Food web reconstruction through phylogenetic transfer of low-rank network representation

Tanya Strydom [1,2,] Salomé Bouskila [1,] Francis Banville [1,3,2] Ceres Barros [4] Dominique Caron [5,2] Maxwell J Farrell [6] Marie-Josée Fortin [6] Victoria Hemming [4] Benjamin Mercier [3,2] Laura J. Pollock [5,2] Rogini Runghen [7] Giulio V. Dalla Riva [8] Timothée Poisot [1,2]

[1] Département de Sciences Biologiques, Université de Montréal, Montréal, Canada [2] Québec Centre for Biodiversity Sciences, Montréal, Canada [3] Université de Sherbrooke, Sherbrooke, Canada [4] Department of Forest Resources Management, University of British Columbia, Vancouver, Canada [5] Department of Biology, McGill University, Montréal, Canada [6] Department of Ecology & Evolutionary Biology, University of Toronto, Toronto, Canada [7] Centre for Integrative Ecology, School of Biological Sciences, University of Canterbury, Canterbury, New Zealand [8] School of Mathematics and Statistics, University of Canterbury, Canterbury, New Zealand

These authors contributed equally to the work

Correspondance to:

Timothée Poisot — `timothee.poisot@umontreal.ca`

Last revision: *October 21, 2021*

Despite their importance in many ecological processes, collecting data and information on ecological interactions is an exceedingly challenging task. For this reason, large parts of the world have a data deficit when it comes to species interactions, and how the resulting networks are structured. As data collection alone is unlikely to be sufficient, community ecologists must adopt predictive methods. Here we develop such a method, relying on graph embedding and transfer learning to assemble a predicted list of trophic interactions between Canadian mammals. This interaction list is derived from the European food web, despite sharing only 5% of common species with Canada. The results of the predictive model are compared against databases of recorded pairwise interactions, showing that we correctly recover over 95% of known interactions. We provide guidance on how this method can be adapted by substituting some approaches or predictors in order to make it more generally applicable.

Introduction

There are two core challenges we are faced with in furthering our understanding of ecological networks across space, particularly at macro-ecologically relevant scales (*e.g.* Trøjelsgaard & Olesen 2016). First, networks within a location are difficult to sample properly (Jordano 2016a, b), resulting in a widespread "Eltonian shortfall" (Hortal *et al.* 2015), *i.e.* a lack of knowledge about inter and intra specific relationships. This first challenge has been, in large part, addressed by the recent emergence of a suite of methods aiming to predict interactions within *existing* networks, many of which are reviewed in Strydom *et al.* (2021a). Second, recent analyses based on collected data (Poisot *et al.* 2021a) or metadata (Cameron *et al.* 2019) highlight that ecological networks are currently studied in a biased subset of space and bioclimates, which impedes our ability to generalize any local understanding of network structure. Meaning that, although the framework to address incompleteness *within* networks exists, there would still be regions for which, due to a *lack* of local interaction data, we are unable to infer potential species interactions. Having a general solution for inferring ~~a~~ *plausible* ~~metaweb~~ interactions (despite the unavailability of interaction data) could be the catalyst for significant breakthroughs in our ability to start thinking about species interaction networks over large spatial scales. In a recent overview of the field of ecological network prediction, Strydom *et al.* (2021a) identified two challenges of interest to the prediction of interactions at large scales. First, there is a relative scarcity of relevant data in most places globally – paradoxically, this restricts our ability to infer interactions to locations where inference is perhaps the least required; second, accurate predictions often demand accurate predictors, and the lack of methods that can leverage small amount of data is a serious impediment to our predictive ability globally.

Here, we present a general method ~~for~~ to recommend possible trophic interactions, relying on the transfer learning of network representations, ~~relying on the~~ specifically by using similarities of species in a biologically/ecologically relevant proxy space (*e.g.* shared morphology or ancestry). Transfer learning is a machine learning methodology that uses the knowledge gained from solving one problem and applying it to a related (destination) problem (Pan & Yang 2010; Torrey & Shavlik 2010). In this instance, we solve the problem of predicting trophic interactions between species, based on knowledge extracted from another species pool for which interactions are known by using phylogenetic structure as a medium for transfer. ~~This allows us to construct a *probabilistic* metaweb for a community for which we have *no* prior trophic interaction data for the desired species pool. Our methodology is outlined in fig. 1, where we provide an~~

~~illustration based on learning the embedding of a metaweb of trophic interactions for European mammals (known interactions; Maiorano *et al.* 2020b, a) and, based on phylogenetic relationships between mammals globally (*i.e.*, phylogenetic tree Upham *et al.* 2019), infer a metaweb for the Canadian mammalian species pool (interactions are treated as unknown in this instance).~~

There is a plurality of measures of species similarities that can be used for metaweb reconstruction (see *e.g.* Morales-Castilla *et al.* 2015); however, phylogenetic proximity has several desirable properties when working at large scales. Gerhold *et al.* (2015) made the point that phylogenetic signal captures diversification of characters (large macro-evolutionary process), but not necessarily community assembly (fine ecological process); Dormann *et al.* (2010) previously found very similar conclusions. Interactions tend reflect a phylogenetic signal because they have a conserved pattern of evolutionary convergence that encompasses a wide range of ecological and evolutionary mechanisms (Cavender-Bares *et al.* 2009; Mouquet *et al.* 2012), and - most importantly - retain this signal even when it is not detectable at the community scale (Hutchinson *et al.* 2017; Poisot & Stouffer 2018). Finally, species interactions at macro-ecological scales seem to respond mostly to macro-evolutionary processes (Price 2003); which is evidenced by the presence of conserved backbones in food webs (Dalla Riva & Stouffer 2016), strong evolutionary signature on prey choice (Stouffer *et al.* 2012), and strong phylogenetic signature in food web intervality (Eklöf & Stouffer 2016). Phylogenetic reconstruction has also previously been used within the context of ecological networks, namely understanding ancestral plant-insect interactions (Braga *et al.* 2021). Taken together, these considerations suggest that phylogenies can reliably be used to transfer knowledge on species interactions.

[Figure 1 about here.]

Our methodology is outlined in fig. 1, where we provide an illustration based on learning the embedding of a metaweb of trophic interactions for European mammals (Maiorano *et al.* 2020; known interactions; **Maiorano2020TetEu?**) and, based on phylogenetic relationships between mammals globally (*i.e.*, phylogenetic tree Upham *et al.* 2019), infer a metaweb for the Canadian mammalian species pool (interactions are treated as unknown in this instance). Following the definition of Dunne (2006), a metaweb is a network analogue to the concept of a regional species pool. Specifically, a metaweb is an inventory of all *possible* interactions within species likely to occurr within a spatially delimited area (the network γ-diversity, in a sense). The metaweb is, therefore, *not* a prediction of the food web at any specific locale within the frontiers of the spatial area it recovers, and will in fact have a different structure (notably by having a larger connectance; see *e.g.* **Wood2015EffSpa?**). These local food webs are expected to be a subset of both the species and the interactions of their metaweb, and have been called "metaweb realizations" (Poisot *et al.* 2015). The difference between a food web at a specific location and the metaweb has to do with a variety of mechanisms, including species co-occurrence, local environmental conditions, and local distribution of functional traits. Nevertheless, the metaweb represents the total of functional, phylogenetic, and macroecological processes (Morales-Castilla *et al.* 2015), and therefore still holds valuable ecological information. Because the metaweb can be down-sampled given appropriate knowledge of local species composition (the network α-diversity, in a sense), it is possible to infer what may drive the structure of food webs at finer spatial scales. This has been done for example for tree-gallers-parasitoid systems (**Gravel2018BriElt?**), fish trophic interactions (Albouy *et al.* 2019), tetrapods trophic interactions (O'Connor *et al.* 2020), and crop-pests networks (Grünig *et al.* 2020). Whereas the original metaweb definition was based on presence/absence, we focus on *probabilistic* metawebs; not only does our method recommed interactions that may exist, it gives each interaction a score that is mathematically equivalent to the chance of success of a Bernoulli trial (see *e.g.* Poisot *et al.* 2016), which allows properly weigh interactions as a function of how likely they are.

Our case study shows that phylogenetic transfer learning is ~~indeed~~ an effective approach to ~~predict the Canadian mammalian metaweb~~the generation of probabilistic metawebs. This showcases that although the components (species) that make up the Canadian and European communities may be *minimally* shared (the overall species overlap is less than 4%), if the medium (proxy space) selected in the transfer step is biologically plausible, we can still effectively learn from the known network and make biologically relevant predictions of interactions. Indeed, as we detail in the result, when validated against known but fracational data of trophic interactions between Canadian mammals, our model achieves a predictive accuracy of approx. 91%. It should be reiterated that the framework presented in fig. 1 is amenable to changes; notably, the measure of similarity may not be phylogeny, and can be replaced by information on foraging (Beckerman *et al.* 2006), cell-level mechanisms (Boeckaerts *et al.* 2021), or a combination of traits and phylogenetic structure (Stock 2021).

Data used for the case study

We use data from the European metaweb assembled by ~~Maiorano *et al.*~~ (~~2020b), following the definition of the metaweb first introduced by Dunne (2006), *i.e.* an inventory of all possible interactions within species from a spatially delimited pool. Notably the metaweb is not a prediction of the food web at any specific locale within the frontiers of the species pool – in fact, these local food webs are expected to have a subset of both the species and the interactions of their metaweb (Poisot *et al.* 2012). This being said, as the metaweb represents the total of functional, phylogenetic, and macroecological processes (Morales-Castilla *et al.* 2015), it is thus still worthy of ecological attention. We deduced~~ **Maiorano2020TetEu?**).

TANYA TODO description of the data

We induced the subgraph corresponding to all mammals by matching species names in the original network to the GBIF taxonomic backbone (GBIF Secretariat 2021) and retaining all those who matched to mammals. This serves a dual purpose 1) to extract only mammals from the European network and 2) to match and standardize species names when aggregating the different data sources further downstream (which is an important consideration when combining datasets (Grenié *et al.* 2021)). All nodes had valid matches to GBIF at this step, and so this backbone is used for all name reconciliation steps as outlined below.

The European metaweb represents the knowledge we want to learn and transfer; the phylogenetic similarity of mammals here represents the information for transfer. We used the mammalian consensus supertree by Upham *et al.* (2019), for which all approximatively 6000 names have been similarly matched to their GBIF valid names. This step allows us to place each node of the mammalian European metaweb in the phylogeny.

The destination problem to which we want to transfer knowledge is the trophic interactions between mammals in Canada. We obtained the list of extant species from the IUCN checklist, and selected the terrestrial and semi-aquatic species (this corresponds to the same selection that was applied by ~~Maiorano *et al.*~~ ~~(2020b~~(**Maiorano2020TetEu?**) in the European metaweb). The IUCN names were, as previously, reconciled against GBIF to have an exact match to the taxonomy.

After taxonomic cleaning and reconciliation as outlined in the following sections, the mammalian European metaweb has 260 species, and the Canadian species pool has 163; of these, 17 (about 4% of the total) are shared, and 89 species from Canada (54%) had at least one congeneric species in Europe. The similarity for both species pools predictably increases with higher taxonomic order, with 19% of shared genera, 47% of shared families, and 75% of shared orders; for the last point, Canada and Europe each had a single unique order (*Didelphimorphia* for Canada, *Erinaceomorpha* for Europe).

In the following sections, we describe the representational learning step applied to European data, the transfer step through phylogenetic similarity, and the generation of a probabilistic metaweb for the destination species pool.

Method description

The crux of the method is the transfer of knowledge of a known network, in order to predict interactions between species from another location. In fig. 1, we give a high-level overview of the approach; in the example around which this manuscript is built (leveraging detailed knowledge about binary trophic interactions between Mammalia in Europe to predict the less known trophic interactions between closely phylogenetically related Mammalia in Canada), we use a series of specific steps for network embedding, trait inference, network prediction and thresholding.

Specifically, our approach can be summarized as follows: from the known network in Europe, we use a truncated Singular Value Decomposition (t-SVD; Halko *et al.* 2011) to generate latent traits representing a low-dimensional embedding of the network; these traits give an unbiased estimate of the node's position in the latent feature spaces. Then, we map these latent traits onto a reference phylogeny (other distance-based measures of species proximity that allow for the inference of features in the latent space can be used, for example the dissimilarity in functional traits). Based on the reconstructed latent traits for species in the destination species pool, a Random Dot Product Graph model (hereafter RDPG; Young & Scheinerman 2007) predicts the interaction between species through a function of the nodes' features through matrix multiplication. Thus, from latent traits and node position, we can infer interactions.

The method we develop is, ecologically speaking, a "black box," *i.e.* an algorithm that can be understood mathematically, but whose component parts are not always directly tied to ecological processes. There is a growing realization in machine learning that (unintentional) black box algorithms are not necessarily a bad thing Holm (2019), as long as their constituent parts can be examined (which is the case with our method). But more importantly, data hold more information that we may thought; as such, even algorithms that are disconnected from the model can make correct guesses most of the time (Halevy *et al.*

2009); in fact, in an instance of ecological forecasting of spatio-temporal systems, model-free approaches (*i.e.* drawing all of their information from the data) outperformed model-informed ones (Perretti *et al.* 2013).

Implementation and code availability

The entire pipeline is implemented in *Julia* 1.6 (Bezanson *et al.* 2017) and is available under the permissive MIT License at `https://osf.io/2zwqm/`. The taxonomic cleanup steps are done using `GBIF.jl` (~~Dansereau & Poisot 2021~~**Dansereau2021SimJl?**). The network embedding and analysis is done using `EcologicalNetworks.jl` (Poisot *et al.* 2019; Banville *et al.* 2021). The phylogenetic simulations are done using `PhyloNetworks.jl` (Solís-Lemus *et al.* 2017) and `Phylo.jl` (Reeve *et al.* 2016). A complete `Project.toml` file specifying the full tree of dependencies is available alongside the code. This material also includes a fully annotated copy of the entire code required to run this project (describing both the intent of the code and discussing some technical implementation details), a vignette for every step of the process, and a series of Jupyter notebooks with the text and code. The pipeline can be executed on a laptop in a matter of minutes, and therefore does not require extensive computational power.

Step 1: Learning the origin network representation

The first step in transfer learning is to learn the structure of the original dataset. In order to do so, we rely on an approach inspired from representational learning, where we learn a *representation* of the metaweb (in the form of the latent subspaces), rather than a list of interactions (species *a* eats *b*). This approach is conceptually different from other metaweb-scale predictions (*e.g.* Albouy *et al.* 2019), in that the metaweb representation is easily transferable. Specifically, we use RDPG to create a number of latent variables that can be combined into an approximation of the network adjacency matrix. RDPG results are known to have strong phylogenetic signal, and to capture the evolutionary backbone of food webs (Dalla Riva & Stouffer 2016). In addition, recent advances show that the latent variables produced this way can be used to predict *de novo* network edges~~(*i.e.* interactions;~~. Interstingly, the latent variables do not need to be prouced by decomposing the network itself; in a recent contribution, Runghen *et al.* (2021) show that deep artificial neural networks are able to reconstruct the left and right subspaces of a RDPG, in order to predict human movement networks from individual/location metadata. This is an exciting opportunity, as it opens up the possibility of using additional predictors.

The latent variables are created by performing a truncated Singular Value Decomposition (t-SVD) on the adjacency matrix. SVD is an appropriate embedding of ecological networks, which has recently been shown to both capture their complex, emerging properties (Strydom *et al.* 2021b) and to allow highly accurate prediction of the interactions within a single network (Poisot *et al.* 2021b). Under SVD, an adjacency matrix $\mathbf{A}$ (where $\mathbf{A}_{m,n} \in \mathbb{B}$ where 1 indicates predation and 0 an absence thereof) is decomposed into three components resulting in $\mathbf{A} = \mathbf{L}\mathbf{\Sigma}\mathbf{R}'$. Here, $\mathbf{\Sigma}$ is a $m \times n$ diagonal matrix and contains only singular (σ) values along its diagonal, $\mathbf{L}$ is a $m \times m$ unitary matrix, and $\mathbf{R}'$ a $n \times n$ unitary matrix. Truncating the SVD removes additional noise in the dataset by omitting non-zero and/or smaller σ values from $\mathbf{\Sigma}$ using the rank of the matrix. Under a t-SVD $\mathbf{A}_{m,n}$ is decomposed so that $\mathbf{\Sigma}$ is a square $r \times r$ diagonal matrix (whith $1 \leq r \leq r_{full}$ where r_{full} is the full rank of $\mathbf{A}$ and r the rank at which we truncate the matrix) containing only non-zero σ values. Additionally, $\mathbf{L}$ is now a $m \times r$ semi unitary matrix and $\mathbf{R}'$ a $n \times r$ semi-unitary matrix.

The specific rank at which the SVD ought to be truncated is a difficult question. The purpose of SVD is to remove the noise (expressed at high dimensions) and to focus on the signal, (expressed at low dimensions). In datasets with a clear signal/noise demarcation, a scree plot of $\mathbf{\Sigma}$ can show a sharp drop at the rank where noise starts (Zhu & Ghodsi 2006). Because the European metaweb is almost entirely known, the amount of noise (uncertainty) is low; this is reflected in fig. 2 (left), where the scree plot shows no important drop, and in fig. 2 (right) where the proportion of variance explained increases smoothly at higher dimensions. For this reason, we default back to a threshold that explains 60% of the variance in the underlying data, corresponding to 12 dimensions - *i.e.* a tradeoff between accuracy and a reduced number of features.

A RDPG estimates the probability of observing interactions between nodes (species) as a function of the nodes' latent variables. The latent variables used for the RDPG, called the left and right subspaces, are defined as $\mathscr{L} = \mathbf{L}\sqrt{\mathbf{\Sigma}}$, and $\mathscr{R} = \sqrt{\mathbf{\Sigma}}\mathbf{R}$ – using the full rank of $\mathbf{A}$, $\mathscr{L}\mathscr{R}' = \mathbf{A}$, and using any smaller rank results in $\mathscr{L}\mathscr{R}' \approx \mathbf{A}$. Using a rank of 1 for the t-SVD provides a first-order approximation of the network.

[Figure 2 about here.]

Because RDPG relies on matrix multiplication, the higher dimensions essentially serve to make specific interactions converge towards 0 or 1; therefore, for reasonably low ranks, there is no guarantee that the values in the reconstructed network will be within the unit range. In order to determine what constitutes an appropriate threshold for probability, we performed the RDPG approach on the European metaweb, and evaluated the probability threshold by treating this as a binary classification problem, specifically assuming that both 0 and 1 in the European metaweb are all true. Given the methodological details given in ~~Maiorano *et al.* (2020b~~(**Maiorano2020TetEu?**) and O'Connor *et al.* (2020), this seems like a reasonable assumption, although one that does not hold for all metawebs. We used the thresholding approach presented in Poisot *et al.* (2021b), and picked a cutoff that maximized Youden's *J* statistic (a measure of the informedness (trust) of predictions; Youden (1950)); the resulting cutoff was 0.22, and gave an accuracy above 0.99.

The left and right subspaces for the European metaweb, accompanied by the threshold for prediction, represent the knowledge we seek to transfer. In the next section, we explain how we rely on phylogenetic similarity to do so.

Steps 2 and 3: Transfer learning through phylogenetic relatedness

In order to transfer the knowledge from the European metaweb to the Canadian species pool, we performed ancestral character estimation using a Brownian motion model, which is a conservative approach in the absence of strong hypotheses about the nature of phylogenetic signal in the network decomposition (Litsios & Salamin 2012). This uses the estimated feature vectors for the European mammals to create a state reconstruction for all species (conceptually something akin to a trait-based mammalian phylogeny using generality and vulnerability traits) and allows us to impute the missing (latent) trait data for the Canadian species that are not already in the European network; as we are focused on predicting contemporary interactions, we only retained the values for the tips of the tree. We assumed that all traits (*i.e.* the feature vectors for the left and right subspaces) were independent, which is a reasonable assumption as every trait/dimension added to the t-SVD has an *additive* effect to the one before it. Note that the Upham *et al.* (2019) tree itself has some uncertainty associated to inner nodes of the phylogeny. In this case study, we have decided to not propagate this uncertainty, as it would complexify the process. The Brownian motion algorithm returns the *average* value of the trait, and its upper and lower bounds. Because we do not estimate other parameters of the traits' distributions, we considered that every species trait is represented as a uniform distribution between these bounds~~; in a situation where the algorithm would return point values for all simulations, one could in theory either estimate the~~

~~parameters of a distribution for each tip, or draw randomly from the outputs. In all cases,~~. The choice of the uniform distribution was made because the algorithm returns a minimum and maximum point estimate for the value, and given this information, the uniform distribution is the one with maximum entropy. Had all mean parameters estimates been positive, the exponential distribution would have been an alternative, but this is not the case for the subspaces of an RDPG. In order to examine the consequences of the choice of distribution, we estimated the variance per latent variable per node to use a normal distribution; as we show in Supp. Mat. 2, this decision results in dramatically over-estimating the number and probability of interactions, and therefore we keep the discussions in the main text to the uniform case. The inferred left and right sub-spaces for the Canadian species pool ($\hat{\mathscr{L}}$ and $\hat{\mathscr{R}}$) have entries that are distributions, representing the range of values for a given species at a given dimension.

These objects represent the transferred knowledge, which we can use for prediction of the Canadian metaweb.

Step 4: Probabilistic prediction of the destination network

The phylogenetic reconstruction of $\hat{\mathscr{L}}$ and $\hat{\mathscr{R}}$ has an associated uncertainty, represented by the breadth of the uniform distribution associated to each of their entries. Therefore, we can use this information to assemble a *probabilistic* metaweb in the sense of Poisot *et al.* (2016), *i.e.* in which every interaction is represented as a single, independent, Bernoulli event of probability p.

[Figure 3 about here.]

Specifically, we have adopted the following approach. For every entry in $\hat{\mathscr{L}}$ and $\hat{\mathscr{R}}$, we draw a value from its distribution. This results in one instance of the possible left ($\hat{\ell}$) and right ($\hat{r}$) subspaces for the Canadian metaweb. These can be multiplied, to produce one matrix of real values. Because the entries in $\hat{\ell}$ and $\hat{r}$ are in the same space where $\mathscr{L}$ and $\mathscr{R}$ were originally predicted, it follows that the threshold ρ estimated for the European metaweb also applies. We use this information to produce one random Canadian metaweb, $N = \hat{\mathscr{L}}\hat{\mathscr{R}}' \geq \rho$. As we can see in (fig. 3) the European and Canadian metawebs are structurally similar (as would be expected given the biogeographic similarities) and the two (left and right) subspaces are distinct *i.e.* capturing predation (generality) and prey (vulnerability) traits.

Because the intervals around some trait values can be broad (in fact, probably broader than what they would actually be, see *e.g.* Garland *et al.* 1999), we repeat the above process 2×10^5 times, which results in a probabilistic metaweb P, where the probability of an interaction (here conveying our degree of trust that it exists given the inferred trait distributions) is given by the number of times where it appears across all random draws N, divided by the number of samples. An interaction with $P_{i,j} = 1$ means that these two species were predicted to interact in all 2×10^5 random draws.

Data cleanup, discovery, validation, and thresholding

Once the probabilistic metaweb for Canada has been produced, we followed a number of data inflation steps to finalize it. This step is external to the actual transfer learning framework but rather serves as a way to augment and validate the predicted metaweb.

[Figure 4 about here.]

First, we extracted the subgraph corresponding to the 17 species shared between the European and Canadian pools and replaced these interactions with a probability of 0 (non-interaction) or 1 (interaction), according to their value in the European metaweb. This represents a minute modification of the inferred network (about 0.8% of all species pairs from the Canadian web), but ensures that we are directly re-using knowledge from Europe.

Second, we looked for all species in the Canadian pool known to the Global Biotic Interactions (GLoBI) database (Poelen *et al.* 2014), and extracted their known interactions. Because GLOBI aggregates observed interactions, it is not a *networks* data source, and therefore the only information we can reliably extract from it is that a species pair *was reported to interact at least once*. This last statement should yet be taken with caution, as some sources in GLOBI (*e.g.* Thessen & Parr 2014) are produced through text analysis, and therefore may not document direct evidence of the interaction. Nevertheless, should the predictive model work, we would expect that a majority of interactions known to GLOBI would also be predicted. After performing this check, we set the probability of all interactions known to GLOBI (366 in total, 33 of which were not predicted by the model, for a success rate of 91%) to 1.

Finally, we downloaded the data from Strong & Leroux (2014), who mined various literature sources to identify trophic interactions in Newfoundland. This dataset documented 25 interactions between mammals, only two of which were not part of our (Canada-level) predictions, resulting in a success rate of

92%. These two interactions were added to our predicted metaweb with a probability of 1. A table listing all interactions in the predicted Canadian metaweb can be found in the supplementary material.

[Figure 5 about here.]

Because the confidence intervals on the inferred trait space are probably over-estimates, we decided to apply a thresholding step to the interactions after the data inflation (fig. 5). Cirtwill & Hambäck (2021) proposed a number of strategies to threshold probabilistic networks. Their methods assume the underlying data to be tag-based sequencing, which represents interactions as co-occurrences of predator and prey within the same tags; this is conceptually identical to our Bernoulli-trial based reconstruction of a probabilistic network. We performed a full analysis of the effect of various cutoffs, and as they either resulted in removing too few interactions, or removing enough interactions that species started to be disconnected from the network, we set this threshold for a probability equivalent to 0 to the largest possible value that still allowed all species to have at least one interaction with a non-zero probability. The need for this slight deviation from the Cirtwill & Hambäck (2021) method highlights the need for additional development on network thresholding.

Results and discussion of the case study

In fig. 5, we examine the effect of varying the cutoff on $P(i \rightarrow j)$ on the number of links, species, and connectance. Determining a cutoff using the maximum curvature, or central difference approximation of the second order partial derivative, as suggested by *e.g.* Cirtwill & Hambäck (2021), results in species being lost, or almost all links being kept. We therefore settled on the value that allowed all species to remain with at least one interaction. This result, in and of itself, suggests that additional methodological developments for the thresholding of probabilistic networks are required.

[Figure 6 about here.]

The t-SVD embedding is able to learn relevant ecological features for the network. fig. 6 shows that the first rank correlates linearly with generality and vulnerability (Schoener 1989), *i.e.* the number of preys and predators. Importantly, this implies that a rank 1 approximation represents the configuration model for the metaweb, *i.e.* a set of random networks generated from a given degree sequence (Park & Newman 2004). Accounting for the probabilistic nature of the degrees, the rank 1 approximation also represents the *soft* configuration model (van der Hoorn *et al.* 2018). Both models are maximum entropy graph models (Garlaschelli *et al.* 2018), with sharp (all network realizations satisfy the specified degree sequence) and soft (network realizations satisfy the degree sequence on average) local constraints, respectively. The (soft) configuration model is an unbiased random graph model widely used by ecologists in the context of null hypothesis significance testing of network structure (*e.g.* Bascompte *et al.* 2003) and can provide informative priors for Bayesian inference of network structure (*e.g.* Young *et al.* 2021). It is noteworthy that for this metaweb, the relevant information was extracted at the first rank. Because the first rank corresponds to the leading singular value of the system, the results of fig. 6 have a straightforward interpretation: degree-based processes are the most important in structuring the mammalian food web.

Discussion

One important aspect in which Europe and Canada differ (despite their comparable bioclimatic conditions) is the degree of the legacy of human impacts, which have been much longer in Europe. Nenzén *et al.* (2014) showed that even at small scales (the Iberian peninsula), mammal food webs retain the signal of both climate change and human activity, even when this human activity was orders of magnitude less important than it is now. Similarly, Yeakel *et al.* (2014) showed that changes in human occupation over several centuries can lead to food web collapse. Megafauna in particular seems to be very sensitive to human arrival (Pires *et al.* 2015). In short, there is well-substantiated support for the idea that human footprint affects more than the risk of species extinction (Marco *et al.* 2018), and can lead to changes in interaction structure. Yet, owing to the inherent plasticity of interactions, there have been documented instances of food webs undergoing rapid collapse/recovery cycles over short periods of time (Pedersen *et al.* 2017). The embedding of a network, in a sense, embeds its macro-evolutionary history, especially as RDPG captures ecological signal (Dalla Riva & Stouffer 2016); at this point, it is important to recall that a metaweb is intended as a catalogue of all possible interactions, which should then be filtered (Morales-Castilla *et al.* 2015). In practice (and in this instance) the reconstructed metaweb will predict interactions that are plausible based on the species' evolutionary history, however some interactions would not be realized due to human impact.

Dallas *et al.* (2017) suggest that most links in ecological networks may be cryptic, *i.e.* uncommon or otherwise hard to observe. This argument essentially echoes Jordano (2016b): the sampling of ecological interactions is difficult because it requires first the joint observation of two species, and then the observation of their interaction. In addition, it is generally expected that weak or rare links would be more common in networks (Csermely 2004), compared to strong, persistent links; this is notably the case in food chains, wherein many weaker links are key to the stability of a system (Neutel *et al.* 2002). In the light of these observations, the results in fig. 4 are not particularly surprising: we expect to see a surge in these low-probability interactions under a model that has a good predictive accuracy. Because the predictions we generate are by design probabilistic, then one can weigh these rare links appropriately. In a sense, that most ecological interactions are elusive can call for a slightly different approach to sampling: once the common interactions are documented, the effort required in documenting each rare interaction may increase exponentially. Recent proposals suggest that machine learning algorithms, in these situations, can act as data generators (Hoffmann *et al.* 2019): in this perspective, high quality observational data can be supplemented with synthetic data coming from predictive models, which increases the volume of information available for inference.

Cirtwill *et al.* (2019) previously made the point that network inference techniques based on Bayesian approaches would perform far better in the presence of an interaction-level informative prior; the desirable properties of such a prior would be that it is expressed as a probability, preferably representing a Bernoulli event, the value of which would be representative of relevant biological processes (probability of predation in this case). We argue that the probability returned at the very last step of our framework may serve as this informative prior; indeed, the output of our analysis can be used in subsequent steps, also possibly involving expert elicitation to validate some of the most strongly recommended interactions. One important *caveat* to keep in mind when working with interaction inference is that interactions can never really be true negatives (in the current state of our methodological framework and data collection limitations); this renders the task of validating a model through the usual application of binary classification statistics very difficult (although see Strydom *et al.* 2021a for a discussion of alternative suggestions). The other way through which our framework can be improved is by substituting the predictors that are used for transfer. For example, in the presence of information on species traits that are known to be predictive of species interactions, one might want to rely on functional rather than phylogenetic distances – in food webs, body size (and allometrically related variables) has been established as such a variable (Brose *et al.* 2006); the identification of relevant functional traits is facilitated by recent methodological developments (Rosado *et al.* 2013). It should be noted that Xing & Fayle (2021) highlight phylogenetic relatedness as one of the core components of network comparison at the global scale. In this case study, we have embedded the original metaweb using t-SVD, because it lends itself to a RDPG reconstruction, which is known to capture the consequences of evolutionary processes (Dalla Riva & Stouffer 2016); this being said, there are others ways to embed graphs (Cai *et al.* 2017; Arsov & Mirceva 2019; Cao *et al.* 2019), which can be used as alternatives.

As Herbert (1965) rightfully pointed out, “[y]ou can’t draw neat lines around planet-wide problems”; in this regard, our approach must contend with two interesting problems. The first is the limit of the metaweb to embed and transfer. If the initial metaweb is too narrow in scope, notably from a taxonomic point of view, the chances of finding another area with enough related species to make a reliable inference decrease. This is notably true if the metaweb is assembled in an area with mostly endemic species. Conversely, the metaweb should be reliably filled, which assumes that the S^2 interactions in a pool of S species have been examined, either through literature surveys or expert elicitation. The second problem is to determine which area should be used to infer the new metaweb in, as this determines the species pool that must be used. In our application, we focused on the mammals of Canada. The upside of this approach is that information at the country level is likely to be required by policy makers and stakeholders for their biodiversity assessment, as each country tends to set goals at the national level (Buxton *et al.* 2021) for which quantitative instruments are designed (Turak *et al.* 2017), with specific strategies often enacted at smaller scales (Ray *et al.* 2021). Yet these national divisions, in large parts of the world, reflect nothing except for the legacy of settler colonialism; indeed, the use of ecological data is not an apolitical act (Nost & Goldstein 2021). Operating under the framework of national boundaries, and relying on data infrastructures designed to answer questions within them, ~~and operating under them~~ must be done under the clear realization that ~~they~~ this logic contributed to the ongoing biodiversity crisis (Adam 2014), can reinforce environmental injustice (Choudry 2013; Domínguez & Luoma 2020), and on Turtle Island especially, will probably end up being replaced by Indigenous principles of land management (Eichhorn *et al.* 2019; No’kmaq *et al.* 2021).

Acknowledgements: We acknowledge that this study was conducted on land within the traditional unceded territory of the Saint Lawrence Iroquoian, Anishinabewaki, Mohawk, Huron-Wendat, and Omàmiwininiwak nations. TP, TS, DC, and LP received funding from the Canadian Institue for Ecology &

Evolution. FB is funded by the Institut de Valorisation des Données. TS, SB, and TP are funded by a donation from the Courtois Foundation. CB was awarded a Mitacs Elevate Fellowship no. IT12391, in partnership with fRI Research, and also acknowledges funding from Alberta Innovates and the Forest Resources Improvement Association of Alberta. M-JF acknowledges funding from NSERC Discovery Grant and NSERC CRC. RR is funded by New Zealand's Biological Heritage Ng Koiora Tuku Iho National Science Challenge, administered by New Zealand Ministry of Business, Innovation, and Employment. BM is funded by the NSERC Alexander Graham Bell Canada Graduate Scholarship and the FRQNT master's scholarship. LP acknowledges funding from NSERC Discovery Grant (NSERC RGPIN-2019-05771). TP acknowledges financial support from NSERC through the Discovery Grants and Discovery Accelerator Supplement programs.

References

Adam, R. (2014). *Elephant treaties: The Colonial legacy of the biodiversity crisis*. UPNE.

Albouy, C., Archambault, P., Appeltans, W., Araújo, M.B., Beauchesne, D., Cazelles, K., *et al.* (2019). The marine fish food web is globally connected. *Nature Ecology & Evolution*, 3, 1153–1161.

Arsov, N. & Mirceva, G. (2019). ~~Network Embedding: An Overview. *arXiv:1911.11726 cs, stat*~~*Network Embedding: An Overview*. Available at: `http://arxiv.org/abs/1911.11726`. Last accessed.

Banville, F., Vissault, S. & Poisot, T. (2021). Mangal.jl and EcologicalNetworks.jl: Two complementary packages for analyzing ecological networks in Julia. *Journal of Open Source Software*, 6, 2721.

Bascompte, J., Jordano, P., Melian, C.J. & Olesen, J.M. (2003). The nested assembly of plant-animal mutualistic networks. *Proceedings of the National Academy of Sciences*, 100, 9383–9387.

Beckerman, A.P., Petchey, O.L. & Warren, P.H. (2006). Foraging biology predicts food web complexity. *Proceedings of the National Academy of Sciences*, 103, 13745–13749.

Bezanson, J., Edelman, A., Karpinski, S. & Shah, V. (2017). Julia: A Fresh Approach to Numerical Computing. *SIAM Review*, 59, 65–98.

Boeckaerts, D., Stock, M., Criel, B., Gerstmans, H., De Baets, B. & Briers, Y. (2021). Predicting bacteriophage hosts based on sequences of annotated receptor-binding proteins. *Scientific Reports*, 11, 1467.

Braga, M.P., Janz, N., Nylin, S., Ronquist, F. & Landis, M.J. (2021). Phylogenetic reconstruction of ancestral ecological networks through time for pierid butterflies and their host plants. *Ecology Letters*, n/a.

Brose, U., Jonsson, T., Berlow, E.L., Warren, P., Banasek-Richter, C., Bersier, L.-F., *et al.* (2006). ~~ConsumerResource~~ Consumer–Resource Body-Size Relationships in Natural Food Webs. *Ecology*, 87, 2411–2417.

Buxton, R.T., Bennett, J.R., Reid, A.J., Shulman, C., Cooke, S.J., Francis, C.M., *et al.* (2021). Key information needs to move from knowledge to action for biodiversity conservation in Canada. *Biological Conservation*, 256, 108983.

Cai, H., Zheng, V.W. & Chang, K.C.-C. (2017). ~~A Comprehensive Survey of Graph Embedding: Problems, Techniques and Applications. *arXiv preprint arXiv:1709.07604*~~*A Comprehensive Survey of Graph Embedding: Problems, Techniques and Applications*. Available at: http://arxiv.org/abs/1709.07604. Last accessed.

Cameron, E.K., Sundqvist, M.K., Keith, S.A., CaraDonna, P.J., Mousing, E.A., Nilsson, K.A., *et al.* (2019). Uneven global distribution of food web studies under climate change. *Ecosphere*, 10, e02645.

Cao, R.-M., Liu, S.-Y. & Xu, X.-K. (2019). Network embedding for link prediction: The pitfall and improvement. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 29, 103102.

Cavender-Bares, J., Kozak, K.H., Fine, P.V.A. & Kembel, S.W. (2009). The merging of community ecology and phylogenetic biology. *Ecology Letters*, 12, 693–715.

Choudry, A. (2013). Saving biodiversity, for whom and for what? Conservation NGOs, complicity, colonialism and conquest in an era of capitalist globalization. In: *NGOization: Complicity, contradictions and prospects*. Bloomsbury Publishing, pp. 24–44.

Cirtwill, A.R., ~~Eklf~~Eklf, A., Roslin, T., Wootton, K. & Gravel, D. (2019). A quantitative framework for investigating the reliability of empirical network construction. *Methods in Ecology and Evolution*, 0.

Cirtwill, A.R. & Hambäck, P. (2021). Building food networks from molecular data: Bayesian or fixed-number thresholds for including links. *Basic and Applied Ecology*, 50, 67–76.

Csermely, P. (2004). Strong links are important, but weak links stabilize them. *Trends in Biochemical Sciences*, 29, 331–334.

Dalla Riva, G.V. & Stouffer, D.B. (2016). Exploring the evolutionary signature of food webs' backbones using functional traits. *Oikos*, 125, 446–456.

~~Dansereau, G. & Poisot, T. (2021). SimpleSDMLayers.jl and GBIF.jl: A Framework for Species Distribution Modeling in Julia.~~ Dallas, T., Park, A.W. & Drake, J.M. (2017). Predicting cryptic links in host-parasite networks. ~~*Journal of Open Source Software*~~*PLOS Computational Biology*, ~~6, 2872.~~ 13, e1005557.

Domínguez, L. & Luoma, C. (2020). Decolonising Conservation Policy: How Colonial Land and Conservation Ideologies Persist and Perpetuate Indigenous Injustices at the Expense of the Environment. *Land*, 9, 65.

Dormann, C.F., Gruber, B., Winter, M. & Herrmann, D. (2010). Evolution of climate niches in European mammals? *Biology Letters*, 6, 229–232.

Dunne, J.A. (2006). The Network Structure of Food Webs. In: *Ecological networks: Linking structure and dynamics* (eds. Dunne, J.A. & Pascual, M.). Oxford University Press, pp. 27–86.

Eichhorn, M.P., Baker, K. & Griffiths, M. (2019). Steps towards decolonising biogeography. *Frontiers of Biogeography*, 12, 1–7.

Eklöf, A. & Stouffer, D.B. (2016). The phylogenetic component of food web structure and intervality. *Theoretical Ecology*, 9, 107–115.

Garland, T., JR., Midford, P.E. & Ives, A.R. (1999). An Introduction to Phylogenetically Based Statistical Methods, with a New Method for Confidence Intervals on Ancestral Values1. *American Zoologist*, 39, 374–388.

Garlaschelli, D., Hollander, F. den & Roccaverde, A. (2018). Covariance structure behind breaking of ensemble equivalence in random graphs. *Journal of Statistical Physics*, 173, 644–662.

GBIF Secretariat. (2021). GBIF Backbone Taxonomy.

Gerhold, P., Cahill, J.F., Winter, M., Bartish, I.V. & Prinzing, A. (2015). Phylogenetic patterns are not proxies of community assembly mechanisms (they are far better). *Functional Ecology*, 29, 600–614.

Grenié, M., Berti, E., Carvajal-Quintero, J.D., Winter, M. & Sagouis, A. (2021). Harmonizing taxon names in biodiversity data: A review of tools, databases, and best practices.

Grünig, M., Mazzi, D., Calanca, P., Karger, D.N. & Pellissier, L. (2020). Crop and forest pest metawebs shift towards increased linkage and suitability overlap under climate change. *Communications Biology*, 3, 1–10.

Halevy, A., Norvig, P. & Pereira, F. (2009). The Unreasonable Effectiveness of Data. *IEEE Intelligent Systems*, 24, 8–12.

Halko, N., Martinsson, P.G. & Tropp, J.A. (2011). Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions. *SIAM Review*, 53, 217–288.

Herbert, F. (1965). *Dune*. ~~First~~1st edn. Chilton Book Company, Philadelphia.

Hoffmann, J., Bar-Sinai, Y., Lee, L.M., Andrejevic, J., Mishra, S., Rubinstein, S.M., *et al.* (2019). Machine learning in a data-limited regime: Augmenting experiments with synthetic data uncovers order in crumpled sheets. *Science Advances*, 5, eaau6792.

Holm, E.A. (2019). In defense of the black box. *Science*, 364, 26–27.

Hortal, J., de Bello, F., Diniz-Filho, J.A.F., Lewinsohn, T.M., Lobo, J.M. & Ladle, R.J. (2015). Seven Shortfalls that Beset Large-Scale Knowledge of Biodiversity. *Annual Review of Ecology, Evolution, and Systematics*, 46, 523–549.

Hutchinson, M.C., Cagua, E.F. & Stouffer, D.B. (2017). Cophylogenetic signal is detectable in pollination interactions across ecological scales. *Ecology*, n/a–n/a.

Jordano, P. (2016a). Chasing Ecological Interactions. *PLOS Biol*, 14, e1002559.

Jordano, P. (2016b). Sampling networks of ecological interactions. *Functional Ecology*, 30, 1883–1893.

Litsios, G. & Salamin, N. (2012). Effects of Phylogenetic Signal on Ancestral State Reconstruction. *Systematic Biology*, 61, 533–538.

Maiorano, L., Montemaggiori, A., Ficetola, G.F., O'Connor, L. & Thuiller, W. (~~2020a~~2020). Data from: Tetra-EU 1.0: A species-level trophic meta-web of European tetrapods.

~~Maiorano, L., Montemaggiori, A., Ficetola, G.F., O'Connor, L. & Thuiller, W. (2020b). TETRA-EU 1.0: A species-level trophic metaweb of European tetrapods. *Global Ecology and Biogeography*, 29, 1452–1457.~~

Marco, M.D., Venter, O., Possingham, H.P. & Watson, J.E.M. (2018). Changes in human footprint drive changes in species extinction risk. *Nature Communications*, 9, 4621.

Morales-Castilla, I., Matias, M.G., Gravel, D. & Araújo, M.B. (2015). Inferring biotic interactions from proxies. *Trends in Ecology & Evolution*, 30, 347–356.

Mouquet, N., Devictor, V., Meynard, C.N., Munoz, F., Bersier, L.-F., Chave, J., *et al.* (2012). Ecophylogenetics: Advances and perspectives. *Biological Reviews*, 87, 769–785.

Nenzén, H.K., Montoya, D. & Varela, S. (2014). The Impact of 850,000 Years of Climate Changes on the Structure and Dynamics of Mammal Food Webs. *PLOS ONE*, 9, e106651.

Neutel, A.-M., Heesterbeek, J.A.P. & de Ruiter, P.C. (2002). Stability in Real Food Webs: Weak Links in Long Loops. *Science*, 296, 1120–1123.

No'kmaq, M., Marshall, A., Beazley, K.F., Hum, J., joudry, shalan, Papadopoulos, A., *et al.* (2021). "Awakening the sleeping giant": Re-Indigenization principles for transforming biodiversity conservation in Canada and beyond. *FACETS*, 6, 839–869.

Nost, E. & Goldstein, J.E. (2021). A political ecology of data. *Environment and Planning E: Nature and Space*, 25148486211043503.

O'Connor, L.M.J., Pollock, L.J., Braga, J., Ficetola, G.F., Maiorano, L., ~~Martinez-Almoyna~~MartinezAlmoyna, C., *et al.* (2020). Unveiling the food webs of tetrapods across Europe through the prism of the Eltonian niche. *Journal of Biogeography*, 47, 181–192.

Pan, S.J. & Yang, Q. (2010). A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, 22, 1345–1359.

Park, J. & Newman, M.E.J. (2004). Statistical mechanics of networks. *Physical Review E*, 70, 066117.

Pedersen, E.J., Thompson, P.L., Ball, R.A., Fortin, M.-J., Gouhier, T.C., Link, H., *et al.* (2017). Signatures of the collapse and incipient recovery of an overexploited marine ecosystem. *Royal Society Open Science*, 4, 170215.

Perretti, C.T., Munch, S.B. & Sugihara, G. (2013). Model-free forecasting outperforms the correct mechanistic model for simulated and experimental data. *Proceedings of the National Academy of Sciences*, 110, 5253–5257.

Pires, M.M., Koch, P.L., Fariña, R.A., de Aguiar, M.A.M., dos Reis, S.F. & Guimarães, P.R. (2015). Pleistocene megafaunal interaction networks became more vulnerable after human arrival. *Proceedings of the Royal Society B: Biological Sciences*, 282, 20151367.

Poelen, J.H., Simons, J.D. & Mungall, C.J. (2014). Global biotic interactions: An open infrastructure to share and analyze species-interaction datasets. *Ecological Informatics*, 24, 148–159.

Poisot, T., Belisle, Z., Hoebeke, L., Stock, M. & Szefer, P. (2019). EcologicalNetworks.jl - analysing ecological networks. *Ecography*.

Poisot, T., Bergeron, G., Cazelles, K., Dallas, T., Gravel, D., MacDonald, A., *et al.* (2021a). Global knowledge gaps in species interaction networks data. *Journal of Biogeography*, n/a.

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

Poisot, T., Cirtwill, A.R., Cazelles, K., Gravel, D., Fortin, M.-J. & Stouffer, D.B. (2016). The structure of probabilistic networks. *Methods in Ecology and Evolution*, 7, 303–312.

Poisot, T., Ouellet, M.-A., Mollentze, N., Farrell, M.J., Becker, D.J., Albery, G.F., *et al.* (2021b). ~~Imputing the mammalian virome with linear filtering and singular value decomposition. *arXiv:2105.14973 q-bio*~~*Imputing the mammalian virome with linear filtering and singular value decomposition*. Available at: `http://arxiv.org/abs/2105.14973`. Last accessed.

Poisot, T. & Stouffer, D.B. (2018). Interactions retain the co-phylogenetic matching that communities lost. *Oikos*, 127, 230–238.

Poisot, T., Stouffer, D.B. & Gravel, D. (2015). Beyond species: Why ecological interaction networks vary through space and time. *Oikos*, 124, 243–251.

Price, P.W. (2003). *Macroevolutionary theory on macroecological patterns*. Cambridge University Press.

Ray, J.C., Grimm, J. & Olive, A. (2021). The biodiversity crisis in Canada: Failures and challenges of federal and sub-national strategic and legal frameworks. *FACETS*, 6, 1044–1068.

Reeve, R., Leinster, T., Cobbold, C.A., Thompson, J., Brummitt, N., Mitchell, S.N., *et al.* (2016). ~~How to partition diversity. *arXiv:1404.6520 q-bio*~~*How to partition diversity*. Available at: `http://arxiv.org/abs/1404.6520`. Last accessed.

Rosado, B.H.P., Dias, A. & de Mattos, E. (2013). Going Back to Basics: Importance of Ecophysiology when Choosing Functional Traits for Studying Communities and Ecosystems. *Natureza & conservaç~ao revista brasileira de conservaç~ao da natureza*, 11, 15–22.

Runghen, R., Stouffer, D.B. & Dalla Riva, G.V. (2021). Exploiting node metadata to predict interactions in large networks using graph embedding and neural networks.

Schoener, T.W. (1989). Food webs from the small to the large. *Ecology*, 70, 1559–1589.

Solís-Lemus, C., Bastide, P. & Ané, C. (2017). PhyloNetworks: A Package for Phylogenetic Networks. *Molecular Biology and Evolution*, 34, 3292–3298.

Stock, M. (2021). Pairwise learning for predicting pollination interactions based on traits and phylogeny. *Ecological Modelling*, 14.

Stouffer, D.B., Sales-Pardo, M., Sirer, M.I. & Bascompte, J. (2012). Evolutionary Conservation of Species' Roles in Food Webs. *Science*, 335, 1489–1492.

Strong, J.S. & Leroux, S.J. (2014). Impact of Non-Native Terrestrial Mammals on the Structure of the Terrestrial Mammal Food Web of Newfoundland, Canada. *PLOS ONE*, 9, e106264.

Strydom, T., Catchen, M.D., Banville, F., Caron, D., Dansereau, G., Desjardins-Proulx, P., *et al.* (2021a). ~~*A Roadmap Toward Predicting Species Interaction Networks (Across Space and Time)* (Preprint). EcoEvoRxiv.~~ A roadmap towards predicting species interaction networks (across space and time). *Philosophical Transactions of the Royal Society B: Biological Sciences*, 376, 20210063.

Strydom, T., Dalla Riva, G.V. & Poisot, T. (2021b). SVD Entropy Reveals the High Complexity of Ecological Networks. *Frontiers in Ecology and Evolution*, 9.

Thessen, A.E. & Parr, C.S. (2014). Knowledge extraction and semantic annotation of text from the encyclopedia of life. *PloS one*, 9, e89550.

Torrey, L. & Shavlik, J. (2010). Transfer learning. In: *Handbook of research on machine learning applications and trends: Algorithms, methods, and techniques*. IGI global, pp. 242–264.

Trøjelsgaard, K. & Olesen, J.M. (2016). Ecological networks in motion: Micro- and macroscopic variability across scales. *Functional Ecology*, 30, 1926–1935.

Turak, E., Brazill-Boast, J., Cooney, T., Drielsma, M., DelaCruz, J., Dunkerley, G., *et al.* (2017). Using the essential biodiversity variables framework to measure biodiversity change at national scale. *Biological Conservation*, SI:measures of biodiversity, 213, 264–271.

Upham, N.S., Esselstyn, J.A. & Jetz, W. (2019). Inferring the mammal tree: Species-level sets of phylogenies for questions in ecology, evolution, and conservation. *PLOS Biology*, 17, e3000494.

van der Hoorn, P., Lippner, G. & Krioukov, D. (2018). Sparse Maximum-Entropy Random Graphs with a Given Power-Law Degree Distribution. *Journal of Statistical Physics*, 173, 806–844.

Xing, S. & Fayle, T.M. (2021). The rise of ecological network meta-analyses: Problems and prospects. *Global Ecology and Conservation*, 30, e01805.

Yeakel, J.D., Pires, M.M., Rudolf, L., Dominy, N.J., Koch, P.L., Guimarães, P.R., *et al.* (2014). Collapse of an ecological network in Ancient Egypt. *PNAS*, 111, 14472–14477.

Youden, W.J. (1950). Index for rating diagnostic tests. *Cancer*, 3, 32–35.

Young, J.-G., Cantwell, G.T. & Newman, M.E.J. (2021). Bayesian inference of network structure from unreliable data. *Journal of Complex Networks*, 8.

Young, S.J. & Scheinerman, E.R. (2007). Random Dot Product Graph Models for Social Networks. In: *Algorithms and Models for the Web-Graph*, Lecture Notes in Computer Science (eds. Bonato, A. & Chung, F.R.K.). Springer, Berlin, Heidelberg, pp. 138–149.

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

figures/figure-concept.png

Figure 1: Overview of the phylogenetic transfer learning (and prediction) of species interactions networks. Starting from an initial, known, network, we learn its representation through a graph embedding step (here, a truncated Singular Value Decomposition; Step 1), yielding a series of latent traits (vulnerability traits representing species at the lower trophic-level and generality traits representing species at higher trophic-levels; *sensu* Schoener (1989)); second, for the destination species pool, we perform ancestral character estimation using a phylogeny (here, using a Brownian model for the latent traits; Step 2); we then sample from the reconstructed distribution of latent traits (Step 3) to generate a probabilistic metaweb at the destination (here, assuming a uniform distribution of traits), and threshold it to yield the final list of interactions (Step 4).

figures/figure-screeplot.png

Figure 2: Left: representation of the screeplot of the singular values from the t-SVD on the European metaweb. The screeplot shows no obvious drop in the singular values that may be leveraged to automatically detect a minimal dimension for embedding, after *e.g.* Zhu & Ghodsi (2006). Right: cumulative fraction of variance explained by each dimension up to the rank of the European metaweb. The grey lines represent cutoffs at 50, 60… 90% of variance explained. For the rest of the analysis, we reverted to an arbitrary threshold of 60% of variance explained, which represented a good tradeoff between accuracy and reduced number of features.

figures/figure-subspaces.png

Figure 3: Visual representation of the left (green/purple) and right (green/brown) subspaces, alongside the adjacency matrix of the food web they encode (greyscale). The European metaweb is on the left, and the imputed Canadian metaweb (before data inflation) on the right. This figure illustrates how much structure the left sub-space captures. As we show in fig. 6, the species with a value of 0 in the left subspace are species without any prey.

figures/figure-validation.png

Figure 4: Left, comparison of the probabilities of interactions assigned by the model to all interactions (grey curve), the subset of interactions found in GLOBI (red), and in the Strong & Leroux (2014) Newfoundland dataset (blue). The model recovers more interaction with a low probability compared to data mining, which can suggest that collected datasets are biased towards more common or easy to identify interactions. Right, distribution of the in-degree and out-degree of the mammals from Canada in the reconstructed metaweb. This figure describes a flat, relatively short food web, in which there are few predators but a large number of preys.

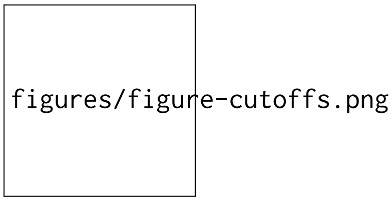

Figure 5: Left: effect of varying the cutoff for probabilities to be considered non-zero on the number of unique links and on $\hat{L}$, the probabilistic estimate of the number of links assuming that all interactions are independent. Right: effect of varying the cutoff on the number of disconnected species, and on network connectance. In both panels, the grey line indicates the cutoff $P(i \rightarrow j) \approx 0.08$ that resulted in the first species losing all of its interactions.

figures/figure-degree.png

Figure 6: Top: biological significance of the first dimension. Left: there is a linear relationship between the values on the first dimension of the left subspace and the generality, *i.e.* the relative number of preys, *sensu* Schoener (1989). Species with a value of 0 in this subspace are at the bottom-most trophic level. Right: there is, similarly, a linear relationship between the position of a species on the first dimension of the right subspace and its vulnerability, *i.e.* the relative number of predators. Taken together, these two figures show that the first-order representation of this network would capture its degree distribution. Bottom: topological consequences of the first dimension. Left: differences in the z-score of the actual configuration model for the reconstructed network, and the prediction based only on the first dimension. Right: distribution of the differences in the left panel.