger and F. L. Parker. Diversity of Planktonic Foraminifera in Deep-Sea Sediments. *Science*, 168(3937):1345–1347, Dec. 1970. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.168.3937.1345. URL http://www.sciencemag.org/content/168/3937/1345. [p7]

P. M. Blau. *Inequality and heterogeneity: A primitive theory of social structure*, volume 7. Free Press New York, 1977. [p7]

L. Ceriani and P. Verme. The origins of the Gini index: extracts from Variabilità e Mutabilità (1912) by Corrado Gini. *The Journal of Economic Inequality*, 10(3):421–443, June 2011. ISSN 1569-1721, 1573-8701. doi: 10.1007/s10888-011-9188-x. URL http://link.springer.com/article/10.1007/s10888-011-9188-x. [p9]

A. Chao and T.-J. Shen. Nonparametric estimation of Shannon's index of diversity when there are unseen species in sample. *Environmental and Ecological Statistics*, 10(4):429–443, Dec. 2003. ISSN 1352-8505, 1573-3009. doi: 10.1023/A:1026096204727. [p14]

A. Chao, C.-H. Chiu, and L. Jost. Unifying Species Diversity, Phylogenetic Diversity, Functional Diversity, and Related Similarity and Differentiation Measures Through Hill Numbers. *Annual Review of Ecology, Evolution, and Systematics*, 45(1):297–324, 2014a. doi: 10.1146/annurev-ecolsys-120213-091540. URL http://dx.doi.org/10.1146/annurev-ecolsys-120213-091540. [p10, 14]

A. Chao, N. J. Gotelli, T. Hsieh, E. L. Sander, K. Ma, R. K. Colwell, and A. M. Ellison. Rarefaction and extrapolation with Hill numbers: a framework for sampling and estimation in species diversity studies. *Ecological Monographs*, 84(1):45–67, 2014b. [p10]

D. Chavarro, P. Tang, and I. Rafols. Interdisciplinarity and research on local issues: evidence from a developing country. *Research Evaluation*, 23(3):195–209, 2014. [p1]

G. Csardi and T. Nepusz. The igraph software package for complex network research. *InterJournal*, Complex Systems:1695, 2006. URL http://igraph.org. [p3]

Z. Daróczy. Generalized information functions. *Information and control*, 16(1):36–51, 1970. [p10]

V. Debastiani and V. Pillar. SYNCSA - R tool for analysis of metacommunities based on functional traits and phylogeny of the community components. *Bioinformatics*, 28:2067–2068, 2012. [p1]

N. Eagle, M. Macy, and R. Claxton. Network Diversity and Economic Development. *Science*, 328 (5981):1029–1031, May 2010. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1186605. URL http://science.sciencemag.org/content/328/5981/1029. [p1]

J. Farchy and H. Ranaivoson. Measuring the Diversity of Cultural Expressions: Applying the Stirling Model of Diversity in Culture. Technical Report 6, Unesco Institute for Statistics, 2011. [p1]

P. Forey, C. Humphries, and R. Vane-Wright. Systematics and conservation evaluation. pages xxvi + 438 pp. Clarendon Press, 1994. [p1]

K. Frenken, F. V. Oort, and T. Verburg. Related variety, unrelated variety and regional economic growth. *Regional Studies*, 41(5):685–697, July 2007. ISSN 0034-3404. doi: 10.1080/00343400601120296. URL http://dx.doi.org/10.1080/00343400601120296. [p1]

C. Gini. *Variabilità e Mutuabilità. Contributo allo Studio delle Distribuzioni e delle Relazioni Statistiche.* Bologna, c. cuppini edition, 1912. [p7]

M. R. Guevara, D. Hartmann, M. Aristarán, M. Mendoza, and C. A. Hidalgo. The research space: using career paths to predict the evolution of the research output of individuals, institutions, and nations. *Scientometrics*, pages 1–15, Sept. 2016. ISSN 0138-9130, 1588-2861. doi: 10.1007/s11192-016-2125-9. URL http://link.springer.com/article/10.1007/s11192-016-2125-9. [p1, 11]

D. Hartmann, M. R. Guevara, C. Jara-Figueroa, M. Aristarán, and C. A. Hidalgo. Linking Economic Complexity, Institutions and Income Inequality. *arXiv:1505.07907 [physics, q-fin]*, 2016. URL http://arxiv.org/abs/1505.07907. arXiv: 1505.07907. [p11]

D. M. A. Haughton and S. Mukerjee. The economic measurement and determinants of diversity. *Social Indicators Research*, 36(3):201–225, Nov. 1995. ISSN 0303-8300, 1573-0921. doi: 10.1007/BF01078814. URL http://link.springer.com/article/10.1007/BF01078814. [p1]

R. Hausmann, C. A. Hidalgo, S. Bustos, M. Coscia, S. C. (M.A.), J. Jimenez, A. Simões, and M. A. Yıldırım. *The Atlas of Economic Complexity: Mapping Paths to Prosperity*. Center for International Development, Harvard University, 2011. ISBN 978-0-615-54662-9. [p3]

J. Havrda and F. Charvát. Quantification method of classification processes. Concept of structural $ a $-entropy. *Kybernetika*, 3(1):30–35, 1967. [p7, 10]

C. Hidalgo. *Why Information Grows: The Evolution of Order, from Atoms to Economies*. Basic Books, June 2015. ISBN 978-0-465-04899-1. [p10]

C. A. Hidalgo and R. Hausmann. The building blocks of economic complexity. *Proceedings of the National Academy of Sciences*, 106(26):10570–10575, June 2009. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.0900943106. URL http://www.pnas.org/content/106/26/10570. [p1]

C. A. Hidalgo, B. Klinger, A.-L. Barabási, and R. Hausmann. The product space conditions the development of nations. *Science*, 317(5837):482–487, July 2007. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1144581. URL http://www.sciencemag.org/content/317/5837/482. [p1, 3, 8, 11]

M. O. Hill. Diversity and Evenness: A Unifying Notation and Its Consequences. *Ecology*, 54(2):427–432, Mar. 1973. ISSN 0012-9658. doi: 10.2307/1934352. URL http://www.esajournals.org/doi/abs/10.2307/1934352. [p7, 10]

T. Jombart, M. Kendall, J. Almagro-Garcia, and C. Colijn. *treescape: Statistical Exploration of Landscapes of Phylogenetic Trees*. 2016. URL http://CRAN.R-project.org/package=treescape. R package version 1.8.16. [p1]

L. Jost. Entropy and diversity. *Oikos*, 113(2):363–375, May 2006. ISSN 1600-0706. doi: 10.1111/j.2006.0030-1299.14714.x. URL http://onlinelibrary.wiley.com/doi/10.1111/j.2006.0030-1299.14714.x/abstract. [p10]

L. Jost. The Relation between Evenness and Diversity. *Diversity*, 2(2):207–232, Feb. 2010. doi: 10.3390/d2020207. URL http://www.mdpi.com/1424-2818/2/2/207. [p8]

K. Keenan, P. McGinnity, T. F. Cross, W. W. Crozier, and P. A. Prodöhl. diveRsity: An R package for the estimation of population genetics parameters and their associated errors. *Methods in Ecology and Evolution*, 4(8):782–788, 2013. doi: 10.1111/2041-210X.12067. URL http://onlinelibrary.wiley.com/doi/10.1111/2041-210X.12067/abstract. R package version 1.9.89. [p1]

R. Kindt and R. Coe. *Tree diversity analysis. A manual and software for common statistical methods for ecological and biodiversity studies*. World Agroforestry Centre (ICRAF), Nairobi (Kenya), 2005. URL http://www.worldagroforestry.org/treesandmarkets/tree_diversity_analysis.asp. ISBN 92-9059-179-X. [p1, 2]

R. Kolde. *pheatmap: Pretty Heatmaps*. 2015. URL http://CRAN.R-project.org/package=pheatmap. R package version 1.0.7. [p2]

L. I. Kuncheva and C. J. Whitaker. Measures of Diversity in Classifier Ensembles and Their Relationship with the Ensemble Accuracy. *Machine Learning*, 51(2):181–207, May 2003. ISSN 0885-6125, 1573-0565. doi: 10.1023/A:1022859003006. [p1]

E. Laliberté and P. Legendre. A distance-based framework for measuring functional diversity from multiple traits. *Ecology*, 91(1):299–305, 2010. ISSN 0012-9658. doi: 10.1890/08-2244.1. URL http://www.esajournals.org.ezproxy.canterbury.ac.nz/doi/abs/10.1890/08-2244.1. [p1]

E. Marcon and B. Hérault. entropart: An R Package to Measure and Partition Diversity. *Journal of Statistical Software*, 67(8):1–26, 2015. doi: 10.18637/jss.v067.i08. [p1, 2, 14]

Z. Marion, J. Fordyce, and B. Fitzpatrick. *hierDiversity: Hierarchical Multiplicative Partitioning of Complex Phenotypes*. 2015. URL http://CRAN.R-project.org/package=hierDiversity. R package version 0.1. [p1]

D. G. Mcdonald and J. Dimmick. The Conceptualization and Measurement of Diversity. *Communication Research*, 30(1):60–79, Jan. 2003. ISSN 0093-6502, 1552-3810. doi: 10.1177/0093650202239026. URL http://crx.sagepub.com/content/30/1/60. [p3]

M. M. MediaLab. Pantheon - Mapping historical cultural production, 2014. URL http://pantheon.media.mit.edu/methods. [p2]

D. Meyer and C. Buchta. *proxy: Distance and Similarity Measures*. 2015. URL http://CRAN.R-project.org/package=proxy. R package version 0.4-15. [p2]

Nature. Diversity challenge. *Nature*, 513(7518):279–279, Sept. 2014. ISSN 0028-0836, 1476-4687. doi: 10.1038/513279a. URL http://www.nature.com/doifinder/10.1038/513279a. [p1]

S. V. Nederland. Sovon Home Page, 2015. URL https://www.sovon.nl/en. [p2]

J. Oksanen, F. G. Blanchet, R. Kindt, P. Legendre, P. R. Minchin, R. B. O'Hara, G. L. Simpson, P. Solymos, M. H. H. Stevens, and H. Wagner. *vegan: Community Ecology Package*. 2016. URL http://CRAN.R-project.org/package=vegan. R package version 2.3-5. [p1]

E. C. Pielou. *Introduction to Mathematical Ecology*. John Wiley & Sons Inc, New York, Jan. 1970. ISBN 978-0-471-68918-8. [p7, 8]

M. Pietrzak, M. Seweryn, and G. Rempala. *divo: Tools for Analysis of Diversity and Similarity in Biological Systems*. 2016. URL http://CRAN.R-project.org/package=divo. R package version 0.1.2. [p1]

R Core Team. *foreign: Read Data Stored by Minitab, S, SAS, SPSS, Stata, Systat, Weka, dBase, ...* 2015. URL http://CRAN.R-project.org/package=foreign. R package version 0.8-65. [p5]

I. Rafols. Knowledge Integration and Diffusion: Measures and Mapping of Diversity and Coherence. In Y. Ding, R. Rousseau, and D. Wolfram, editors, *Measuring Scholarly Impact*, pages 169–190. Springer International Publishing, 2014. ISBN 978-3-319-10376-1 978-3-319-10377-8. URL http://link.springer.com/chapter/10.1007/978-3-319-10377-8_8. [p1, 12]

I. Rafols and M. Meyer. Diversity and network coherence as indicators of interdisciplinarity: case studies in bionanoscience. *Scientometrics*, 82(2):263–287, June 2009. ISSN 0138-9130, 1588-2861. doi: 10.1007/s11192-009-0041-y. URL http://link.springer.com/article/10.1007/s11192-009-0041-y. [p11]

I. Rafols, A. L. Porter, and L. Leydesdorff. Science overlay maps: A new tool for research policy and library management. *Journal of the American Society for Information Science & Technology*, 61(9):1871–1887, Sept. 2010. ISSN 15322882. doi: 10.1002/asi.21368. URL http://libproxy.mit.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=53286068&site=eds-live. [p1, 3, 4, 11]

C. R. Rao. Diversity and dissimilarity coefficients: A unified approach. *Theoretical Population Biology*, 21(1):24–43, Feb. 1982. ISSN 0040-5809. doi: 10.1016/0040-5809(82)90004-1. URL http://www.sciencedirect.com/science/article/pii/0040580982900041. [p3, 7]

S. A. Rhoades. Herfindahl-Hirschman Index, The. *Federal Reserve Bulletin*, 79:188, 1993. URL http://heinonline.org/HOL/Page?handle=hein.journals/fedred79&id=376&div=&collection=. [p7]

A. Rényi. On Measures of Entropy and Information. The Regents of the University of California, 1961. URL http://projecteuclid.org/euclid.bsmsp/1200512181. [p7]

R. Scherer and P. Pallmann. *simboot: Simultaneous inference for diversity indices*. 2014. URL http://CRAN.R-project.org/package=simboot. R package version 0.2-5. [p1]

SCImago. Scimago Journal & Country Rank, 2007. URL http://www.scimagojr.com/. [p2]

C. E. Shannon. A mathematical theory of communication. *The Bell System technical Journal*, 27: 379–423,623–656, Oct. 1948. [p7, 10]

E. H. Simpson. Measurement of diversity. *Nature*, 163, 1949. [p7]

A. Stirling. On the economics and analysis of diversity. *Science Policy Research Unit (SPRU), Electronic Working Papers Series, Paper*, 28:1–156, 1998. [p3, 12]

A. Stirling. A general framework for analysing diversity in science, technology and society. *Interface The Journal of Royal Society*, 4(15):707–719, Aug. 2007. [p2, 3, 7, 11, 12, 15]

M. Tennekes. *treemap: Treemap Visualization*. 2016. URL http://CRAN.R-project.org/package=treemap. R package version 2.4-1. [p3]

C. Tsallis. Possible generalization of Boltzmann-Gibbs statistics. *Journal of statistical physics*, 52(1-2): 479–487, 1988. [p7, 10]

H. Tuomisto. An updated consumer's guide to evenness and related indices. *Oikos*, 121(8):1203–1218, Aug. 2012. ISSN 1600-0706. doi: 10.1111/j.1600-0706.2011.19897.x. URL http://onlinelibrary.wiley.com/doi/10.1111/j.1600-0706.2011.19897.x/abstract. [p8]

C. S. Wagner, J. D. Roessner, K. Bobb, J. T. Klein, K. W. Boyack, J. Keyton, I. Rafols, and K. Börner. Approaches to understanding and measuring interdisciplinary scientific research (IDR): A review of the literature. *Journal of Informetrics*, 5(1):14–26, Jan. 2011. ISSN 1751-1577. doi: 10.1016/j.joi.2010.06.004. URL http://www.sciencedirect.com/science/article/pii/S1751157710000581. [p1]

C. Wang, M. Genkin, G. Berry, L. Chen, and M. Brashearswork. *Blaunet: Calculate and Analyze Blau Status for Measuring Social Distance*. 2016. URL http://CRAN.R-project.org/package=Blaunet. R package version 2.0.4. [p1]

J. Wang, B. Thijs, and W. Glänzel. Interdisciplinarity and Impact: Distinct Effects of Variety, Balance, and Disparity. *PLoS ONE*, 10(5):e0127298, May 2015. doi: 10.1371/journal.pone.0127298. URL http://dx.doi.org/10.1371/journal.pone.0127298. [p12]

A. Z. Yu, S. Ronen, K. Hu, T. Lu, and C. A. Hidalgo. Pantheon 1.0, a manually verified dataset of globally famous biographies. *Scientific data*, 3, 2016. [p2]

Miguel R. Guevara
Computer Science Department, Universidad de Playa Ancha, and
Department of Informatics, Universidad Técnica Federico Santa María
Valparaíso
Chile
miguel.guevara@upla.cl

Dominik Hartmann
Chair of Innovation Management and Innovation Economics, University of Leipzig
Grimmaische Straße 12, 04109, Leipzig
Fraunhofer Center for International Management and Knowledge Economy
Neumarkt 9-19, 04109, Leipzig
Germany
dominik.hartmann@uni-leipzig.de

Marcelo Mendoza
Department of Informatics, Universidad Técnica Federico Santa María
Av. Vicuna Mackeña 3939, San Joaquín, Santiago
Chile
mmendoza@inf.utfsm.cl