February 2016.

Supplementary Material for *Important, Unique, Central: Species' Relevance in Food Webs*

Giulio Valentino Dalla Riva
University of Canterbury, New Zealand
gvd16@uclive.ac.nz

Carey E. Priebe
John Hopkins University, MD, USA

SUPPLEMENTARY METHODS

The Random Dot Product Graph model: context and details

In our paper we introduce the Random Dot Product Graph model (RDPG) for food webs. This model can be seen as a special case of the stochastic block graph model, originally developed for the analysis of undirected social networks (Holland *et al.* 1983) and subsequently generalised to directed graphs (Wang & Wong 1987). Under the stochastic blockmodel assumptions, each of the nodes of a network are assigned to one of K distinct blocks. The probability of a link within and between the K blocks are given by the model parameters. However, in practice we do not observe the assignment of the nodes to the blocks; rather, we observe the interactions between nodes and try to estimate the assignment. This approach has been by some recent models for food webs (Allesina & Pascual 2009, baskerville2011spatial). Recently, it has been proved that a consistent block estimator (i.e., an estimator such that the proportion of nodes assigned to the wrong group converges in probability to zero as

the number of the species grows to infinity) based on the spectral partitioning of the normalized Laplacian of the adjacency matrix exist (Rohe *et al.* 2011). The RDPG model can be read as a particular case of the stochastic blockmodels (Fishkind *et al.* 2013) where each species is assigned to different block and the probability of the interaction from block (species) i to block (species) j depends on the distance between i and j in the metric space, i.e., it is given by the dot product of the two position vectors. We observe the realized interactions and estimate the position of the species in the metric space. Notice that, as the interaction probabilities are given by the pairwise distances, and the observed graph is an outcome of a stochastic process defined by those probabilities, any transformation of the metric space preserving the distance structure begets an equivalent model.

In particular, the RDPG model we consider in the paper is a generalization of the RDPG model to binary directed graphs: as the interactions are no more symmetric, we have to consider a pair of metric spaces, an *inward* and an *outward* one. We use an estimator based on the spectral partitioning of adjacency (non Laplacian) matrix of a graph (Sussman *et al.* 2012) to estimate species positions in the underlying metric space.

With some abuse of notation, we let A denote both a food web and its adjacency matrix. Let the three matrices L, Σ, R denote a singular value decomposition of the adjacency matrix A (a food web with S species). Thus, the matrices L, Σ, R satisfy $A = L \times \Sigma \times R^t$; L and R are real, orthogonal $S \times S$ matrices; Σ is an $S \times S$ diagonal matrix whose entries are the singular values of A, in a non-decreasing order. Fixed a model dimension d, we define three new matrices:

1. L', an $S \times d$ matrix given by the first d columns of L;
2. R', an $S \times d$ matrix given by the first d columns of R;
3. $(\Sigma')^{1/2}$, an $d \times d$ diagonal matrix defined by the square root of the first d entries of Σ, i.e., the square root of the first d greatest singular values of A.

Then, we let $\hat{L}$ denote the rescaled matrix $L' \times (\Sigma')^{1/2}$ and we let $\hat{R}$ denote the rescaled matrix $(\Sigma')^{1/2} \times R'$. The two matrices $\hat{l}$ and $\hat{R}$ capture the d leading traits for the community of species as prey and predators: the rows of $\hat{L}$ define the species' vulnerability functional traits (*outward*) and the rows of $\hat{R}$ define the foraging functional traits (*inward*) of the species in A. We call the column binding of $\hat{L}$ and $\hat{R}$ the species functional traits as both prey and predators (*total*).

Notice that although $\hat{L}$ and $\hat{R}$ are not uniquely defined, as any orthogonal transformation of those matrices preserve their dot product (and hence the distance between the estimated stochastic food web's backbone and its observed adjacency matrix), the species' relative position and the induced pairwise distance structure in the abstract functional space are uniquely defined. The measures we introduce in the paper are

based on the invariant structure of the food web, rather then on the absolute position of the species.

Different approaches are available for the choice of a suitable dimension d range. This akin to a dimensionality reduction problem, discussed for the Principal Component Analysis scenario in (Jolliffe 2002). The available methods include *a priori* selection procedures (e.g., the visual analysis of the singular values scree plot (Cattell 1966), hard singular values thresholding (Chatterjee & others 2014,Gavish & Donoho (2014)) or the maximization of a profile likelihood function (Zhu & Ghodsi 2006)) and *a posteriori* maximization of a goodness-of-fit criterion, as we exampled in (Dalla Riva & Stouffer 2015). In the datasets analysed all the previous methods provided compatible results.

The RDPG based measures Let $X(A)$ denote the matrix of (outward, inward or total) functional traits of the species in the food web A. As we saw above, the pairwise distance structure induced by $X(A)$ is uniquely defined, and hence it is possible to investigate the distribution of the species in the functional traits space just estimated. In particular, we define the (outward, inward or total) *uniqueness* of a species i in the food web A as the mean distance between i and every other species j in the (outward, inward or total) functional traits space. Let $d\left(p,q\right)$ denote the d dimensional euclidean distance between the point p and q; let $\langle f(i,j)\rangle_j$ be the mean of the function f over all the species j except i. That is, $\langle f(i,j)\rangle_j = \frac{1}{S}\sum_{j\neq i}(f(i,j))$. Then, the **uniqueness** of species i is defined as:

$$\text{uniqueness}(i) := \langle d\left(X(A)_i, X(A)_j\right)\rangle_j \tag{1}$$

Let M denote a $N \times M$ matrix: we denote $M^{d(i)}$ the matrix obtained by dropping the i-th column and row from M; te denote $M^{r(i)}$ the matrix obtained by removing just the i-th row from M. Let W denote another $N \times M$ matrix. We define $M_{proc}(W)$ as the Procrustes transformation (i.e., a combination of translation, rotation and uniform rescaling) of M of minimal distance to W; we will drop the argument (W) from the notation whenever it is clear from the context. Finally, we denote $||M||_F$ the squared sum of entries of M, i.e., the Frobenius norm of the matrix M . In particular, $||M_{proc}(W) - W||_F$ is also called the Procrustes distance between M and W (Dryden & Mardia 1998).

We compute the matrix

$$\left[X\left(A^{d(i)}\right)\right]_{proc}\left(X\left(A\right)^{r(i)}\right),$$

that is, the Procrustes transformation of $X\left(A^{d(i)}\right)$ of minimal distance to $X(A)^{r(i)}$ and we denote it $\hat{X}\left(A^{d(i)}\right)$. We define the rank d **strain** of the species i as the sum

of squared entries of the differences between $X(A)^{r(i)}$ and $\hat{X}\left(A^{d(i)}\right)$. Being $X(A)$ the matrix of either the inward, outward or total d dimensional functional traits, we will speak of species' *inward*, *outward* or *total* strain, respectively. In formula:

$$\text{strain}(i) := ||X(A)^{r(i)} - \hat{X}\left(A^{d(i)}\right)||_F \tag{2}$$

Computing the strain of species i reduces to computing the ordinary Procrustes distance between the matrices $X(A)^{r(i)}$ and $X\left(A^{d(i)}\right)$, which is implemented by the *procOPA* function in the R package *shapes* (Dryden 2013,R Core Team (2014)).

Finally, we identify the (outward, inward, total) diversity of the food web A as the volume of the convex hull containing all the species in the (outward, inward, total) functional traits space. Conceptually borrowing from the functional traits literature, we define the **contribution** of the species i to the food web (outward, inward, total) **functional diversity** as the volume difference of the convex hulls of $X\left(A\right)$ and $X\left(A\right)^{r(i)}$. We compute the convex hulls and their volumes using Qhull (Barber *et al.* 1996), through the *R* package *geometry* (Habel *et al.* 2014).

Keystone Centralities We compare our novel measures with six graph centralities measures that have been adopted to idenfity keystone species in ecological networks. In particular, for each species i in the food web, we consider:

- the betweennessi (BC, Freeman 1977) of species i, given by the number $s_{jij'}$ of shortest paths connecting every pair j, j' of species in the food web traversing species ii, weighted by the total number $s_{jj'}$ of paths between j and j':

$$BC(i) := \frac{s_{jij'}}{s_{jj'}}$$

- the closeness (CC, Bavelas 1950) of species i, defined as the reciprocal of the sum of the path distances, $d_p(\cdot)$, from i to every other species j in the food web:

$$CC(i) := \left(\sum_{j \neq i} d_p(j,i)\right)^{-1}$$

- the degree (DC) of species i, measuring the number of interactions involving the species i (both as a predator or as a prey):

$$DC(i) := |\left\{j \in A | i \to j \text{ or } j \to i\right\}|$$

- the eigenvector centrality (EC, Bonacich 1987) of species i, that is a graph centrality satisfying the request that the score of each species i in the food web is proportional to the sum of the centrality scores of the species interacting with i. The values of EC are computed as the entries of the first eigenvector of A.

- the information centrality (IC, Stephenson & Zelen 1989) of species i, that is the harmonic mean of the resistance distances (Klein & Randić 1993) toward the species i. Let I_{ji} denote the resistance distance from j to i, then:

$$IC(i) := \frac{S}{\sum_{j \neq i} (I_{ji})^{-1}}$$

- the subgraph centrality (SC, Estrada & Rodriguez-Velazquez 2005) of species i, which counts the number of returning loops starting from species i, discounted exponentially by their size. It is possible to give a closed expression for IC in terms of the exponential of the adjacency matrix A:

$$SC(i) := \left[e^{A}\right]_{ii}$$

The above centralities can be computed directly from the adjacency matrix of the considered food web and are implemented in the *R* package *igraph* (Csardi & Nepusz 2006). For a discussion of the interpretation of the previous graph centralities in an ecological network context, see (Jordán 2009,Jordán *et al.* (2009)).

DATA

We analysed five different food webs. We present them graphically using the x axis to show the species' omnivory index (adding some noise to avoid nodes overlap), the y axis to show the species trophic level, the size of the nodes to shows species total degree and the color of the nodes their ranking based on the total strain (deep blue for lower values, light yellow for higher values).

The original images and the interaction data are available online at the webpage of the paper.

Caribbean Sea food web

We analyse the Caribbean sea food web as compiled by (Opitz 1996). To observe how our model responds to leveld of taxonomic definitions, we consider both the original food web, where taxa are defined to the level of species, and a coarser version where we cluster species into families.

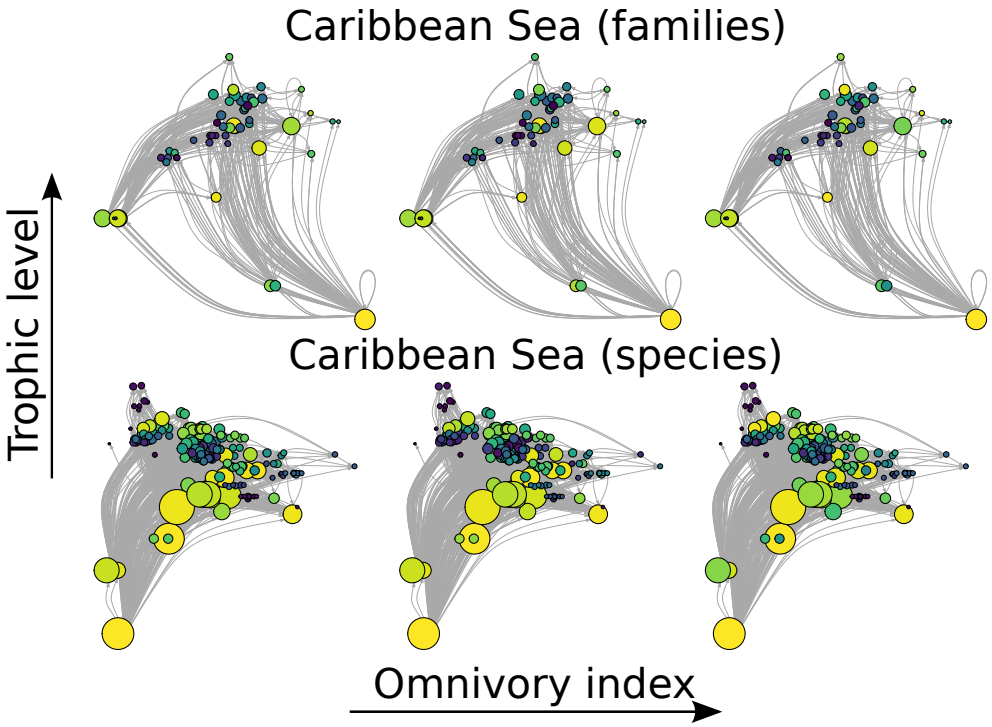


Figure 1: Caribbean Sea food web: (above) nodes represent families, (below) nodes represent species

Serengeti (Baskerville) food web

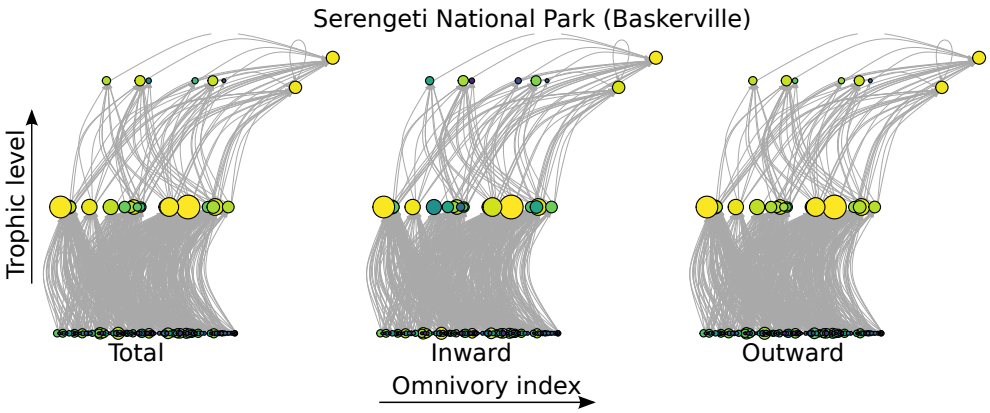


Figure 2: Serengeti National Park food web (by Baskerville)

The Serengeti National Park food web was compiled by (Baskerville *et al.* 2011), notice the increased definition of the plant guilds with respect to (Visser *et al.* 2011). The higher degree of omnivory in this web is only apparent and is due to the normalisation effect.

Serengeti (de Visser) food web

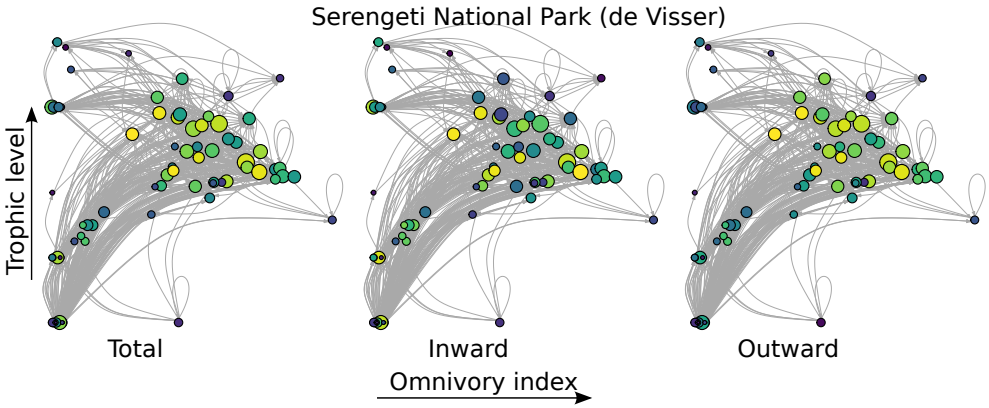


Figure 3: Serengeti National Park food web (de Visser)

The Serengeti National Park food web by (Visser *et al.* 2011)

Weddell Sea food web

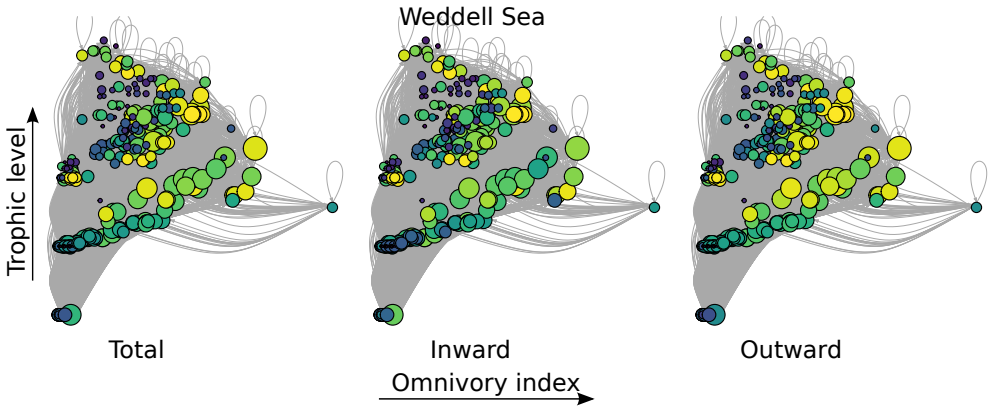


Figure 4: Weddell Sea food web

The Weddell Sea food web is a large marine food web compiled by (Jennings *et al.* 2002)

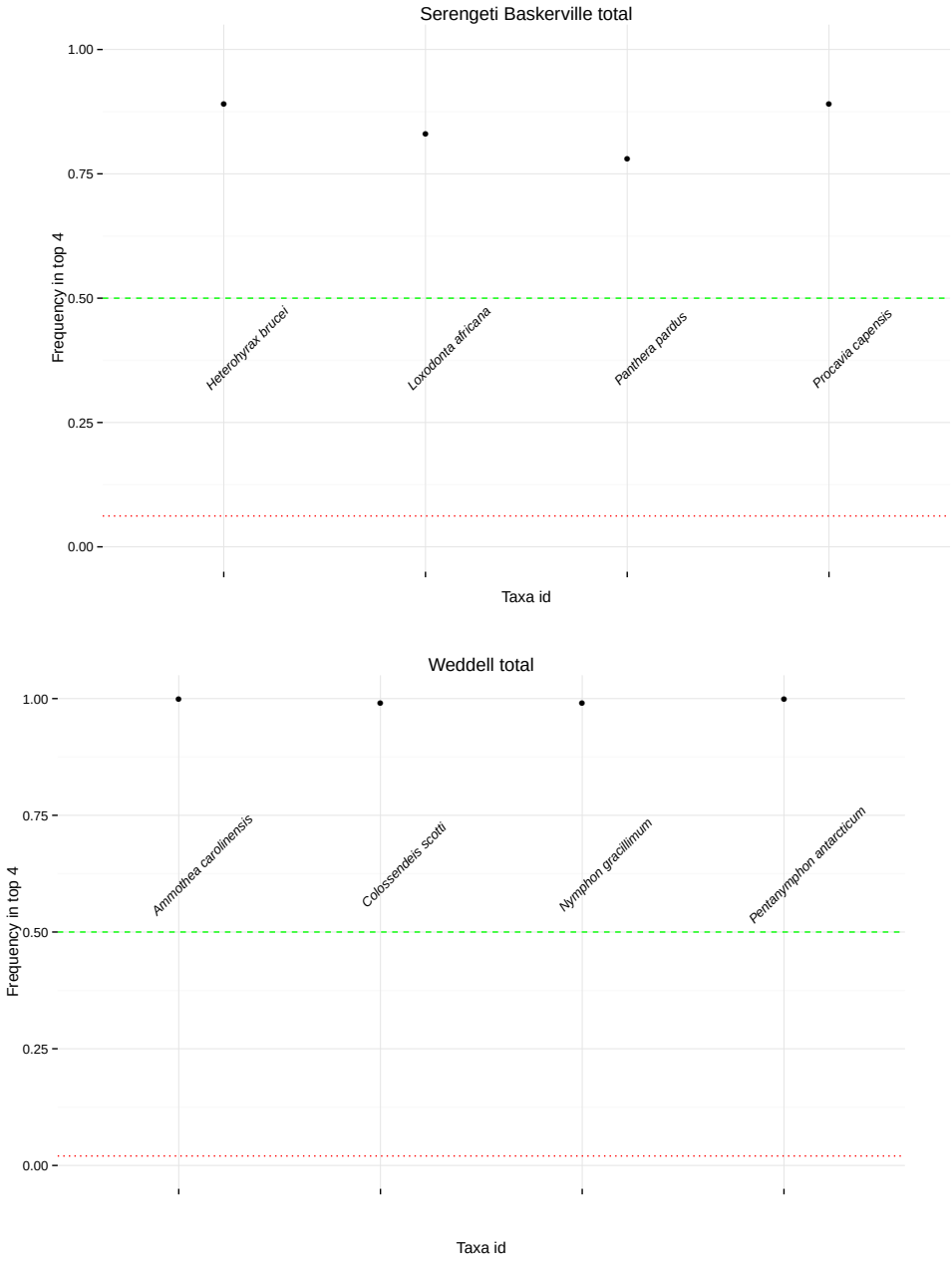
RESULTS

Weighted networks

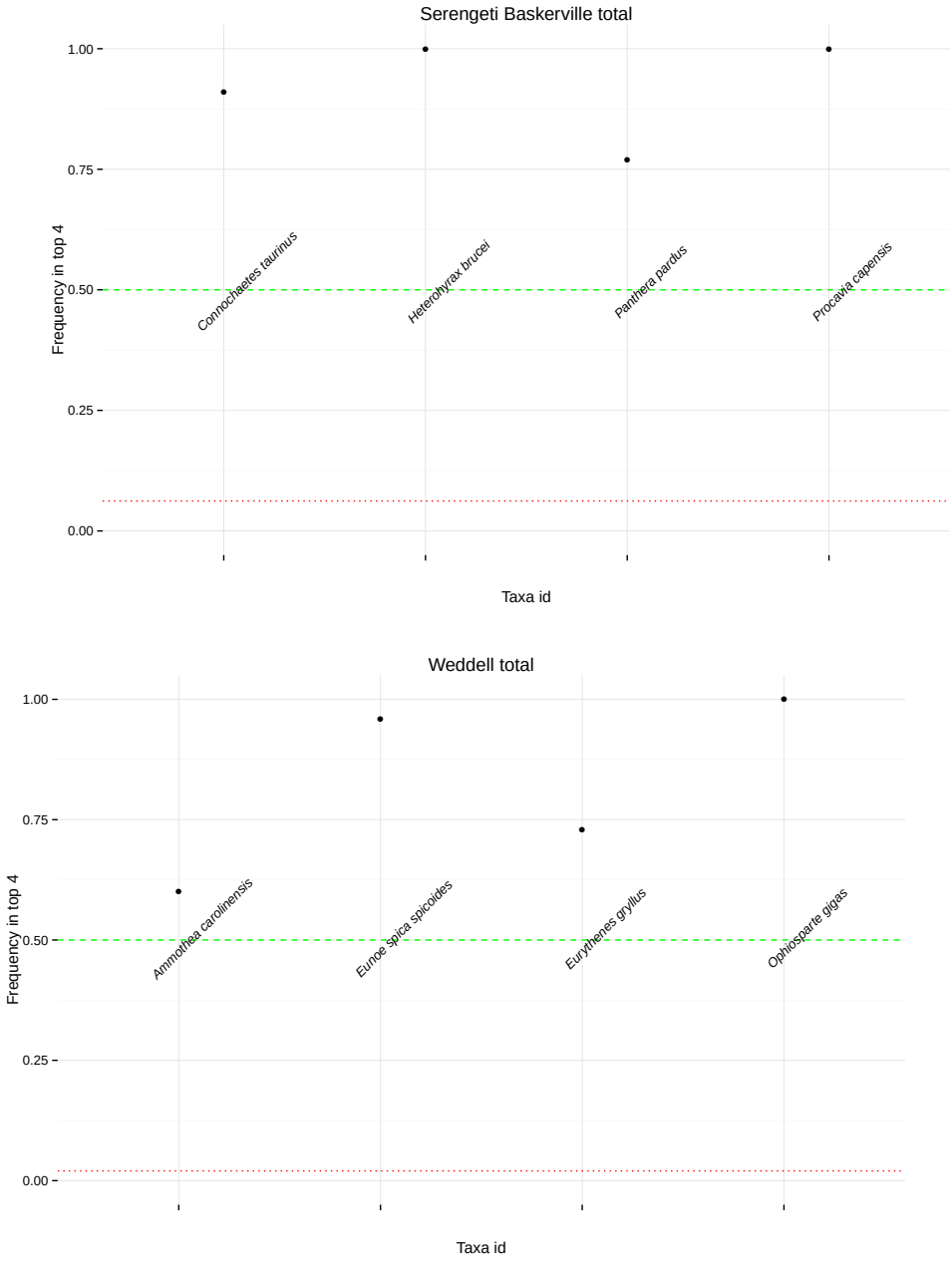
In the manuscript we presented species strain from binary food webs' data only. In fact, obtaining estimates for the amount of energy flowing between a pair of species is often difficult. Therefore, we usually have a more reliable knowledge of the topological structure of an interaction network rather than of its weighted version. However, the components of the species' diets are not equally important. Thus, if the species' ranking we estimated from the topological data were extremely sensitive to the interactions' weight its applicability would be limited. On the other hand, if the topological species' ranking were not affected at all by the specification of interactions weights, that would rise doubt about its ecological meaning.

To test the extent to which our *strain* and *mean distance* measures are robust to the specification of interactions' weights, we compared the ranking of the species based on topological data with the rankings we obtained by simulating interactions weights. To do so, we sampled the interactions weights from a Log-Normal distribution, truncated so that their minimum value was 10^{-6} and normalised so that the maximum value was 1.

The results we obtained show that the correlation between the topological and the weighted rankings were significant and positive for more than 95% of the simulations. However, the amount of variation in the weighted ranking explained by the topological ranking had a large variance (i.e., it spanned the range from almost null to almost one). Yet, the set of species with higher *strain* and the set of species with higher *mean distance* as estimated from the topological data was consistent across the simulated weighted networks, indicating that our measures are able to identigy the species with distinctively high ecological importance. Here we just show part of the results (restricted to the total functional trait space and the the Baskerville's Serengeti and the Weddell sea food webs), to support the notion that the four species with highest strain and uniqueness as computed from the binary webs are consistently in the set of four species with highest strain and uniqueness as computed from weighted webs.



Frequency of presence in the set of four species with the highest strain as estimated by the simulated weighted networks for the four species with the highest strain as estimated by the topological networks.



Frequency of presence in the set of four species with the highest mean distance as estimated by the simulated weighted networks for the four species with the highest mean distance as estimated by the topological networks.

We present here additional results similar to the ones presented in the main paper. All the original graphic files, the source code used to produce them, and the data on which they are based are available online at http://gvdr.github.io.

Strain

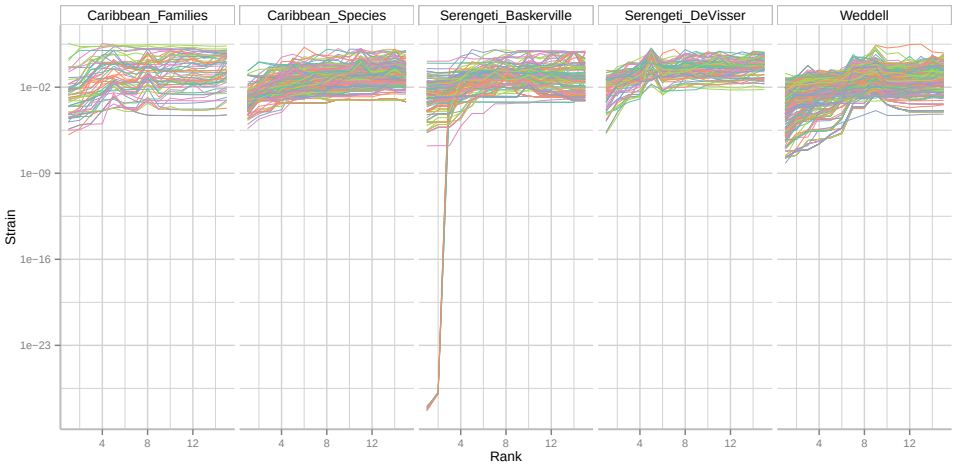


Figure 5: Total strain (log transformed) as function of the model dimension (rank).

We computed the (inward, outward and total) species strain for all the species in the five food webs.

Strain distribution

We show here the distribution of the 3-dimensional total, inward and outward strain for the five foodwebs.

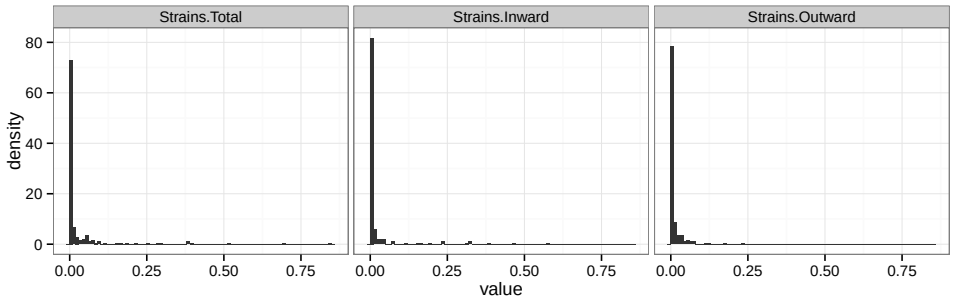


Figure 6: Caribbean Sea food web, nodes represent species

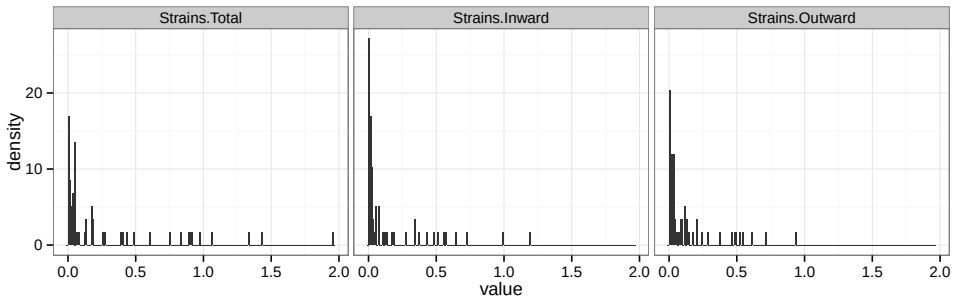


Figure 7: Caribbean Sea food web, nodes represent families

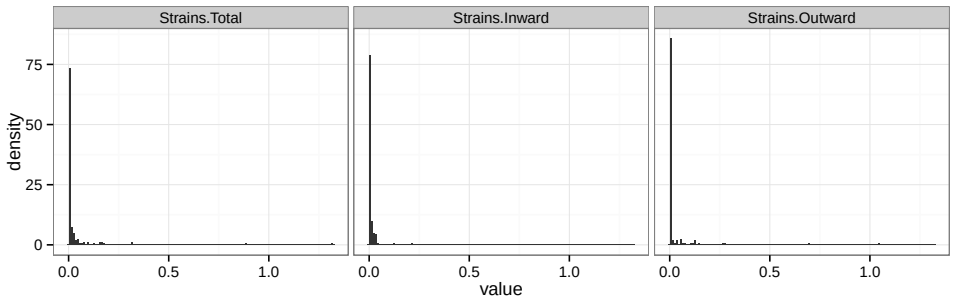


Figure 8: Serengeti National Park food web (Baskerville)

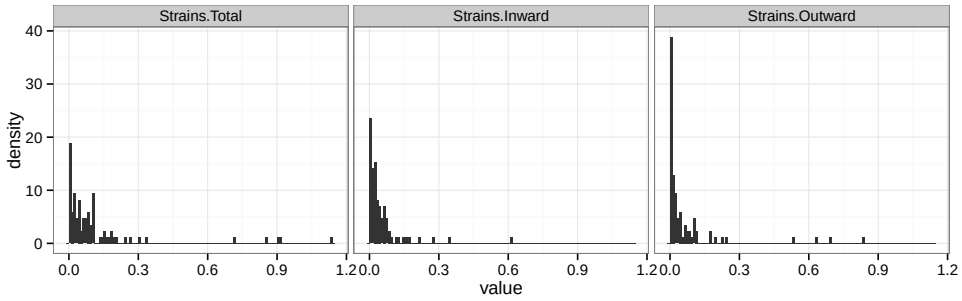


Figure 9: Serengeti National Park food web (de Visser)

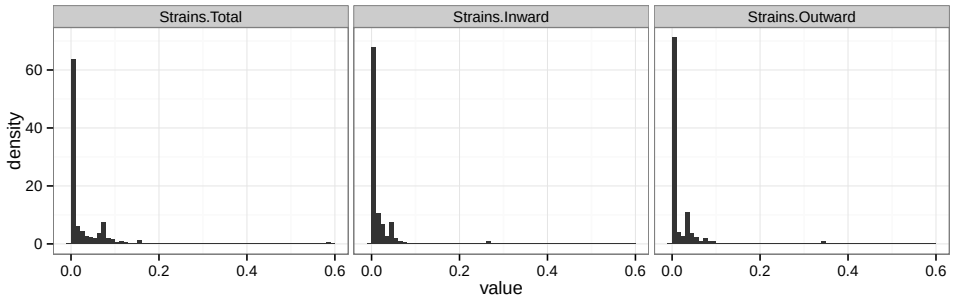


Figure 10: Weddell Sea food web

Strain correlation

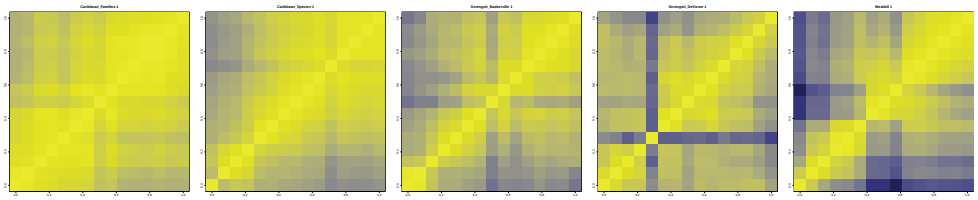


Figure 11: Correlation of the species' total strain as a function of the model dimension

The choice of the model dimension does not substantially affect the ranking of species' strain, supporting the notion that the measure is robust to model parameters. Deep blue for lower coefficients of determination, light yellow for higher values. All correlations are significant.

Uniqueness correlation

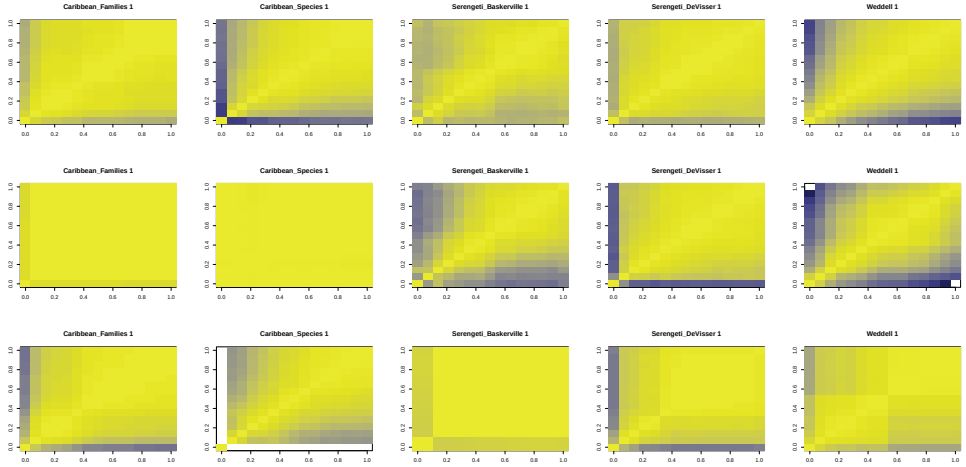


Figure 12: Correlation of the species' total uniqueness as a function of the model dimension

The choice of the model dimension does not substantially affect the ranking of species' uniqueness, supporting the notion that the measure is robust to model parameters. The rows show inward uniqueness, outward uniqueness and total uniqueness respectively. Deep blue for lower coefficients of determination, light yellow for higher values. Blank cells represent non significant correlations.

Strain vs. Uniqueness

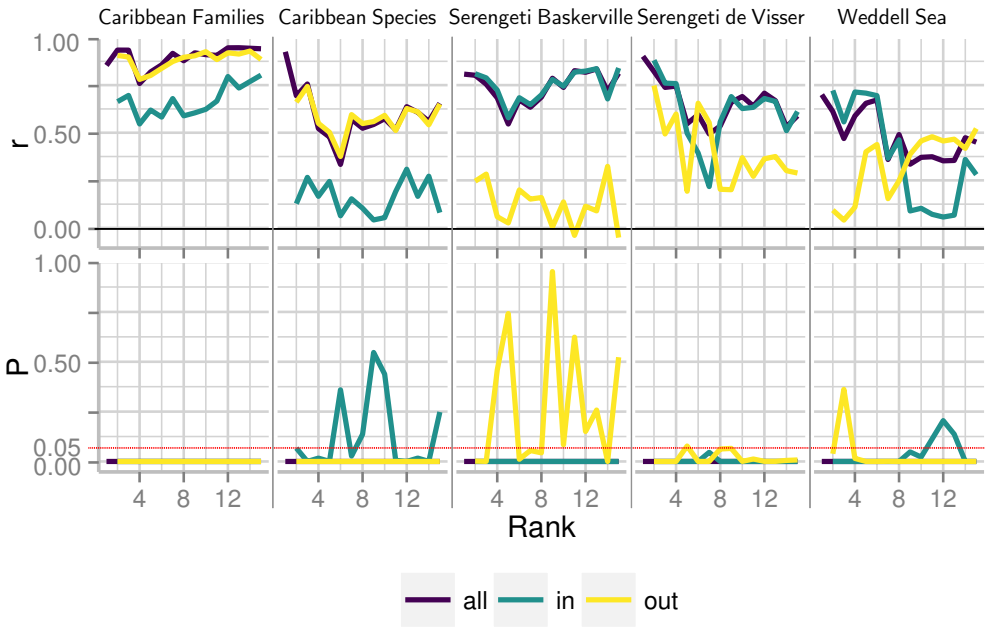


Figure 13: Correlation between species' strain and uniqueness

There's a significant correlation between species' strain (inward, outward or total) and uniqueness (inward, outward or total). The correlation does not grow with the model dimension (and for the Caribbean Sea food web (species level), the Serengeti de Visser food web and the Weddell Sea food web it decreases).

Strain vs. Keystone centralities

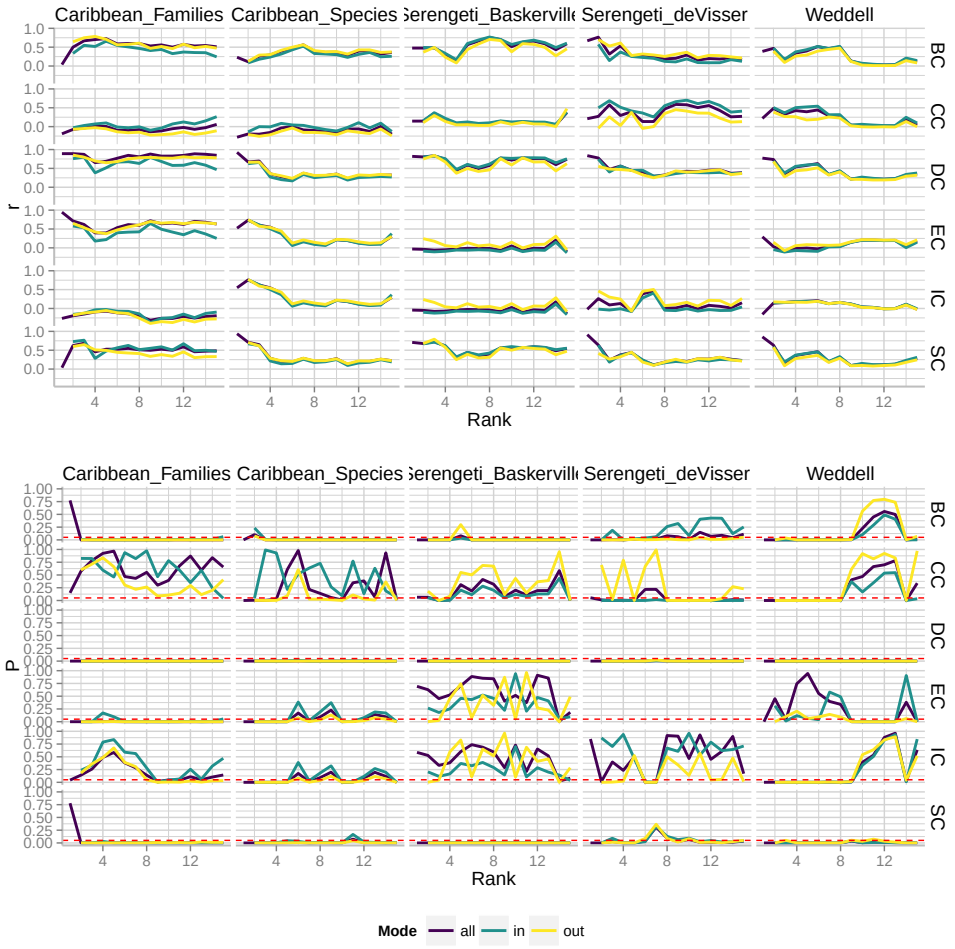


Figure 14: Correlation between species' strain and six classic centrality measures

There's a significant correlation between species' strain (inward, outward or total) and the six centrality measures we considered. We observe more often significant correlations at the lower model dimensions and the coefficients of determination do not grow with model dimensions. This suggest the conclusion the the food web are inherently low dimensional.

Uniqueness vs. Keystone centralities

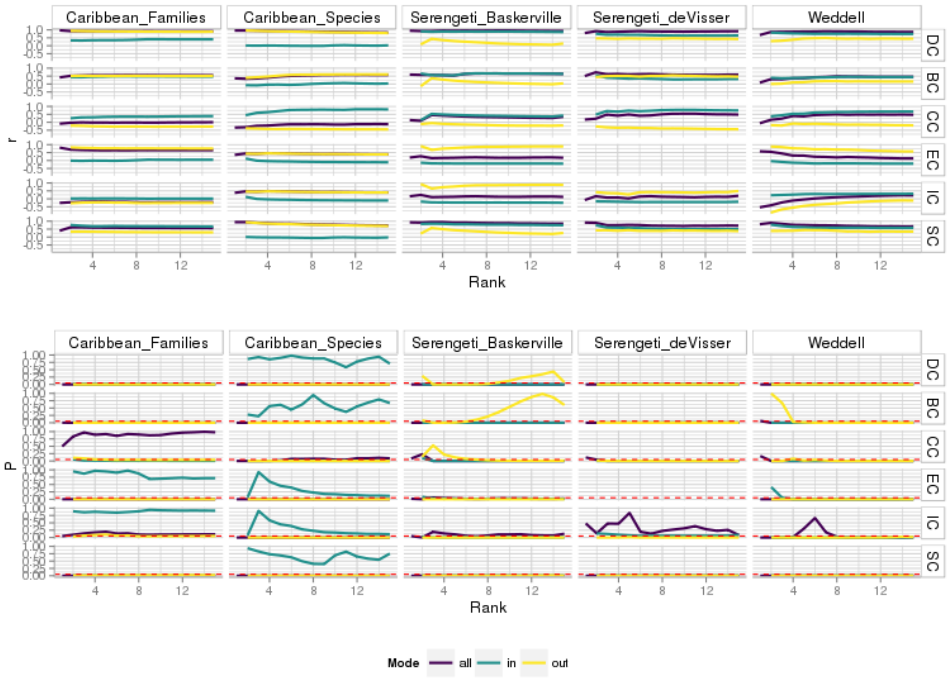


Figure 15: Correlation between species' uniqueness and six classic centrality measures

As for the species' strain, there's a significant correlation between species' uniqueness (inward, outward or total) and the six centrality measures we considered. We observe more often significant correlations at the lower model dimensions and the coefficients of determination do not grow with model dimensions.

BIBLIOGRAPHY

1.Allesina, S. & Pascual, M. (2009). Food web models: A plea for groups. *Ecology letters*, 12, 652–662.

2.Barber, C.B., Dobkin, D.P. & Huhdanpaa, H. (1996). The quickhull algorithm for convex hulls. *ACM Transactions on Mathematical Software (TOMS)*, 22, 469–483.

3.Baskerville, E.B., Dobson, A.P., Bedford, T., Allesina, S., Anderson, T.M. & Pascual, M. (2011). Spatial guilds in the serengeti food web revealed by a bayesian group model. *PLoS Computational Biology*, Vol. 7, e1002321.

4.Bavelas, A. (1950). Communication patterns in task-oriented groups. *Journal of the acoustical society of America*, 22.

5.Bonacich, P. (1987). Power and centrality: A family of measures. *American journal of sociology*, 1170–1182.

6.Cattell, R.B. (1966). The scree test for the number of factors. *Multivariate behavioral research*, 1, 245–276.

7.Chatterjee, S. & others. (2014). Matrix estimation by universal singular value thresholding. *The Annals of Statistics*, 43, 177–214.

8.Csardi, G. & Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal*, Complex Systems, 1695.

9.Dalla Riva, G.V. & Stouffer, D.B. (2015). Exploring the evolutionary signature of food webs' backbones using functional traits. *Oikos*, Online early view.

10.Dryden, I.L. (2013). *Shapes: Statistical shape analysis*.

11.Dryden, I.L. & Mardia, K.V. (1998). *Statistical shape analysis*. Wiley Chichester.

12.Estrada, E. & Rodriguez-Velazquez, J.A. (2005). Subgraph centrality in complex networks. *Physical Review E*, Vol. 71, 056103.

13.Fishkind, D.E., Sussman, D.L., Tang, M., Vogelstein, J.T. & Priebe, C.E. (2013). Consistent adjacency-spectral partitioning for the stochastic block model when the model parameters are unknown. *SIAM Journal on Matrix Analysis and Applications*, 34, 23–39.

14.Freeman, L.C. (1977). A set of measures of centrality based on betweenness. *Sociometry*, 40, 35–41.

15.Gavish, M. & Donoho, D.L. (2014). The optimal hard threshold for singular values is. *Information Theory, IEEE Transactions on*, 60, 5040–5053.

16.Habel, K., Grasman, R., Stahel, A., Stahel, A. & Sterratt, D.C. (2014). *Geometry: Mesh generation and surface tesselation*.

17.Holland, P.W., Laskey, K.B. & Leinhardt, S. (1983). Stochastic blockmodels: First steps. *Social networks*, 5, 109–137.

18.Jennings, S., Greenstreet, S., Hill, L., Piet, G., Pinnegar, J. & Warr, K. (2002). Long-term trends in the trophic structure of the north sea fish community: Evidence

from stable-isotope analysis, size-spectra and community metrics. *Marine Biology*, Vol. 141, 1085–1097.

19.Jolliffe, I.T. (2002). *Principal component analysis*. Springer-Verlag.

20.Jordán, F. (2009). Keystone species and food webs. *Philosophical Transactions of the Royal Society B: Biological Sciences*, Vol. 364, 1733–1741.

21.Jordán, F., Liu, W.-C. & Mike, Á. (2009). Trophic field overlap: A new approach to quantify keystone species. *Ecological Modelling*, Vol. 220, 2899–2907.

22.Klein, D.J. & Randić, M. (1993). Resistance distance. *Journal of Mathematical Chemistry*, 12, 81–95.

23.Opitz, S. (1996). Trophic interactions in caribbean coral reefs. *The WorldFish Center Working Papers*.

24.R Core Team. (2014). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

25.Rohe, K., Chatterjee, S. & Yu, B. (2011). Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics*, 1878–1915.

26.Stephenson, K. & Zelen, M. (1989). Rethinking centrality: Methods and examples. *Social Networks*, Vol. 11, 1–37.

27.Sussman, D.L., Tang, M., Fishkind, D.E. & Priebe, C.E. (2012). A consistent adjacency spectral embedding for stochastic blockmodel graphs. *Journal of the American Statistical Association*, 107, 1119–1128.

28.Visser, S.N. de, Freymann, B.P. & Olff, H. (2011). The serengeti food web: Empirical quantification and analysis of topological changes under increasing human impact. *Journal of Animal Ecology*, Vol. 80, 484–494.

29.Wang, Y.J. & Wong, G.Y. (1987). Stochastic blockmodels for directed graphs. *Journal of the American Statistical Association*, 82, 8–19.

30.Zhu, M. & Ghodsi, A. (2006). Automatic dimensionality selection from the scree plot via the use of profile likelihood. *Computational Statistics & Data Analysis*, Vol. 51, 918–930.