THE STRUCTURE OF PROBABILISTIC NETWORKS

T. POISOT, A.~~CIRTWILL~~ R. CIRTWILL , K. CAZELLES , D. GRAVEL, M.-J. FORTIN, AND D.B. STOUFFER

ABSTRACT

1. There is a growing realization among community ecologists that interactions between species vary in space and time. Yet, our current numerical framework to analyze the structure of interactions, largely based on graph-theoretical approaches, is unsuited to this type of data. Since the variation of species interactions holds much information, there is a need to develop new metrics to exploit it.
2. We present analytical expressions of key network metrics, using a probabilistic framework. Our approach is based on modeling each interaction as a Bernoulli event, and using basic calculus to express the expected value, and when mathematically tractable, its variance. We provide a free and open-source implementation of these measures.
3. We show that our approach allows to overcome limitations of both neglecting the variation of interactions (over-estimation of rare events) and using simulations (extremely high computational demand). We present a few case studies that highlight how these measures can be used.
4. We conclude this contribution by discussing how the sampling and data representation of ecological network can be adapted to better allow the application of a fully probabilistic numerical framework.

Keywords: ecological networks, connectance, degree distribution, nestedness, modularity

Date: Mar. 2015.

INTRODUCTION

Ecological networks are an efficient way to 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). The key result of this line of research was linking this structure to community or ecosystem-level properties such as stability (McCann 2014), coexistence (Bastolla *et al.* 2009; Haerter *et al.* 2014), or ecosystem functioning (Duffy 2002~~; Thébault Loreau 2003; Poisot 2012~~). To a large extent, the description of ecological networks resulted in the emergence of questions about how functions ~~emerged from~~ and properties of communities emerged from their structure, and this stimulated the development of a rich methodological literature, defining a wide array of structural properties.

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* from this network. 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. 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), on the other hand, is an emergent property that is not directly predictable from the degree of all nodes. Though the difference may appear to be semantics, establishing a 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 or other higher-order moments.

In the recent years, the interpretation of the properties of network structure (as indicators of the action of ecological or evolutionary processes) has been somewhat complicated by the observation that network structure varies through space and time. This happens because, contrary to a long-standing assumption of network studies, species from the same pool do not interact in a consistent way (Poisot *et al.* 2012). Empirical and theoretical studies suggest that the network is not the right unit to understand this variation; rather, network variation is an emergent property of the response of ecological interactions to environmental factors and chance events (~~???~~see Poisot *et al.* 2015 for a review). Interactions can vary because of local mismatching in phenology (Olesen *et al.* 2011; Vizentin-Bugoni *et al.* 2014; Maruyama *et al.* 2014), populations fluctuations preventing the interaction (Canard *et al.* 2014), or a combination of both (Chamberlain *et al.* 2014; Olito & Fox 2014). For example, Olito & Fox (2014) show 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) show that within a meta-community, not all interactions are equally variable: some are highly consistent, whereas others are extremely rare. These empirical results all point to the fact that species interactions cannot always be adequately ~~modeled~~ represented as yes-no events; since it is well established that they do vary, it is necessary to represent them as probabilities. To the question of *Do these two species interact?*, we should substitute the question of *How likely is it that they will interact?*. This also requires the considerable methodological adjustment of re-writing measures of network structure to account for the fact that interactions are not consistent; in this paper, we re-develop a unified toolkit of measures to characterize the structure of probabilistic interaction networks.

The current way of dealing with probabilistic interactions are either to ignore variability entirely or to generate random networks. Probabilistic metrics are a mathematically rigorous alternative to both. When ignoring the probabilistic nature of interactions (henceforth *binary* networks), every non-zero element of the network is assumed to be 1. This leads to over-representation of some rare events, and increases the number of interactions; as a result, this changes the estimated value of different network properties, in a way that is not understood at all. Issues are most likely to arise for connectances where the topological (Chagnon 2015) or permutational (Poisot & Gravel 2014) space of random network is small, leading to over-replication or uncharacterized biases. An alternative is to consider only the interactions above a given threshold, which leads to an under-representation of rare events and decreases the effective number of interactions (in addition to the problem that there is no robust criterion to decide on a treshold). More importantly, this introduces the risk of removing species that establish a lot of interactions that each have a low probability. Taken together, 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 — current measures neglect the variability of interactions, and are therefore discarding valuable ecological information. Because the methodological corpus available to describe ecological networks had first been crafted at a time when it was assumed that interactions were invariants, it is unsuited to address the questions that probabilistic networks allow us to ask.

In this paper, we show that several direct and emergent core properties of ecological networks (both bipartite and unipartite) can be re-formulated in a probabilistic context (Yeakel *et al.* 2012; ~~???~~Poisot *et al.* 2015); we conclude by showing how this methodology can be applied to exploit the information contained in the variability of networks, and to reduce the computational burden of current methods in network analysis. ~~We also provide a free and open-source (MIT license) implementation of this suite of measures in a library for the `julia` language, available at `http://github.com/PoisotLab/ProbabilisticNe`~~

Suite of probabilistic network metrics

Throughout this paper, we use the following notation. $\mathbf{A}$ is a matrix wherein A_{ij} is $\mathrm{P}(ij)$, *i.e.* the probability that species i establishes an interaction with species j. If $\mathbf{A}$ 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 $\mathbf{A}$ 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~~. The only bipartite-only measure is nestedness. The unipartite transformation of a bipartite matrix $\mathbf{A}$ is the block matrix

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

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 will also assume that all interactions are independent (so that ~~$\text{P}(ij|kl) = \text{P}(ij)\text{P}(kl)$~~ $\text{P}(ij \cap kl) = \text{P}(ij)\text{P}(kl)$ for any species), and can be represented as a series of Bernoulli trials (so that $0 \leq \text{P}(ij) \leq 1$). A Bernoulli trial is simply the realization of a probability event, giving 1 with probability $\text{P}(ij)$, and 0 else. The latter condition allows us to derive estimates for the *variance* ($\text{var}(X) = p(1-p)$), and expected values ($\text{E}(X) = p$). We can therefore estimate the variance of most properties, using the fact that the variance of additive independent events is the sum of their individual variances, and that the variance of multiplicative independent events is

$$\text{var}(X_1 X_2 ... X_n) = \prod_i \left(\text{var}(X_i) + [\text{E}(X_i)]^2\right) - \prod_i [\text{E}(X_i)]^2 \tag{2}$$

As all X_i are Bernouilli random variables ,

$$\text{var}(X_1 X_2 ... X_n) = \prod_i p_i - \prod_i p_i^2 \tag{3}$$

As a final note, all of the measures described below can be applied on the binary (0/1) versions of the networks ~~and will give the exact value of the non-probabilistic measure~~ in which case they effectively are 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 network, whether they are represented in a probabilistic or binary way. Nonetheless, this approach is different from using *weighted* networks, in that it answers a completely different question. Probabilistic networks describe the probability that any interaction will happen, whereas weighted networks describe the effect of the interaction when it happens. Although there are several measures for *quantitative* networks (Bersier *et al.* 2002), in which interactions happen but with different outcomes, these are not relevant for probabilistic networks, which require to account for the fact that interactions are probabilistic event, *i.e.* they display a variance that will cascade up to the network level. Actually, the weight of each interaction is best viewed as a second modeling step, focusing only on the non-zero 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 modeled first, and its abundance second, using a (possibly) different set of predictors (Boulangeat *et al.* 2012).

Direct 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

$$\hat{L} = \sum_{i,j} A_{ij}, \tag{4}$$

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

$$\hat{k}_i = \sum_j (A_{ij} + A_{ji}) \tag{5}$$

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 as expected, $\sum \hat{k}_i = 2\hat{L}$.

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~~ Emergent 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

$$n_{ij}^{\widehat{(k)}} = \left(\mathbf{A}^k\right)_{ij}, \tag{6}$$

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 as $m_g = A_{ig}A_{gj}$ 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

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

which can also be noted

$$\hat{p}^{(2)}_{ij} = 1 - \prod_{g}(1 - A_{ig}Agj).\tag{8}$$

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 ~~$(1-A_{ij})\hat{p}^{(2)}_{ij}$. One~~ $\hat{p}^{(2)*}_{ij} = (1-A_{ij})\hat{p}^{(2)}_{ij}$. These results can be expanded to any length k in $[2, n-1]$. First one can, by the same logic, generate the expression for having at least one path of length ~~3~~k:

$$\hat{p}^{(3)(k)}{}_{ij} = (1-A_{ij})(1-{}^{(2)}_{ij})1 - \prod(1-\mathbf{m})\prod_{x,y}(1-A_{iy})_{(g_1,g_2...,g_{k-1})}(1 - A_{xjig_1}A_{g_1g_2}...A_{g_{k-1}j}),\tag{9}$$

where ~~$\mathbf{m}$ is the vector of all $A_{ix}A_{xy}A_{yj}$ for $x \notin (i,j)$, $y \neq x$. This gives the probability of having at least one path from i to j, passing through any pair of nodes x and y,~~ $(g_1, g_2..., g_{k-1})$ are all the $(k-1)$-permutations of $1, 2, ..., n\backslash(i,j)$. Then having a path of length k without having any ~~shorter path. In theory, this approach can be generalized up to an arbitrary path length, but it becomes rapidly untractable.~~ smaller path is

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

Unipartite projection of bipartite networks. The unipartite projection of a bipartite network is obtained by linking any two nodes of one mode that are connected through at least one node of the other mode; for example, ~~to~~ 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)~~. They define nestedness~~ ; this measure is a correction of NODF (Almeida-Neto *et al.* 2008) for ties in species degree. 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

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

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, *i.e.* 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. Assuming a vector **s** which, for each node in the network, holds the value of the module it belongs to (an integer in $[1, c]$), Newman (2004) proposed a general measure of modularity, which is

$$Q = \sum_{m=1}^{c} \left(e_{mm} - a_m^2\right)$$

, where c is the number of modules,

$$e_{mm} = \sum_{ij} \frac{\mathbf{A}_{ij}}{2c} \delta(\mathbf{s}_i, \mathbf{s}_j)$$

, and

$$a_m = \sum_{n} e_{mn}$$

,

with δ being Kronecker's function, returning 1 if its arguments are equal, 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. We derive the expected value of centrality according to Katz (1953). This ~~measures~~ 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

$$C_i = \sum_{j=1}^{n} \sum_{k=1}^{\cancel{\infty}\, n-1} \alpha^k (\mathbf{A}^k)_{ji}. \quad (12)$$

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 ~~generality). DG: to the degreeor generality?~~ 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

$$C_i = \frac{C_i}{\mathbf{C}}. \tag{13}$$

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, *e.g.*, predators will go extinct. Alternatively, when prior information about traits are available, this can allows 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

$$\prod_j (1 - A_{ij}). \tag{14}$$

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

$$\hat{PP} = \sum_i \left(\prod_j (1 - A_{ij}) \right). \tag{15}$$

and its variance is

$$\mathrm{var}(\hat{PP}) = \sum_i \left(\prod_j (1 - A_{ij}^2) - \prod_j (1 - A_{ij})^2 \right) \tag{16}$$

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 *e.g.* a top predator may have no preys, which do not mean it will not become a primary producer. For this reason, the trophic position of the species may better be measured on 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

$$\hat{TP} = \sum_i \left(\prod_{j \neq i} (1 - A_{ji}) \right) \tag{17}$$

Note that we exclude self-interactions, as top-predators 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. Note that from a methodological point of view, this can 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

$$\prod_{j \neq i} (1 - A_{ij})(1 - A_{ji}) \tag{18}$$

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

$$\hat{FS} = \sum_i \prod_{j \neq i} (1 - A_{ij})(1 - A_{ji}) \tag{19}$$

The variance of the number of species with no interactions is

$$\text{(20)} \qquad \text{var}(\hat{FS}) = \sum_i \left(A_{ij}(1 - A_{ij})A_{ji}(1 - A_{ji}) + A_{ij}(1 - A_{ij})A_{ji}^2 + A_{ji}(1 - A_{ji})A_{ij}^2 \right)$$

Self-loops. Self-loops (the existence of an interaction of a species onto itself) is only meaningful in unipartite networks. The expected proportion of species with self-loops is very simply defined as Tr($\mathbf{A}$), that is, the sum of all diagonal elements. The variance is Tr($\mathbf{A} \diamond (1 - \mathbf{A})$), where $\diamond$ is the element-wise 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 for example one predator sharing two preys. 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 (one predator, two prey) where i is the predator, j and k are the prey, is

$$\text{(21)} \qquad \text{P}(i, j, k \in \text{app. comp}) = A_{ij}(1 - A_{ji})A_{ik}(1 - A_{ki})(1 - A_{jk})(1 - A_{kj})$$

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

$$\text{(22)} \qquad \text{P}(i, j, k \in \text{omniv.}) = A_{ij}(1 - A_{ji})A_{ik}(1 - A_{ki})A_{jk}(1 - A_{kj})$$

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

$$\text{(23)} \qquad \hat{N}_\text{m} = \sum_i \sum_{j \neq i} \sum_{k \neq j} P(i, j, k \in \text{m})$$

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 *et al.* (2003). 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.

Applications

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 ~~free species~~ 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 as explained in Jordano & Bascompte 2013) break the assumption of independence between interactions.

Comparison of probabilistic networks. In this sub-section, we apply the above measures to a bacteria–phage interaction network. Poullain *et al.* (2008) have measured the probability that 24 phages can infect 24 strains of bacteria of the *Pseudomonas fluorescens* species (group SBW25). ~~Each probability has been observed~~ The (probabilistic) adjacency matrix was constructed by estimating the probability of each phage–bacteria interaction though 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 result (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, ~~transforming the probabilistic matrix into a binary one~~ treating all interactions as having the same probability, *i.e.* removing the information about variability, (i) overestimates nestedness by ≈ 0.2, ~~and~~ (ii) overestimates the number of links by ~~115.~~ 115, and (iii) underestimate 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 runtime between the probabilistic and Bernoulli trials approaches can be extremely important.

Using Bernoulli trials had the effect of slightly over-estimating 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 – though 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 Nov. 2014.

We use the following null models. First (Type I, Fortuna & Bascompte (2006)), any interaction between plant and animals happens with the fixed probability $\text{P} = Co$. This model controls for connectance, but removes the effect of degree distribution. Second, (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 *et al.* 2013), that use the row-wise and column-wise 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. ~~For each null model i, the difference $\Delta_N^{(i)}$ in nestedness N is expressed as $\Delta_N^{(i)} = N - \mathcal{N}^{(i)}(N)$, where $\mathcal{N}^{(i)}(N)$ is the nestedness of null model i.~~ Our results are presented in Figure 1.

~~group style=columns=2, horizontal sep=2cm, xmin=0, xmax=0.6, ymin=0, ymax=0.6black!10, no markerscoordinates (0,0) (0.6,0.6); only markstable x = d1, y = d2figures/app2.dat; at (axis cs:0.1,0.55)**A**; black!10, no markerscoordinates (0,0) (0.6,0.6); only markstable x = d3i, y = d3ofigures/app2.dat; at (axis cs:0.1,0.55)**B**;~~

~~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.~~

There are two striking results. First, empirical data are consistently *more* nested than the null expectation, as evidenced by the fact that all Δ_N values are strictly positive. Second, this underestimation is *linear* between null models I and II~~(in that it does not depends on how nested the empirical network is)~~, although null model II is always closer to the nestedness of the empirical network (which makes sense, since null model II incorporates the higher order constraint of respecting 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 since ~~Bernoulli simulations~~ networks generated using Bernoulli trials revealed a very low variance in ~~the simulated~~ 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 ~~simulations~~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 – ~~resp.~~ respectively 0.129 and 0.123, compared to resp. 0.219 and 0.156 for model I and II. Although the error is overall sensitive to model type (Kruskal-Wallis $\chi^2 = 35.80$, 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~~ different models have different systematic biases, with models of the type III performing overall 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

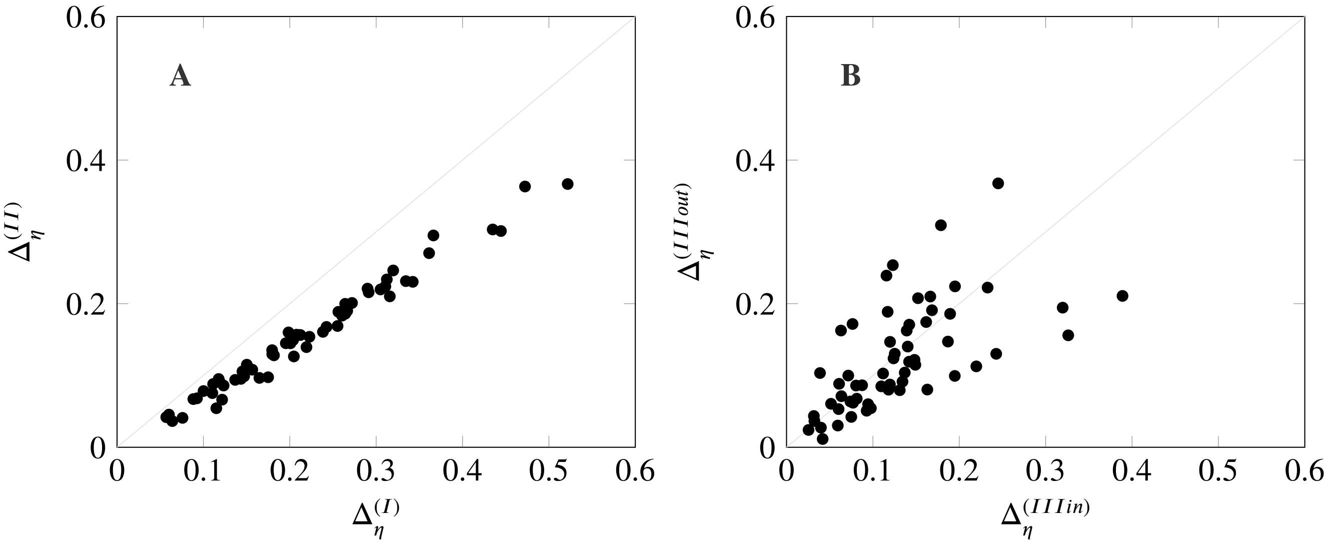

FIGURE 1. Results of the null model analysis of 59 plant-pollination networks. **A**. There is a consistent tendency for (i) both models I and II to estimate less nestedness than in the empirical network, although null model II yields more accurate estimates. **B**. Models III in and III out also estimate 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.

average of a row-wise and column-wise nestedness. These depend on the species degree, and as such should be well predicted by models III. The novelty of this approach is that, instead of having to evaluate 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 are. Our approach is essentially 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-analyze the data from Trøjelsgaard *et al.* (2015), to investigate how spatial information can be used to derive probability of interactions. In the original dataset, fourteen locations have been sampled to describe the local plant-pollination network. There is both species and interaction variability across sampling locations. We define the overall probability of an interaction in the following way,

$$\text{P}(i \rightarrow j) = \frac{\mathbf{N}_{ij}}{\mathbf{O}_{ij}}\,, \tag{24}$$

where $\mathbf{O}_{i,j}$ is the number of sampling locations in which both pollinator i and plant j co-occur, and $\mathbf{N}_{i,j}$ 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).

Based on this information, we compare the connectance, nestedness, and modularity, of each sampled network, to the expected values if interactions are well predicted by the probability given above. The results are presented in Figure 2.

~~IMPLICATIONS FOR DATA COLLECTION~~DISCUSSION

Understanding the structure of ecological networks, and whether it relates to ecosystem properties, is emergent as a key challenge 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. By developing these metrics, we allow future analyses of network structure to account for this phenomenon. There are two main considerations highlighted by this methodological development. First, in what way are probabilistic data independent; second, what are the implications for data collection.

Non-independance 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 do 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 the ecological reality, where different interactions are known to 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-independance of interactions cannot currently be robustly accounted for: analytical expectations for non-independant Bernoulli events require to know the full dependence structure. Not only does it severely limits 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.

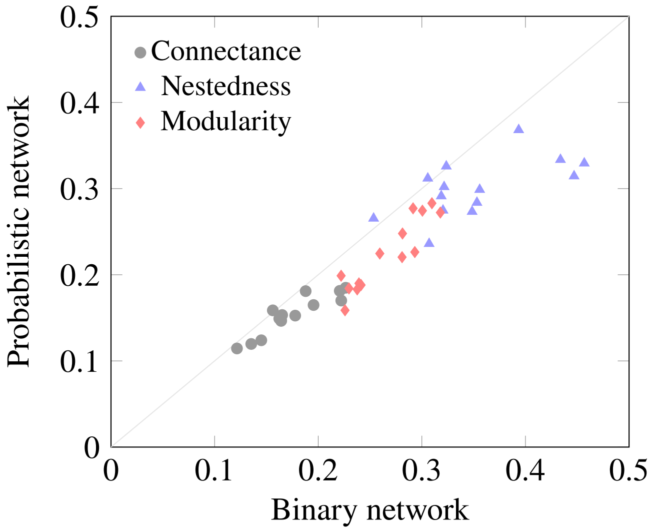

FIGURE 2. Local network structure infered from the locally observed interactions (x-axis) or the spatial probabilistic model (y-axis) in the Canaria Island dataset. Although the binary networks slightly under-estimate the properties studied here, there is a positive and linear relationship between the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

Estimates of interaction probabilities. Estimating interaction probabilities based on species abundances (Olito & Fox 2014; Canard *et al.* 2014) do not ~~, for example,~~ 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; even a simple criteria, such as the frequency of interactions when the two species are put together, is a way of estimating probability. Using the approach ~~outline by (**???**~~outlined by Poisot *et al.* (2015), both 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. This effort requires improved communications between scientists collecting data and scientists developing methodology to analyze them.

Another way to obtain approximation of the probability of interactions is to use spatially replicated sampling. Some studies (Tylianakis *et al.* 2007; Olito & Fox 2014; Carstensen *et al.* 2014; 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 since interactions are harder to observe than species (Poisot *et al.* 2012; Gilarranz *et al.* 2014), yet it enables the re-analysis of existing datasets in a probabilistic context.

~~Understanding the structure of ecological networks, and whether it relates to ecosystem properties, is emergent as a keychallenge 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. By developing these metrics, we allow future analyses of network structure to account for this phenomenon~~

Implications for data collection. An important development is that, when estimating probabilities from observational data, it becomes possible to have an estimate of how robust the sampling is. How completely a networks is sampled is a key, yet an often overlooked one, driver of some measures of structure (Nielsen & Bascompte 2007; Chacoff *et al.* 2011). 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 structure of error), the confidence interval around a probability p estimated from n samples is

$$\epsilon = z\sqrt{\frac{1}{n}p(1-p)}$$

For a 95% confidence interval, $z \approx 1.96$. If an interaction is estimated to happen at $p = 0.3$, 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. This points out to a fundamental issue with the sampling of networks: a correct estimate of the probability of interaction from observational data is tremendously difficult to achieve, and the development of predictive models should be a research priority since it partly alleviates this difficulty.

Implementation. We provide these measures in a free and open-source (MIT license) library for the julia language, available at `http://github.com/PoisotLab/ProbabilisticNetwork.jl`. The code can be cited using the following DOI: **TODO**. A user guide, and API reference, can be found at `http://probabilisticnetworkjl.readthedocs.org/en/latest/`. The code library undergoes automated testing and coverage analysis, the results of which can be accessed from the *GitHub* page given above.

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.

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.

Bascompte, J., Jordano, P., Melián, C.J. & Olesen, J.M. (2003). The nested assembly of plantanimal mutualistic networks. *Proc. Natl. Acad. Sci. U. S. A.*, **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.

Bersier, L.F., Bana\vsek-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. *Ecol. Lett.*, **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. *The 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. (2011). Evaluating sampling completeness in a desert plant-pollinator network. *J. Anim. Ecol.*, no–no.

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*, 1–12.

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. *Ecol. Lett.*, **9**, 281–286.

Gilarranz, L.J., Sabatino, M., Aizen, M.A. & Bascompte, J. (2014). Hot spots of mutualistic networks. *J Anim Ecol*, n/a–n/a.

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

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

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. *Am. Nat.*, **129**, 657–677.

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

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

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D. & Alon, U. (2002). Network motifs: simple building blocks of complex networks. *Science*, **298**, 824–7.

Mirchandani, P.B. (1976). Shortest distance and reliability of probabilistic networks. *Comput. Oper. Res.*, **3**, 347–355.

Newman, M.E.J. (2004). Fast algorithm for detecting community structure in networks. *Phys. Rev. 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. *Proc. R. Soc. B*, **278**, 725–732.

Olito, C. & Fox, J.W. (2014). Species traits and abundances predict metrics of plantpollinator network structure, but not pairwise interactions. *Oikos*, n/a–n/a.

~~Poisot, T. (2012). L'ABC de la spécialisation: apparition, biodiversité, conservation. *Prisme À Idées*, **4**, 49–52.~~

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. *Ecol Lett*, **15**, 1353–1361.

Poisot, T., Lounnas, M. & Hochberg, M.E. (2013). The structure of natural microbial enemy-victim networks. *Ecol. Process.*, **2**, 13.

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. van. (2012). Indirect commensalism promotes persistence of secondary consumer species. *Biology Letters*, 960–963.

Stouffer, D.B. & Bascompte, J. (2011). Compartmentalization increases food-web persistence. *Proc. Natl. Acad. Sci. U.S.A.*, **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. *Proc. R. Soc. B Biol. Sci.*, **274**, 1931–40.

~~Thébault, E. Loreau, M. (2003). Food-web constraints on biodiversityecosystem functioning relationships. *Proc. Natl. Acad. Sci. U. S. A.*, **100**, 14949–14954.~~

Trøjelsgaard, K., Jordano, P., Carstensen, D.W. & Olesen, J.M. (2015). Geographical variation in mutualistic networks: similarity, turnover and partner fidelity. *Proc. R. Soc. B*, **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. *Proc. R. Soc. B*, **281**, 20132397.

Yeakel, J.D., Guimarães, P.R., Novak, M., Fox-Dobbs, K. & Koch, P.L. (2012). Probabilistic patterns of interaction: the effects of link-strength variability on food web structure. *J. R. Soc. Interface*, rsif20120481.