

THE STRUCTURE OF PROBABILISTIC NETWORKS

T. POISOT, A.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

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). To a large extent, 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 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

1 *et al.* 2012). Empirical and theoretical studies suggest that the network is not the right unit to under-
2 stand this variation; rather, network variation is an emergent property of the response of ecological
3 interactions to environmental factors and chance events (see Poisot *et al.* 2015 for a review). Inter-
4 actions can vary because of local mismatching in phenology (Olesen *et al.* 2011; Vizentin-Bugoni *et al.*
5 2014; Maruyama *et al.* 2014), populations fluctuations preventing the interaction (Canard *et al.*
6 2014), or a combination of both (Chamberlain *et al.* 2014; Olito & Fox 2014). For example, Olito &
7 Fox (2014) show that accounting for neutral (population-size driven) and trait-based effects allows
8 the prediction of the cumulative change in network structure, but not of the change at the level of in-
9 dividual interactions. In addition, Carstensen *et al.* (2014) show that within a meta-community, not
10 all interactions are equally variable: some are highly consistent, whereas others are extremely rare.
11 These empirical results all point to the fact that species interactions cannot always be adequately
12 represented as yes-no events; since it is well established that they do vary, it is necessary to represent
13 them as probabilities. To the question of *Do these two species interact?*, we should substitute the
14 question of *How likely is it that they will interact?*. This also requires the considerable methodologi-
15 cal adjustment of re-writing measures of network structure to account for the fact that interactions are
16 not consistent; in this paper, we re-develop a unified toolkit of measures to characterize the structure
17 of probabilistic interaction networks.

18 The current way of dealing with probabilistic interactions are either to ignore variability entirely or
19 to generate random networks. Probabilistic metrics are a mathematically rigorous alternative to both.
20 When ignoring the probabilistic nature of interactions (henceforth *binary* networks), every non-zero
21 element of the network is assumed to be 1. This leads to over-representation of some rare events, and
22 increases the number of interactions; as a result, this changes the estimated value of different network
23 properties, in a way that is not understood at all. Issues are most likely to arise for connectances
24 where the topological (Chagnon 2015) or permutational (Poisot & Gravel 2014) space of random
25 network is small, leading to over-replication or uncharacterized biases. An alternative is to consider
26 only the interactions above a given threshold, which leads to an under-representation of rare events
27 and decreases the effective number of interactions (in addition to the problem that there is no robust
28 criterion to decide on a threshold). More importantly, this introduces the risk of removing species that

1 establish a lot of interactions that each have a low probability. Taken together, these considerations
 2 highlight the need to amend our current methodology for the description of ecological networks, in
 3 order to give more importance to the variation of individual interactions — current measures neglect
 4 the variability of interactions, and are therefore discarding valuable ecological information. Because
 5 the methodological corpus available to describe ecological networks had first been crafted at a time
 6 when it was assumed that interactions were invariants, it is unsuited to address the questions that
 7 probabilistic networks allow us to ask.

8 In this paper, we show that several direct and emergent core properties of ecological networks (both
 9 bipartite and unipartite) can be re-formulated in a probabilistic context (Yeakel *et al.* 2012; Poisot *et*
 10 *al.* 2015); we conclude by showing how this methodology can be applied to exploit the information
 11 contained in the variability of networks, and to reduce the computational burden of current methods
 12 in network analysis.

SUITE OF PROBABILISTIC NETWORK METRICS

14 Throughout this paper, we use the following notation. \mathbf{A} is a matrix wherein A_{ij} is $P(ij)$, *i.e.* the
 15 probability that species i establishes an interaction with species j . If \mathbf{A} represents a unipartite network
 16 (*e.g.* a food web), it is a square matrix and contains the probabilities of each species interacting with
 17 all others, including itself. If \mathbf{A} represents a bipartite network (*e.g.* a pollination network), it will
 18 not necessarily be square. We call S the number of species, and R and C respectively the number of
 19 rows and columns. $S = R = C$ in unipartite networks, and $S = R + C$ in bipartite networks.

20 Note that all of the measures defined below can be applied on a bipartite network that has been made
 21 unipartite. The only bipartite-only measure is nestedness. The unipartite transformation of a bipartite
 22 matrix \mathbf{A} is the block matrix

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

4

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
2 centrality to be relevant in bipartite networks, this matrix should be made symmetric: $\mathbf{B}_{ij} = \mathbf{B}_{ji}$.
3 We will also assume that all interactions are independent (so that $P(ij|kl) = P(ij)P(kl)$ for any
4 species), and can be represented as a series of Bernoulli trials (so that $0 \leq P(ij) \leq 1$). A Bernoulli
5 trial is simply the realization of a probability event, giving 1 with probability $P(ij)$, and 0 else. The
6 latter condition allows us to derive estimates for the *variance* ($\text{var}(X) = p(1 - p)$), and expected
7 values ($E(X) = p$). We can therefore estimate the variance of most properties, using the fact that the
8 variance of additive independent events is the sum of their individual variances, and that the variance
9 of multiplicative independent events is

$$(2) \quad \text{var}(X_1 X_2 \dots X_n) = \prod_i (\text{var}(X_i) + [E(X_i)]^2) - \prod_i [E(X_i)]^2$$

10 As a final note, all of the measures described below can be applied on the binary (0/1) versions
11 of the networks in which case they effectively are the non-probabilistic version of the measure as
12 usually calculated. This property is particularly desirable as it allows our framework to be used on
13 any network, whether they are represented in a probabilistic or binary way. Nonetheless, this ap-
14 proach is different from using *weighted* networks, in that it answers a completely different question.
15 Probabilistic networks describe the probability that any interaction will happen, whereas weighted
16 networks describe the effect of the interaction when it happens. Although there are several measures
17 for *quantitative* networks (Bersier *et al.* 2002), in which interactions happen but with different out-
18 comes, these are not relevant for probabilistic networks, which require to account for the fact that
19 interactions are probabilistic event, *i.e.* they display a variance that will cascade up to the network
20 level. Actually, the weight of each interaction is best viewed as a second modeling step, focusing
21 only on the non-zero cases (*i.e.* the interactions that are realized); this is similar to the method now
22 frequently used in species distribution models, where the species presence is modeled first, and its
23 abundance second, using a (possibly) different set of predictors (Boulangeat *et al.* 2012).

24 **Direct properties.**

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

$$(3) \quad \hat{L} = \sum A_{ij},$$

5 and $\hat{Co} = \hat{L}/(R \times C)$. Likewise, the variance of the number of interactions is $\text{var}(\hat{L}) = \sum (A_{ij}(1 -$
 6 $A_{ij}))$.

7 *Node degree.* The degree distribution of a network is the distribution of the number of interactions
 8 established (number of successors) and received (number of predecessors) by each node. The ex-
 9 pected degree of species i is

$$(4) \quad \hat{k}_i = \sum_j (A_{ij} + A_{ji})$$

10 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
 11 that as expected, $\sum \hat{k}_i = 2\hat{L}$.

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

14 **Emergent properties.**

15 *Path length.* Networks can be used to describe indirect interactions between species through the use
 16 of paths. The existence of a path of length 2 between species i and j means that they are connected
 17 through at least one additional species k . In a probabilistic network, unless some elements are 0, all

1 pairs of species i and j are connected through a path of length 1, with probability A_{ij} . The expected
 2 number of paths of length k between species i and j is given by

$$(5) \quad n_{ij}^{(k)} = (\mathbf{A}^k)_{ij},$$

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

4 It is possible to calculate the probability of having at least one path of length k between the two
 5 species: this can be done by calculating the probability of having no path of length k , then taking
 6 the running product of the resulting array of probabilities. For the example of length 2, species i and
 7 j are connected through g with probability $A_{ig}A_{gj}$, and so this path does not exist with probability
 8 $1 - A_{ig}A_{gj}$. 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
 9 1976). The probability of not having any path of length 2 is $\prod(1 - \mathbf{m})$. Therefore, the probability of
 10 having a path of length 2 between i and j is

$$(6) \quad \hat{p}_{ij}^{(2)} = 1 - \prod(1 - \mathbf{m}).$$

11 In most situations, one would be interested in knowing the probability of having a path of length 2
 12 *without* having a path of length 1; this is simply expressed as $(1 - A_{ij})\hat{p}_{ij}^{(2)}$. One can, by the same
 13 logic, generate the expression for having at least one path of length 3:

$$(7) \quad \hat{p}_{ij}^{(3)} = (1 - A_{ij})(1 - \hat{p}_{ij}^{(2)}) \left(1 - \prod(1 - \mathbf{m})\right) \prod_{x,y} ((1 - A_{ix})(1 - A_{xy})),$$

14 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
 15 at least one path from i to j , passing through any pair of nodes x and y , without having any shorter
 16 path. In theory, this approach can be generalized up to an arbitrary path length, but it becomes rapidly
 17 untractable.

1 *Unipartite projection of bipartite networks.* The unipartite projection of a bipartite network is ob-
 2 tained by linking any two nodes of one mode that are connected through at least one node of the
 3 other mode; for example, to plants are connected if they share at least one pollinator. It is readily
 4 obtained using the formula in the *Path length* section. This yields either the probability of an edge
 5 in the unipartite projection (of the upper or lower nodes), or if using the matrix multiplication, the
 6 expected number of such nodes.

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

13 Nestedness, in a probabilistic network, is defined as

$$(8) \quad \eta^{(R)} = \sum_{i < j} \frac{\sum_k A_{ik} A_{jk}}{\min(g_i, g_j)},$$

14 where g_i is the expected generality of species i . The reciprocal holds for $\eta^{(C)}$ when using v_i (the
 15 vulnerability) instead of g_i .

16 The values returned are within $[0; 1]$, with $\eta = 1$ indicating complete nestedness.

17 *Modularity.* Modularity represents the extent to which networks are compartmentalized, *i.e.* the
 18 tendency for subsets of species to be strongly connected together, while they are weakly connected
 19 to the rest of the network (Stouffer & Bascompte 2011). Modularity is measured as the proportion of
 20 interactions between nodes of an arbitrary number of modules, as opposed to the random expectation.
 21 Assuming a vector \mathbf{s} which, for each node in the network, holds the value of the module it belongs to
 22 (an integer in $[1, c]$), Newman (2004) proposed a general measure of modularity, which is

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

1 , where c is the number of modules,

$$e_{mm} = \sum_{ij} \frac{A_{ij}}{2c} \delta(s_i, s_j)$$

2 , and

$$a_m = \sum_n e_{mn}$$

3 ,

4 with δ being Kronecker's function, returning 1 if its arguments are equal, and 0 otherwise. This
 5 formula can be *directly* applied to probabilistic networks. Modularity takes values in $[0; 1]$, where 1
 6 indicates perfect modularity.

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

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

$$(9) \quad C_i = \sum_{j=1}^n \sum_{k=1}^{\infty} \alpha^k (\mathbf{A}^k)_{ji}.$$

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

$$(10) \quad C_i = \frac{C_i}{\mathbf{C}}.$$

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

7 *Species with no outgoing links.* Estimating the number of species with no outgoing links (successors)
 8 can be useful when predicting whether, *e.g.*, predators will go extinct. Alternatively, when prior
 9 information about traits are available, this can allows predicting the invasion success of a species in a
 10 novel community. A species has no successors if it manages *not* to establish any outgoing interaction,
 11 which for species i happens with probability

$$(11) \quad \prod_j (1 - A_{ij}).$$

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

$$(12) \quad \hat{P}P = \sum_i \left(\prod_j (1 - A_{ij}) \right).$$

13 and its variance is

$$(13) \quad \text{var}(\hat{P}P) = \sum_i \left(\prod_j (1 - A_{ij}^2) - \left(\prod_j (1 - A_{ij}) \right)^2 \right)$$

1 Note that in a non-probabilistic context, species with no outgoing links would be considered primary
 2 producers. This is not the case here: if interactions are probabilistic events, then *e.g.* a top predator
 3 may have no preys, which do not mean it will not become a primary producer. For this reason, the
 4 trophic position of the species may better be measured on the binary version of the matrix.

5 *Species with no incoming links.* Using the same approach as for the number of species with no out-
 6 going links, the expected number of species with no incoming links is therefore

$$(14) \quad T\hat{P} = \sum_i \left(\prod_{j \neq i} (1 - A_{ji}) \right)$$

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

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

13 A species has no interactions with probability

$$(15) \quad \prod_{j \neq i} (1 - A_{ij})(1 - A_{ji})$$

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

$$(16) \quad F\hat{S} = \sum_i \prod_{j \neq i} (1 - A_{ij})(1 - A_{ji})$$

1 The variance of the number of species with no interactions is

$$(17) \quad \text{var}(\hat{FS}) = \sum_i \left(A_{ij}(1 - A_{ij})A_{ji}(1 - A_{ji}) + A_{ij}(1 - A_{ij})A_{ji}^2 + A_{ji}(1 - A_{ji})A_{ij}^2 \right)$$

2 *Self-loops.* Self-loops (the existence of an interaction of a species onto itself) is only meaningful in
 3 unipartite networks. The expected proportion of species with self-loops is very simply defined as
 4 $\text{Tr}(\mathbf{A})$, that is, the sum of all diagonal elements. The variance is $\text{Tr}(\mathbf{A} \diamond (1 - \mathbf{A}))$, where \diamond is the
 5 element-wise product operation (Hadamard product).

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

9 The probability that three species form an apparent competition motif (one predator, two prey) where
 10 i is the predator, j and k are the prey, is

$$(18) \quad P(i, j, k \in \text{app. comp}) = A_{ij}(1 - A_{ji})A_{ik}(1 - A_{ki})(1 - A_{jk})(1 - A_{kj})$$

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

$$(19) \quad P(i, j, k \in \text{omniv.}) = A_{ij}(1 - A_{ji})A_{ik}(1 - A_{ki})A_{jk}(1 - A_{kj})$$

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

$$(20) \quad \hat{N}_m = \sum_i \sum_{j \neq i} \sum_{k \neq j} P(i, j, k \in m)$$

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

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

12 Within our framework, these measures can be applied to probabilistic networks. The expected values
 13 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}
 14 is measured requires to alter the matrices \mathbf{A}_1 and \mathbf{A}_2 . To measure β_{OS} , one must remove all unique
 15 species; to measure β_{WN} , one must expand the two matrices so that they have the same species at the
 16 same place, and give a weight of 0 to the added interactions.

17 APPLICATIONS

18 In this section, we contrast the use of probabilistic measures to the current approaches of either
 19 using binary networks, or working with null models through simulations. When generating ran-
 20 dom networks, what we call *Bernoulli trials* from here on, a binary network is generated by doing a
 21 Bernoulli trial with probability A_{ij} , for each element of the matrix. This generates networks that have
 22 only 0/1 interactions, and are realizations of the probabilistic network. This is problematic because
 23 higher order structures involving rare events will be under-represented in the sample, and because
 24 most naive approaches (*i.e.* not controlling for species degree) are likely to generate species with no
 25 interactions, especially in sparsely connected networks frequently encountered in ecology (Milo *et*
 26 *al.* 2003; Poisot & Gravel 2014) – on the other hand, non-naive approaches (*e.g.* based on swaps

1 or quasi-swaps as explained in Jordano & Bascompte 2013) break the assumption of independence
2 between interactions.

3 **Comparison of probabilistic networks.** In this sub-section, we apply the above measures to a
4 bacteria–phage interaction network. Poullain *et al.* (2008) have measured the probability that 24
5 phages can infect 24 strains of bacteria of the *Pseudomonas fluorescens* species (group SBW25).
6 The (probabilistic) adjacency matrix was constructed by estimating the probability of each phage–
7 bacteria interaction through independent infection assays, and can take values of 0, 0.5 (interaction is
8 variable), and 1.0. We have generated a “Binary” network by setting all interactions with a proba-
9 bility higher than 0 to unity, to simulate the results that would have been obtained in the absence of
10 estimates of interaction probability.

11 Measuring the structure of the Binary, Bernoulli trials, and Probabilistic network gives the following
12 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

13 As these results show, treating all interactions as having the same probability, *i.e.* removing the
14 information about variability, (i) overestimates nestedness by ≈ 0.2 , (ii) overestimates the number
15 of links by 115, and (iii) underestimate the number of motifs (we have limited our analysis to the
16 two following motifs: one consumer sharing two resources, and two consumers competing for one
17 resource). For the number of links, both the probabilistic measures and the average and variance of
18 10^4 Bernoulli trials were in strong agreement (they differ only by the second decimal place). For
19 the number of motifs, the difference was larger, but not overly so. It should be noted that, especially

1 for computationally demanding operations such as motif-counting, the difference in runtime between
2 the probabilistic and Bernoulli trials approaches can be extremely important.

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

14 **Null-model based hypothesis testing.** In this section, we analyse 59 pollination networks from the
15 literature using two usual null models of network structure, and two models with intermediate con-
16 straints. These data cover a wide range a situations, from small to large, and from densely to sparsely
17 connected networks. They provide a good demonstration of the performance of probabilistic metrics.
18 Data come from the *InteractionWeb Database*, and were queried on Nov. 2014.

19 We use the following null models. First (Type I, Fortuna & Bascompte (2006)), any interaction
20 between plant and animals happens with the fixed probability $P = C_o$. This model controls for con-
21 nectance, but removes the effect of degree distribution. Second, (Type II, Bascompte *et al.* (2003)),
22 the probability of an interaction between animal i and plant j is $(k_i/R + k_j/C)/2$, the average of
23 the richness-standardized degree of both species. In addition, we use the models called Type III in
24 and out (Poisot *et al.* 2013), that use the row-wise and column-wise probability of an interaction
25 respectively, as a way to understand the impact of the degree distribution of upper and lower level
26 species.

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

7 We measured the nestedness of the 59 (binary) networks, then generated the random networks under
8 the four null models, and calculated the expected nestedness using the probabilistic measure. Our
9 results are presented in [Figure 1](#).

10 There are two striking results. First, empirical data are consistently *more* nested than the null expect-
11 ation, as evidenced by the fact that all Δ_N values are strictly positive. Second, this underestimation
12 is *linear* between null models I and II, although null model II is always closer to the nestedness of the
13 empirical network (which makes sense, since null model II incorporates the higher order constraint
14 of respecting the degree distribution of both levels). That the nestedness of the null model probability

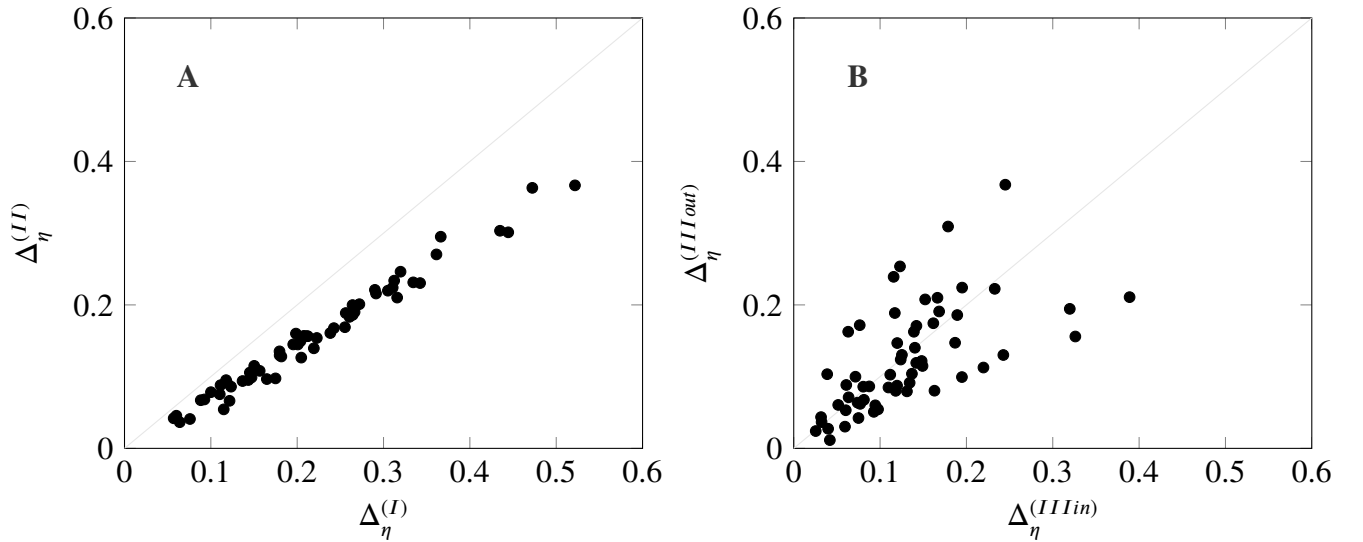


FIGURE 1. Results of the null model analysis of 59 plant-pollination networks. **A.** There is a consistent tendency for (i) both models I and II to estimate less nestedness than in the empirical network, although null model II yields more accurate estimates. **B.** Models III in and III out also estimate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

1 matrix is so strongly determined by the nestedness of the empirical networks calls for a closer eval-
2 uation of how the results of null models are interpreted (especially since networks generated using
3 Bernoulli trials revealed a very low variance in their nestedness).

4 There is a strong, and previously unaccounted for, circularity in this approach: empirical networks
5 are compared to a null model which, as we show, has a systematic bias *and* a low variance (in the
6 properties of the networks it generates), meaning that differences in nestedness that are small (thus
7 potentially ecologically irrelevant) have a good chance of being reported as significant. Interestingly,
8 models III in and III out made overall *fewer* mistakes at estimating nestedness – respectively 0.129
9 and 0.123, compared to resp. 0.219 and 0.156 for model I and II. Although the error is overall
10 sensitive to model type (Kruskal-Wallis $\chi^2 = 35.80$, d.f. = 3, $p \leq 10^{-4}$), the three pairs of models
11 that were significantly different after controlling for multiple comparisons are I and II, I and III in,
12 and I and III out (model II is not different from either models III in or out).

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

25 DISCUSSION

26 Understanding the structure of ecological networks, and whether it relates to ecosystem properties,
27 is emergent as a key challenge for community ecology. A proper estimation of this structure requires

1 tools that address all forms of complexity, the most oft-neglected yet pervasive of which is the fact
2 that interactions are variable. By developing these metrics, we allow future analyses of network
3 structure to account for this phenomenon. There are two main considerations highlighted by this
4 methodological development. First, in what way are probabilistic data independent; second, what
5 are the implications for data collection.

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

17 **Estimates of interaction probabilities.** Estimating interaction probabilities based on species abun-
18 dances (Olito & Fox 2014; Canard *et al.* 2014) do not yield independent probabilities: changing the
19 abundance of one species changes all probabilities in the network. They are not Bernoulli events
20 either, as the sum of all probabilities derived this way sums to unity. On the other hand, “cafeteria
21 experiments” (in which individuals from two species are directly exposed to one another to observe
22 whether or not an interaction occurs) give truly independent probabilities of interactions; even a sim-
23 ple criteria, such as the frequency of interactions when the two species are put together, is a way of
24 estimating probability. Using the approach outlined by Poisot *et al.* (2015), both sources of infor-
25 mation (species abundance, trait distribution, and the outcome of experiments) can be combined to
26 estimate the probability that interactions will happen in empirical communities. This effort requires

1 improved communications between scientists collecting data and scientists developing methodology
2 to analyze them.

3 Another way to obtain approximation of the probability of interactions is to use spatially replicated
4 sampling. Some studies (Tylianakis *et al.* 2007; Olito & Fox 2014; Carstensen *et al.* 2014; Trø-
5 jelsgaard *et al.* 2015) surveyed the existence of interactions at different locations, and a simple ap-
6 proach of dividing the number of observations of an interaction by the number of co-occurrence of
7 the species involved will provide a (somewhat crude) estimate of the probability of this interaction.
8 This approach requires extensive sampling, especially since interactions are harder to observe than
9 species (Poisot *et al.* 2012; Gilarranz *et al.* 2014), yet it enables the re-analysis of existing datasets
10 in a probabilistic context.

11 **Implications for data collection.** An important development is that, when estimating probabilities
12 from observational data, it becomes possible to have an estimate of how robust the sampling is. How
13 completely a networks is sampled is a key, yet an often overlooked one, driver of some measures of
14 structure (Nielsen & Bascompte 2007; Chacoff *et al.* 2011). The probabilistic approach allows to
15 estimate the *confidence interval* of the interaction probability, knowing the number of samples used
16 for the estimation. Assuming normally distributed observational error (this can be generalized for
17 other structure of error), the confidence interval around a probability p estimated from n samples is

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

18 For a 95% confidence interval, $z \approx 1.96$. If an interaction is estimated to happen at $p = 0.3$, its 95%
19 confidence interval is $[0; 0.74]$ when estimated from four samples, $[0.01; 0.58]$ when estimated from
20 ten, and $[0.21; 0.38]$ when estimated from a hundred. This points out to a fundamental issue with
21 the sampling of networks: a correct estimate of the probability of interaction from observational data
22 is tremendously difficult to achieve, and the development of predictive models should be a research
23 priority since it partly alleviates this difficulty.

24 **Implementation.** We provide these measures in a free and open-source (MIT license) library for
25 the julia language, available at <http://github.com/PoisotLab/ProbabilisticNetwork.jl>.

1 The code can be cited using the following DOI: **TODO**. A user guide, and API reference, can be
2 found at <http://probabilisticnetworkjl.readthedocs.org/en/latest/>. The code library
3 undergoes automated testing and coverage analysis, the results of which can be accessed from the
4 *GitHub* page given above.

5 **Acknowledgements:** This work was funded by a CIEE working group grant to TP, DG, and DBS. TP
6 is funded by a starting grant from the Université de Montréal, and a Discovery grant from NSERC.

7 REFERENCES

- 8 Almeida-Neto, M., Guimarães, P., Guimarães, P.R., Loyola, R.D. & Ulrich, W. (2008). A consistent
9 metric for nestedness analysis in ecological systems: reconciling concept and measurement. *Oikos*,
10 **117**, 1227–1239.
- 11 Bascompte, J., Jordano, P., Melián, C.J. & Olesen, J.M. (2003). The nested assembly of plant-animal
12 mutualistic networks. *Proc. Natl. Acad. Sci. U. S. A.*, **100**, 9383–9387.
- 13 Bastolla, U., Fortuna, M.A., Pascual-García, A., Ferrera, A., Luque, B. & Bascompte, J. (2009). The
14 architecture of mutualistic networks minimizes competition and increases biodiversity. *Nature*, **458**,
15 1018–1020.
- 16 Bersier, L.F., Banavsek-Richter, C. & Cattin, M.F. (2002). Quantitative descriptors of food-web
17 matrices. *Ecology*, **83**, 2394–2407.
- 18 Boulangéat, I., Gravel, D. & Thuiller, W. (2012). Accounting for dispersal and biotic interactions to
19 disentangle the drivers of species distributions and their abundances. *Ecol. Lett.*, **15**, 584–593.
- 20 Canard, E.F., Mouquet, N., Mouillot, D., Stanko, M., Miklisova, D. & Gravel, D. (2014). Empirical
21 evaluation of neutral interactions in host-parasite networks. *The American Naturalist*, **183**, 468–479.
- 22 Carstensen, D.W., Sabatino, M., Trøjelsgaard, K. & Morellato, L.P.C. (2014). Beta Diversity of
23 Plant-Pollinator Networks and the Spatial Turnover of Pairwise Interactions. *PLoS ONE*, **9**, e112903.
- 24 Chacoff, N.P., Vázquez, D.P., Lomáscolo, S.B., Stevani, E.L., Dorado, J. & Padrón, B. (2011). Eval-
25 uating sampling completeness in a desert plant-pollinator network. *J. Anim. Ecol.*, no–no.

- 1 Chagnon, P.-L. (2015). Characterizing topology of ecological networks along gradients: The limits
2 of metrics' standardization. *Ecological Complexity*, **22**, 36–39.
- 3 Chamberlain, S.A., Cartar, R.V., Worley, A.C., Semmler, S.J., Gielens, G., Elwell, S., Evans, M.E.,
4 Vamosi, J.C. & Elle, E. (2014). Traits and phylogenetic history contribute to network structure across
5 Canadian plantpollinator communities. *Oecologia*, 1–12.
- 6 Duffy, J.E. (2002). Biodiversity and ecosystem function: the consumer connection. *Oikos*, **99**, 201–
7 219.
- 8 Dunne, J.A. (2006). The Network Structure of Food Webs. *Ecological networks: Linking structure*
9 *and dynamics* (eds J.A. Dunne & M. Pascual), pp. 27–86. Oxford University Press.
- 10 Fortuna, M.A. & Bascompte, J. (2006). Habitat loss and the structure of plantanimal mutualistic
11 networks. *Ecol. Lett.*, **9**, 281–286.
- 12 Gilarranz, L.J., Sabatino, M., Aizen, M.A. & Bascompte, J. (2014). Hot spots of mutualistic net-
13 works. *J Anim Ecol*, n/a–n/a.
- 14 Golubski, A.J. & Abrams, P.A. (2011). Modifying modifiers: what happens when interspecific in-
15 teractions interact? *J. Anim. Ecol.*, **80**, 1097–1108.
- 16 Haerter, J.O., Mitarai, N. & Sneppen, K. (2014). Phage and bacteria support mutual diversity in a
17 narrowing staircase of coexistence. *ISME Journal*.
- 18 Ims, R.A., Henden, J.-A., Thingnes, A.V. & Killengreen, S.T. (2013). Indirect food web interactions
19 mediated by predatorrodent dynamics: relative roles of lemmings and voles. *Biology Letters*, **9**,
20 20130802.
- 21 Jordano, P. (1987). Patterns of mutualistic interactions in pollination and seed dispersal: connectance,
22 dependence asymmetries, and coevolution. *Am. Nat.*, **129**, 657–677.
- 23 Jordano, P. & Bascompte, J. (2013). *Mutualistic Networks*. Princeton Univ Press.
- 24 Katz, L. (1953). A new status index derived from sociometric analysis. *Psychometrika*, **18**, 39–43.
- 25 Koleff, P., Gaston, K.J. & Lennon, J.J. (2003). Measuring beta diversity for presence-absence data.
26 *J. Anim. Ecol.*, **72**, 367–382.

- 1 Maruyama, P.K., Vizentin-Bugoni, J., Oliveira, G.M., Oliveira, P.E. & Dalsgaard, B. (2014). Mor-
2 phological and Spatio-Temporal Mismatches Shape a Neotropical Savanna Plant-Hummingbird Net-
3 work. *Biotropica*, **46**, 740–747.
- 4 McCann, K.S. (2014). Diversity and Destructive Oscillations: Camerano, Elton, and May. *Bulletin*
5 *of the Ecological Society of America*, **95**, 337–340.
- 6 Milo, R., Kashtan, N., Itzkovitz, S., Newman, M.E.J. & Alon, U. (2003). On the uniform generation
7 of random graphs with prescribed degree sequences. *ArXivcond-Mat0312028*.
- 8 Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D. & Alon, U. (2002). Network motifs:
9 simple building blocks of complex networks. *Science*, **298**, 824–7.
- 10 Mirchandani, P.B. (1976). Shortest distance and reliability of probabilistic networks. *Comput. Oper.*
11 *Res.*, **3**, 347–355.
- 12 Newman, M.E.J. (2004). Fast algorithm for detecting community structure in networks. *Phys. Rev.*
13 *E*, **69**, 066133.
- 14 Nielsen, A. & Bascompte, J. (2007). Ecological networks, nestedness and sampling effort. *Ecology*,
15 **95**, 1134–1141.
- 16 Olesen, J.M., Bascompte, J., Dupont, Y.L., Elberling, H., Rasmussen, C. & Jordano, P. (2011). Miss-
17 ing and forbidden links in mutualistic networks. *Proc. R. Soc. B*, **278**, 725–732.
- 18 Olito, C. & Fox, J.W. (2014). Species traits and abundances predict metrics of plantpollinator network
19 structure, but not pairwise interactions. *Oikos*, n/a–n/a.
- 20 Poisot, T. & Gravel, D. (2014). When is an ecological network complex? Connectance drives degree
21 distribution and emerging network properties. *PeerJ*, **2**, e251.
- 22 Poisot, T., Canard, E., Mouillot, D., Mouquet, N. & Gravel, D. (2012). The dissimilarity of species
23 interaction networks. *Ecol Lett*, **15**, 1353–1361.
- 24 Poisot, T., Lounnas, M. & Hochberg, M.E. (2013). The structure of natural microbial enemy-victim
25 networks. *Ecol. Process.*, **2**, 13.

- 1 PoisoT, T., Stouffer, D.B. & Gravel, D. (2015). Beyond species: why ecological interaction networks
2 vary through space and time. *Oikos*, **124**, 243–251.
- 3 Poullain, V., Gandon, S., Brockhurst, M.A., Buckling, A. & Hochberg, M.E. (2008). The Evolution of
4 Specificity in Evolving and Coevolving Antagonistic Interactions Between a Bacteria and Its Phage.
5 *Evolution*, **62**, 1–11.
- 6 Sanders, D. & Veen, F.J.F. van. (2012). Indirect commensalism promotes persistence of secondary
7 consumer species. *Biology Letters*, 960–963.
- 8 Stouffer, D.B. & Bascompte, J. (2011). Compartmentalization increases food-web persistence. *Proc.*
9 *Natl. Acad. Sci. U.S.A.*, **108**, 3648–3652.
- 10 Stouffer, D.B., Camacho, J., Jiang, W. & Amaral, L.A.N. (2007). Evidence for the existence of a
11 robust pattern of prey selection in food webs. *Proc. R. Soc. B Biol. Sci.*, **274**, 1931–40.
- 12 Trøjelsgaard, K., Jordano, P., Carstensen, D.W. & Olesen, J.M. (2015). Geographical variation in
13 mutualistic networks: similarity, turnover and partner fidelity. *Proc. R. Soc. B*, **282**, 20142925.
- 14 Tylianakis, J.M., Tscharnkte, T. & Lewis, O.T. (2007). Habitat modification alters the structure of
15 tropical hostparasitoid food webs. *Nature*, **445**, 202–205.
- 16 Vizentin-Bugoni, J., Maruyama, P.K. & Sazima, M. (2014). Processes entangling interactions in
17 communities: forbidden links are more important than abundance in a hummingbirdplant network.
18 *Proc. R. Soc. B*, **281**, 20132397.
- 19 Yeakel, J.D., Guimarães, P.R., Novak, M., Fox-Dobbs, K. & Koch, P.L. (2012). Probabilistic patterns
20 of interaction: the effects of link-strength variability on food web structure. *J. R. Soc. Interface*,
21 rsif20120481.