Motif-based tests for bipartite networks

Sarah Ouadah, Pierre Latouche, Stephane S. Robin

Motif-based tests for bipartite networks

Sarah Ouadah[1,2], Pierre Latouche[2], Stéphane Robin[1,3]

([1]) UMR MIA-Paris, AgroParisTech, INRAE, Université Paris-Saclay, 75005 Paris, France

([2]) MAP5, UMR CNRS 8145, Université de Paris, 75006 Paris, France

([3]) CESCO, UMR 7204, MNHN - CNRS - UPMC, Paris, France

Abstract

Bipartite networks are a natural representation of the interactions between entities from two different types. The organization (or topology) of such networks gives insight to understand the systems they describe as a whole. Here, we rely on motifs which provide a meso-scale description of the topology. Moreover, we consider the bipartite expected degree distribution (B-EDD) model which accounts for both the density of the network and possible imbalances between the degrees of the nodes. Under the B-EDD model, we prove the asymptotic normality of the count of any given motif, considering sparsity conditions. We also provide close-form expressions for the mean and the variance of this count. This allows to avoid computationally prohibitive resampling procedures. Based on these results, we define a goodness-of-fit test for the B-EDD model and propose a family of tests for network comparisons. We assess the asymptotic normality of the test statistics and the power of the proposed tests on synthetic experiments and illustrate their use on ecological data sets.

Keywords: bipartite networks; network motifs; goodness-of-fit; network comparison; expected degree distribution

1 Introduction

Bipartite interaction networks are used to represent a diverse range of interactions in various fields such as biology, ecology, sociology or economics. For instance, in ecology, bipartite graphs depict interactions between two groups of species such as plants and pollinators [see e.g. Simmons et al., 2019b, Doré et al., 2020] or host and parasites [see e.g. Vacher et al., 2008, D'Bastiani et al., 2020], in agroethnology, they may involve interactions between farmers and crop species [see Thomas et al., 2015] and in economics, country-product trades as signals of the 2007-2008 financial crisis [see Saracco et al., 2016]. Formally, a bipartite interaction network can be viewed as a bipartite graph, the nodes of which being individuals pertaining to two different groups, and an edge between two nodes being present if these two individuals interact. In the sequel, the two types of nodes will be referred to as top nodes and bottom nodes, respectively. Characterizing the general organization of such a network, namely its topology, is key to understand the behavior of the system as a whole.

The topology of a network can be studied at various scales. Micro-scale analyses typically focus on the degree of each node, the betweenness of each edge or on the closeness between each pair of nodes. On the opposite, macro-scale analysis focus on global properties of the network such as its density or its modularity. The reader may refer to Newman

[2003] or Simmons et al. [2019b] for a general discussion. In this paper, we are mostly interested in the meso-scale description of the network that is provided by the frequency of motifs [Milo et al., 2002].
A motif is defined as a given subgraph depicting the interactions between a small number of nodes; the count of a motif consists in the number of occurrences of this subgraph in the observed network. Figures 7 and 8 display the set of all bipartite motifs involving up to 6 top or bottom nodes. Counting the occurrences of a motif is a computationally challenging task [see Milo et al., 2002, Picard et al., 2008, for simple– i.e. non-bipartite – networks]; efficient tools have been recently proposed by Simmons et al. [2019a,b] for bipartite networks.

Whatever the description scale, the analysis must account for a series of characteristics of the network at hand (such as its dimension or its density) to make the results comparable. A convenient way to account for such peculiarities is to define a null model capable to fit the network characteristics. We consider here a bipartite and exchangeable version of the expected degree distribution model proposed by Chung and Lu [2002] for simple binary graphs. The bipartite expected degree distribution (B-EDD) model simply states that each (top or bottom) node is associated with an expected degree and that a pair of nodes is connected with a probability that is proportional to the product of their respective expected degrees.
The B-EDD model can obviously accommodate to the network dimension (number of top and bottom nodes), for its density but also for some existing imbalances between the degrees of the nodes. Such imbalances play an important role in many fields: in ecology they are related to the opposition between generalist insects (capable of pollinate a large number of plant species) and specialist insects (interacting with a limited number of plant species) [Simmons et al., 2019b].
In addition to its interpretation, this model is attractive because we can calculate the expected frequency of motifs under B-EDD such as their variance.

The distribution of motif counts in simple graphs has been widely studied, especially for simple motifs like triangles [see e.g. Nowicki and Wierman, 1988, Stark, 2001, Picard et al., 2008]. In this paper, we prove the asymptotic normality of the count of any given motif under the B-EDD model, under sparsity conditions. One important feature of the B-EDD model is that the mean and the variance of the count have close form expressions. The strategy to derive these moments is related to the one introduced by Picard et al. [2008] for simple networks.
This property has a major practical impact as the expectation and the variance of a motif count could not be evaluated via resampling, because of the computational cost of motif counting event for networks with intermediate size. The knowledge of the asymptotic distribution of the motif counts opens a series of possible applications, including goodness-of-fit tests for the B-EDD model and a series of tests for network comparison in the B-EDD framework.

The paper is organized as follows. Section 2 is devoted to the definition and properties of motifs in the B-EDD model and Section 3 to tests for bipartite networks. More specifically, we establish the asymptotic normality of motif frequencies in Section 3.1 and propose a goodness-of-fit test for the B-EDD model and comparison tests for two bipartite networks in Section 3.2 and Section 3.3, respectively. The accuracy of the normal approximation for finite graphs and the power of the proposed tests are assessed via a simulation study in Section 4. Finally, proofs are given in Section 5.

2 Motifs in the bipartite expected degree model

We consider a bipartite graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with N nodes. The set of nodes is $\mathcal{V} = (\mathcal{V}^t, \mathcal{V}^b)$, where $\mathcal{V}^t = [\![1, m]\!]$ (resp. $\mathcal{V}^b = [\![1, n]\!]$) stands for the set of top (resp. bottom) nodes, and the set of edges is $\mathcal{E} \subset \mathcal{V}^t \times \mathcal{V}^b$, meaning than an edge can only connect a top node with a bottom node. The total number of nodes is therefore $N = n + m$. We denote by G the corresponding $m \times n$ incidence matrix where the entry G_{ij} of G is 1 if $(i, j) \in \mathcal{E}$, and 0 otherwise.

2.1 Bipartite expected degree model

The bipartite expected degree (B-EDD) model is defined as follows:

$$\begin{aligned} \{U_i\}_{1\leq i\leq m} \text{ iid}, \qquad & U_1 \sim \mathcal{U}_{[0,1]}, \\ \{V_j\}_{1\leq j\leq n} \text{ iid}, \qquad & V_1 \sim \mathcal{U}_{[0,1]}, \\ \{G_{ij}\}_{1\leq i\leq m, 1\leq j\leq n} \text{ indep. } | \{U_i\}_{1\leq i\leq m}, \{V_j\}_{1\leq j\leq n}, \qquad & G_{ij}|U_i, V_j \sim \mathcal{B}\left(\rho g(U_i)h(V_j)\right), \end{aligned} \tag{1}$$

where $g, h : [0, 1] \mapsto \mathbb{R}^+$, such that $\int g(u)\mathrm{d}u = \int h(v)\mathrm{d}v = 1$ and $1 \leq \rho \leq 1$.

The parameter ρ controls the density of the graph ($\mathbb{E}G_{ij} = \rho$) whereas the function g (resp. h) encodes the heterogeneity of the expected degrees of the top (resp. bottom) nodes. More specifically, denoting $K_i = \sum_{1\leq j\leq n} G_{ij}$ the degree of the top node i, we have that $\mathbb{E}(K_i \mid U_i) = n\rho g(U_i)$. The symmetric property holds for bottom nodes.

Remark 1. *Lovász and Szegedy [2006] and Diaconis and Janson [2008] introduced a generic model for exchangeable random graphs called the W-graph, which is based on a graphon function $\Phi : [0, 1]^2 \mapsto [0, 1]$. The B-EDD model is a natural extension of the W-graph for bipartite graphs with a product-form graphon function $\Phi(u, v) = \rho g(u)h(v)$. The B-EDD model is obviously exchangeable is the sens that the distribution of the incidence matrix G is preserved under permutation of the top nodes and/or the bottom nodes.*

Remark 2. *The B-EDD model can also be seen has an exchangeable bipartite version of the expected degree sequence model studied in Chung and Lu [2002] and of the configuration model from Newman [2003]. Under these two models, the degree of each node is fixed which makes them non exchangeable.*

2.2 Bipartite motifs in the B-EDD model

Bipartite motifs. We are interested in the distribution of the count of motifs (or subgraphs) in bipartite graphs arising from the B-EDD model. A bipartite motif s is defined by its number of top nodes p_s, its number of bottom nodes q_s and a $p_s \times q_s$ incidence matrix A^s. Figures 7 and 8 display the 44 bipartite motifs involving between two and six nodes, from which we see that

$$A^2 = \begin{pmatrix} 1 & 1 \end{pmatrix}, \qquad A^5 = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \qquad A^{15} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix}.$$

An important characteristic of a graph motif s is its number of automorphisms r_s [Stark, 2001], that is the number of non-redundant permutations of its incidence matrix (see, e.g. section 2.4 in Picard et al. [2008]):

$$r_s = \left| \left\{ A^s_{\sigma^t,\sigma^b} = \left(A^s_{\sigma^t(u),\sigma^b(v)} \right)_{1\leq u\leq p_s, 1\leq v\leq q_s} : \sigma^t \in \sigma\left([\![1, p_s]\!]\right), \sigma^b \in \sigma\left([\![1, q_s]\!]\right) \right\} \right|. \tag{2}$$

Note that, because pairs of permutations (σ^t, σ^b) yielding the same matrix $A^s_{\sigma^t,\sigma^b}$ are not counted twice, we obviously have that $r_s \leq (p_s!) \times (q_s!)$. In many cases, r_s turns out to be much smaller: in particular, $r_s = 1$ for star-motifs, which will be defined later. We further denote by d^s_u the degree of the top node u ($1 \leq u \leq p_s$) within motif s, that is $d^s_u = \sum_{1 \leq v \leq q_s} A^s_{u,v}$. The degree of the bottom node v within s is defined similarly as $e^s_v = \sum_{1 \leq u \leq p_s} A^s_{u,v}$.

Motif occurrence. Counting the occurrences of motif s in $\mathcal{G}$ simply consists in considering all possible of p_s (resp. q_s) top (resp. bottom) nodes among the m (resp. n) and check for each possible automorphism of s if an occurrence is observed. More formally, let us define the set $\mathcal{P}_s$ of possible positions for motif s as the Cartesian product of the set of the $\binom{m}{p_s}\binom{n}{q_s}$ possible locations with the set of the r_s (top, bottom) permutations giving rise to each of the automorphisms of s. So, a *position* results from the combination of a *location* with a *permutation*. Because the graph is bipartite, any position α from $\mathcal{P}_s$ decomposes as $\alpha = (\alpha^t, \alpha^b)$ where α^t stands for an ordered list of top nodes and α^b for an ordered list of bottom nodes. The number of positions for motif s in $\mathcal{G}$ is precisely

$$c_s := |\mathcal{P}_s| = r_s \binom{m}{p_s}\binom{n}{q_s}. \tag{3}$$

Now, for a given position $\alpha = (\alpha^t, \alpha^b) \in \mathcal{P}_s$, we define $Y_s(\alpha)$ as the indicator for motif s to occur in position α:

$$Y_s(\alpha) = \prod_{i \in \alpha^t, j \in \alpha^b} G_{ij}^{A^s_{ij}}. \tag{4}$$

Remark 3. *Note that the occurrence defined by Equation* (4) *corresponds to an* induced *occurrence, which means that we consider that a motif s is observed at position α as soon as all the present edges that are specified by its incidence matrix A^s are observed, even if additional edges are also observed. In other words, we do not check for the absent edges specified by A^s.*

Remark 4. *As opposed to an induced occurrence, an* exact *occurrence is observed when both the presence and the absence of edges are satisfied. The indicator variable corresponding to an exact occurrence writes $\prod_{i \in \alpha^t, j \in \alpha^b} G_{ij}^{A^s_{ij}} (1 - G_{ij})^{1 - A^s_{ij}}$. Counting induced and exact occurrences in a graph is actually equivalent, as these counts are related in a deterministic manner. For example, each exact occurrence of motif 6 corresponds to two induced occurrences of motif 5.*

Motif probability. The B-EDD model is an exchangeable bipartite graph model in the sense that, for any pair of permutations $(\sigma^t \in \sigma\left(\llbracket 1, m \rrbracket\right), \sigma^b \in \sigma\left(\llbracket 1, n \rrbracket\right))$, we have that $\mathbb{P}\{G = \{g_{ij}\}_{1 \leq i \leq m, 1 \leq j \leq n}\} = \mathbb{P}\{G = \{g_{\sigma^t(i)\sigma^b(j)}\}_{1 \leq \sigma^t(i) \leq m, 1 \leq \sigma^b(j) \leq n}\}$ [see e.g. Lovász and Szegedy, 2006, Diaconis and Janson, 2008, for simple graphs]. For any exchangeable graph model, we may define ϕ_s as the probability for motif s to occur in position $\alpha = (\alpha^t, \alpha^b)$:

$$\phi_s := \mathbb{P}\left(Y_s(\alpha) = 1\right).$$

Importantly, because the model is exchangeable, this probability does not depend on α.

Star motifs. We define a star as a bipartite motif s for which either $q_s = 1$ or $p_s = 1$ (or both). More specifically, we name top stars (resp bottom stars) motifs for which $p_s = 1$ (resp. $q_s = 1$). The top stars in Figures 7 and 8 are motifs 1, 2, 7, 17 and 44, and the bottom stars are motifs 1, 3, 4, 8 and 18. Observe that $r_s = 1$ for all star motifs, that $d_v^s = 1$ for all v in all top star motifs, and that $e_u^s = 1$ for all u in all bottom star motifs.
Because they will play a central role in the sequel, we adopt a specific notation for the probability of star motifs, denoting γ_d the occurrence probability of the top star with degree d and λ_d for the occurrence probability of the bottom star with degree d. As a consequence, we have that

$$\begin{array}{lllll} \gamma_1 = \phi_1, & \gamma_2 = \phi_2, & \gamma_3 = \phi_7 & \gamma_4 = \phi_{17}, & \gamma_5 = \phi_{44}, \\ \lambda_1 = \phi_1, & \lambda_2 = \phi_3, & \lambda_3 = \phi_4 & \lambda_4 = \phi_8, & \lambda_5 = \phi_{18}. \end{array} \tag{5}$$

2.3 Moments of motif counts

Expected count. Let us now denote by N_s the count, that is the number of occurrences of a motif s in a graph $\mathcal{G}$. We simply have that

$$N_s = \sum_{\alpha \in \mathcal{P}_s} Y_s(\alpha)$$

As a consequence, the expected count of s in $\mathcal{G}$ is $\mathbb{E}(N_s) = c_s \phi_s$. We also define the normalized frequency of motif s as

$$F_s = N_s / c_s,$$

which is an unbiased estimate of ϕ_s.

Illustration. As an illustration, we consider two of the networks studied by Simmons et al. [2019a], which include both plant-pollinator and seed dispersal networks extracted from the Web of Life database (`www.web-of-life.es`). More specifically, we consider the two largest networks of each type, which were first published by Robertson [1929] and Silva [2002], respectively. The plant-pollinator network involves 546 plant species and 1044 insects and the seed dispersal network 207 plant species and 110 seed dispersers (birds or insects). Table 1 gives the counts and the frequency of the star motifs with up to four branch. For the sake of clarity, we will limit ourselves to motifs up to five nodes in the illustrations. Observe that both the counts N_s and the number of possible positions c_s range over huge order of magnitudes.

Main property of motif probabilities under B-EDD. The tests we propose rely on the comparison between the observed count (or normalized frequency) of a motif, with its theoretical counterpart under a B-EDD model. More specifically, the motif probabilities have a close form expression under the B-EDD model.

Proposition 1. *Under the B-EDD model* (1)*, we have that*

$$\overline{\phi}_s = \prod_{u=1}^{p_s} \gamma_{d_u^s} \prod_{v=1}^{q_s} \lambda_{e_v^s} \Big/ (\phi_1)^{d_+^s} \, . \tag{6}$$

where $d_+^s := \sum_u d_u^s = \sum_v e_v^s$ *stands for the total number of edges in* s*.*

plant-pollinator: $m = 546, n = 1044$ [Robertson, 1929]

	edge	top stars			bottom stars		
s	1	2	7	17	3	4	8
c_s	$4.76\ 10^5$	$2.48\ 10^8$	$8.62\ 10^{10}$	$2.24\ 10^{13}$	$1.08\ 10^8$	$1.64\ 10^{10}$	$1.86\ 10^{12}$
N_s	$1.53\ 10^4$	$2.61\ 10^5$	$3.04\ 10^6$	$2.72\ 10^7$	$3.07\ 10^5$	$6.82\ 10^6$	$1.48\ 10^8$
F_s	$3.20\ 10^{-2}$	$1.05\ 10^{-3}$	$3.52\ 10^{-5}$	$1.21\ 10^{-6}$	$2.84\ 10^{-3}$	$4.16\ 10^{-4}$	$7.99\ 10^{-5}$

seed dispersal: $m = 207, n = 110$ [Silva, 2002]

	edge	top stars			bottom stars		
s	1	2	7	17	3	4	8
c_s	$2.28\ 10^4$	$1.24\ 10^6$	$4.47\ 10^7$	$1.20\ 10^9$	$2.35\ 10^6$	$1.60\ 10^8$	$8.17\ 10^9$
N_s	$1.12\ 10^3$	$6.50\ 10^3$	$4.07\ 10^4$	$2.32\ 10^5$	$1.24\ 10^4$	$1.31\ 10^5$	$1.23\ 10^6$
F_s	$4.92\ 10^{-2}$	$5.23\ 10^{-3}$	$9.11\ 10^{-4}$	$1.94\ 10^{-4}$	$5.28\ 10^{-3}$	$8.16\ 10^{-4}$	$1.50\ 10^{-4}$

Table 1: Coefficients c_s, counts N_s and frequency F_s of all star motifs. Top: plant-pollinator network, bottom: seed dispersal network. The motif number s refers to Figure 7.

Proof. This follows from the fact that, under B-EDD, the edges are independent conditionally on the latent coordinates U_i and V_j defined in (1), which are all independent with respect to one other. Consider an arbitrary position $\alpha = (\alpha^t, \alpha^b)$; for the sake of clarity, we identify the elements of α^t with $[\![1, p_s]\!]$ and the elements of α^b with $[\![1, q_s]\!]$. We have

$$\begin{aligned}
\overline{\phi}_s &= \mathbb{E}_{(U_i)_{1\le i\le p_s},(V_j)_{1\le j\le q_s}}\left(\mathbb{P}\left\{\prod_{1\le i\le p_s,1\le v\le q_s} G_{ij}^{A^s_{ij}} = 1 \,\middle|\, (U_i)_{1\le i\le p_s},(V_j)_{1\le j\le q_s}\right\}\right)\\
&= \mathbb{E}_{(U_i)_{1\le i\le p_s},(V_j)_{1\le j\le q_s}}\left(\prod_{1\le i\le p_s,1\le j\le q_s\,:\,A^s_{ij}=1} \rho g(U_i)h(V_j)\right)\\
&= \mathbb{E}_{(U_i)_{1\le i\le p_s},(V_j)_{1\le j\le q_s}}\left(\rho^{d^s_+}\prod_{1\le i\le p_s} g(U_i)^{d^s_i}\prod_{1\le j\le q_s} h(V_i)^{e^s_j}\right)\\
&= \rho^{d^s_+}\prod_{1\le i\le p_s}\left(\int g(u)^{d^s_i}\mathrm{d}u\right)\prod_{1\le j\le q_s}\left(\int h(v)^{e^s_j}\mathrm{d}v\right).
\end{aligned}$$

The result then results from the fact that

$$\gamma_d = \rho^d\int g(u)^d\mathrm{d}u, \qquad \lambda_d = \rho^d\int h(v)^d\mathrm{d}v, \qquad \rho = \phi_1. \tag{7}$$

■

An important consequence of Proposition 1 is that, under B-EDD, the motif probability of any motif can be expressed in terms of probabilities of star motifs. Figure 1 provides an intuition of this: a motif can be decomposed in terms of top and bottom stars arising from each of its nodes.

In the sequel, to distinguish the motif probability ϕ_s under an arbitrary exchangeable model from the probability under the B-EDD model, we will denote by $\overline{\phi}_s$ the probability of motif s under B-EDD. Figure 7 provides the list of all $\overline{\phi}_s$ expressions.

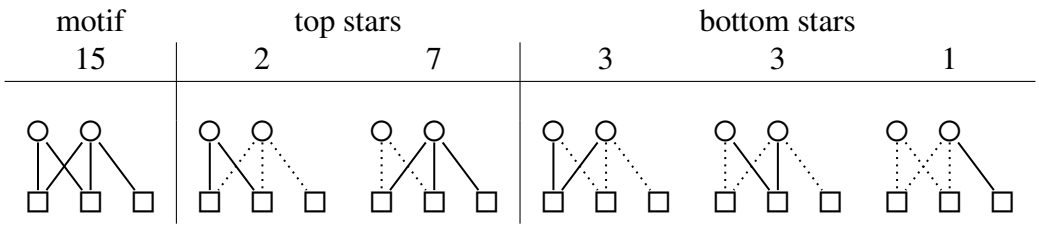


Figure 1: Decomposition of motif 15 as an overlap of 2 top stars (motifs 2 and 7) and 3 bottom stars (motifs 3, 3 and 1). Because each edge is accounted for twice, we get $\overline{\phi}_{15} = \phi_2\phi_7\phi_3\phi_3\phi_1/\phi_1^5 = \phi_2\phi_7\phi_3^2/\phi_1^4$.

Probability estimate under B-EDD. Proposition 1 suggests a natural plug-in estimator for the B-EDD motif probability $\overline{\phi}_s$:

$$\overline{F}_s = \frac{\prod_{u=1}^{p_s} \Gamma_{d_u^s} \prod_{v=1}^{q_s} \Lambda_{e_v^s}}{F_1^{d_+^s}}, \tag{8}$$

where Γ_d (resp Λ_d) denotes the normalized frequency of the top (resp. bottom) star motif with degree d. Obviously, Γ_d (resp Λ_d) is an unbiased estimated of γ_d (resp. λ_d).

Variance of the count. We now consider the variance of the count, that is

$$\begin{aligned} \mathbb{V}(N_s) &= \mathbb{E}(N_s^2) - \mathbb{E}(N_s)^2, \\ \text{where} \qquad N_s^2 &= \sum_{\alpha,\beta \in \mathcal{P}_s} Y_s(\alpha) Y_s(\beta) \\ &= \sum_{\alpha \in \mathcal{P}_s} Y_s(\alpha) + \sum_{\alpha,\beta \in \mathcal{P}_s : |\alpha \cap \beta| = 0} Y_s(\alpha) Y_s(\beta) + \sum_{\alpha,\beta \in \mathcal{P}_s : \alpha \neq \beta, |\alpha \cap \beta| > 0} Y_s(\alpha) Y_s(\beta). \end{aligned} \tag{9}$$

When positions α and β are equal, the product $Y_s(\alpha)Y_s(\beta)$ is simply given by $Y_s(\alpha)$, the indicator of the presence of s at position α. Then, when positions α and β do not overlap ($|\alpha \cap \beta| = 0$), the product $Y_s(\alpha)Y_s(\beta)$ simply indicates that two occurrences of motif s occur in position α and β, which are independent under the B-EDD model. When positions α and β are different and do overlap ($|\alpha \cap \beta| > 0$), the product $Y_s(\alpha)Y_s(\beta)$ becomes the indicator of a super-motif, that is a motif made of two overlapping automorphisms of s. We denote by $\mathcal{S}_2(s)$ the set of super-motifs generated by the overlaps of two occurrences of the motif s; Figure 2 provides some examples of super-motifs.

An expression similar to (9) can be derived for the covariance between two counts:

$$\begin{aligned} \mathbb{C}\text{ov}(N_s, N_t) &= \mathbb{E}(N_s N_t) - \mathbb{E}(N_s)\mathbb{E}(N_t), \\ \text{where} \qquad N_s N_t &= \sum_{\alpha \in \mathcal{P}_s, \beta \in \mathcal{P}_t} Y_s(\alpha) Y_t(\beta) \\ &= \sum_{\alpha \in \mathcal{P}_s, \beta \in \mathcal{P}_t : |\alpha \cap \beta| = 0} Y_s(\alpha) Y_t(\beta) + \sum_{\alpha \in \mathcal{P}_s, \beta \in \mathcal{P}_t : \alpha \neq \beta, |\alpha \cap \beta| > 0} Y_s(\alpha) Y_t(\beta). \end{aligned} \tag{10}$$

Again, the last term corresponds to occurrences of super-motifs resulting from an overlap between an occurrence of motif s and an occurrence of motif t. We denote by $\mathcal{S}_2(s,t)$ the set of these super-motifs. We use the strategy described in Picard et al. [2008] to determine the sets of super-motifs $\mathcal{S}_2(s)$ and $\mathcal{S}(s,s')$. Observe that these sets do not depend on the

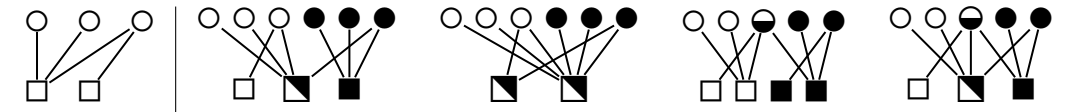
observed networks, so, to alleviate the computational burden, they can be determined and stored once for all.

Eq. (9) shows that $\mathbb{E}(N_s^2)$ only depends on $\mathbb{E}(Y_s(\alpha)Y_s(\beta))$, which is ϕ_s^2 when positions α and β do not overlap and the probability of the corresponding super-motif when they overlap. As a consequence, we have that

$$\mathbb{E}(N_s^2) = \kappa_{m,n,s}\phi_s + \kappa'_{m,n,s}\phi_s^2 + \sum_{S \in \mathcal{S}_2(s)} \kappa''_{m,n,s,S}\phi_S, \tag{11}$$

where the $\kappa_{m,n,s}$, $\kappa'_{m,n,s}$, $\kappa''_{m,n,s,S}$ are constants, which depend on the dimensions of the graph, on the motif s and on the super-motif S. The order of magnitude of $\kappa_{m,n,s,S}$ for large m and n will be studied in Section 5.1.2.

Because super-motifs are actually motifs, their respective occurrence probability $\overline{\phi}_S$ under B-EDD are given by Proposition 1 as well, so the expectation and the variance of N_s under B-EDD can be expressed as functions of the $\overline{\phi}_s$ and $\{\overline{\phi}_S\}_{S \in \mathcal{S}_2(s)}$. An estimate $\overline{F}_S$ of each $\overline{\phi}_S$ can be obtained using Eq. (8) in the same way.



$\lvert\alpha^t \cap \beta^t\rvert$	0	0	1	1
$\lvert\alpha^b \cap \beta^b\rvert$	1	2	0	1

Figure 2: Some super-motifs from $\mathcal{S}_2(s)$ for motif $s = 9$ (top left) with $p_s = 3$ top nodes and $q_s = 2$ bottom nodes. $|\alpha^t \cap \beta^t|$ (resp. $|\alpha^b \cap \beta^b|$): number of top (resp. bottom) nodes shared by the overlapping positions α and β. Black: nodes from α, white: nodes from β, black/white: nodes from $\alpha \cap \beta$. There are actually $|\mathcal{S}_2(9)| = 396$ such super-motifs of motif 9.

Remark 5. *The estimate defined in* (8) *is only based on empirical quantities (the counts of stars motifs) and does not depend on any parameter estimation. Especially, the functions g and h do not need to be estimated as the frequency of star motifs provides all necessary information about the degree distributions. As a consequence, we may define plug-in estimates of the occurrence probability, the expected count and the variance of the count of any motif under B-EDD.*

Illustration. Table 2 compares the empirical frequencies F_s of a selection of motifs with their respective estimated probability $\overline{F}_s$. The probability estimates are computed according to Equation 8, using the star motifs frequencies Γ_d and Λ_e given in Table 1. Observe that the difference between the observed frequency F_s and their estimated expectation under the B-EDD model $\overline{F}_s$ are of the same order of magnitude, if not smaller, than their estimated standard deviations.

3 Tests for bipartite networks

Asymptotic framework. We consider a sequence of B-EDD random graphs defined as follows.

$\{\mathcal{G}_N\}_{N \geq 2}$ is a sequence of independent graphs, where $\mathcal{G}_N$ is a B-EDD random graph with $m = \lfloor \lambda N \rfloor$ top nodes with $\lambda \in (0, 1)$, $n = N - m$ bottom nodes and parameters ρ_N, h

plant-pollinator

s	5	6	10	15	16
F_s	$9.21\ 10^{-5}$	$1.00\ 10^{-5}$	$8.12\ 10^{-6}$	$3.32\ 10^{-7}$	$4.47\ 10^{-8}$
$\overline{F}_s$	$9.29\ 10^{-5}$	$8.41\ 10^{-6}$	$8.23\ 10^{-6}$	$2.82\ 10^{-7}$	$2.62\ 10^{-8}$
$\sqrt{\widehat{\mathbb{V}}(F_s)}$	$1.26\ 10^{-5}$	$1.61\ 10^{-6}$	$1.58\ 10^{-6}$	$6.54\ 10^{-8}$	$7.60\ 10^{-9}$

seed dispersal

s	5	6	10	15	16
F_s	$5.13\ 10^{-4}$	$1.15\ 10^{-4}$	$5.07\ 10^{-5}$	$1.79\ 10^{-5}$	$5.96\ 10^{-6}$
$\overline{F}_s$	$5.61\ 10^{-4}$	$1.30\ 10^{-4}$	$6.02\ 10^{-5}$	$2.26\ 10^{-5}$	$8.59\ 10^{-6}$
$\sqrt{\widehat{\mathbb{V}}(F_s)}$	$2.25\ 10^{-4}$	$7.24\ 10^{-5}$	$3.25\ 10^{-5}$	$1.59\ 10^{-5}$	$7.38\ 10^{-6}$

Table 2: Empirical frequency F_s, estimated probability $\overline{F}_s$ and estimated standard-deviation of the frequency according to the B-EDD model for a selection of motifs. All estimates are derived from the star motifs frequencies given in Table 1.

and g, where the sequence $\{\rho_N\}_{N\geq 2}$ satisfies $\rho_N = \Theta(m^{-a}n^{-b})$ with $a, b > 0$. All quantities computed on $\mathcal{G}_N$ should be indexed by N as well but for the sake of clarity, we will drop that index in the rest of the paper.

3.1 Asymptotic normality of motif frequencies

This section is devoted to the asymptotic normality of motif frequencies under the B-EDD model. More precisely, our first main result states the asymptotic normality of the following statistic W_s relying on F_s the empirical frequency of a given motif s in $\mathcal{G}$:

$$W_s = \frac{F_s - \overline{F}_s}{\sqrt{\widehat{\mathbb{V}}(F_s)}}, \tag{12}$$

where $\overline{F}_s$ denotes the estimator of $\overline{\phi}_s$ defined in (8) and $\widehat{\mathbb{V}}(F_s)$ the one of $\mathbb{V}(F_s)$ obtained by the plug-in of $\overline{F}_S$ (S being any super-motif generated by two occurrences of s) in the expressions of $\mathbb{V}(N_s)$ given in (9)-(11).

Theorem 1. *If $a + b < 2/d_+^s$, then for all non-star motif s and under the B-EDD model, the statistic W_s is asymptotically normal as $m \sim n \to \infty$:*

$$W_s \xrightarrow{D} \mathcal{N}(0, 1).$$

The proof is based on three results given hereafter in Proposition 2, Lemma 1 and Lemma 2.

Sketch of proof. Let first consider the following decomposition of the numerator of W_s :

$$F_s - \overline{F}_s := L_s + C_s \qquad \text{where } L_s = F_s - \overline{\phi}_s \text{ and } C_s = \overline{\phi}_s - \overline{F}_s.$$

Under the null B-EDD model, we show that, (i) $L_s/\sqrt{\mathbb{V}(F_s)}$ is asymptotically normal in Proposition 2, it is the leader term, (ii) $C_s/\sqrt{\mathbb{V}(F_s)}$ is negligible in Lemma 1, it is the

reminder term. Then, we conclude using Slutsky Theorem Lemma 2 which states that $\widehat{\mathbb{V}}(F_s)/\mathbb{V}(F_s) \to 1$ in probability. ■

Remark 6. *Like $\overline{F}_s$, W_s is only based on empirical quantities, that is i) the empirical frequency of motif s and ii) the empirical frequencies of the stars motifs forming s. The expected frequencies of the supermotifs of s involved in $\widehat{\mathbb{V}}(F_s)$ also depend only on empirical star frequencies.*

Remark 7. *Gao and Lafferty [2017] proved a similar result as Theorem 1 in the EDD model, for a test statistic which is a linear combination of edges, vees and triangles empirical frequencies in the case of simple graphs, and under a specific condition on the graph density. Though their result is not comparable to ours since triangles can not occur in bipartite graphs and we do not account for stars motifs. Although they seem similar, a fair comparison between Theorem 1 and the result from Gao and Lafferty [2017] is not easy (i) because the model is not the same (we consider bipartite graphs whereas they consider simple graphs) and (ii) because they only consider vees (which are star-motifs) and triangles (which do not occur in bipartite graphs).*

In the following proposition, the asymptotic normality of the statistic ruling the law of W_s is stated under the null. This statistic involves the empirical frequency of a given non star motif s and its theoretical expectation and variance. The proof of its asymptotic normality mostly relies on tools of martingale theory. We show that we can exhibit conditional martingale difference sequences relative to a specific filtration. This filtration is generated by the sequence of graphs $\mathcal{G}_N$ (see a proper definition of the filtration in Section 5.1.1). So, we could apply the central limit theorem of Hall and Heyde [2014].

Proposition 2. *If $a + b < 2/d_+^s$, then for all star motif s and under the B-EDD model, we have, as $m \sim n \to \infty$,*

$$\frac{F_s - \overline{\phi}_s}{\sqrt{\mathbb{V}(F_s)}} \xrightarrow{D} \mathcal{N}(0,1).$$

The complete proof is given in Section 5.2, it relies especially on Lemma 6 and Lemma 7.

Sketch of proof. We first consider the decomposition $L_s = F_s - \overline{\phi}_s = M_s + R_s$ with M_s being the difference between F_s and its expectation conditionally to the considered filtration and U, V, and R_s the difference between the latter conditional expectation and $\overline{\phi}_s$; the proper definitions are given in Section 5.2.1. Lemma 6 shows that, under the null B-EDD model, the reminder term $R_s/\sqrt{\mathbb{V}(F_s)}|U,V \to 0$ a.s. as $m \sim n \to \infty$. Lemma 7 shows that, under the B-EDD model, the leader term $M_s/\sqrt{\mathbb{V}(F_s)}|U,V$ is asymptotically normal with variance $\mathbb{V}(N_s|U,V)/\mathbb{V}(N_s)$. Slutsky theorem implies the asymptotic normality of $L_s/\sqrt{\mathbb{V}(F_s)}$ conditional on (U,V). Then, Lemma 4 shows that $\mathbb{V}(N_s|U,V)/\mathbb{V}(N_s)$ tends to 1 in probability for all (U,V), which allows deconditionning. ■

The two following lemmas combined with Proposition 2 permit to conclude to Theorem 1. Their proofs are given in sections 5.3 and 5.4 respectively.

Lemma 1. *If $a + b < 2/d_+^s$, then for all non-star motif s and under the B-EDD model, we have, as $m \sim n \to \infty$,*

$$\frac{\overline{F}_s - \overline{\phi}_s}{\sqrt{\mathbb{V}(F_s)}} \to 0 \text{ a.s.}$$

Lemma 2. *If $a + b < 2/d_+^s$, then for all star motif s and under the B-EDD model, we have, as $m \sim n \to \infty$,*

$$\widehat{\mathbb{V}}(F_s)/\mathbb{V}(F_s) \to 1 \text{ a.s.}$$

3.2 Goodness-of-fit tests for the B-EDD model

We consider a bipartite network $\mathcal{G}$ and we want to test if it arises from the B-EDD model:

$$\begin{cases} H_0 : \mathcal{G} \text{ follows a B-EDD model,} \\ H_1 : \mathcal{G} \text{ does not follow a B-EDD model.} \end{cases}$$

To this aim, we consider the test statistic $W_s = (F_s - \overline{F}_s)/\sqrt{\widehat{\mathbb{V}}(F_s)}$ defined in (12). The idea is thus to compare the frequency of a motif observed in the network with its expected value under the B-EDD model.

Remark 8. *We can consider more specific hypothesis. Suppose we want to test the top node heterogeneity under B-EDD, more specifically H_0 : $\mathcal{G}$ follows a B-EDD model and g is constant. Then, according to* (7)*, we have that $\gamma_d = \rho^d$ under H_0, so a similar statistic to W_s can be designed by considering $\overline{F}_s = \prod_{u=1}^{p_s} F_1^{d_u^s} \prod_{v=1}^{q_s} \Lambda_{e_v^s}/F_1^{d_+^s}$. In the same manner, a statistic can be designed to test the bottom node heterogeneity.*

Illustration. Table 3 gives the test statistics W_s for goodness of fit to the B-EDD model for the same motifs as in Table 2. According to Theorem 1, these statistics should be compared with the quantiles of standard normal distribution $\mathcal{N}(0, 1)$. Almost no motif frequency displays a significant deviation from its expectation under the B-EDD model. Only motif 16 in the plant-pollinator network displays a higher frequency than expected under B-EDD (with p-value 7.5 10^{-3}).

plant-pollinator

s	5	6	10	15	16
W_s	-6.45 10^{-2}	9.96 10^{-1}	-6.63 10^{-2}	7.52 10^{-1}	2.43

seed dispersal

s	5	6	10	15	16
W_s	-2.14 10^{-1}	-2.14 10^{-1}	-2.93 10^{-1}	-2.95 10^{-1}	-3.56 10^{-1}

Table 3: Test statistics W_s for the goodness-of-fit of B-EDD for the same motifs as in Table 2.

3.3 Tests for the comparison of two bipartite networks

This section is devoted to network comparison test. More specifically, considering two networks assumed to arise from two B-EDD models, we want to test if they arise from the same B-EDD model, or for, instance, from two different B-EDD model with same function g. The rational behind the tests we propose is to compare the frequency of a motif observed in one network with its expected value according to the parameters of the other network. To this aim, we need to introduce specific notations.

Notations. The B-EDD model is parametrized with the (m, n, ρ, g, h) but all moments depend on $(m, n, \rho, \gamma, \lambda)$, where γ (resp. λ) stands for the sequence of occurrence probability of all the top (resp. bottom) star motifs. In the sequel we denote by E_s the expected frequency of motif s:

$$E_s(m, n, \rho, \gamma, \lambda) := \phi_s,$$

so its plug-in estimate is $E_s(m, n, F_1, \Gamma, \Lambda) = \overline{F}_s$. Similarly, we denote the variance of the frequency by $V_s(m, n, \rho, \gamma, \lambda) := \mathbb{V}(F_s)$ and its plug-in estimate $V_s(m, n, F_1, \Gamma, \Lambda) := \widehat{\mathbb{V}}_s(F_s)$.

A global test. We consider two bipartite networks $\mathcal{G}^A$ and $\mathcal{G}^B$ supposed to arise from B-EDD models with respective dimensions and parameters $(m^A, n^A, \rho^A, \gamma^A, \lambda^A)$ and $(m^B, n^B, \rho^B, \gamma^B, \lambda^B)$. We want to test

$$\begin{cases} H_0 : \left\{(\rho^A, g^A, h^A) = (\rho^B, g^B, h^B)\right\}, \\ H_1 : \left\{\rho^A \neq \rho^B \text{ or } g^A \neq g^B \text{ or } h^A \neq h^B\right\}. \end{cases}$$

This is to test that, although the two networks may have different dimensions (m, n), they have the same density (ρ), the same top node heterogeneity (g) and the same bottom node heterogeneity (h).

Test statistics. The test statistic is based on F_s^A and F_s^B the empirical frequencies of motif s in $\mathcal{G}^A$ and $\mathcal{G}^B$ respectively. The superscript A (resp. B) is added to all quantities observed in $\mathcal{G}^A$ (resp. $\mathcal{G}^B$).

$$W_s = \frac{\left(F_s^A - E_s(m^A, n^A, F_1^B, \Gamma^B, \Lambda^B)\right) - \left(F_s^B - E_s(m^B, n^B, F_1^A, \Gamma^A, \Lambda^A)\right)}{\sqrt{V_s(m^A, n^A, F_1^B, \Gamma^B, \Lambda^B) + V_s(m^B, n^B, F_1^A, \Gamma^A, \Lambda^A)}}. \tag{13}$$

Theorem 2. *If both m^A/m^B and n^A/n^B tends to constants, if $a + b < 2/d_+^s$, then for all non-star motif s and under H_0, the statistic W_s is asymptotically normal as $m^A \sim n^A \sim m^B \sim n^B \to \infty$:*

$$W_s \xrightarrow{D} \mathcal{N}(0, 1).$$

Proof. We decompose

$$\begin{aligned} F_s^A - E_s(m^A, n^A, F_1^B, \Gamma^B, \Lambda^B) &= L_s^A + C_s^A \\ \text{where} \quad L_s^A &= F_s^A - E_s(m^A, n^A, \phi_1^B, \gamma^B, \lambda^B) \\ \text{and} \quad C_s^A &= E_s(m^A, n^A, \phi_1^B, \gamma^B, \lambda^B) - E_s(m^A, n^A, F_1^B, \Gamma^B, \Lambda^B). \end{aligned}$$

Because (m^A, n^A) go to infinity at the same speed as (m^B, n^B), under H_0, $L_s^A/V_s(m^A, n^A, \phi_1^B, \gamma^B, \lambda^B)$ is asymptotically normal according to Proposition 2, whereas $C_s^A/V_s(m^A, n^A, \phi_1^B, \gamma^B, \lambda^B)$ tends to zero according to Lemma 1. Using the same arguments for the symmetric term, we get that and the negligible one $\left(C_s^A/V_s(m^A, n^A, \phi_1^B, \gamma^B, \lambda^B), C_s^B/V_s(m^A, n^A, \phi_1^A, \gamma^A, \lambda^A)\right)$, replacing $V_s(m^A, n^A, \phi_1^B, \gamma^B, \lambda^B)$ and $V_s(m^B, n^B, \phi_1^A, \gamma^A, \lambda^A)$ with their plug-in estimate $V_s(m^A, n^A, F_1^B, \Gamma^B, \Lambda^B)$ and $V_s(m^B, n^B, F_1^A, \Gamma^A, \Lambda^A)$. We conclude using Lemma 2 and Slutsky Theorem. ■

Testing equal top nodes heterogeneity. Suppose we want to test that, although the two networks may have different dimensions, different densities, and different bottom node heterogeneity, they have the same top node heterogeneity, that is

$$\begin{cases} H_0 : \left\{g^A = g^B\right\}, \\ H_1 : \left\{g^A \neq g^B\right\}. \end{cases}$$

Since we allow the two networks to have different densities, one might normalize the probabilities of star motifs given in (5) as follows:

$$\begin{array}{lllll} \widetilde{\gamma}_1 = 1, & \widetilde{\gamma}_2 = \phi_2/\phi_1^2 & \widetilde{\gamma}_3 = \phi_7/\phi_1^3 & \widetilde{\gamma}_4 = \phi_{17}/\phi_1^4, & \widetilde{\gamma}_5 = \phi_{44}/\phi_1^4, \\ \widetilde{\lambda}_1 = 1, & \widetilde{\lambda}_2 = \phi_3/\phi_1^2, & \widetilde{\lambda}_3 = \phi_4/\phi_1^3 & \widetilde{\lambda}_4 = \phi_8/\phi_1^4, & \widetilde{\lambda}_5 = \phi_{18}/\phi_1^4. \end{array}$$

This allows to see that we can rewrite $E_s(m, n, \rho, \gamma, \lambda) = \phi_s$ as an expression of g on which relies the test we consider. According to (6) and to the definition of ϕ_s under the B-EDD model, we get:

$$E_s(m, n, \rho, g, h) = \rho^{d_+^s} \prod_{u=1}^{p_s} \tilde{\gamma}_{d_u^s} \prod_{v=1}^{q_s} \tilde{\lambda}_{e_v^s} = \rho^{d_+^s} \prod_{u=1}^{p_s} \prod_{v=1}^{q_s} g_{d_u^s} h_{e_v^s},$$

where $g_d = \int g(u)^d \mathrm{d}u$ and $h_e = \int h(v)^e \mathrm{d}v$. We may consider the following test statistic:

$$W_s^g = \frac{\left(F_s^A - E_s(m^A, n^A, F_1^A, \widetilde{\Gamma}^B, \widetilde{\Lambda}^A)\right) - \left(F_s^B - E_s(m^B, n^B, F_1^B, \widetilde{\Gamma}^A, \widetilde{\Lambda}^B)\right)}{\sqrt{V_s(m^A, n^A, F_1^A, \widetilde{\Gamma}^B, \widetilde{\Lambda}^A) + V_s(m^B, n^B, F_1^B, \widetilde{\Gamma}^A, \widetilde{\Lambda}^B)}},$$

where $\widetilde{\Gamma}$ and $\widetilde{\Lambda}$ are the plug-in estimates of $\widetilde{\gamma}$ and $\widetilde{\lambda}$ respectively. Similar statistics can be designed to test $\rho^A = \rho^B$, $h^A = h^B$ or any combination.

Illustration. Both the plant-pollinator and the seed dispersal networks involve plants species. Although these species are not the same, one may be interested in comparing if the level of heterogeneity across plants (encoded in the function g) is the same in both networks. From an ecological point of view, this amounts to test if there is the same the degree of imbalance between specialists and generalists among plants regarding pollination and seed dispersion, that are two of the main reproduction means.
Table 4 provides the results of the network comparison test presented above. No significant difference is observed, suggesting that, although generalist and specialist plants may exist for both types of interactions, the degree of imbalance between them is comparable.

4 Simulation study

We designed a simulation study to illustrate Theorem 1 and to assess the performance of the goodness-of-fit test and the comparison test described in Section 3.2 and Section 3.3 respectively. More specifically, our purpose is to illustrate the asymptotic normality of the test statistics and evaluate the power of the tests for various graph sizes, densities and sparsity regimes.

s	5	6	10	15	16
F_s^A	9.21 10^{-5}	1.00 10^{-5}	8.12 10^{-6}	3.32 10^{-7}	4.47 10^{-8}
$\widehat{\mathbb{E}}_0 F_s^A$	1.96 10^{-4}	3.75 10^{-5}	1.74 10^{-5}	4.25 10^{-6}	1.33 10^{-6}
F_s^B	5.13 10^{-4}	1.15 10^{-4}	5.07 10^{-5}	1.79 10^{-5}	5.96 10^{-6}
$\widehat{\mathbb{E}}_0 F_s^B$	2.66 10^{-4}	2.92 10^{-5}	2.85 10^{-5}	1.50 10^{-6}	1.69 10^{-7}
$F_s^B - F_s^A$	-4.21 10^{-4}	-1.05 10^{-4}	-4.26 10^{-5}	-1.76 10^{-5}	-5.91 10^{-6}
$\widehat{\mathbb{E}}_0(F_s^B - F_s^A)$	-6.96 10^{-5}	8.37 10^{-6}	-1.11 10^{-5}	2.75 10^{-6}	1.16 10^{-6}
$\sqrt{\widehat{\mathbb{V}}_0(F_s^A) + \widehat{\mathbb{V}}_0(F_s^B)}$	2.25 10^{-4}	7.24 10^{-5}	3.26 10^{-5}	1.59 10^{-5}	7.38 10^{-6}
W_s	-1.56	-1.56	-0.97	-1.28	-0.96

Table 4: Network comparison test for $H_0 = \{g^A = g^B\}$ as defined in Section 3.3 for the same motifs as in Table 2. Networks: A = plant-pollinator, B = seed dispersal. $\widehat{\mathbb{E}}_0(\cdot)$ is a shorthand for the notation $E_s(\cdots)$ (idem for $\widehat{\mathbb{V}}_0(\cdot)$ and $V_s(\cdots)$).

4.1 Asymptotic normality

Simulation design. We simulated series of networks with parameters $(m, n, \rho, \mu_g, \mu_h)$ varying according to the following design:

Network dimension: We simulated networks with equal dimensions $m = n$, with values in $\{50, 100, 200, 500, 1000, 2000\}$;

Sparsity regime: We considered equal parameters $a = b$ in $\{1/3, 1/4, 1/5, 1/6\}$;

Network density: The resulting density is $\rho = \rho_0 m^{-a} n^{-b}$, ρ_0 being fixed so that $\rho = .01$ when $m = n = 100$;

Degree imbalance: We considered the functions $g(u) = \mu_g u^{\mu_g - 1}$ and $h(v) = \mu_h v^{\mu_h - 1}$; observe that $\mu_g = 1$ means that g is constant so no imbalance does exist top nodes (resp. for μ_h, h and bottom nodes). We set $\mu_g = 2$, $\mu_h = 3$.

For each configuration, $S = 100$ networks were sampled and the test applied.

Results. The results are displayed in Figure 3 and Figure 4. In Figure 3, the QQ-plots of the W_s statistic (black dots) defined in (12) and the $\widetilde{W}_s$ statistic (blue dots) defined in (14) hereafter, are given for four motifs in a network with dimension $m = n = 1000$ and sparsity regime $a = b = 1/3$. Remember that the larger the power a, the sparser the graph. We observe that normality of W_s holds for motifs 6 and 15, but not for motifs 5 and 10.
Actually, the latter case is due to the fluctuations of $\overline{F}_s$. More specifically, for non-star motifs, $\overline{F}_s$ is not an unbiased estimate of $\overline{\phi}_s$ and it is not independent from F_s. As a consequence, for finite dimensions m and n, we both have that $\mathbb{E}(\overline{F}_s) \neq \phi_s = \mathbb{E}(F_s)$ and $\mathbb{V}(F_s - \overline{F}_s) \neq \mathbb{V}(F_s)$. Both the bias of $\overline{F}$: $\mathbb{B}(\overline{F}_s) = \mathbb{E}(\overline{F}) - \phi_s$ and the variance of the numerator of W_s: $\mathbb{V}(F_s - \overline{F}_s)$ can be estimated using the delta method, which requires the covariance given in Equation (10). This enables us to define a corrected version $\widetilde{W}_s$ of the test statistic W_s:

$$\widetilde{W}_s := \widehat{\mathbb{V}}\left(F_s - \overline{F}_s\right)^{-1/2} \left(F_s - \overline{F}_s + \widehat{\mathbb{B}}(\overline{F}_s)\right), \tag{14}$$

where the bias $\widehat{\mathbb{B}}(\overline{F}_s)$ and $\widehat{\mathbb{V}}\left(F_s - \overline{F}_s\right)$ are both plug-in estimates.

Illustration. We provide in Table 5 the values of corrected corrected statistics $\widetilde{W}_s$ for the plant-pollinator and the seed dispersal networks, to be compared with Table 3. Observe that the correction does not yield in different conclusions, in terms of fit to the B-EDD model for both networks.

s	5	6	10	15	16
plant-pollinator	-0.05	1.03	-0.03	0.79	2.49
seed dispersal	-0.17	-0.14	-0.20	-0.19	-0.22

Table 5: Corrected test statistics $\widetilde{W}_s$ for the goodness-of-fit of B-EDD for the same motifs as in Table 2.

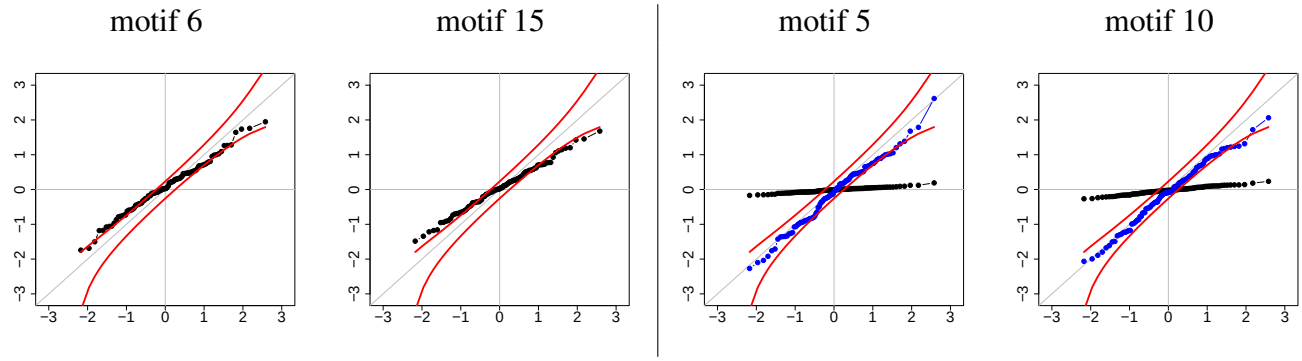


Figure 3: Qq-plots of the test statistics W_s for 4 motifs in a network with dimension $m = n = 2000$ and sparsity regime $a = 1/3$ (black dots). Blue dots: qq-plot for the corrected statistic $\widetilde{W}_s$ defined in Equation (14). Red line: 95% confidence interval for a qq-plot with sample size $S = 100$.

Figure 4 displays the QQ-plots of the corrected test statistics $\widetilde{W}_s$ gathered according to the order of magnitude of the expected motif frequencies. All network sizes, sparsity regimes and non-star motifs are thus considered here together. As expected, the normality becomes more accurate when the motifs frequency increases.

4.2 Power of the goodness-of-fit test

Simulation design. In order to illustrate the power of the goodness-of-fit test, we simulated a series of networks from a mixture of a B-EDD model and a latent block model (LBM) [Govaert and Nadif, 2008], characterizing the presence of clusters of rows and columns in incidence matrices. Thus, a mixing weight α varying from 0 to 1 was considered so that $\alpha = 0$ corresponds to a B-EDD that is H_0. In details, the following simulation setup was investigated:

Network dimension and density: We considered dimensions similar to the pollination and seed dispersal binary networks studied in Simmons et al. [2019b], that is $m = n \in \{10^1, \dots, 10^3\}$. To mimic the sparsity of the same networks, we fitted the density via a linear regression and obtained $\log_{10}(\rho) = 0.3457 - 0.3958 \log_{10}(mn)$;

B-EDD model: We used the same functions g and h as in Section 4.1, with $\mu_g = 2$, $\mu_h = 3$;

LBM model: We considered 2 groups in rows and 2 groups in columns, all groups with proportion $1/2$ and all connection probabilities $\gamma_{k\ell} = C\gamma_{\min}$ for all $1 \leq k, \ell \leq 2$,

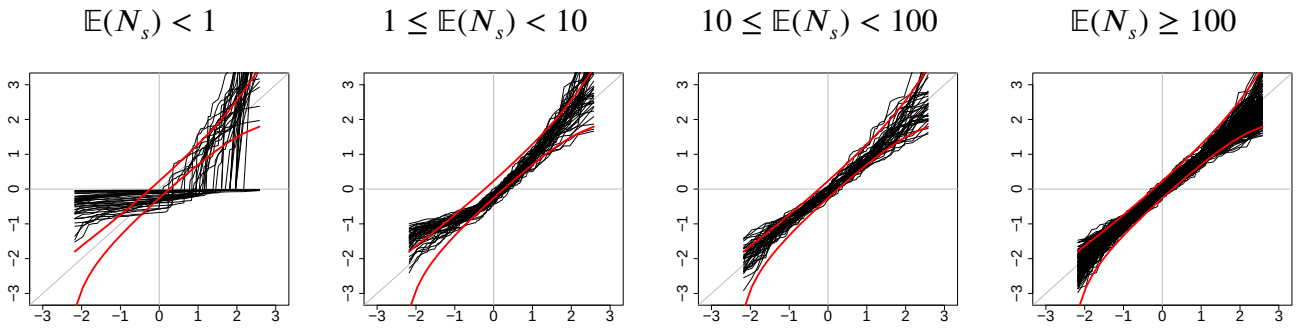


Figure 4: Qq-plots of the corrected test statistics $\widetilde{W}_s$. The plot displays the results of the simulation design (i.e for all network size n, sparsity regime a and non-star motifs s. The qq-plots are gathered accorded to the order of magnitude of the expected count $\mathbb{E}(N_s)$, from the smallest (top left) to the largest (bottom right). Red line: same legend as Figure 3.

except $\gamma_{22} = C\gamma_{\max}$, with C set such that $C(\gamma_{\max} + 3\gamma_{\min})/4 = 1$. Two regimes were considered: $\gamma_{\max} = 0.95$ (scenario I: easy) and $\gamma_{\max} = 0.5$ (scenario II: hard);

Connection probability: We sampled the $\{U_i\}_{1\leq i\leq m}$ and $\{V_j\}_{1\leq j\leq n}$ all independently and uniformly over $[0, 1]$, and set the $\{Z_i\}_{1\leq i\leq m}$ and $\{W_j\}_{1\leq j\leq n}$ as $Z_i = \mathbb{I}\{U_i > .5\} + 1$ and $W_j = \mathbb{I}\{V_j > .5\} + 1$. Finally, the edges were sampled with probability

$$\mathbb{P}\{G_{ij} = 1 \mid U_i, V_j\} = \rho\left((1-\alpha)g(U_i)h(V_j) + \alpha\gamma_{Z_iW_j}\right).$$

For each configuration, $S = 500$ networks were sampled and the test applied. Again the test corrected statistic $\widetilde{W}_s$ was used.

Results. The results are given in Figure 5. For illustration purposes, we only present the results we obtained for $m = n$ ranging from 50 to 500. Moreover, for the sake of clarity, we only consider motifs 5, 6, 10, and 15 which constitute a representative panel of the set of motifs with size 4 and 5.
As the network dimensions increase, we can clearly observe that the tests become more powerful. For small networks with $m = n = 50$ and $m = n = 100$, the LBM regime with $\gamma_{\max} = 0.95$ is easier and leads to tests associated with motifs 5 and 6 with higher power. These differences vanish for larger values of n and m. Overall, we found that motifs 5 and 6 lead to more powerful tests. These results illustrate that the methodology proposed is relevant and that the goodness-of-fit tests for different motifs can be used to detect the departure from a B-EDD model.

4.3 Power of the network comparison test

Simulation design. We also studied the power of the test for network comparison introduced in Section 3.3. To this aim, we simulated series of networks A with parameters $(m_A, n_A, \rho^A, \mu_g^A, \mu_h^A)$ varying according to the same design as in Section 4.1, where μ_g^A was set to 2.
We focused on the test of $H_0 = \{g^A = g^B\}$ so, for each network A, we simulated a sequence of networks B with same dimensions $(m_B = m_A, n_B = n_A)$, but a with a different parameter μ_g^B. More specifically, setting $\mu_g^* = 1$ (absence of degree imbalance between top nodes), we sampled networks B with $\mu_g^B = (1-\alpha)\mu^A + \alpha\mu_g^*$, with $\alpha = 0, 0.1, 0.2, \ldots 1$, so that $\alpha = 0$

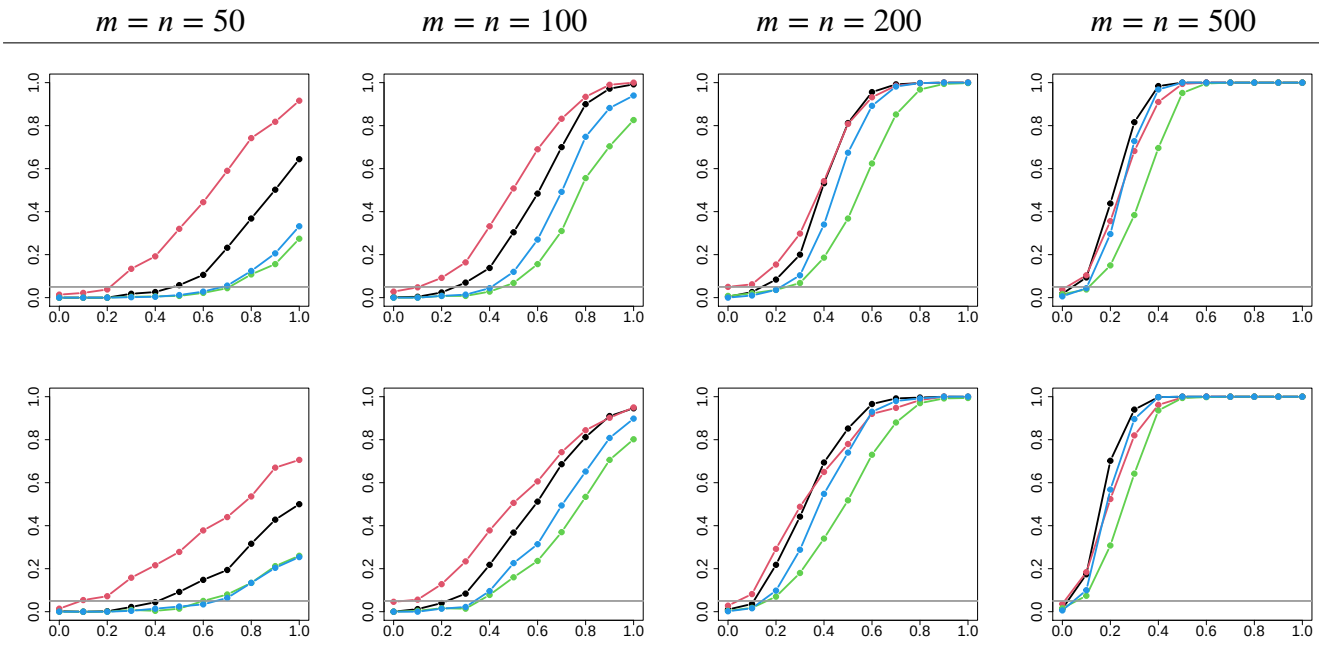


Figure 5: Empirical power of the goodness-of-fit tests, averaged over $S = 500$ simulations. Top: scenario I (easy: $\gamma_{\max} = 0.95$); bottom: scenario II (hard: $\gamma_{\max} = 0.5$). From left to right: $m = n = 50, 100, 200, 500$. Color = motif: black=5, red=6, green=10, blue=15.

corresponds to H_0.
Regarding the two remaining parameters ρ^B and μ_h^B, we considered two scenarios:

I (easy): $\rho^B = \rho^B$, $\mu_h^B = \mu_h^A$, so that the two networks only differ with respect to μ_g;

II (hard): $\rho^B = \rho^A/2$, $\mu_h^B = 2$, so that the two network differ in all parameters, but only the difference in μ_g is tested.

The 'hard' scenario is designed to assess the ability of the proposed test statistic to accommodate to differences in density and bottom node imbalance between the two networks, when testing the equality of their top node imbalance. For each configuration, $S = 500$ pairs of networks (A, B) were sampled and compared.
Following the simulation results presented in Section 4.1, we used the delta-method to derive a corrected version $\widetilde{W}_s$ of the test statistic W_s defined in Equation (13). Similarly to Section 4.1, the performances of the uncorrected test statistic W_s become similar to these of the corrected version $\widetilde{W}_s$ for large networks (results not shown).

Illustration. Again, to illustrate the effect of the proposed correction, we provide in Table 6 the values of corrected statistics $\widetilde{W}_s$ testing $H_0 = \{g^A = g^B\}$, network A being plant-pollinator and network B being seed dispersal. These results can be compared with Table 4: The correction yields in (moderately) higher absolute values, suggesting a gain of power.

Results. The results are displayed in Figure 6. We only present the results for $m_A = n_A = m_B = n_B$ ranging for 50 to 500. Moreover, as in the previous section, we only consider motifs 5, 6, 10 and 15.
As expected, the test becomes more powerful when the networks dimensions increase. More interestingly, for small networks, the smaller motifs (5 and 6, with size 4) turn out to yield

s	5	6	10	15	16
$\widetilde{W}_s$	-2.71	-1.90	-1.76	-1.34	-0.96

Table 6: Corrected test statistics $\widetilde{W}_s$ for $H_0 = \{g^A = g^B\}$ for the same motifs as in Table 2 and same networks as in Table 4.

a higher power. The difference vanishes when the dimensions increase.
These conclusions hold under the two scenarios, which shows that the proposed test statistic does accommodate for departures that may exist between two networks, not being the departure under study (scenario II 'hard'). Still, the power is always better under scenario I: obviously, the test performs better when focusing on the only difference that actually exists (scenario I 'easy').

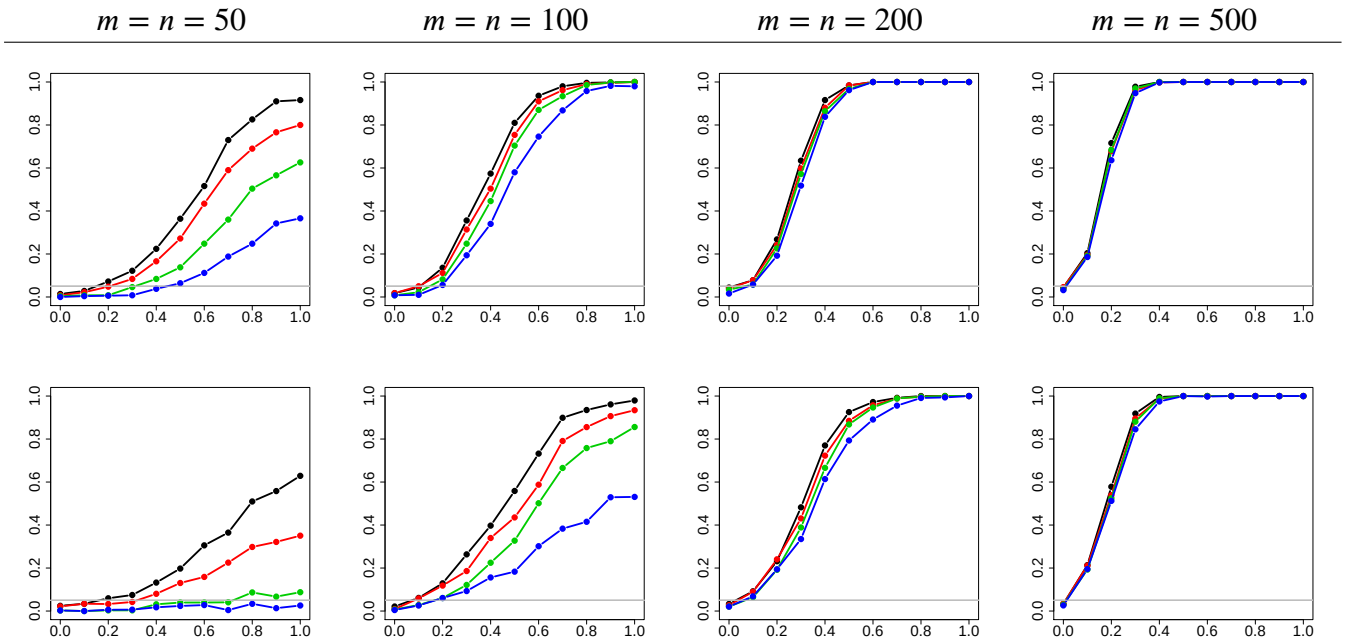


Figure 6: Empirical power of the network comparison test for $H_0 = \{g^A = g^B\}$, averaged over $S = 500$ simulations. Top: scenario I (easy); bottom: scenario II (hard). From left to right: $m = n = 50, 100, 200, 500$. Color = motif: same legend as Figure 5.

5 Proofs

5.1 Definitions and technical lemmas

In this section, we introduce notations and useful technical lemmas for establishing proofs of Proposition 2 in Section 5.2, Lemma 1 in Section 5.3 and Lemma 2 in Section 5.4.

5.1.1 Definitions

Let remind that we consider a bipartite graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with N nodes. The set of nodes is $\mathcal{V} = (\mathcal{V}^t, \mathcal{V}^b)$, where $\mathcal{V}^t = [\![1, m]\!]$ (resp. $\mathcal{V}^b [\![1, n]\!]$) stands for the set of top (resp. bottom)

nodes, and the set of edges is $\mathcal{E} \subset \mathcal{V}^t \times \mathcal{V}^b$, meaning than an edge can only connect a top node with a bottom node. The total number of nodes is therefore $N = n + m$. We denote by G the corresponding $m \times n$ incidence matrix where the entry G_{ij} of G is 1 if $(i, j) \in \mathcal{E}$, and 0 otherwise.

Let consider now a collection of bipartite graphs $(\mathcal{G}_\ell)_{\ell \in [\![1,N]\!]} = (\mathcal{V}_\ell, \mathcal{E}_\ell)$ with ℓ nodes. In the following, we introduce notations for subsets of interest and a filtration we will use to construct differences of martingales involving motif counts.

Subsets definitions. Let introduce the following subsets definitions:

- $\mathcal{V}_\ell = \{(k_1, \dots, k_\ell) \subset \mathcal{V}^t \cup \mathcal{V}^b$ with at least one top node and one bottom node$\}$, $\ell \in [\![2, N]\!]$, it is the set of nodes of $\mathcal{G}_\ell$ meaning the ℓ selected nodes among $\mathcal{G}$, and k_ℓ denotes the ℓ-th and last selected one; we will use k_ℓ several times hereafter;
- $V_\ell^t = \mathcal{V}_\ell \cap \mathcal{V}^t$ and $V_\ell^b = \mathcal{V}_\ell \cap \mathcal{V}^b$, these are the sets of top and bottom nodes in $\mathcal{V}_\ell$;
- $\mathcal{P}_{s,\ell} = \left\{(i_1, \dots, i_{p_s}) \subset V_\ell^t\right\} \times \left\{(j_1, \dots, j_{q_s}) \subset V_\ell^b\right\}$, $\ell \in [\![p_s + q_s, N]\!]$, it is the positions set of motif s in $\mathcal{G}_\ell$;
- $T_\ell = \{k_\ell \in \mathcal{V}^t\}$ is an event;
- $\mathcal{Q}_{s,\ell} = \begin{cases} \{\mathcal{P}_{s,\ell-1} \setminus i_{p_s}\} \cup \{i_{p_s} = k_\ell\} & \text{if } T_\ell, \\ \{\mathcal{P}_{s,\ell-1} \setminus j_{q_s}\} \cup \{j_{q_s} = k_\ell\} & \text{otherwise,} \end{cases}$
  it is the positions set of motif s in $\mathcal{G}_\ell$ with the particularity that k_ℓ the last node added to $\mathcal{V}_\ell$ is part of motif s.

Filtration. The filtration $(\mathcal{F}_\ell)_{\ell \in [\![2,N]\!]}$ is defined by the σ-algebra $\mathcal{F}_\ell = \sigma\left(\mathcal{G}_\ell\right)$.

5.1.2 Technical lemmas

We present here three lemmas which are key arguments in the proofs of Proposition 2, Lemma 1 and Lemma 2.

The following lemma gives the order of magnitude of the variance of a count. Before, its statement let give the order of magnitude of the expected count of a motif s with p_s top nodes and q_s bottom nodes. It writes $\mathbb{E}(N_s) = c_s \phi_s$, with

$$
\begin{aligned}
c_s &= \Theta(m^{p_s} n^{q_s}) \quad \text{(normalizing coefficient specific to } s\text{)} && (15)\\
\rho &= \Theta(m^{-a} n^{-b}), \quad \text{with} \quad a, b > 0 \quad \text{(graph density)} && (16)\\
\phi_s &= \Theta(\rho^{d_+^s}) = \Theta(m^{-a d_+^s} n^{-b d_+^s}) \quad \text{(expected frequency of } s\text{)}, && (17)
\end{aligned}
$$

where d_+^s stands for the total number of edges in s and c_s being defined in (3).

Lemma 3. *We have,*

$$
\mathbb{V}(N_s) = \Theta\left(\max(m^{2p_s - 2ad_+^s - 1} n^{2q_s - 2bd_+^s}, m^{2p_s - 2ad_+^s} n^{2q_s - 2bd_+^s - 1}, m^{2p_s - ad_+^s - 1} n^{2q_s - bd_+^s - 1})\right).
$$

Proof. Let observe that, for $\alpha, \beta \in \mathcal{P}_{s,N}$,

$$
N_s^2 = \sum_\alpha Y_s(\alpha) + \sum_{\alpha \cap \beta \neq \emptyset} Y_s(\alpha) Y_s(\beta) + \sum_{\alpha \cap \beta = \emptyset} Y_s(\alpha) Y_s(\beta).
$$

Thus, a general form for the variance is the following:

$$
\mathbb{V}(N_s) = \mathbb{E}(N_s) + \sum_{t \in S_2(s)} \mathbb{E}(N_t) + \left(|\mathcal{O}_s| - c_s^2\right) \phi_s^2, \tag{18}
$$

where $\mathcal{O}_s = \left\{\alpha, \beta \in \mathcal{P}_{s,N} \,:\, \alpha \cap \beta = \emptyset\right\}$ and $\mathcal{S}_2(s)$ denotes the set of supermotifs of s which are formed by two overlapping occurrences of s.

Let evaluate the orders of the three added terms of assertion (18). Considering that $\rho = \Theta(m^{-a}n^{-b})$, the first term of (18) is $\Theta(m^{p_s - ad_+^s} n^{q_s - bd_+^s})$. Then denoting $(a)_b = a(a-1)\dots(a-b)$, we see that

$$\begin{aligned} |\mathcal{O}_s| - c_s^2 &= \frac{(m_{2p_s-1})}{(p_s!)^2}\frac{(n)_{2q_s-1}}{(q_s!)^2} - \frac{(m_{p_s-1})^2}{(p_s!)^2}\frac{(n)_{q_s-1}^2}{(q_s!)^2} \\ &= \Theta\left(\frac{(-1)^{2p_s-1}p_s^2 m^{2p_s-1}n^{2q_s} + (-1)^{2q_s-1}q_s^2 m^{2p_s}n^{2q_s-1}}{(p_s!)^2(q_s!)^2}\right) \\ &= \Theta\left(\max(m^{2p_s-1}n^{2q_s}, m^{2p_s}n^{2q_s-1})\right). \end{aligned} \tag{19}$$

Thus the third term is $\Theta\left(\max(m^{2p_s-2ad_+^s-1}n^{2q_s-2bd_+^s}, m^{2p_s-2ad_+^s}n^{2q_s-2bd_+^s-1})\right)$.

Let focus now on the second term. When $t \in \mathcal{S}_k(s)$, it can result of an overlap of (i) only top nodes, (ii) only bottom nodes, or (iii) both. For each case we have

(i) $p_t < 2p_s, q_t = 2q_s, d_+^t = 2d_+^s$ so $\mathbb{E}N_t = O(m^{2p_s-1}n^{2q_s}\rho^{2d_+^s}) = O(m^{2p_s-2ad_+^s-1}n^{2q_s-2bd_+^s})$;

(ii) $p_t = 2p_s, q_t < 2q_s, d_+^t = 2d_+^s$ so $\mathbb{E}N_t = O(m^{2p_s}n^{2q_s-1}\rho^{2d_+^s}) = O(m^{2p_s-2ad_+^s}n^{2q_s-2bd_+^s-1})$;

(iii) $p_t < 2p_s, q_t < 2q_s, d_+^s < d_+^t < 2d_+^s$ so $\mathbb{E}N_t = O(m^{2p_s-ad_+^s-1}n^{2q_s-bd_+^s-1})$.

Combining the orders of the three terms of assertion (18), we get that the order of magnitude of the variance of a count is

$$\mathbb{V}(N_s) = \Theta\left(\max(m^{2p_s-2ad_+^s-1}n^{2q_s-2bd_+^s}, m^{2p_s-2ad_+^s}n^{2q_s-2bd_+^s-1}, m^{2p_s-ad_+^s-1}n^{2q_s-bd_+^s-1})\right).$$

■

The last argument of proof of Proposition 2, Lemma 7 and Lemma 1 relies on the following result.

Lemma 4. *We have, as $m \sim n \to \infty$,*

$$\mathbb{V}(N_s|U,V)/\mathbb{V}(N_s) \to 1 \text{ in probability.}$$

Proof. First let us write that

$$\begin{aligned} \mathbb{E}(N_s|U,V) &= \sum_{\alpha\in\mathcal{P}_s} \mathbb{P}\left(Y_s(\alpha) = 1|U_{\alpha^t}, V_{\alpha^b}\right) \\ \mathbb{E}(N_s^2|U,V) &= \sum_{\alpha,\beta\in\mathcal{P}_s} \mathbb{P}\left(Y_s(\alpha)Y_s(\beta) = 1|U_{\alpha^t}, V_{\alpha^b}, U_{\beta^t}, V_{\beta^b}\right). \end{aligned}$$

The proof relies on showing the convergence in probability of the two above expectations towards $\sum_{\alpha\in\mathcal{P}_s}\mathbb{P}\left(Y_s(\alpha)=1\right)$ and $\sum_{\alpha,\beta\in\mathcal{P}_s}\mathbb{P}\left(Y_s(\alpha)Y_s(\beta)=1\right)$, respectively. Let us now introduce the equivalence relation $\mathfrak{R}_s$ and the set R_s defined as follows:

$$\mathfrak{R}_s \,:\, (\sigma_t,\sigma_b) \sim (\tilde{\sigma_t},\tilde{\sigma_b}) \Leftrightarrow A^s_{\sigma_t,\sigma_b} = A^s_{\tilde{\sigma_t},\tilde{\sigma_b}} \text{ and } R_s = \left(\sigma\left([\![1,p_s]\!]\right)\otimes\sigma\left([\![1,q_s]\!]\right)\right)/\mathfrak{R}_s.$$

Then, we can exhibit the two following quantities which are two-samples U-Statistics (see Section 12.2, p.165 in Van der Vaart [2000]):

$$\frac{r_s}{c_s}\sum_{\alpha\in\mathcal{L}_s} k_1\left(U_{\alpha^t}, V_{\alpha^b}\right) \quad \text{and} \quad \frac{r_s}{c_s}\sum_{\alpha\in\mathcal{L}_s} k_2\left(U_{\alpha^t}, V_{\alpha^b}, U_{\beta^t}, V_{\beta^b}\right),$$

with r_S and c_s being defined in (2), (3), respectively, $\mathcal{L}_s$ denoting the location relative to a given position for motif s and where

$$k_1\left(U_{\alpha^t}, V_{\alpha^b}\right) = \sum_{\sigma \in R_s} \mathbb{P}\left(Y_s(\sigma(\alpha)) = 1 | U_{\sigma_t(\alpha^t)}, V_{\sigma_b(\alpha^b)}\right)$$

$$k_2\left(U_{\alpha^t}, V_{\alpha^b}, U_{\beta^t}, V_{\beta^b}\right) = \sum_{(\sigma_\alpha, \sigma_\beta) \in R_s} \mathbb{P}\left(Y_s(\sigma_\alpha(\alpha)) Y_s(\sigma_\beta(\beta)) = 1 | U_{\sigma^t_\alpha(\alpha^t)}, V_{\sigma^b_\alpha(\alpha^b)}, U_{\sigma^t_\beta(\beta^t)}, V_{\sigma^b_\beta(\beta^b)}\right),$$

with $k_1(\cdot)$ and $k_2(\cdot)$ being permutation symmetric kernels in $(U_i)_i$ and $(V_j)_j$ separetely. We conclude by applying the central limit theorem for two-sample U-Statistics (see Theorem 12.6 in Van der Vaart [2000]) which holds under the assumption that the kernel of the U-statistic has a finite moment of order two. Here, as it concerns probabilities this assumption is obviously fulfilled. ■

In proofs of Lemma 2, Lemma 6 and Lemma 7, we need to know the cardinal order of the sets $\mathcal{Q}_{s,\ell}^{\otimes k} \setminus \mathcal{O}_{s,\ell}^{(k)}, k = 2, 4$ which contains only dependent k-uplets of positions of motif s on the event T_ℓ for which the last node added to $\mathcal{V}_\ell$ is a top node. Recall that $\mathcal{Q}_{s,\ell}$ is the positions set of motif s in the subgraph of $\mathcal{G}$ with nodes in $\mathcal{V}_\ell$ and the particularity that k_ℓ the last node added to $\mathcal{V}_\ell$ is part of motif s. The definition of the other set of interest is the following:

$$\mathcal{O}_{s,\ell}^{(k)} = \left\{\alpha_1, \dots \alpha_k \in \mathcal{Q}_{s,\ell} \,:\, (\alpha_1^t \setminus k_\ell) \times \alpha_1^b \cap \dots \cap (\alpha_k^b \setminus k_\ell) \times \alpha_k^b = \emptyset\right\}.$$

Lemma 5. *We have, on T_ℓ,*

$$|\mathcal{Q}_{s,\ell}|^k - |\mathcal{O}_{s,\ell}^{(k)}| = \Theta\left(\ell_t^{k(p_s-1)-1} \ell_b^{kq_s}\right),$$

with ℓ_t and ℓ_b denoting respectively top and bottom nodes in $\mathcal{V}_\ell$.

Proof. Let observe that

$$\begin{aligned}
|\mathcal{Q}_{s,\ell}| &= \binom{\ell_t - 1}{p_s - 1}\binom{\ell_b}{q_s}, \\
|\mathcal{O}_{s,\ell}^{(k)}| &= \binom{\ell_t - 1}{p_s - 1}^k \binom{\ell_b}{q_s \dots q_s\, \ell_b - kq_s} + \binom{\ell_t}{p_s \dots p_s\, \ell_t - kp_s}\binom{\ell_b}{q_s}^k \\
&\quad + \binom{\ell_t}{p_s \dots p_s\, \ell_t - kp_s}\binom{\ell_b}{q_s \dots q_s\, \ell_b - kq_s}.
\end{aligned}$$

The leader term of order $\Theta\left(\ell_t^{k(p_s-1)} \ell_b^{kq_s}\right)$ obviously vanishes and imply the lost of one order (the calculation omitted here are simply based on the same arguments as in (19)). ■

5.2 Proof of Proposition 2

For establishing the proof of Proposition 2, we first consider a decomposition of $L_s = F_s - \overline{\phi}_s$ in Section 5.2.1 , then we focus on the reminder term of this decomposition in Lemma 6 and finally show the asymptotic normality of the leading term in Lemma 7.

5.2.1 Decomposition of L_s

Let use the sets introduced in Section 5.1.1 to express L_s as follows:

$$\begin{aligned} L_s(U,V) = F_s - \phi_s(U,V) &= \frac{1}{c_s} \sum_{\alpha=(\alpha^t,\alpha^b)\in\mathcal{P}_{s,N}} \{Y_s(\alpha) - \overline{\phi}_s(U_{\alpha^t}, V_{\alpha^b})\} \\ &= \frac{1}{c_s} \sum_{\ell=1}^{N} \sum_{\alpha\in\mathcal{Q}_{s,\ell}} \{Y_s(\alpha) - \overline{\phi}_s(U_{\alpha^t}, V_{\alpha^b})\}, \end{aligned}$$

with the random variables U, V of the B-EDD model (1). Then let decompose L_s as the sum of two expressions, the first one corresponding to a martingale difference sequence relative to the filtration $(\mathcal{F}_\ell)_{\ell\in[\![2,N]\!]}$, the second one being a term of rest:

$$L_s(U,V) := M_s(U,V) + R_s(U,V),$$

where

$$M_s(U,V) = \frac{1}{c_s} \sum_{\ell=1}^{N} \sum_{\alpha\in\mathcal{Q}_{s,\ell}} \{Y_s(\alpha) - \mathbb{E}(Y_s(\alpha)|\mathcal{F}_{\ell-1}; U, V)\}$$

$$R_s(U,V) = \frac{1}{c_s} \sum_{\ell=1}^{N} \sum_{\alpha\in\mathcal{Q}_{s,\ell}} \{\mathbb{E}(Y_s(\alpha)|\mathcal{F}_{\ell-1}; U, V) - \phi_s(U_{\alpha^t}, V_{\alpha^b})\}.$$

Observe that by construction, $M_{s,\ell} = \sum_{\alpha\in\mathcal{Q}_{s,\ell}} \{Y_s(\alpha) - \mathbb{E}(Y_s(\alpha)|\mathcal{F}_{\ell-1}; U, V)\}$ is a conditional martingale difference with respect to $(\mathcal{F}_\ell)_{\ell\in[\![2,N]\!]}$:

$$\mathbb{E}(M_{s,\ell}(U,V)|\mathcal{F}_{\ell-1}; U, V) = 0.$$

5.2.2 Study of R_s

Lemma 6. *Under the B-EDD model and condition $a + b < 2/d_+^s$,*

$$R_s(U,V)/\sqrt{\mathbb{V}(F_s)}|U,V \to 0 \text{ a.s. as } m \sim n \to \infty,$$

where $R_s(U,V) = \frac{1}{c_s} \sum_{\ell=1}^{N} \sum_{\alpha\in\mathcal{Q}_{s,\ell}} \{\mathbb{E}(Y_s(\alpha)|\mathcal{F}_{\ell-1}; U, V) - \overline{\phi}_s(U_{\alpha^t}, V_{\alpha^b})\}$.

Proof. The proof consists in showing the two following assertions:

(A1) $\mathbb{E}\left(R_s(U,V)/\sqrt{\mathbb{V}(F_s)}|U,V\right) = 0$;

(A2) $\mathbb{V}\left(c_s R_s(U,V)/\sqrt{\mathbb{V}(N_s)}|U,V\right) \to 0$ almost surely as n tends to infinity under condition $a + b < 2/d_+^s$.

Let show assertion (A1):

$$\begin{aligned} \mathbb{E}\left(R_s(U,V)|U,V\right) &= \mathbb{E}\left(\frac{1}{c_s} \sum_{\ell=1}^{N} \sum_{\alpha\in\mathcal{Q}_{s,\ell}} \{\mathbb{E}(Y_s(\alpha)|\mathcal{F}_{\ell-1}; U, V) - \overline{\phi}_s(U_{\alpha^t}, V_{\alpha^b})\}|U,V\right) \\ &= \mathbb{E}\left(\frac{1}{c_s} \sum_{\ell=1}^{N} \sum_{\alpha\in\mathcal{Q}_{s,\ell}} \mathbb{E}(Y_s(\alpha)|\mathcal{F}_{\ell-1}; U, V)\right) - \frac{1}{c_s} \sum_{\ell=1}^{N} \sum_{\alpha\in\mathcal{Q}_{s,\ell}} \overline{\phi}_s(U_{\alpha^t}, V_{\alpha^b}) \\ &= \frac{1}{c_s} \sum_{\alpha=(\alpha^t,\alpha^b)\in\mathcal{P}_{s,N}} \mathbb{E}(Y_s(\alpha)|U,V) - \frac{1}{c_s} \sum_{\alpha=(\alpha^t,\alpha^b)\in\mathcal{P}_{s,N}} \overline{\phi}_s(U_{\alpha^t}, V_{\alpha^b}) = 0. \end{aligned}$$

Let focus now on assertion (A2). Let first observe that,

$$\begin{aligned}
&\mathbb{V}\left(R_s(U,V)|U,V\right)\\
&= \mathbb{V}\left(\frac{1}{c_s}\sum_{\ell=1}^{N}\sum_{\alpha\in\mathcal{Q}_{s,\ell}}\{\mathbb{E}(Y_s(\alpha)|\mathcal{F}_{\ell-1};U,V)-\overline{\phi}_s(U_{\alpha^t},V_{\alpha^b})\}|U,V\right)\\
&= \sum_{\ell=1}^{N}\mathbb{V}\left(\frac{1}{c_s}\sum_{\alpha\in\mathcal{Q}_{s,\ell}}\mathbb{E}(Y_s(\alpha)|\mathcal{F}_{\ell-1};U,V)|U,V\right),
\end{aligned}$$

by independance of successive choices of $\mathcal{G}_\ell$. Using definition (4) of the indicator motif, we see that

$$\begin{aligned}
&\sum_{\ell=1}^{N}\mathbb{V}\left(\frac{1}{c_s}\sum_{\alpha\in\mathcal{Q}_{s,\ell}}\mathbb{E}(Y_s(\alpha)|\mathcal{F}_{\ell-1};U,V)|U,V\right)\\
&= \sum_{\ell=1}^{N}\mathbb{V}\left(\frac{1}{c_s}\sum_{\alpha\in\mathcal{Q}_{s,\ell}}\mathbb{E}\left(\prod_{i\in\alpha^t,j\in\alpha^b}G_{ij}^{A^s_{ij}}|\mathcal{F}_{\ell-1};U,V\right)|U,V\right).
\end{aligned}$$

Then according to measurability with respect to $\mathcal{F}_{\ell-1}$ and the position (top or bottom) of k_ℓ the last selected node, we get

$$\begin{aligned}
&\sum_{\ell=1}^{N}\mathbb{V}\left(\frac{1}{c_s}\sum_{\alpha\in\mathcal{Q}_{s,\ell}}\mathbb{E}\left(\prod_{i\in\alpha^t,j\in\alpha^b}G_{ij}^{A^s_{ij}}|\mathcal{F}_{\ell-1};U,V\right)|U,V\right)\\
&= \mathbb{P}(T_\ell)\sum_{\ell=1}^{N}\mathbb{V}\left(\frac{1}{c_s}\sum_{\alpha\in\mathcal{Q}_{s,\ell}}\left(\prod_{j\in\alpha^b}\mathbb{E}(G_{k_\ell j}|U,V)^{A^s(k_\ell j)}\right)\left(\prod_{i\in\alpha^t\setminus k_\ell,j\in\alpha^b}G_{ij}^{A^s_{ij}}\right)|U,V\right)\\
&+ (1-\mathbb{P}(T_\ell))\sum_{\ell=1}^{N}\mathbb{V}\left(\frac{1}{c_s}\sum_{\alpha\in\mathcal{Q}_{s,\ell}}\left(\prod_{i\in\alpha^t}\mathbb{E}(G_{ik_\ell}|U,V)^{A^s(ik_\ell)}\right)\left(\prod_{i\in\alpha^t,j\in\alpha^b\setminus k_\ell}G_{ij}^{A^s_{ij}}\right)|U,V\right),
\end{aligned}$$

and using the usual notation of the conditional expectation of G_{ij}'s, we have

$$\begin{aligned}
&\mathbb{V}\left(R_s(U,V)|U,V\right)\\
&= \mathbb{P}(T_\ell)\sum_{\ell=1}^{N}\mathbb{V}\left(\frac{1}{c_s}\sum_{\alpha\in\mathcal{Q}_{s,\ell}}\left(\prod_{j\in\alpha^b}\phi_1(U_{k_\ell},V_j)^{A^s(k_\ell j)}\right)\left(\prod_{i\in\alpha^t\setminus k_\ell,j\in\alpha^b}G_{ij}^{A^s_{ij}}\right)|U,V\right)\\
&+ (1-\mathbb{P}(T_\ell))\sum_{\ell=1}^{N}\mathbb{V}\left(\frac{1}{c_s}\sum_{\alpha\in\mathcal{Q}_{s,\ell}}\left(\prod_{i\in\alpha^t}\phi_1(U_i,V_{k_\ell})^{A^s(ik_\ell)}\right)\left(\prod_{i\in\alpha^t,j\in\alpha^b\setminus k_\ell}G_{ij}^{A^s_{ij}}\right)|U,V\right).
\end{aligned}$$

Then, considering the fact that $\mathbb{V}(\sum_i a_i X_i) \le \left(\sum_i a_i\sqrt{\mathbb{V}(X_i)}\right)^2$, we get

$$\begin{aligned}
&\mathbb{V}\left(R_s(U,V)|U,V\right)\\
&\le 2\sum_{\ell=1}^{N}\left(\frac{1}{c_s}\sum_{\alpha\in\mathcal{Q}_{s,\ell}}\left(\prod_{j\in\alpha^b}\phi_1(U_{k_\ell},V_j)^{A^s(k_\ell j)}\right)\sqrt{\mathbb{V}\left(\prod_{i\in\alpha^t\setminus k_\ell,j\in\alpha^b}G_{ij}^{A^s_{ij}}|U,V\right)}\right)^2.
\end{aligned}$$

From now, we will work on the set $\mathcal{Q}_{s,\ell}^{\otimes 2} \setminus \mathcal{O}_{s,\ell}^{(2)}$ which contains only dependent pairs of positions. It follows from the Bernoulli conditional distribution of G_{ij} combined with the fact that $a\sqrt{b} < \sqrt{ab}$ when $a < 1$, that

$$\begin{aligned}
&\mathbb{V}\left(R_s(U,V)|U,V\right) \\
&\leq 2\sum_{\ell=1}^{N}\left(\frac{1}{c_s}\sum_{\alpha\in\mathcal{Q}_{s,\ell}^{\otimes 2}\setminus\mathcal{O}_{s,\ell}^{(2)}}\sqrt{\prod_{j\in\alpha^b}\phi_1(U_{k_\ell},V_j)^{A_s(k_\ell j)}\prod_{i\in\alpha^t\setminus k_\ell, j\in\alpha^b}\phi_1(U_i,V_j)^{A_s(ij)}}\right)^2 \\
&\leq 2\sum_{\ell=1}^{N}\frac{1}{c_s^2}\left(\sum_{\alpha\in\mathcal{Q}_{s,\ell}^{\otimes 2}\setminus\mathcal{O}_{s,\ell}^{(2)}}\sqrt{\overline{\phi}_s(U_{\alpha^t},V_{\alpha^b})}\right)^2 \\
&\leq \frac{2}{c_s^2}\sum_{\ell=1}^{N}\left(|\mathcal{Q}_{s,\ell}|^2-|\mathcal{O}_{s,\ell}^{(2)}|\right)\max_{\alpha\in\mathcal{Q}_{s,\ell}^{\otimes 2}\setminus\mathcal{O}_{s,\ell}^{(2)}}\overline{\phi}_s(U_{\alpha^t},V_{\alpha^b}).
\end{aligned}$$

In order to evaluate the right-hand side term of the above inequality, recall that $c_s = \Theta(m^{p_s}n^{q_s})$ by (15), $\phi_s = \Theta\left(m^{-ad_+^s}n^{-bd_+^s}\right)$ by (17) and $|\mathcal{Q}_{s,\ell}|^2 - |\mathcal{O}_{s,\ell}^{(2)}| = \Theta\left(\ell_t^{2p_s-3}\ell_b^{2q_s}\right)$ by Lemma 5, ℓ_t and ℓ_b denoting respectively top and bottom nodes in $\mathcal{V}_\ell$. Thus, we get

$$\begin{aligned}
&\frac{2}{c_s^2}\sum_{\ell=1}^{N}\left(|\mathcal{Q}_{s,\ell}|^2-|\mathcal{O}_{s,\ell}|\right)\max_{\alpha\in\mathcal{Q}_{s,\ell}^{\otimes 2}\setminus\mathcal{O}_{s,\ell}^{(2)}}\overline{\phi}_s(U_{\alpha^t},V_{\alpha^b}) \\
&= \Theta\left(m^{-2p_s}n^{-2q_s}\right)\sum_{\ell=\ell_t+\ell_b=1}^{N}\Theta\left(\ell_t^{2p_s-3}\ell_b^{2q_s}\ell_t^{-ad_+^s}\ell_b^{-bd_+^s}\right) \\
&= \Theta\left(m^{-2p_s}n^{-2q_s}\right)\sum_{\ell=\ell_t+\ell_b=1}^{N}\Theta\left(\ell^{2p_s+2q_s-ad_+^s-bd_+^s-3}\right) \\
&= \Theta\left(N^{-ad_+^s-bd_+^s-3}\right).
\end{aligned}$$

By taking the normalization $\sqrt{\mathbb{V}(F_s)} = \sqrt{\mathbb{V}(N_s)}/c_s$ which order is

$$\Theta\left(\max\left(N^{-2ad_+^s-2bd_+^s-1}, N^{-ad_+^s-bd_+^s-2}\right)\right)$$

by Lemma 2 and (15), we conclude to $\mathbb{V}\left(\frac{R_s(U,V)}{\sqrt{\mathbb{V}(F_s)}}|U,V\right) \to 0$ almost surely as n tends to infinity under condition $a + b < 2/d_+^s$. ■

5.2.3 Study of M_s

Lemma 7. *Under the B-EDD model and condition $a + b < 2/d_+^s$,*

$$M_s(U,V)/\sqrt{\mathbb{V}(F_s)}|U,V \xrightarrow{D} \mathcal{N}\left(0, \frac{\mathbb{V}(N_s|U,V)}{\mathbb{V}(N_s)})\right), \text{ as } m\sim n\to\infty,$$

where $M_s(U,V) = \frac{1}{c_s}\sum_{\ell=1}^{N}\sum_{\alpha\in\mathcal{Q}_{s,\ell}}\{Y_s(\alpha) - \mathbb{E}(Y_s(\alpha)|\mathcal{F}_{\ell-1};U,V)\}$.

Proof. We will apply the following martingale central limit theorem to the conditional martingale difference sequence $M_{s,\ell}(U,V) = \sum_{\alpha \in \mathcal{Q}_{s,\ell}} \{Y_s(\alpha) - \mathbb{E}(Y_s(\alpha)|\mathcal{F}_{\ell-1}; U, V)\}$ with respect to $(\mathcal{F}_\ell)_{\ell \in [\![2,N]\!]}$.

Theorem 3 ([Hall and Heyde, 2014])**.** *Suppose that for every $n \in \mathbb{N}$ and $k_n \to \infty$ the random variables $X_{n,1}, \ldots, X_{n,k_n}$ are a martingale difference sequence relative to an arbitrary filtration $\mathcal{F}_{n,1} \subset \mathcal{F}_{n,1} \subset \ldots \subset \mathcal{F}_{n,k_n}$. If*

1. $\sum_{i=1}^{k_n} \mathbb{E}(X_{n,i}^2|\mathcal{F}_{n,i-1}) \to 1$ *in probability,*
2. $\sum_{i=1}^{k_n} \mathbb{E}(X_{n,i}^2 \mathbb{I}\{|X_{n,i}| > \epsilon\}|\mathcal{F}_{n,i-1}) \to 0$ *in probability for every $\epsilon > 0$,*

then $\sum_{i=1}^{k_n} X_{n,i} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1)$.

Here $X_{n,i}$ and $\mathcal{F}_{n,i}$ would be $M_{s,\ell}(U,V)/(c_s\sqrt{\mathbb{V}(F_s)})$ and $\mathcal{F}_\ell$ respectively, and we have to verify the two following conditions:

(C1) $\frac{1}{\mathbb{V}(N_s)} \sum_{\ell=1}^N \mathbb{E}(M_{s,\ell}^2(U,V)|\mathcal{F}_{\ell-1}; U, V) \to \frac{\mathbb{V}(N_s|U,V)}{\mathbb{V}(N_s)}$ in probability,

(C2) $\frac{1}{\mathbb{V}(N_s)} \sum_{\ell=1}^N \mathbb{E}\left(M_{s,\ell}^2(U,V) \mathbb{I}\left\{ \frac{|M_{s,\ell}(U,V)|}{\sqrt{\mathbb{V}(N_s)}} > \epsilon \right\} |\mathcal{F}_{\ell-1}; U, V \right) \to 0$ in probability for every $\epsilon > 0$.

Let verify condition (C1). First observe that it follows from properties of martingale differences, meaning variance decomposition, null conditional expectation and conditional orthogonality of differences, that

$$
\begin{aligned}
&\mathbb{V}\left(\sum_{\ell=1}^N M_{s,\ell}(U,V)|U,V \right) \\
&= \mathbb{E}\left[\mathbb{V}\left(\sum_{\ell=1}^N M_{s,\ell}(U,V)|\mathcal{F}_{\ell-1}; U, V \right) \right] + \mathbb{V}\left[\mathbb{E}\left(\sum_{\ell=1}^N M_{s,\ell}(U,V)|\mathcal{F}_{\ell-1}; U, V \right) \right] \\
&= \mathbb{E}\left[\mathbb{V}\left(\sum_{\ell=1}^N M_{s,\ell}(U,V)|\mathcal{F}_{\ell-1}; U, V \right) \right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\left(\sum_{\ell=1}^N M_{s,\ell}(U,V) \right)^2 |\mathcal{F}_{\ell-1}; U, V \right] \right] \\
&= \mathbb{E}\left(\sum_{\ell=1}^N \mathbb{E}\left(M_{s,\ell}^2(U,V)|\mathcal{F}_{\ell-1}; U, V \right) \right),
\end{aligned}
$$

and further notice that $\mathbb{V}\left(\sum_{\ell=1}^N M_{s,\ell}(U,V)|U,V \right) = \mathbb{V}\left(c_s M_s(U,V)|U,V \right)$. Since $M_s = L_s - R_s$ (see Section 5.2.1),

$$
\begin{aligned}
&\mathbb{E}\left(\frac{1}{\mathbb{V}(N_s)} \sum_{\ell=1}^N \mathbb{E}(M_{s,\ell}^2(U,V)|\mathcal{F}_{\ell-1}; U, V) \right) \\
&= \mathbb{V}\left(c_s \frac{L_s(U,V) - R_s(U,V)}{\sqrt{\mathbb{V}(N_s)}} |U,V \right) \to \mathbb{V}(N_s|U,V)/\mathbb{V}(N_s), \text{ as } n \to \infty,
\end{aligned}
$$

in probability and under condition $a + b < 2/d_+^s$, because $\mathbb{V}\left(c_s \frac{L_s(U,V)}{\sqrt{\mathbb{V}(N_s)}} |U,V \right) = \frac{\mathbb{V}(N_s|U,V)}{\mathbb{V}(N_s)}$ and $\mathbb{V}\left(c_s \frac{R_s(U,V)}{\sqrt{\mathbb{V}(N_s)}} |U,V \right) \to 0$ a.s. under condition $a + b < 2/d_+^s$ by Lemma 6.

Now, let verify condition (C2). First, by applying the Cauchy-Schwartz inequality, we get

$$\begin{aligned}
&\frac{1}{\mathbb{V}(N_s)}\sum_{\ell=1}^{N}\mathbb{E}\left(M_{s,\ell}^2(U,V)\mathbb{I}\left\{\frac{|M_{s,\ell}(U,V)|}{\sqrt{\mathbb{V}(N_s)}}>\epsilon\right\}|\mathcal{F}_{\ell-1};U,V\right)\\
\leq\ &\sum_{\ell=1}^{N}\mathbb{E}\left(\frac{M_{s,\ell}^4(U,V)}{\mathbb{V}(N_s)^2}|\mathcal{F}_{\ell-1};U,V\right)^{1/2}\times\sum_{\ell=1}^{N}\mathbb{P}\left(\frac{|M_{s,\ell}(U,V)|}{\sqrt{\mathbb{V}(N_s)}}>\epsilon|\mathcal{F}_{\ell-1};U,V\right)^{1/2},
\end{aligned}$$

then applying Bienaymé-Tchebychev inequality implies that

$$\begin{aligned}
&\sum_{\ell=1}^{N}\mathbb{E}\left(\frac{M_{s,\ell}^4(U,V)}{\mathbb{V}(N_s)^2}|\mathcal{F}_{\ell-1};U,V\right)^{1/2}\times\sum_{\ell=1}^{N}\mathbb{P}\left(\frac{|M_{s,\ell}(U,V)|}{\sqrt{\mathbb{V}(N_s)}}>\epsilon|\mathcal{F}_{\ell-1};U,V\right)^{1/2}\\
\leq\ &\frac{1}{\mathbb{V}(N_s)^2}\sum_{\ell=1}^{N}\mathbb{E}(M_{s,\ell}^4(U,V)|\mathcal{F}_{\ell-1};U,V)\times\frac{1}{\epsilon^2\mathbb{V}(N_s)}\sum_{\ell=1}^{N}\mathbb{E}(M_{s,\ell}^2(U,V)|\mathcal{F}_{\ell-1};U,V),
\end{aligned}$$

and by condition (C1), we get

$$\begin{aligned}
&\frac{1}{\mathbb{V}(N_s)}\sum_{\ell=1}^{N}\mathbb{E}\left(M_{s,\ell}^2(U,V)\mathbb{I}\left\{\frac{|M_{s,\ell}(U,V)|}{\sqrt{\mathbb{V}(N_s)}}>\epsilon\right\}|\mathcal{F}_{\ell-1};U,V\right)\\
\leq\ &\frac{1}{\epsilon^2\mathbb{V}(N_s)^2}\sum_{\ell=1}^{N}\mathbb{E}(M_{s,\ell}^4(U,V)|\mathcal{F}_{\ell-1};U,V)\times\frac{\mathbb{V}(N_s|U,V)}{\mathbb{V}(N_s)}.
\end{aligned}$$

Then, we use the following notation for expressing $M_{s,\ell}$:

$$M_{s,\ell}=\sum_{\alpha\in\mathcal{Q}_{s,\ell}}\{Y_s(\alpha)-\mathbb{E}(Y_s(\alpha)|\mathcal{F}_{\ell-1};U,V)\}=N_{s,\ell}-\mathbb{E}(N_{s,\ell}|\mathcal{F}_{\ell-1};U,V).$$

By the binomial formula we thus have

$$\begin{aligned}
\mathbb{E}(M_{s,\ell}^4(U,V)|\mathcal{F}_{\ell-1};U,V)\ &=\ \mathbb{E}\left(\left(N_{s,\ell}-\mathbb{E}(N_{s,\ell}|\mathcal{F}_{\ell-1};U,V)\right)^4|\mathcal{F}_{\ell-1};U,V\right)\\
&=\ \mathbb{E}\left(N_{s,\ell}^4|\mathcal{F}_{\ell-1};U,V\right)-4\mathbb{E}\left(N_{s,\ell}^3\mathbb{E}(N_{s,\ell}|\mathcal{F}_{\ell-1};U,V)|\mathcal{F}_{\ell-1};U,V\right)\\
&\quad+6\mathbb{E}\left(N_{s,\ell}^2\mathbb{E}(N_{s,\ell}|\mathcal{F}_{\ell-1};U,V)^2|\mathcal{F}_{\ell-1};U,V\right)\\
&\quad-4\mathbb{E}\left(N_{s,\ell}\mathbb{E}(N_{s,\ell}|\mathcal{F}_{\ell-1};U,V)^3|\mathcal{F}_{\ell-1};U,V\right)+\mathbb{E}(N_{s,\ell}|\mathcal{F}_{\ell-1};U,V)^4.
\end{aligned}$$

Using the same arguments as in the proof of Lemma 6, observe that

$$\begin{aligned}
E\left(N_{s,\ell}|\mathcal{F}_{\ell-1};U,V\right)\ &\leq\ 2\sum_{\alpha\in\mathcal{Q}_{s,\ell}}\left(\prod_{j\in\alpha^b}\phi_1(U_{k_\ell},V_j)^{A^s(k_\ell j)}\right)\mathbb{E}\left(\left(\prod_{i\in\alpha^t,j\in\alpha^b\setminus k_\ell}G_{ij}^{A_{ij}^s}\right)|U,V\right)\\
&\leq\ 2\sum_{\alpha\in\mathcal{Q}_{s,\ell}}\overline{\phi}_s(U_{\alpha^t},V_{\alpha^b}),
\end{aligned}$$

and we have,

$$\mathbb{E}\left(N_{s,\ell}^k|\mathcal{F}_{\ell-1};U,V\right)=\mathbb{E}\left(N_{s,\ell}|\mathcal{F}_{\ell-1};U,V\right)+\sum_{t\in\mathcal{S}_k(s)}\mathbb{E}\left(N_{t,\ell}|\mathcal{F}_{\ell-1};U,V\right)+\mathbb{E}\left(N_{s,\ell}|\mathcal{F}_{\ell-1};U,V\right)^k,$$

where $S_k(s)$ denotes here the set of supermotifs of s which are here formed by k overlapping occurrences of s. Finally, we get

$$\begin{aligned}
&\frac{1}{\mathbb{V}(N_s)}\sum_{\ell=1}^{N}\mathbb{E}\left(M_{s,\ell}^2(U,V)\mathbb{I}\left\{\frac{|M_{s,\ell}(U,V)|}{\sqrt{\mathbb{V}(N_s)}}>\epsilon\right\}|\mathcal{F}_{\ell-1};U,V\right)\\
\leq\ &\frac{1}{\epsilon^2\mathbb{V}(N_s)^2}\sum_{\ell=1}^{N}\mathbb{E}\left(M_{s,\ell}^4(U,V)|\mathcal{F}_{\ell-1};U,V\right)\times\frac{\mathbb{V}(N_s|U,V)}{\mathbb{V}(N_s)}\\
\leq\ &\frac{2}{\epsilon^2\mathbb{V}(N_s)^2}\sum_{\ell=1}^{N}|\mathcal{Q}_{s,\ell}|^4\times\left(\max_{\alpha\in\mathcal{Q}_{s,\ell}}\overline{\phi}_s(U_{\alpha^t},V_{\alpha^b})\right)^4\times\frac{\mathbb{V}(N_s|U,V)}{\mathbb{V}(N_s)}.
\end{aligned}$$

Condition (C2) holds since $\mathbb{V}(N_s)^2=\Theta\left(\max\left(N^{4p_s+4q_s-4ad_+^s-4bd_+^s-2},N^{4p_s+4q_s-2ad_+^s-2bd_+^s-4}\right)\right)$ by Lemma 3, $|\mathcal{Q}_{s,\ell}|^4=\Theta\left(\ell_t^{4p_s-4}\ell_b^{4q_s}\right)$ (see the proof of Lemma (5)), $\phi_s^4=\Theta\left(N^{-4ad_+^s-4bd_+^s}\right)$ by (17) and $\mathbb{V}(N_s|U,V)/\mathbb{V}(N_s)=\Theta(1)$ by Lemma 4. ■

5.3 Proof of Lemma 1

Proof. Let show that $(\overline{F}_s-\overline{\phi}_s)/\sqrt{\mathbb{V}(F_s)}\to 0$ a.s. as $n\to\infty$ under the B-EDD model and condition $a+b<2/d_+^s$ ruling the graph density. Recall (8) the definition of $\overline{F}_s$:

$$\overline{F}_s=\frac{\prod_{u=1}^{p_s}\Gamma_{d_u^s}\prod_{v=1}^{q_s}\Lambda_{e_v^s}}{F_1^{d_+^s}},$$

where Γ_d (resp Λ_d) denote the normalized empirical frequencies of the top (resp bottom) star motif with degree d and F_1 the one of the edge.

Let begin with a Taylor expansion of order 1 of $\overline{F}_s$ in parameters (γ,λ,ϕ_1) denoting the top star motif, bottom star motif and edge probabilities respectively:

$$\begin{aligned}
\overline{F}_s(\Gamma,\Lambda,F_1) &= \overline{F}_s(\gamma,\lambda,\phi_1)+\partial\overline{F}_s(\gamma,\lambda,\phi_1)\left((\Gamma,\Lambda,F_1)-(\gamma,\lambda,\phi_1)\right)+o\left((\Gamma,\Lambda,F_1)-(\gamma,\lambda,\phi_1)\right)\\
&= \overline{\phi}_s+\overline{\phi}_s\partial\log(\overline{F}_s(\gamma,\lambda,\phi_1))\left((\Gamma,\Lambda,F_1)-(\gamma,\lambda,\phi_1)\right)+o\left((\Gamma,\Lambda,F_1)-(\gamma,\lambda,\phi_1)\right)\\
&= \overline{\phi}_s+\overline{\phi}_s\left\{\sum_{u=1}^{p_s}\frac{1}{\gamma_{d_u^s}}(\Gamma_{d_u^s}-\gamma_{d_u^s})+\sum_{v=1}^{q_s}\frac{1}{\lambda_{e_v^s}}(\Lambda_{e_v^s}-\lambda_{e_v^s})-\frac{d_+}{\phi_1}(F_1-\phi_1)\right\}\\
&\quad +o\left(\Gamma-\gamma,\Lambda-\lambda,F_1-\phi_1\right).
\end{aligned}$$

Given the two following observations: i) the asymptotic normality of $(F_s-\overline{\phi}_s)/\sqrt{\mathbb{V}(F_s)}$ holds for any motif s, including star motifs, under the B-EDD model and condition $a+b<2/d_+^s$ by Proposition 2, ii) the empirical frequencies of motifs converge to the expected ones by the law of large numbers, we get

$$\begin{aligned}
&\frac{\overline{F}_s-\overline{\phi}_s}{\sqrt{\mathbb{V}(F_s)}}\\
&=\sum_{u=1}^{p_s}\Theta\left(\frac{\phi_s}{\gamma_{d_u^s}}\sqrt{\frac{\mathbb{V}(\Gamma_{d_u^s})}{\mathbb{V}(F_s)}}\right)+\sum_{v=1}^{q_s}\Theta\left(\frac{\phi_s}{\lambda_{e_v^s}}\sqrt{\frac{\mathbb{V}(\Lambda_{e_v^s})}{\mathbb{V}(F_s)}}\right)+\Theta\left(\frac{\phi_s}{\phi_1}\sqrt{\frac{\mathbb{V}(F_1)}{\mathbb{V}(F_s)}}\right)+o(1)\\
&=\sum_{u=1}^{p_s}\Theta\left(\frac{\phi_s c_s}{\gamma_{d_u^s}c_\gamma}\sqrt{\frac{\mathbb{V}(N_{\Gamma_{d_u^s}})}{\mathbb{V}(N_s)}}\right)+\sum_{v=1}^{q_s}\Theta\left(\frac{\phi_s c_s}{\lambda_{e_v^s}c_\lambda}\sqrt{\frac{\mathbb{V}(N_{\Lambda_{e_v^s}})}{\mathbb{V}(N_s)}}\right)+\Theta\left(\frac{\phi_s c_s}{\phi_1 c_1}\sqrt{\frac{\mathbb{V}(N_1)}{\mathbb{V}(N_s)}}\right)+o(1).
\end{aligned}$$

Here and only here, N_{Γ_d} (resp. N_{Λ_d}) and c_γ (resp. c_λ) denote, by abuse of notation, the count of top stars (resp. bottom stars) of degree d and their number of positions in the graph. Considering only non-star motifs s, according to the orders of magnitude of c_s, ϕ_s and $\mathbb{V}(N_s)$ given in (15), (17) and Lemma 3 respectively, we conclude to $(\overline{F}_s - \overline{\phi}_s)/\sqrt{\mathbb{V}(F_s)} \to 0$ a.s. as $n \to \infty$ because $-2d(a+b) < 0$, with $d = d_u^s, e_u^s$ or 1. ■

5.4 Proof of Lemma 2

Proof. Let show that $\hat{\mathbb{V}}(F_s)/\mathbb{V}(F_s) \to 1$ a.s., as $n \to \infty$. First, observe that according to (18), we can write:

$$\mathbb{V}(N_s) = \sum_{t \in \{s\} \cup S_2(s)} \mathbb{E}(N_t) - \mathbb{E}(N_s)^2 = c_s\phi_s + \sum_{t \in S_2(s)} c_t\phi_t - c_s^2\phi_s^2,$$

where $S_2(s)$ denotes here the set of super-motifs of s which are formed by two overlapping occurrences of s. Then considering $\hat{\mathbb{V}}(N_s)$ its plug-in version, meaning $\overline{F}_s$ replaces ϕ_s, we get

$$\hat{\mathbb{V}}(N_s) - \mathbb{V}(N_s) = c_s(\overline{F}_s - \phi_s) + \sum_{t \in S_2(s)} c_t(\overline{F}_t - \phi_t) - c_s^2(\overline{F}_s^2 - \phi_s^2).$$

Now we use Lemma 1 stating that, under the B-EDD model and condition $a + b < 2/d_+^s$, $\overline{F}_s - \phi_s = o(\sqrt{\mathbb{V}(F_s)})$ for all motif s and the continuous mapping theorem, to obtain that

$$\begin{aligned} \hat{\mathbb{V}}(N_s) - \mathbb{V}(N_s) &= c_s o(\sqrt{\mathbb{V}(F_s)}) + \sum_{tS_2(s)} c_t o(\sqrt{\mathbb{V}(F_t)}) - c_s^2 o(\mathbb{V}(F_s)) \\ &= o(\sqrt{\mathbb{V}(N_s)}) + \sum_{t \in S_2(s)} o(\sqrt{\mathbb{V}(N_t)}) - o(\mathbb{V}(N_s)). \end{aligned} \tag{20}$$

Let discuss now the order of $(\hat{\mathbb{V}}(N_s) - \mathbb{V}(N_s))/\mathbb{V}(N_s)$. The first and last terms of (20) divided by $\mathbb{V}(N_s)$ obviously vanish. When $t \in S_2(s)$, we refer to the order of magnitude of $\mathbb{V}(N_s)$ given in Lemma 3 and its proof (see (i)-(ii)-(iii)) to get that $\sqrt{\mathbb{V}(N_t)}/\mathbb{V}(N_s)$ vanishes under condition $a + b < (p_s + q_s)/d_+^s$. We can finally conclude to $\hat{\mathbb{V}}(F_s)/\mathbb{V}(F_s) \to 1$ a.s., as $n \to \infty$ under condition of Theorem 1.

■

References

F. Chung and L. Lu. Connected components in random graphs with given expected degree sequences. *Annals of combinatorics*, 6(2):125–145, 2002.

E. D'Bastiani, K. M. Campião, W. A. Boeger, and S. B. L Araújo. The role of ecological opportunity in shaping host–parasite networks. *Parasitology*, 147(13):1452–1460, 2020.

P. Diaconis and S. Janson. Graph limits and exchangeable random graphs. *Rend. Mat. Appl.*, 7(28):33–61, 2008.

M. Doré, C. Fontaine, and E. Thébault. Relative effects of anthropogenic pressures, climate, and sampling design on the structure of pollination networks at the global scale. *Global Change Biology*, 2020.

C. Gao and J. Lafferty. Testing for global network structure using small subgraph statistics. Technical Report 1710.00862, arXiv, 2017.

G. Govaert and M. Nadif. Block clustering with bernoulli mixture models: Comparison of different approaches. *Computational Statistics & Data Analysis*, 52(6):3233–3245, 2008.

P. Hall and C. C Heyde. *Martingale limit theory and its application*. Academic press, 2014.

L. Lovász and B. Szegedy. Limits of dense graph sequences. *Journal of Combinatorial Theory, Series B*, 96(6):933 – 957, 2006. ISSN 0095-8956.

R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824–827, 2002.

M. EJ Newman. The structure and function of complex networks. *SIAM review*, 45(2): 167–256, 2003.

K. Nowicki and J. C Wierman. Subgraph counts in random graphs using incomplete u-statistics methods. *Discrete Mathematics*, 72(1-3):299–310, 1988.

F. Picard, J.-J. Daudin, M. Koskas, S. Schbath, and S. Robin. Assessing the exceptionality of network motifs,. *J. Comp. Biol.*, 15(1):1–20, 2008.

C. Robertson. Flowers and insects: lists of visitors to four hundred and fifty-three flowers. carlinville, il, usa, c. robertson. *National Center for Ecological Analysis and Synthesis Interaction Web Database*, 456, 1929.

F. Saracco, R. Di Clemente, A. Gabrielli, and T. Squartini. Detecting early signs of the 2007–2008 crisis in the world trade. *Scientific reports*, 6(1):1–11, 2016.

WR Silva. Patterns of fruit-frugivore interactions in two atlantic forest bird communities of south-eastern brazil: implications for conservation. *Seed dispersal and frugivory: ecology, evolution and conservation*, pages 423–435, 2002.

B. I. Simmons, M. JM. Sweering, M. Schillinger, L. V. Dicks, W. J. Sutherland, and R. Di Clemente. bmotif: A package for motif analyses of bipartite networks. *Methods in Ecology and Evolution*, 10(5):695–701, 2019a.

B.I. Simmons, A. Cirtwill, N. Baker, L.V. Dicks, D.B. Stouffer, and W.J. Sutherland. Motifs in bipartite ecological networks: uncovering indirect interactions. *Oikos*, 128(2):154–170, 2019b.

D. Stark. Compound poisson approximations of subgraph counts in random graphs. *Random Structures & Algorithms*, 18(1):39–60, 2001.

M. Thomas, N. Verzelen, P. Barbillon, O. T. Coomes, S. Caillon, D. McKey, M. Elias, E. Garine, C. Raimond, E. Dounias, et al. A network-based method to detect patterns of local crop biodiversity: validation at the species and infra-species levels. In *Advances in Ecological Research*, volume 53, pages 259–320. Elsevier, 2015.

C. Vacher, D. Piou, and M. L. Desprez-Loustau. Architecture of an antagonistic tree/fungus network: the asymmetric influence of past evolutionary history. *PloS one*, 3(3):e1740, 2008.

A. W. Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

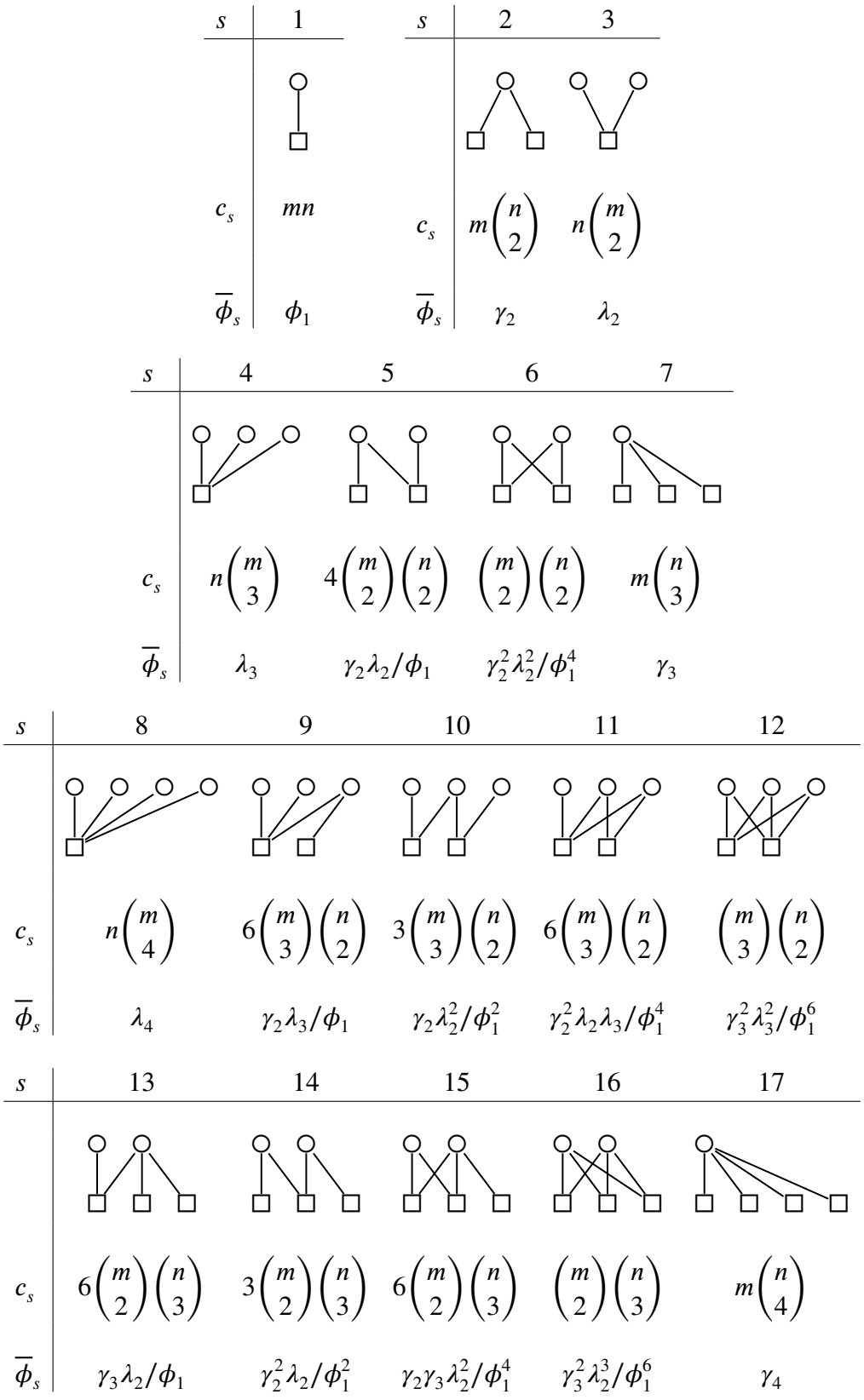


Figure 7: Bipartite motifs of size 2, 3, 4 and 5 as given in Simmons et al. [2019b].

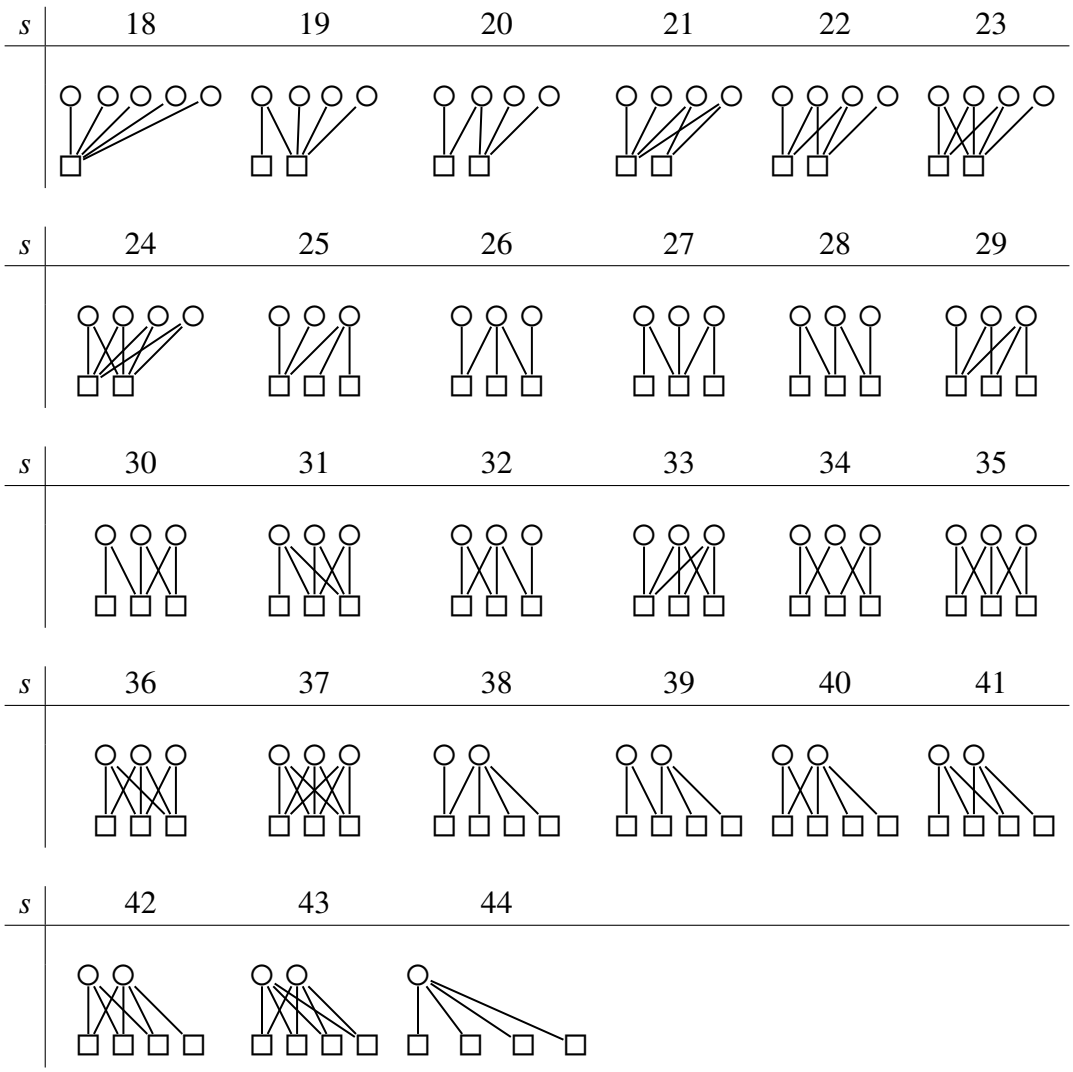


Figure 8: Bipartite motifs of size 6 as given in Simmons et al. [2019b].