THE STRUCTURE OF PROBABILISTIC NETWORKS

THE MAGICAL MISTERY STOUFFER GROUP

INTRODUCTION

Ecological networks are an efficient way to represent the interactions between individual, populations, or species. Historically, their study has focused on (i) linking their structure to community or ecosystem-level properties such as stability (McCann 2014), the maintenance of species richness (Bastolla *et al.* 2009; Haerter *et al.* 2014), ecosystem functioning (Duffy 2002; Thébault & Loreau 2003), and (ii) describing the overall structure of networks, with a particular attention on food webs (Dunne 2006) and plant-pollinator interactions (Jordano 1987; Bascompte *et al.* 2003). To a large extent, the description of network structure enabled questions about how it ties into functional properties, and it is no surprise that the methodology to describe networks is large.

Most measures of network structure function in the following way. Given a network as input, they return a *property* based on one or several *units* within this network. Some of the properties are *direct* properties (they only require knowledge of the unit on which they are applied), and some others are *emerging* properties (they require knowledge of higher-order structures). For example, connectance, the proportion of realized 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, whereas the degree distribution is an emerging property of all nodes. Establishing a difference between direct and emerging properties is important when interpreting their values: direct properties are conceptually equivalent to means, whereas emerging properties are conceptually equivalent to variances.

In the recent years, the interpretation of the values 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; 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

Date: Working paper – Oct. 2014.

unit to understand this variation; rather, network variation is an emerging property of the response of ecological interactions to environmental factors and chance events (Poisot *et al.* 2014). Interactions can vary because of local mis-matching in phenology (Olesen *et al.* 2011), populations fluctuations preventing the interaction (Canard *et al.* 2014), or a combination of both (Chamberlain *et al.* 2014; Olito & Fox 2014). Olito & Fox (2014) show that accounting for neutral (population-size driven) and trait-based effects allows predicting the cumulative change in network structure, but not the change at the level of individual interactions.

Taken together, these considerations highlight the need to amend our current methodology on ecological network to give more importance to the variation at the interaction level. Because the methodology to describe networks has first been crafted at a time when assuming that interactions did not vary, it is unsuited to address the questions that probabilistic networks allows asking. In this paper, we show that several direct and emerging core properties of ecological networks (both bipartite and unipartite) can be re-formulated in a probabilistic context; we conclude by showing how this methodology can be applied to exploit the information contained in the variability and networks, and reduce the computational burden of current methods in network analysis.

METRICS

Throughout this section, we will assume the following notation. $\mathbf{A}$ is a matrix wherein A_{ij} is $\text{P}(ij)$, *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 the probabilities of each species interacting with itself. If $\mathbf{A}$ represents a bipartite network (*e.g.* a pollination network), it will most likely not be square. We call S the number of species, and R and C respectively the number of rows and columns. $S = R + C$ in unipartite networks, and $S = R + C$ in bipartite networks.

Note that all of the measures defined below can be applied on a bipartite network that has been made unipartite; the unipartite transformation of a bipartite matrix $\mathbf{A}$ 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 filled with 0s, etc.

We assume that all interactions are independent (so that $\text{P}(ij|kl) = \text{P}(ij)\text{P}(kl)$ for any species), and can be represented as Bernoulli trials (so that $0 \leq \text{P}(ij) \leq 1$). The later condition allows to derive estimates for the *variance* of the measures, since (i) the variance of a single event X_i of probability p is $\text{var}(X) = p(1-p)$, its expected value is $\text{E}(X) = p$, (ii) the variance of additive independent events is the sum of their individual variances, and (iii) 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 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. And ain't that nice?

Direct properties.

Connectance and number of interactions. Connectance is the proportion of realized upon possible interactions, 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 indpendent, the expected value of L, is

$$\hat{L} = \sum A_{ij}, \tag{3}$$

and $\hat{Co} = \hat{L}/(R \times C)$.

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 and received by each node. The expected degree of species i is

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

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

Average 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 $\sum_j A_{ji}(1-A_{ji})$.

Emerging 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 mean 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

$$n_{ij}^{\hat{(2)}} = \left(\mathbf{A}^k\right)_{ij}, \tag{5}$$

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 between the two species: this can be done by calculating the probability of having 0 paths, then multiplying the resulting array of probabilities. For the example of length 2, species i and j are connected through k with probability $A_{ik}A_{kj}$, and so this path does not exist with probability $1 - A_{ik}A_{kj}$. For any pair i, j, let $\mathbf{m}$ be the vector such as $m_k = A_{ik}A_{kj}$ for all $k \notin (i,j)$. The probability of not having any path of length 2 is $\prod(1-\mathbf{m})$. Therefore, the probability of having a path of length 2 between i and j is

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

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 $(1-A_{ij})\hat{p}_{ij}^{(2)}$. One can, by the same logic, generate the expression for having at least one path of length 3:

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

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 j, without having any shorter path. In theory, this approach can be generalized up to an arbitrary path length, but it becomes rapidly untractable.

Nestedness. We use the formula for nestedness proposed by Bastolla et al. (2009). They define nestedness for each margin of the matrix, 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

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

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.

Katz centrality. Although a rough estimate of centrality is the node degree, as described above, it is often needed to measure centrality within the context of a larger neighborhood. In addition, we derive the expected value of centrality according to Katz (1953). This measures generalizes to directed acyclic graphs. Although eigenvector centrality is often used in ecology, it cannot be measured on probabilistic graphs. Eigenvector centrality requires that the matrix has its largest eigenvalues real, which is not the case for *all* probabilistic matrices. Katz's centrality is nonetheless a useful replacement, because it uses the paths of all lengths between two species instead of focusing on the shortest path.

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

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

The parameter $\alpha \in [0;1]$ regulates how important long paths are. When $\alpha = 0$, only first-order paths count. When $\alpha = 1$, all paths are equally important. As C_i is sensitive to the size of the matrix, we suggest to normalise it so that

$$C_i = \frac{C_i}{\mathbf{C}}. \tag{10}$$

This results in the *expected relative centrality* of each node in the probabilistic network. Note that when using only $k = 1$, and $\alpha = 1$, the raw value of Katz's centrality is the species generality.

Number of primary producers. Primary producers, in a food web, are species with no successors, including themselves. Biologically, they are autotrophic organisms, or organisms whose preys or substrates have been remove from the network. A species is a primary producer if it manages *not* to establish any outgoing interaction, which for species i happens with probability

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

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

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

The variance in the number of expected primary producers is

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

Number of top predators. Top-predators can loosely be defined as species that have no predecessors in the network: they are establishing links with other species, but no species are establishing links with them. Using the same approach than for the number of primary producers, the expected number of top-predators is therefore

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

Note that we exclude the 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 interaction (or whether any species will have at least one interaction) is useful to predict whether species will be able to integrate themselves in an existing network, for example.

A species has no interactions with probability

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

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

$$\hat{FS} = \sum_i \prod_{j \neq i} (1 - A_{ij})(1 - A_{ji}) \tag{16}$$

The variance of the number of species with no interactions is

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

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.

Self-predation. Self-predation (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 $\text{Tr}(\mathbf{A})$, that is, the sum of all diagonal elements. The variance is $\text{Tr}(\mathbf{A} \diamond (1-\mathbf{A}))$, where $\diamond$ is the element-wise product operation.

Motifs. Motifs are sets of pre-determined interactions between a fixed number of specie (Milo *et al.* 2002), such as for example one predator sharing two preys. As there is an arbitrarily large number of motifs, we will illustrate the formulae with only two examples.

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

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

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

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

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

$$\hat{N}_{\mathrm{m}} = \sum_{i} \sum_{j\neq i} \sum_{k\neq j} P(i,j,k \in \mathrm{m}) \tag{20}$$

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 computer than written.

Applications

In this section, we will provide an overview of the applications of probabilistic network measures. The current way of dealing with probabilistic interactions is (i) to ignore it entirely or (ii) to generate random networks. Probabilistic metrics are an alternative to that. When ignoring the probabilistic nature of interactions, what we call *Binary* from here on, 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.

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 is problematic because (i) higher order structures involving rare events will be under-represented in the sample, and (ii) naive approaches are likely to generate free species, especially in sparsely connected networks frequently encountered in ecology (Milo *et al.* 2003; Poisot & Gravel 2014).

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 though three independant infection assays, and can take values of 0, 0.5, and 1.0.

Measure	Binary	Bernoulli trials	Probabilistic
links	336	221.58 ± 57.57	221.52 ± 57.25
η	0.73	0.528	0.512

Measure	Binary	Bernoulli trials	Probabilistic
$\eta^{(R)}$	0.72	0.525	0.507
$\eta^{(C)}$	0.75	0.531	0.518

- connectance
- nestedness

As shown in **??**, transforming the probabilistic matrix into a binary one (i) overestimates nestedness by ≈ 0.02, and (ii) overestimates the number of links by 115. 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).

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). 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 the data of (**???**) using two "classical" null models of network structure. Robertson's data are amongst the hardest to analyse with the standard null models: the network is unusually large (1429 animals and 456 plants), and has a low connectance (0.02). Generating networks with all species is therefore both statistically difficult and computationally costly, providing a good demonstration of the performance of probabilistic metrics.

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.

Note that this type of 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, measure some of their properties, the distribution of which is compared to the empirical network.

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. Retrieved October 10, 2014,

Bascompte, J., Jordano, P., Melián, C.J. & Olesen, J.M. (2003). The nested assembly of plant–animal mutualistic networks. *Proceedings of the National Academy of Sciences of the United States of America*, **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. Retrieved October 10, 2014,

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.

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 plant–pollinator communities. *Oecologia*, 1–12. Retrieved September 11, 2014,

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 plant-animal mutualistic networks. *Ecology Letters*, **9**, 281–286.

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

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

Katz, L. (1953). A new status index derived from sociometric analysis. *Psychometrika*, **18**, 39–43. Retrieved October 9, 2014,

McCann, K.S. (2014). Diversity and Destructive Oscillations: Camerano, Elton, and May. *Bulletin of the Ecological Society of America*, **95**, 337–340. Retrieved October 7, 2014,

Milo, R., Kashtan, N., Itzkovitz, S., Newman, M.E.J. & Alon, U. (2003). On the uniform generation of random graphs with prescribed degree sequences. *arXiv:cond-mat/0312028*. Retrieved October 9, 2014,

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

Olesen, J.M., Bascompte, J., Dupont, Y.L., Elberling, H., Rasmussen, C. & Jordano, P. (2011). Missing and forbidden links in mutualistic networks. *Proceedings of the Royal Society B: Biological Sciences*, **278**, 725–732. Retrieved October 7, 2014,

Olito, C. & Fox, J.W. (2014). Species traits and abundances predict metrics of plant–pollinator network structure, but not pairwise interactions. *Oikos*, n/a–n/a. Retrieved September 10, 2014,

Poisot, T. & Gravel, D. (2014). When is an ecological network complex? Connectance drives degree distribution and emerging network properties. *PeerJ*, **2**, e251. Retrieved September 13, 2014,

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

Poisot, T., Stouffer, D.B. & Gravel, D. (2014). Beyond species: why ecological interaction networks vary through space and time. *Oikos*.

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. Retrieved October 10, 2014, Thébault, E. & Loreau, M. (2003). Food-web constraints on biodiversity–ecosystem functioning relationships. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 14949–14954.