Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2016, **7**, 303–312 doi: 10.1111/2041-210X.12468

The structure of probabilistic networks

Timothée Poisot[1,2,3]*, Alyssa R. Cirtwill[3], Kévin Cazelles[2,4], Dominique Gravel[2,4], Marie-Josée Fortin[5] and Daniel B. Stouffer[3]

[1]*Département des Sciences Biologiques, Université de Montréal, 90 Avenue Vincent d'Indy, Montréal, QC H3C 3J7,Canada;* [2]*Québec Centre for Biodiversity Sciences, Montréal, QC, Canada;* [3]*School of Biological Sciences, Centre for Integrative Ecology, University of Canterbury, Christchurch, New Zealand;* [4]*Département de Biologie, Chimie, et Géographie, Université du Québec à Rimouski, Rimouski, QC, Canada; and* [5]*Department of Ecology & Evolutionary Biology, University of Toronto, 25 Harbord Street, Toronto, ON M5S 3G5, Canada*

Summary

1. There is a growing realization among community ecologists that interactions between species vary across space and time and that this variation needs to be quantified. Our current numerical framework to analyse the structure of species interactions, based on graph-theoretical approaches, usually do not consider the variability of interactions. As this variability has been show to hold valuable ecological information, there is a need to adapt the current measures of network structure so that they can exploit it.

2. We present analytical expressions of key measures of network structured, adapted so that they account for the variability of ecological interactions. We do so by modelling each interaction as a Bernoulli event; using basic calculus allows expressing the expected value, and when mathematically tractable, its variance. When applied to non-probabilistic data, the measures we present give the same results as their non-probabilistic formulations, meaning that they can be generally applied.

3. We present three case studies that highlight how these measures can be used, in re-analysing data that experimentally measured the variability of interactions, to alleviate the computational demands of permutation-based approaches, and to use the frequency at which interactions are observed over several locations to infer the structure of local networks. We provide a free and open-source implementation of these measures.

4. We discuss how both sampling and data representation of ecological networks can be adapted to allow the application of a fully probabilistic numerical network approach.

Key-words: connectance, degree distribution, ecological networks, modularity, nestedness, species interactions

Introduction

Ecological networks efficiently represent biotic interactions between individuals, populations or species. Historically, their study focused on describing their structure, with a particular attention on food webs (Dunne 2006) and plant-pollinator interactions (Jordano 1987; Bascompte *et al*. 2003). This established that network structure is linked to community or ecosystem-level properties such as stability (McCann 2014), coexistence (Bastolla *et al*. 2009; Haerter, Mitarai & Sneppen 2014) or ecosystem functioning (Duffy 2002). The description of ecological networks resulted in the emergence of questions about how functions and properties of communities emerged from their structure, and this stimulated the development of a wide array of measures for key network properties (Bersier, Banašek-Richter & Cattin 2002; Banašek-Richter, Cattin & Bersier 2004; Jordano & Bascompte 2013).

Given a network (i.e. a structure where nodes, most often species, are linked by edges, representing ecological interactions) as input, measures of network structure return a *property* based on one or several *units* (e.g. nodes, links or groups thereof) from this network, either directly measured, or after an optimization process. Some of the properties are *direct* properties (they only require knowledge of the unit on which they are applied), whereas others are *emergent* (they require knowledge of, and describe, higher-order structures). For example, connectance, the realized proportion of potential interactions, is a direct property of a network, since it can be derived from the number of nodes and edges only. The degree of a node (how many interactions it is involved in) is a direct property of the node. The nestedness of a network (that is, the extent to which specialists and generalists overlap), is an emergent property, as it is not directly predictable from the degree of all nodes. The difference is no mere semantics: the difference between direct and emergent properties is important when interpreting their values. Direct properties are conceptually equivalent to means, in that they tend to be the first moment of network units, whereas emergent properties are conceptually equivalent to variances, higher-order moments or probability distributions.

*Correspondence author. E-mail: tim@poisotlab.io

The interpretation of the measures of network structure as indicators of the action of ecological or evolutionary processes must now account for the numerous observations that network structure varies through space and time. In addition to the already well-established variation in the composition of the local species pool (Havens 2015), networks vary because species do not interact in a consistent way (Poisot *et al.* 2012). Empirical evidence suggests that the network is not the right unit to understand this variation; rather, network variation emerges from the response of interactions to environmental factors and chance events (see Poisot, Stouffer & Gravel 2015; for a review). Interactions can vary for multiple (non-exclusive) reasons. Local mismatching in phenology creates *forbidden links* (Olesen *et al.* 2011; Maruyama *et al.* 2014; Vizentin-Bugoni, Maruyama & Sazima 2014). Local variations in abundance prevent the species from encountering one another (Canard *et al.* 2014). The joint action of neutral, phenologic and behavioural effects creates complex and hard to predict responses (Chamberlain *et al.* 2014; Olito & Fox 2015; Trøjelsgaard *et al.* 2015). For example, Olito & Fox (2015) showed that accounting for neutral (population-size driven) and trait-based effects allows the prediction of the cumulative change in network structure, but not of the change at the level of individual interactions. In addition, Carstensen *et al.* (2014) showed that not all interactions are equally variable within a metacommunity: some are highly consistent, whereas others are extremely rare. These results suggest that species interactions, because they vary, cannot be adequately represented as yes–no events; it is therefore necessary to represent them as probabilities. We should replace the question of *Do these two species interact?* by *How likely is it that they will interact?*.

Yet the current way of dealing with probabilistic interactions is either to ignore variability entirely, or to generate networks with yes/no interactions based on the measured probabilities. Both approaches incur a net loss of information, and measures of network structure that explicitly account for interaction variability are a much needed mathematically rigorous alternative. When ignoring the probabilistic nature of interactions (henceforth *binary* networks), every non-zero element of the network is explicitly assumed to occur with probability 1. This over-represents rare events and increases the number of interactions; as a result, this changes the estimated value of different network properties, in a way that remains poorly understood. The generation of random binary networks based on probabilities also suffers from biases, especially in the range of connectance within which most ecological systems lie. These biases are (i) pseudo-replication when the permutational space is small (Poisot & Gravel 2014), and (ii) systematic biases in the emergent properties at low connectances (Chagnon 2015). An alternative is to consider only the interactions above a given threshold, which unfortunately leads to under-representation of rare events and decreases the effective number of interactions. The use of thresholds also notably introduces the risk of removing species that have a lot of interactions that individually have a low probability of occurring. These considerations highlight the need to amend our current methodology for the description of ecological networks, in order to give more importance to the variation of individual interactions.

Yet the extant methodological corpus is well accepted, and the properties it describes are well understood. Rather than suggesting measures, we argue that it is more productive to re-express those we already have, in a way that does not lose information when applied to probabilistic networks. We contribute to this effort by redeveloping a unified toolkit of measures to characterize the structure of probabilistic interaction networks. Several direct and emergent core properties of ecological networks (both bipartite and unipartite) can be reformulated in a probabilistic context. We illustrate this toolkit through several case studies and discuss how the current challenges in the (i) measurement and (ii) analysis of probabilistic interaction networks.

Suite of probabilistic network metrics

We use the following notation throughout the paper. **A** is a matrix where each element A_{ij} gives P(ij), that is the probability that species i establishes an interaction with species j. If **A** represents a unipartite network (e.g. a food web), it is a square matrix and contains the probabilities of each species interacting with all others, including itself. If **A** represents a bipartite network (e.g. a pollination network), it will not necessarily be square. We call S the number of species, and R and C, respectively, the number of rows and columns. $S = R = C$ in unipartite networks, and $S = R + C$ in bipartite networks. Note that all of the measures defined below can be applied on a bipartite network that has been made unipartite.

The unipartite transformation of a bipartite matrix **A** is the block matrix **B**:

$$\mathbf{B} = \begin{pmatrix} 0_{(R,R)} & \mathbf{A} \\ 0_{(C,R)} & 0_{(C,C)} \end{pmatrix}, \qquad \text{eqn 1}$$

where $0_{(C,R)}$ is a matrix of C rows and R columns (noted $C \times R$) filled with 0s, etc. Note that for centrality to be relevant in bipartite networks, this matrix should be made symmetric: $\mathbf{B}_{ij} = \mathbf{B}_{ji}$. We assume that all interactions are independent (so that P($ij \cap kl$) = P(ij)P(kl) for any species) and can be represented as a series of Bernoulli trials (so that $0 \leq$ P(ij) ≤ 1). A Bernoulli trial is the realization of a probabilistic event that gives 1 with probability P(ij) and 0 otherwise. The latter condition allows us to derive estimates for both the *variance* (var(X) = $p(1-p)$) and expected values (E(X) = p) of the network measures. The variance of additive independent events is the sum of their individual variances, and the variance of multiplicative independent events is

$$\mathrm{var}(X_1 X_2 \ldots X_n) = \prod_i \Big(\mathrm{var}(X_i) + [\mathrm{E}(X_i)]^2\Big) - \prod_i [\mathrm{E}(X_i)]^2. \qquad \text{eqn 2}$$

As all X_i are Bernoulli random variables,

$$\mathrm{var}(X_1 X_2 \ldots X_n) = \prod_i p_i - \prod_i p_i^2. \qquad \text{eqn 3}$$

As a final note, all of the measures described below can be applied on the binary (0/1) versions of the networks in which

case they converge on the non-probabilistic version of the measure as usually calculated. This property is particularly desirable as it allows our framework to be used on any unweighted network represented in a probabilistic or binary way. The approach outlined here differs from using *weighted* networks, in that it answers a different ecological question. Probabilistic networks describe the probability that any interaction will happen, whereas weighted networks describe some measure of the effect of the interaction when it happens (Berlow *et al.* 2009); weighted networks therefore assume that the interaction happen. Although there are several measures for weighted ecological networks (Bersier, Banašek-Richter & Cattin 2002), in which interactions happen but with different outcomes, these are not relevant for probabilistic networks; they do not account for the fact that interactions display a variance that will cascade up to the network level. Instead, the weight of each interaction is best viewed as a second modelling step focusing on the nonzero cases (i.e. the interactions that are realized); this is similar to the method now frequently used in species distribution models, where the species presence is modelled first, and its abundance second, using a (possibly) different set of ecological predictors (Boulangeat, Gravel & Thuiller 2012).

Direct network properties

CONNECTANCE AND NUMBER OF INTERACTIONS

Connectance (or network density) is the proportion of possible interactions that are realized, defined as $Co = L/(R \times C)$, where L is the total number of interactions. As all interactions in a probabilistic network are assumed to be independent, the expected value of L, is

$$\hat{L} = \sum_{i,j} A_{ij}, \qquad \text{eqn 4}$$

and $\hat{Co} = \hat{L}/(R \times C)$. Likewise, the variance of the number of interactions is $\text{var}(\hat{L}) = \sum(A_{ij}(1 - A_{ij}))$.

NODE DEGREE

The degree distribution of a network is the distribution of the number of interactions established (number of successors) and received (number of predecessors) by each node. The expected degree of species i is

$$\hat{k}_i = \sum_j (A_{ij} + A_{ji}). \qquad \text{eqn 5}$$

The variance of the degree of each species is $\text{var}(\hat{k}_i) = \sum_j (A_{ij}(1 - A_{ij}) + A_{ji}(1 - A_{ji}))$. Note also that $\sum \hat{k}_i = 2\hat{L}$, as expected.

GENERALITY AND VULNERABILITY

By simplification of the above, generality $\hat{g}_i$ and vulnerability $\hat{v}_i$ are given by, respectively, $\sum_j A_{ij}$ and $\sum_j A_{ji}$, with their variances $\sum_j A_{ij}(1 - A_{ij})$ and $\sum_j A_{ji}(1 - A_{ji})$.

Emergent network properties

PATH LENGTH

Networks can be used to describe indirect interactions between species through the use of paths. The existence of a path of length 2 between species i and j means that they are connected through at least one additional species k. In a probabilistic network, unless some elements are 0, all pairs of species i and j are connected through a path of length 1, with probability A_{ij}. The expected number of paths of length k between species i and j is given by

$$\hat{n_{ij}^{(k)}} = \left(\mathbf{A}^k\right)_{ij}, \qquad \text{eqn 6}$$

where $\mathbf{A}^k$ is the matrix multiplied by itself k times.

It is possible to calculate the probability of having at least one path of length k between the two species: this can be done by calculating the probability of having no path of length k, then taking the running product of the resulting array of probabilities. For the example of length 2, species i and j are connected through g with probability $A_{ig}A_{gj}$, and so this path does not exist with probability $1 - A_{ig}A_{gj}$. For any pair i, j, let $\mathbf{m}$ be the vector such that $m_g = A_{ig}A_{gj}$ for all $g \notin (i,j)$ (Mirchandani 1976). The probability of not having any path of length 2 is $\prod(1-\mathbf{m})$. Therefore, the probability of having a path of length 2 between i and j is

$$\hat{p}_{ij}^{(2)} = 1 - \prod(1 - \mathbf{m}), \qquad \text{eqn 7}$$

which can also be noted

$$\hat{p}_{ij}^{(2)} = 1 - \prod_g (1 - A_{ig}Agj). \qquad \text{eqn 8}$$

In most situations, one would be interested in knowing the probability of having a path of length 2 *without* having a path of length 1; this is simply expressed as $\hat{p}_{ij}^{(2)*} = (1 - A_{ij})\hat{p}_{ij}^{(2)}$. These results can be expanded to any length k in $[2,n-1]$. Firstly, one can, by the same logic, generate the expression for having at least one path of length k:

$$\hat{p}_{ij}^{(k)} = 1 - \prod_{(g_1,g_2\ldots,g_{k-1})} (1 - A_{ig_1}A_{g_1g_2}\ldots A_{g_{k-1}j}) \qquad \text{eqn 9}$$

where $(g_1, g_2\ldots, g_{k-1})$ are all the $(k-1)$-permutations of $1,2,\ldots,n\setminus(i,j)$. Then, having a path of length k without having any smaller path is

$$\hat{p}_{ij}^{(k)*} = (1 - A_{ji})(1 - \hat{p}^{(2)})\ldots(1 - \hat{p}^{(k-1)})\hat{p}^{(k)}. \qquad \text{eqn 10}$$

UNIPARTITE PROJECTION OF BIPARTITE NETWORKS

The unipartite projection of a bipartite network is obtained by linking any two nodes of one mode ('side' of the network) that are connected through at least one node of the other mode; for example, two plants are connected if they share at least one pollinator. It is readily obtained using the formula in the *Path length* section. This yields either the probability of an edge in the unipartite projection (of the upper or lower nodes), or if

using the matrix multiplication, the expected number of such nodes.

NESTEDNESS

Nestedness is an important measure of (bipartite) network structure that tells the extent to which the interactions of specialists and generalists overlap. We use the formula for nestedness proposed by Bastolla *et al.* (2009); this measure is a modification of NODF (Almeida-Neto *et al.* 2008) for ties in species degree that removes the constraint of decreasing fill. Nestedness for each margin of the matrix is defined as $\eta^{(R)}$ and $\eta^{(C)}$ for, respectively, rows and columns. As per Almeida-Neto *et al.* (2008), we define a global statistic for nestedness as $\eta = (\eta^{(R)} + \eta^{(C)})/2$.

Nestedness, in a probabilistic network, is defined as

$$\hat{\eta^{(R)}} = \sum_{i<j} \frac{\sum_k A_{ik} A_{jk}}{\min(g_i, g_j)}, \tag{11}$$

where g_i is the expected generality of species i. The reciprocal holds for $\eta^{(C)}$ when using v_i (the vulnerability) instead of g_i. The values returned are within [0;1], with $\eta = 1$ indicating complete nestedness.

MODULARITY

Modularity represents the extent to which networks are compartmentalized, that is the tendency for subsets of species to be strongly connected together, while they are weakly connected to the rest of the network (Stouffer & Bascompte 2011). Modularity is measured as the proportion of interactions between nodes of an arbitrary number of modules, as opposed to the random expectation. The modularity as derived by Newman (2004) can be expressed as

$$\mathcal{Q} = \sum \left[\left(\frac{\mathbf{A}}{2\sum \mathbf{A}} - \frac{\sum_{i\cdot} \mathbf{A} \sum_{\cdot j} \mathbf{A}}{2\sum \mathbf{A}^2} \right) \delta \right] \qquad \text{eqn 12}$$

where $\sum_{i\cdot} \mathbf{A}$ and $\sum_{\cdot j} \mathbf{A}$ are the sums of rows and columns of $\mathbf{A}$, and δ is a matrix, wherein δ_{ij} is 1 if i and j belong to the same module, and 0 otherwise. This formula can be *directly* applied to probabilistic networks. Modularity takes values in [0;1], where 1 indicates perfect modularity.

CENTRALITY

Although node degree is a rough first-order estimate of centrality, other measures are often needed. Here, we derive the expected value of centrality according to Katz (1953). This measure generalizes to directed acyclic graphs (whereas other do not). For example, although eigenvector centrality is often used in ecology, it cannot be measured on probabilistic graphs. Eigenvector centrality requires the matrix's largest eigenvalues to be real, which is not the case for all probabilistic matrices. The measure proposed by Katz is a useful replacement, because it accounts for the paths of all length between two species instead of focusing on the shortest path.

As described above, the expected number of paths of length k between i and j is $(\mathbf{A}^k)_{ij}$. Based on this, the expected centrality of species i is

$$C_i = \sum_{j=1}^{n} \sum_{k=1}^{n-1} \alpha^k (\mathbf{A}^k)_{ji}. \qquad \text{eqn 13}$$

The parameter $\alpha \in [0;1]$ regulates how important long paths are. When $\alpha = 0$, only first-order paths are accounted for (and the centrality is equal to the degree). When $\alpha = 1$, paths of all length are equally important. As C_i is sensitive to the size of the matrix, we suggest normalizing by $\mathbf{C} = \sum C$ so that

$$< C_i > = \frac{C_i}{\mathbf{C}}. \qquad \text{eqn 14}$$

This results in the *expected relative centrality* of each node in the probabilistic network, which sums to unity.

SPECIES WITH NO OUTGOING LINKS

Estimating the number of species with no outgoing links (successors) can be useful when predicting whether, for example, predators will go extinct. Alternatively, when prior information about traits is available, this can allow predicting the invasion success of a species in a novel community.

A species has no successors if it manages *not* to establish any outgoing interaction, which for species i happens with probability

$$\prod_j (1 - A_{ij}). \qquad \text{eqn 15}$$

The number of expected such species is therefore the sum of the above across all species,

$$\hat{PP} = \sum_i \left(\prod_j (1 - A_{ij}) \right), \qquad \text{eqn 16}$$

and its variance is

$$\text{var}(\hat{PP}) = \sum_i \left(\prod_j (1 - A_{ij}^2) - \prod_j (1 - A_{ij})^2 \right). \qquad \text{eqn 17}$$

Note that in a non-probabilistic context, species with no outgoing links would be considered primary producers. This is not the case here: if interactions are probabilistic events, then even a top predator may have no preys, and this clearly does not imply that it will become a primary producer in the community. For this reason, the trophic position of the species may be measured better with the binary version of the matrix.

SPECIES WITH NO INCOMING LINKS

Using the same approach as for the number of species with no outgoing links, the expected number of species with no incoming links is therefore

$$\hat{TP} = \sum_i \left(\prod_{j \neq i} (1 - A_{ji}) \right). \qquad \text{eqn 18}$$

Note that we exclude self-interactions, as top-predators in food webs can, and often do, engage in cannibalism.

NUMBER OF SPECIES WITH NO INTERACTIONS

Predicting the number of species with no interactions (or whether any species will have at least one interaction) is useful when predicting whether species will be able to integrate into an existing network, for example. From a methodological point of view, this can also be a helpful *a priori* measure to determine whether null models of networks will have a lot of species with no interactions, and so will require intensive sampling.

A species has no interactions with probability

$$\prod_{j\neq i}(1-A_{ij})(1-A_{ji}). \qquad \text{eqn 19}$$

As for the above, the expected number of species with no interactions (*free species*) is the sum of this quantity across all i:

$$\hat{FS} = \sum_i \prod_{j\neq i}(1-A_{ij})(1-A_{ji}). \qquad \text{eqn 20}$$

The variance of the number of species with no interactions is

$$\begin{aligned}\operatorname{var}(\hat{FS}) = \sum_i \Big(& A_{ij}(1-A_{ij})A_{ji}(1-A_{ji}) \\ & + A_{ij}(1-A_{ij})A_{ji}^2 + A_{ji}(1-A_{ji})A_{ij}^2. \Big)\end{aligned} \qquad \text{eqn 21}$$

SELF-LOOPS

Self-loops (the existence of an interaction of a species onto itself) are only meaningful in unipartite networks. The expected proportion of species with self-loops is very simply defined as $\operatorname{Tr}(\mathbf{A})$, that is the sum of all diagonal elements. The variance is $\operatorname{Tr}(\mathbf{A}\diamond(1-\mathbf{A}))$, where $\diamond$ is the elementwise product operation (Hadamard product).

MOTIFS

Motifs are sets of pre-determined interactions between a fixed number of species (Milo *et al.* 2002; Stouffer *et al.* 2007), such as apparent competition with one predator sharing two prey. As there are an arbitrarily large number of motifs, we will illustrate the approach with only two examples.

The probability that three species form an apparent competition motif where i is the predator, j and k are the prey, is

$$\begin{aligned}\mathrm{P}(i,j,k \in \text{app. comp}) = & A_{ij}(1-A_{ji})A_{ik}(1-A_{ki}) \\ & (1-A_{jk})(1-A_{kj}).\end{aligned} \qquad \text{eqn 22}$$

Similarly, the probability that these three species form an omnivory motif, in which i and j consume k and i consumes j, is

$$\mathrm{P}(i,j,k \in \text{omniv.}) = A_{ij}(1-A_{ji})A_{ik}(1-A_{ki})A_{jk}(1-A_{kj}). \qquad \text{eqn 23}$$

The probability of the number of *any* three-species motif m in a network is given by

$$\hat{N}_{\mathrm{m}} = \sum_i \sum_{j\neq i} \sum_{k\neq j} \mathrm{P}(i,j,k \in \mathrm{m}). \qquad \text{eqn 24}$$

It is indeed possible to have an expression of the variance of this value, or of the variance of any three species forming a given motif, but their expressions become rapidly untractable and are better computed than written.

NETWORK COMPARISON

The dissimilarity of a pair of (ecological) networks can be measured using the framework set forth by Koleff, Gaston & Lennon (2003) using β-diversity measures. Measures of β diversity compute the dissimilarity between two networks based on the cardinality of three sets, a, c and b, which are, respectively, the shared items, items unique to superset (network) 1 and items unique to superset 2 (the identity of which network is 1 or 2 matters for asymmetric measures). Supersets can be the species within each network, or the interactions between species. Following Poisot *et al.* (2012), the dissimilarity of two networks can be measured as either β_{WN} (all interactions), or β_{OS} (interactions involving only common species), with $\beta_{OS} \leq \beta_{WN}$.

Within our framework, these measures can be applied to probabilistic networks. The expected values of $\bar{a}$, $\bar{c}$, and $\bar{b}$ are, respectively, $\sum \mathbf{A}_1 \diamond \mathbf{A}_2$, $\sum \mathbf{A}_1 \diamond (1-\mathbf{A}_2)$, and $\sum (1-\mathbf{A}_1) \diamond \mathbf{A}_2$. Whether β_{OS} or β_{WN} is measured requires to alter the matrices $\mathbf{A}_1$ and $\mathbf{A}_2$. To measure β_{OS}, one must remove all unique species; to measure β_{WN}, one must expand the two matrices so that they have the same species at the same place, and give a weight of 0 to the added interactions.

IMPLEMENTATION

We provide these measures of probabilistic network structure in a free and open-source (MIT licensed) library for the `julia` language, available at http://github.com/PoisotLab/EcologicalNetwork.jl. The code can be cited using the following DOI: 10.5281/zenodo.28317 (version 1.0.1). A user guide, including examples, resides at http://ecologicalnetworkjl.readthedocs.org/.

Case studies

In this section, we contrast the use of probabilistic measures to the current approaches of either using binary networks, or working with null models through simulations. When generating random networks, what we call *Bernoulli trials* from here on, a binary network is generated by doing a Bernoulli trial with probability A_{ij}, for each element of the matrix. This generates networks that have only 0/1 interactions and are realizations of the probabilistic network. This is problematic because higher-order structures involving rare events will be under-represented in the sample, and because most naive approaches (i.e. not controlling for species

degree) are likely to generate species with no interactions, especially in sparsely connected networks frequently encountered in ecology (Milo *et al.* 2003; Poisot & Gravel 2014; Chagnon 2015) – on the other hand, non-naive approaches (e.g. based on swaps or quasi-swaps) break the assumption of independence between interactions.

COMPARISON OF PROBABILISTIC NETWORKS

In this subsection, we apply the above probabilistic measures to a bacteria–phage interaction network. Poullain *et al.* (2008) measured the probability that 24 phage can infect 24 strains of bacteria of the *Pseudomonas fluorescens* species (group SBW25). The (probabilistic) adjacency matrix was constructed by estimating the probability of each phage–bacteria interaction through independent infection assays and can take values of 0, 0·5 (interaction is variable) and 1·0. We have generated a 'Binary' network by setting all interactions with a probability higher than 0 to unity, to simulate the results that would have been obtained in the absence of estimates of interaction probability.

Measuring the structure of the Binary, Bernoulli trials, and Probabilistic network gives the following results (average, and variance when there is an analytical expression):

Measure	Binary	Bernoulli trials	Probabilistic
links	336	221·58 ± 57·57	221·52 ± 57·25
η	0·73	0·528	0·512
$\eta^{(R)}$	0·72	0·525	0·507
$\eta^{(C)}$	0·75	0·531	0·518
one consumer, two resources motif	4784	2089	2110
two consumers, one resource motif	4718	2116	2120

As these results show, treating all interactions as having the same probability, that is removing the information about variability, (i) overestimates nestedness by ≈0·2, (ii) overestimates the number of links by 115 and (iii) overestimates the number of motifs (we have limited our analysis to the two following motifs: one consumer sharing two resources, and two consumers competing for one resource). For the number of links, both the probabilistic measures and the average and variance of 10^4 Bernoulli trials were in strong agreement (they differ only by the second decimal place). For the number of motifs, the difference was larger, but not overly so. It should be noted that, especially for computationally demanding operations such as motif counting, the difference in run-time between the probabilistic and Bernoulli trials approaches can be extremely important.

Using Bernoulli trials had the effect of slightly overestimating nestedness. The overestimation is statistically significant from a purely frequentist point of view, but significance testing is rather meaningless when the number of replicates is this large and can be increased arbitrarily; what is important is that the relative value of the error is small enough that Bernoulli trials are able to adequately reproduce the probabilistic structure of the network. It is not unexpected that Bernoulli trials are this close to the analytical expression of the measures; due to the experimental design of the Poullain *et al.* (2008) study, probabilities of interactions are bound to be high, and so variance is minimal (most elements of **A** have a value of either 0 or 1, and so their individual variance is 0 – although their confidence interval varies as a function of the number of observations from which the probability is derived). Still, despite overall low variance, the binary approach severely mis-represents the structure of the network.

NULL-MODEL-BASED HYPOTHESIS TESTING

In this section, we analyse 59 pollination networks from the literature using two usual null models of network structure, and two models with intermediate constraints. These data cover a wide range a situations, from small to large, and from densely to sparsely connected networks. They provide a good demonstration of the performance of probabilistic metrics. Data come from the *InteractionWeb Database* and were queried on November 2014.

We use the following null models. Firstly [Type I, Fortuna & Bascompte (2006)], any interaction between plant and animals happens with the fixed probability $P = Co$. This model controls for connectance, but removes the effect of degree distribution. Secondly, [Type II, Bascompte *et al.* (2003)], the probability of an interaction between animal i and plant j is $(k_i/R + k_j/C)/2$, the average of the richness-standardized degree of both species. In addition, we use the models called Type III in and out (Poisot, Lounnas & Hochberg 2013), which use the row-wise and columnwise probability of an interaction, respectively, as a way to understand the impact of the degree distribution of upper and lower level species.

Note that these null models will take a binary network and, through some rules, turn it into a probabilistic one. Typically, this probabilistic network is used as a template to generate Bernoulli trials and measure some of their properties, the distribution of which is compared to the empirical network. This approach is computationally inefficient (Poisot & Gravel 2014), especially using naive models (Milo *et al.* 2003), and as we show in the previous section, can yield biased estimates of the true average of nestedness (and presumably other properties).

We measured the nestedness of the 59 (binary) networks, then generated the random networks under the four null models and calculated the expected nestedness using the probabilistic measure. Our results are presented in Fig. 1.

There are two striking results. Firstly, empirical data are consistently *more* nested than the null expectation, as evidenced by the fact that all Δ_η values are strictly positive. Secondly, this underestimation is *linear* between null models I and II, although null model II is always closer to the nestedness of the empirical network (which makes sense, as null model II incorporates the higher-order constraint of approximating the degree distribution of both levels). That the nestedness of the null-model probability matrix is so strongly determined by the nestedness of the empirical networks calls for a closer evaluation of how the results of null models are interpreted (especially

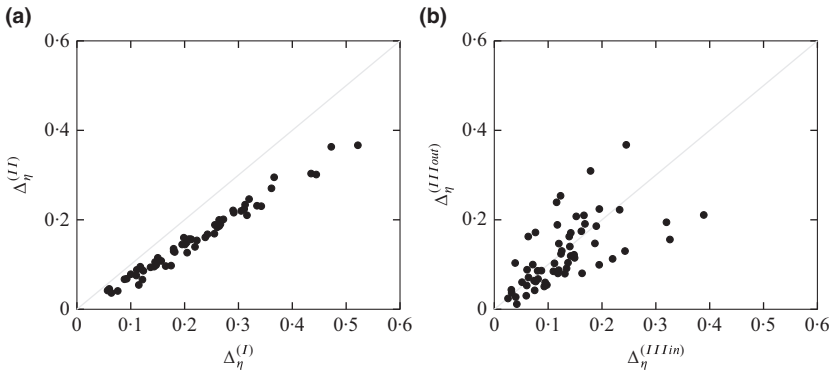

Fig. 1. Results of the null-model analysis of 59 plant-pollination networks. (a) There is a consistent tendency for (i) both models I and II to estimate less nestedness than in the empirical network, although null model II yields more accurate estimates. (b) Models III in and III out also estimate less nestedness than the empirical network, but neither has a systematic bias. For each null model i, the difference $\Delta_\eta^{(i)}$ in nestedness η is expressed as $\Delta_\eta^{(i)} = \eta - \mathcal{N}^{(i)}(\eta)$, where $\mathcal{N}^{(i)}(\eta)$ is the nestedness of null model i.

since networks generated using Bernoulli trials revealed a very low variance in their nestedness).

There is a strong, and previously unaccounted for, circularity in this approach: empirical networks are compared to a null model which, as we show, has a systematic bias *and* a low variance (in the properties of the networks it generates), meaning that differences in nestedness that are small (thus potentially ecologically irrelevant) have a good chance of being reported as significant. Interestingly, models III in and III out made overall *fewer* mistakes at estimating nestedness – respectively 0·129 and 0·123, compared with resp. 0·219 and 0·156 for models I and II. Although the error is overall sensitive to model type (Kruskal–Wallis $\chi^2 = 35{\cdot}80$, d.f. = 3, $P \leq 10^{-4}$), the three pairs of models that where significantly different after controlling for multiple comparisons are I and II, I and III in, and I and III out (model II is not different from either models III in or out).

In short, this analysis reveals that (i) the null expectation of a network property under randomization scenarios can be obtained through the analysis of the probabilistic matrix, instead of the analysis of simulated Bernoulli networks; (ii) different models have different systematic biases, with models of the type III performing better for nestedness than any other models. This can be explained by the fact that nestedness of a network, as expressed by Bastolla *et al.* (2009), is the average of a row-wise and columnwise nestedness. These depend on the species degree, and as such should be well predicted by models III. The true novelty of the approach outlined here is that, rather than having to calculate the measure for thousands of replicates, an *unbiased* estimate of its mean can be obtained in a fraction of the time using the measures described here. This is particularly important since, as demonstrated by Chagnon (2015), the generation of null randomization is subject to biases in the range of connectance where most ecological networks fall. Our approach aims to provide a bias-free, time-effective way of estimating the expected value of a network property.

SPATIAL VARIATION PREDICTS LOCAL NETWORK STRUCTURE

In this final application, we re-analyse data from a previous study by Trøjelsgaard *et al.* (2015), to investigate how spatial information can be used to derive probability of interactions. In the original data set, 14 locations have been sampled to describe the local plant-pollination network. This data set exhibits both species and interaction variability across sampling locations. We define the overall probability of an interaction in the following way,

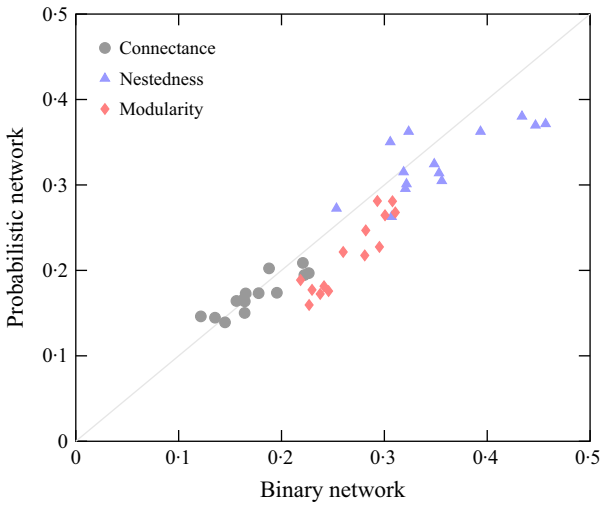

Fig. 2. Local network structure inferred from the locally observed interactions (x-axis) or the spatial probabilistic model (y-axis) in the Canaria Island data set. Although the binary networks slightly underestimate the properties studied here, there is a positive and linear relationship between the empirical structure, and the structure predicted based on probabilities of interactions derived from occurrence information.

$$\mathrm{P}(i \to j) = \frac{\mathbf{N}_{ij}}{\mathbf{O}_{ij}}, \qquad \text{eqn 25}$$

where $\mathbf{O}_{ij}$ is the number of sampling locations in which both pollinator i and plant j co-occur, and $\mathbf{N}_{ij}$ is the number of sampling locations in which they interact. This takes values between 0 (no co-occurence *or* no interactions) and 1 (interaction observed every time there is co-occurrence, including single observations of an interacting species pair). This represents a simple probabilistic model, in which it is assumed that our ability to observe the interaction is a proxy of how frequent it is.

Based on this information, we compare the connectance, nestedness and modularity, of each sampled (binary) network, to the expected values if interactions are well predicted by the probability given above. The results are presented in Fig. 2. There is a clear linear, positive correlation (coeff. 0·89 for

connectance, 0·76 for η and 0·92 for modularity) between the observed network properties (binary matrices) and the predictions based on the probabilistic model. This analysis, although simple, suggests that the *local* structure of ecological networks can represent the outcome of a filtering of species interactions, the signature of which can be detected at the regional level by a variation in the probabilities of interactions. Note, however, that this approach *does not* allow predicting the structure of any arbitrary species pool, as it cannot know the probability of an interaction between two species that never co-occured.

Discussion

Understanding the structure of ecological networks, and whether it relates to emergent ecosystem properties, is a strong research agenda for community ecology. A proper estimation of this structure requires tools that address all forms of complexity, the most oft-neglected yet pervasive of which is the fact that interactions are variable. Through the suite of measures we present here, we allow future analyses of network structure to account for this phenomenon. There are two main considerations highlighted by this methodological development. Firstly, in what way probabilistic data are actually independent? Secondly, what are the implications for data collection?

NON-INDEPENDENCE OF INTERACTIONS

We developed and presented a set of measures to quantify the expected network structure, using the probability that each interaction is observed or happens, in a way that does not require time-consuming simulations. Our framework is set up in such a way that the probabilities of interactions are considered to be independent. This is an over-simplification of what we understand of ecological reality, where interactions have effects on one another (Golubski & Abrams 2011; Sanders & Veen 2012; Ims *et al.* 2013). Yet we feel that, as a first approximation, this assumption is reasonable. There is a strong methodological argument for which the non-independence of interactions cannot currently be robustly accounted for: analytical expectations for non-independent Bernoulli events require knowledge of the full dependence structure. Not only does this severely limit the ability to provide measures of network structure, it requires a far more extensive sampling that what is needed to obtain an estimate of the probability of interactions one by one.

ESTIMATES OF INTERACTION PROBABILITIES

Estimating interaction probabilities based on species abundances (Canard *et al.* 2014; Olito & Fox 2015) do not yield independent probabilities: changing the abundance of one species changes all probabilities in the network. They are not Bernoulli events either, as the sum of all probabilities derived this way sums to unity. On the other hand, 'cafeteria experiments' (in which individuals from two species are directly exposed to one another to observe whether or not an interaction occurs) give truly independent probabilities of interactions–although this approach is limited to systems with a small number of species– and that are amenable to microcosms or mesocosms experiments. Using the approach outlined by Poisot, Stouffer & Gravel (2015), different sources of information (species abundance, trait distribution and the outcome of experiments) can be combined to estimate the probability that interactions will happen in empirical communities.

Another way to obtain approximation of the probability of interactions is to use spatially and temporally replicated sampling (assuming that replicates are done in environments that can be assumed to be comparably homogeneous); in this context, it is not the interactions that are repeatedly sampled, but the network as a whole. Some studies (Tylianakis, Tscharntke & Lewis 2007; Carstensen *et al.* 2014; Olito & Fox 2015; Trøjelsgaard *et al.* 2015) surveyed the existence of interactions at different locations, and a simple approach of dividing the number of observations of an interaction by the number of co-occurence of the species involved will provide a (somewhat crude) estimate of the probability of this interaction. This approach requires extensive sampling, especially as interactions are harder to observe than species (Poisot *et al.* 2012; Gilarranz *et al.* 2015), yet it enables the re-analysis of existing data sets in a probabilistic context.

IMPLICATIONS FOR DATA COLLECTION

An important outcome is that, when estimating probabilities from observational data, it becomes possible to have an estimate of how robust the sampling is. How completely a network is sampled is a key, yet often-overlooked, driver of some measures of structure (Nielsen & Bascompte 2007; Chacoff *et al.* 2012; Fründ, McCann & Williams 2015). The probabilistic approach allows to estimate the *confidence interval* of the interaction probability, knowing the number of samples used for the estimation. Assuming normally distributed observational error (this can be generalized for other error distributions), the confidence interval around a probability p estimated from n samples is

$$\epsilon = z\sqrt{\frac{1}{n}p(1-p)}. \qquad \text{eqn 26}$$

For a 95% confidence interval, $z \approx 1.96$. If an interaction is estimated to happen at $p = 0{\cdot}3$, its 95% confidence interval is [0;0·74] when estimated from four samples, [0·01;0·58] when estimated from ten, and [0·21;0·38] when estimated from a hundred. Note that the above formula tends to perform poorly when $n < 30$, and does not apply when $p \in \{0,1\}$; it nevertheless provides an *estimate* of how robust the probability estimate is.

The quantification, and integration, of uncertainty in the probability of interaction, is a subject that remains to be worked out. To develop a coarse understanding of how it affects the estimate of network properties, one can (for example) sample the interaction probability within its 95% confidence interval. This points to a fundamental issue with the sampling of networks: a precise estimate of the probability of interactions from observational data is tremendously difficult to achieve. Although the development of predictive models partly alleviates

this difficulty, estimating confidence intervals around the probability of an interaction guide empirical research efforts to (i) either collect additional replicates or (ii) provide additional data to improve the performance of predictive models.

Acknowledgements

This work was funded by a CIEE working group grant to TP, DG and DBS. TP is funded by a starting grant from the Université de Montréal, and a Discovery Grant from NSERC. DBS acknowledges support from a Marsden Fund Fast-Start grant (UOC-1101) and Rutherford Discovery Fellowship, both administered by the Royal Society of New Zealand. The idea of network measures as direct/emergent properties of network units was first discussed during the *Web of Life* meeting, held in Montpellier in 2012.

Data accessibility

Data used in this article have already been archived. Data from Application 1 come from Poullain *et al.* (2008). Data from Application 2 are available from the Interaction Web Database (https://www.nceas.ucsb.edu/interactionweb/resources.html). Data from Application 3 are from Hadfield *et al.* (2014) (10.5061/dryad.jf3tj), and were retrieved from (http://mangal.io/data/dataset/4/). Accessibility of the code is mentioned in main text.

References

Almeida-Neto, M., Guimarães, P., Guimarães, P.R., Loyola, R.D. & Ulrich, W. (2008). A consistent metric for nestedness analysis in ecological systems: reconciling concept and measurement. *Oikos*, **117**, 1227–1239.

Banašek-Richter, C., Cattin, M.-F. & Bersier, L.-F. (2004). Sampling effects and the robustness of quantitative and qualitative food-web descriptors. *Journal of Theoretical Biology*, **226**, 23–32.

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

Bastolla, U., Fortuna, M.A., Pascual-García, A., Ferrera, A., Luque, B.& Bascompte, J. (2009). The architecture of mutualistic networks minimizes competition and increases biodiversity. *Nature*, **458**, 1018–1020.

Berlow, E.L., Dunne, J.A, Martinez, N.D., Stark, P.B., Williams, R.J. & Brose, U. (2009). Simple prediction of interaction strengths in complex food webs. *Proceedings of the National Academy of Sciences USA*, **106**, 187–191.

Bersier, L.F., Banašek-Richter, C. & Cattin, M.-F. (2002). Quantitative descriptors of food-web matrices. *Ecology*, **83**, 2394–2407.

Boulangeat, I., Gravel, D. & Thuiller, W. (2012). Accounting for dispersal and biotic interactions to disentangle the drivers of species distributions and their abundances. *Ecology Letters*, **15**, 584–593.

Canard, E.F., Mouquet, N., Mouillot, D., Stanko, M., Miklisova, D. & Gravel, D. (2014). Empirical evaluation of neutral interactions in host-parasite networks. *American Naturalist*, **183**, 468–479.

Carstensen, D.W., Sabatino, M., Trøjelsgaard, K. & Morellato, L.P.C. (2014). Beta diversity of plant-pollinator networks and the spatial turnover of pairwise interactions. *PLoS ONE*, **9**, e112903.

Chacoff, N.P., Vázquez, D.P., Lomáscolo, S.B., Stevani, E.L., Dorado, J. & Padrón, B. (2012). Evaluating sampling completeness in a desert plant-pollinator network. *Journal of Animal Ecology*, **81**, 190–200.

Chagnon, P.-L. (2015). Characterizing topology of ecological networks along gradients: the limits of metrics' standardization. *Ecological Complexity*, **22**, 36–39.

Chamberlain, S.A., Cartar, R.V., Worley, A.C., Semmler, S.J., Gielens, G., Elwell, S., Evans, M.E., Vamosi, J.C. & Elle, E. (2014). Traits and phylogenetic history contribute to network structure across Canadian plantpollinator communities. *Oecologia*, **176**, 545–556.

Duffy, J.E. (2002). Biodiversity and ecosystem function: the consumer connection. *Oikos*, **99**, 201–219.

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

Fortuna, M.A. & Bascompte, J. (2006). Habitat loss and the structure of plantanimal mutualistic networks. *Ecology Letters*, **9**, 281–286.

Fründ, J., McCann, K.S. & Williams, N.M. (2015). Sampling bias is a challenge for quantifying specialization and network structure: lessons from a quantitative niche model. *Oikos*. doi: 10.1111/oik.02256.

Gilarranz, L.J., Sabatino, M., Aizen, M.A. & Bascompte, J. (2015). Hot spots of mutualistic networks. *Journal of Animal Ecology*, **84**, 407–413.

Golubski, A.J. & Abrams, P.A. (2011). Modifying modifiers: what happens when interspecific interactions interact? *Journal of Animal Ecology*, **80**, 1097–1108.

Hadfield, J.D., Krasnov, B.R., Poulin, R. & Nakagawa, S. (2014). A tale of two phylogenies: comparative analyses of ecological interactions. *American Naturalist*, **183**, 174–187.

Haerter, J.O., Mitarai, N. & Sneppen, K. (2014). Phage and bacteria support mutual diversity in a narrowing staircase of coexistence. *ISME Journal*, **8**, 2317–2326.

Havens, K. (2015). Scale and structure in natural food webs. *Science*, **257**, 1107–1109.

Ims, R.A., Henden, J.-A., Thingnes, A.V. & Killengreen, S.T. (2013). Indirect food web interactions mediated by predatorrodent dynamics: relative roles of lemmings and voles. *Biology Letters*, **9**, 20130802.

Jordano, P. (1987). Patterns of mutualistic interactions in pollination and seed dispersal: connectance, dependence asymmetries, and coevolution. *The American Naturalist* **129**, 657–677.

Jordano, P. & Bascompte, J. (2013). *Mutualistic Networks*. Princeton University Press, Princeton, NJ.

Katz, L. (1953). A new status index derived from sociometric analysis. *Psychometrika*, **18**, 39–43.

Koleff, P., Gaston, K.J. & Lennon, J.J. (2003). Measuring beta diversity for presence-absence data. *Journal of Animal Ecology*, **72**, 367–382.

Maruyama, P.K., Vizentin-Bugoni, J., Oliveira, G.M., Oliveira, P.E. & Dalsgaard, B. (2014). Morphological and Spatio-Temporal Mismatches Shape a Neotropical Savanna Plant-Hummingbird Network. *Biotropica*, **46**, 740–747.

McCann, K.S. (2014). Diversity and Destructive Oscillations: Camerano, Elton, and May. *Bulletin of the Ecological Society of America*, **95**, 337–340.

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D. & Alon, U. (2002). Network motifs: simple building blocks of complex networks. *Science*, **298**, 824–827.

Milo, R., Kashtan, N., Itzkovitz, S., Newman, M.E.J. & Alon, U. (2003). On the uniform generation of random graphs with prescribed degree sequences. ArXivcond-Mat0312028

Mirchandani, P.B. (1976). Shortest distance and reliability of probabilistic networks. *Computers & Operations Research*, **3**, 347–355.

Newman, M.E.J. (2004). Fast algorithm for detecting community structure in networks. *Physical Review E*, **69**, 066133.

Nielsen, A. & Bascompte, J. (2007). Ecological networks, nestedness and sampling effort. *Ecology*, **95**, 1134–1141.

Olesen, J.M., Bascompte, J., Dupont, Y.L., Elberling, H., Rasmussen, C. & Jordano, P. (2011). Missing and forbidden links in mutualistic networks. *Proceedings of the Royal Society B: Biological Sciences*, **278**, 725–732.

Olito, C. & Fox, J.W. (2015). Species traits and abundances predict metrics of plantpollinator network structure, but not pairwise interactions. *Oikos*, **124**, 428–436.

Poisot, T. & Gravel, D. (2014). When is an ecological network complex? Connectance drives degree distribution and emerging network properties. *PeerJ*, **2**, e251.

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

Poisot, T., Lounnas, M. & Hochberg, M.E. (2013). The structure of natural microbial enemy-victim networks. *Ecological Processes* **2**, 1–9.

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

Poullain, V., Gandon, S., Brockhurst, M.A., Buckling, A. & Hochberg, M.E. (2008). The evolution of specificity in evolving and coevolving antagonistic interactions between a bacteria and its phage. *Evolution*, **62**, 1–11.

Sanders, D. & Veen, F.J.F. (2012). Indirect commensalism promotes persistence of secondary consumer species. *Biology Letters*, **8**, 960–963.

Stouffer, D.B. & Bascompte, J. (2011). Compartmentalization increases food-web persistence. *Proceedings of the National Academy of Sciences USA*, **108**, 3648–3652.

Stouffer, D.B., Camacho, J., Jiang, W. & Amaral, L.A.N. (2007). Evidence for the existence of a robust pattern of prey selection in food webs. *Proceedings of the Royal Society B: Biological Sciences*, **274**, 1931–1940.

Trøjelsgaard, K., Jordano, P., Carstensen, D.W. & Olesen, J.M. (2015). Geographical variation in mutualistic networks: similarity, turnover and partner fidelity. *Proceedings of the Royal Society B: Biological Sciences*, **282**, 20142925.

Tylianakis, J.M., Tscharntke, T. & Lewis, O.T. (2007). Habitat modification alters the structure of tropical hostparasitoid food webs. *Nature*, **445**, 202–205.

Vizentin-Bugoni, J., Maruyama, P.K. & Sazima, M. (2014). Processes entangling interactions in communities: forbidden links are more important than abundance in a hummingbirdplant network. *Proceedings of the Royal Society B: Biological Sciences*, **281**, 20132397.

Received 17 July 2015; accepted 20 August 2015
Handling Editor: Jana Vamosi

Supporting Information

Additional Supporting Information may be found in the online version of this article.

Data S1. Data for the figure in case study 2.

Data S2. Data for the figure in case study 3.