Motifs and bipartite networks

S. Robin

Sorbonne université

Ecole EcoNet, April 2024, Montpellier

Outline

Outline

Bipartite network

Two types of actors.

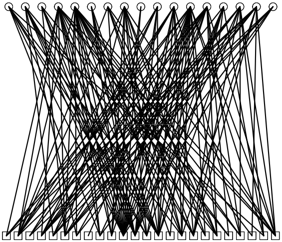
- Mutualistic: plant-pollinator
- Antagonistic: host-parasite

Topological analysis:
understanding the network organisation

Local: node or edge properties (degree, betweenness)

Global: density, connected components, nestedness

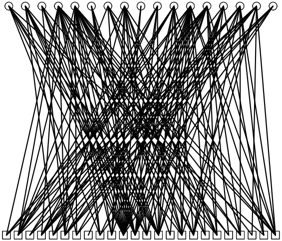
Zackenberg network: [SROB16]



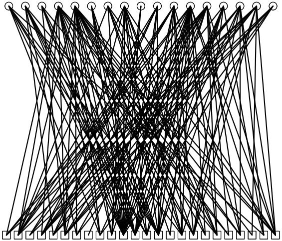
Bipartite network: notations

Species.

- $i = 1, \dots m$ insects = rows = bottom nodes
- $j = 1, \dots n$ plants = columns = top nodes

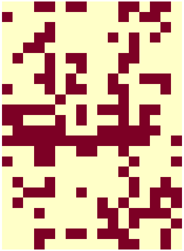


Bipartite network: notations



Species.

- $i = 1, \dots m$ insects = rows = bottom nodes
- $j = 1, \dots n$ plants = columns = top nodes



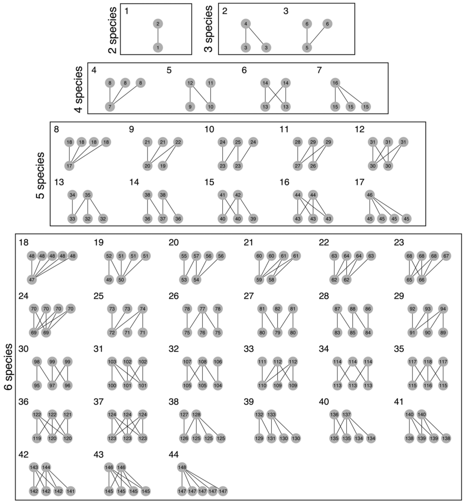
Interactions.

- $A_{ij} = 1$ if insect i interacts with plant j, 0 otherwise
$$A_{ij} = 1 \quad \Leftrightarrow \quad i \sim j$$
- adjacency matrix : $m \times n$
$$A = [A_{ij}]_{1 \leq i \leq m, 1 \leq j \leq n}$$

Bipartite motifs

'Meso-scale' analysis. [SCB+19]

- Motifs ='building-blocks'
- between local (several nodes) and global (sub-graph)



Bipartite motifs

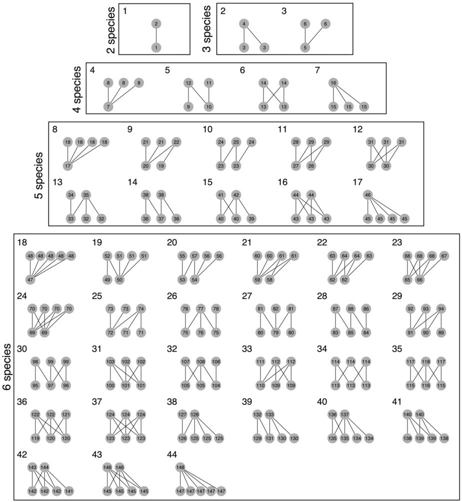
'Meso-scale' analysis. [SCB+19]

- Motifs ='building-blocks'
- between local (several nodes) and global (sub-graph)

Interest.

- Generic description of a network
- Enables network comparison
- Even when the nodes are different

(+ 'species-role': out of the scope here)



Bipartite motifs

'Meso-scale' analysis. [SCB+19]

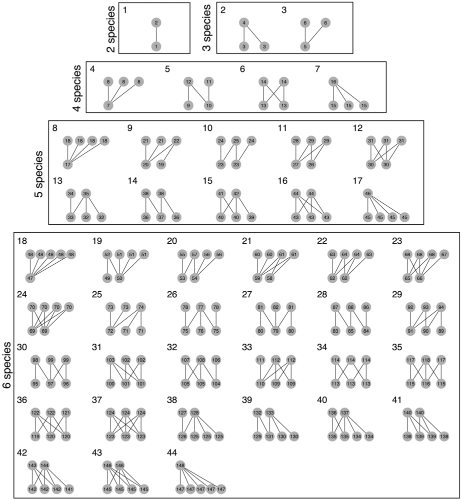
- Motifs ='building-blocks'
- between local (several nodes) and global (sub-graph)

Interest.

- Generic description of a network
- Enables network comparison
- Even when the nodes are different

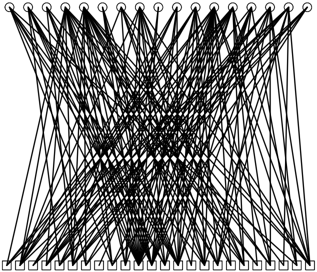
(+ 'species-role': out of the scope here)

Existing tool. `bmotif` package [SSS+19]: counts motif occurrences



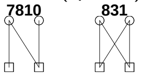
Example

Plant-pollinator network [SROB16]

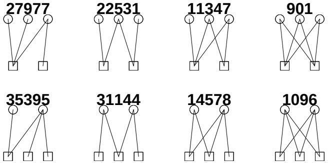


Motif counts.

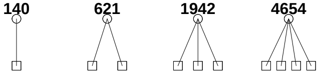
4 nodes (species)



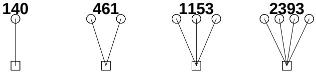
5 nodes



top 'stars' (plants)



bottom 'stars' (pollinators)



Outline

Need for a null model

Motif counts obviously depend on

- the size of the network: $n \times m$
- the density of the network
- the imbalance between bottom-node degrees (specialist vs generalist insects)
- the imbalance between top-node degrees (specialist vs generalist plants)

Need for a null model

Motif counts obviously depend on

- the size of the network: $n \times m$
- the density of the network
- the imbalance between bottom-node degrees (specialist vs generalist insects)
- the imbalance between top-node degrees (specialist vs generalist plants)

Bipartite expected degree distribution (BEDD) model: (in words)

- Consider m insects ($i = 1, \dots m$):
  each plant i has a specific propensity to interact (degree of generalism)

- Consider n plants ($j = 1, \dots n$):
  each plant j has a specific propensity to interact (idem)

- The probability for insect i and plant j to interact is proportional to the product of their respective propensities.

BEDD model

Bipartite expected degree distribution (BEDD) model: (more formaly)

- $\rho =$ network density
- $g =$ top node degree imbalance ($\int g = 1$)
- $h =$ bottom node degree imbalance ($\int h = 1$)

$$\{U_i\}_{i=1,\dots m} \text{ iid } \sim \mathcal{U}[0,1] \qquad \{V_j\}_{j=1,\dots n} \text{ iid } \sim \mathcal{U}[0,1]$$

$$\mathbb{P}\{i \sim j \mid U_i, V_j\} = \rho \; g(U_i) \; h(V_j)$$

(Bipartite version of the EDD model [CL02])

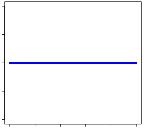
BEDD model

$$\mathbb{P}\{i \sim j \mid U_i, V_j\} = \rho\, g(U_i)\, h(V_j)$$

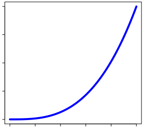
$$\mathbb{E}(D_i \mid U_i) = n\, \rho\, g(U_i)$$

$$\mathbb{E}(D_j \mid V_j) = m\, \rho\, g(V_i)$$

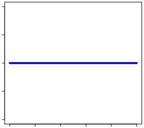
$h_0(v) =$

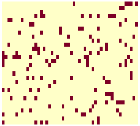


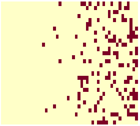
$h(v) =$



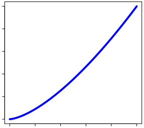
$g_0(u) =$

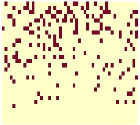


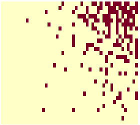




$g(u) =$







Properties of the BEDD model

Assumptions.

- No preferred or avoided specific connexion
- Graph-exchangeable model: insects and plants can be permuted simultaneously

Properties of the BEDD model

Assumptions.

- No preferred or avoided specific connexion
- Graph-exchangeable model: insects and plants can be permuted simultaneously

Properties.

- Expected degree for insect i given U_i : $n\, g(U_i)$.
- Expected degree for plant j given V_j : $m\, h(V_j)$.
- 'Nested' structure by construction

Properties of the BEDD model

Assumptions.

- No preferred or avoided specific connexion
- Graph-exchangeable model: insects and plants can be permuted simultaneously

Properties.

- Expected degree for insect i given U_i : $n\,g(U_i)$.
- Expected degree for plant j given V_j : $m\,h(V_j)$.
- 'Nested' structure by construction

Sufficient statistics to fit BEDD:

- Insect degrees + plant degrees
- or, equivalently, star (single edge, top, bottom) frequencies

Outline

Counting motifs

Number of positions.

- Choose p nodes among m
- Choose q nodes among n
- Try all *automorphisms*

$$c_s := \left(\begin{array}{c} m \\ p \end{array} \right) \times \left(\begin{array}{c} n \\ q \end{array} \right) \times r_s$$

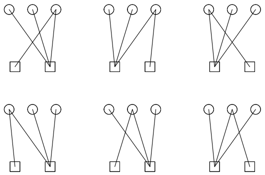
Counting motifs

Number of positions.

- Choose p nodes among m
- Choose q nodes among n
- Try all *automorphisms*

$$c_s := \binom{m}{p} \times \binom{n}{q} \times r_s$$

Automorphisms= non-redundant permutations



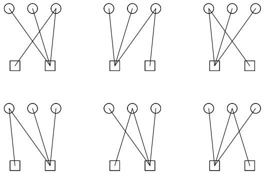
Counting motifs

Number of positions.

- Choose p nodes among m
- Choose q nodes among n
- Try all *automorphisms*

$$c_s := \binom{m}{p} \times \binom{n}{q} \times r_s$$

Automorphisms= non-redundant permutations



Motif count. Try all positions $\alpha = 1, \dots c_s$, define

$$Y_{s\alpha} = 1 \text{ if match,} \qquad 0 \text{ otherwise,}$$

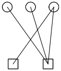
then count the number of matches:

$$N_s = \sum_\alpha Y_{s\alpha}$$

$\rightarrow$ Motif frequency: $F_s := N_s / c_s$

Motif probability

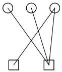
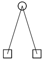
Occurrence probability $\overline{\phi}_s = \mathbb{P}\{Y_{s\alpha} = 1\}$. Under the B-EDD model [OLR22]:

$\overline{\phi}_s := \mathbb{P}_{BEDD} \Bigg($  $\Bigg) = \frac{\qquad}{\qquad}$

Motif probability

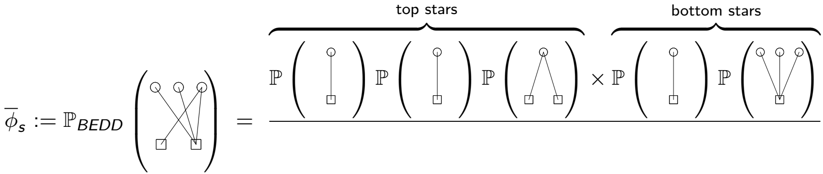
Occurrence probability $\overline{\phi}_s = \mathbb{P}\{Y_{s\alpha} = 1\}$. Under the B-EDD model [OLR22]:

top stars

$\overline{\phi}_s := \mathbb{P}_{BEDD}$ () $=$ $\mathbb{P}$ () $\mathbb{P}$ () $\mathbb{P}$ ()

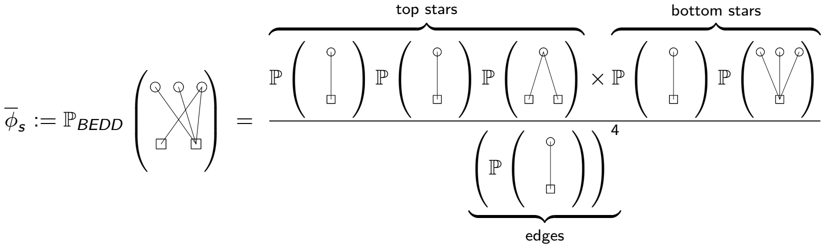
Motif probability

Occurrence probability $\overline{\phi}_s = \mathbb{P}\{Y_{s\alpha} = 1\}$. Under the B-EDD model [OLR22]:



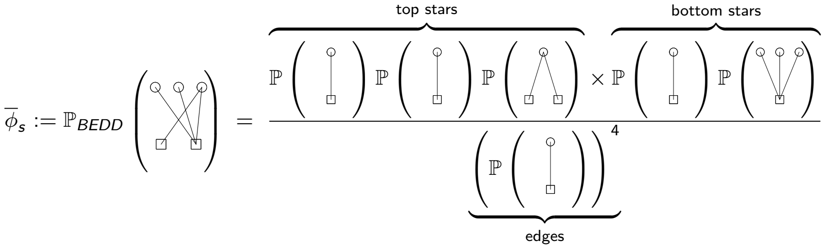
Motif probability

Occurrence probability $\overline{\phi}_s = \mathbb{P}\{Y_{s\alpha} = 1\}$. Under the B-EDD model [OLR22]:



Motif probability

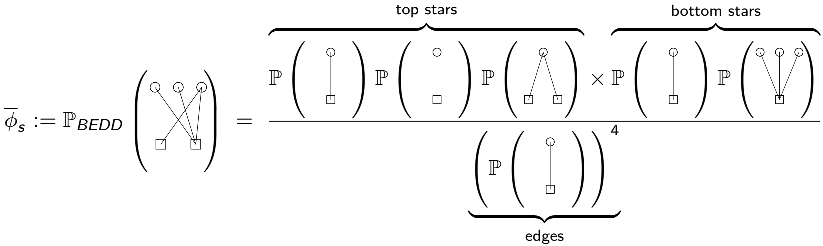
Occurrence probability $\overline{\phi}_s = \mathbb{P}\{Y_{s\alpha} = 1\}$. Under the B-EDD model [OLR22]:



$$= \frac{\phi_1^2 \phi_2 \phi_1 \phi_4}{\phi_1^4} \qquad = \frac{\phi_2 \phi_4}{\phi_1}$$

Motif probability

Occurrence probability $\overline{\phi}_s = \mathbb{P}\{Y_{s\alpha} = 1\}$. Under the B-EDD model [OLR22]:



$$= \frac{\phi_1^2 \phi_2 \phi_1 \phi_4}{\phi_1^4} \qquad = \frac{\phi_2 \phi_4}{\phi_1}$$

Estimated probability.

$$\overline{\phi}_s := \frac{\phi_2 \phi_4}{\phi_1} \qquad \rightarrow \qquad \overline{F}_s := \frac{F_2 F_4}{F_1}$$

where F_1, F_2, F_4 = observed frequencies of edges, top stars and bottom stars.

Moments of the count

- Mean: $\mathbb{E}_{BEDD}(N_s) = c_s \times \overline{\phi}_s$

Moments of the count

- Mean: $\mathbb{E}_{BEDD}(N_s) = c_s \times \overline{\phi}_s$
- Variance: Same game, requires to evaluate $\mathbb{E}_{BEDD}(N_s^2) = \mathbb{E}_{BEDD}\left(\sum_\alpha Y_{s\alpha}\right)^2$

  $\rightarrow$ Need to account for overlap between positions (*super-motifs*: [PDK+08])

Moments of the count

- Mean: $\mathbb{E}_{BEDD}(N_s) = c_s \times \overline{\phi}_s$

- Variance: Same game, requires to evaluate $\mathbb{E}_{BEDD}(N_s^2) = \mathbb{E}_{BEDD}\left(\sum_\alpha Y_{s\alpha}\right)^2$

  $\rightarrow$ Need to account for overlap between positions (*super-motifs*: [PDK+08])

- Covariance: Same game to compute $\mathbb{C}\mathrm{ov}(N_s, N_{s'})$

Distribution of the count

Asymptotic normality for non-star motifs. Under BEDD (and sparsity conditions):

$$(F_s - \overline{F}_s) \Big/ \sqrt{\widehat{\mathbb{V}}(F_s)} \quad \overset{m,n\to\infty}{\longrightarrow} \quad \mathcal{N}(0,1)$$

Proof [OLR22]:

$$\text{decompose} \qquad F_s - \overline{F}_s = \underbrace{(F_s - \phi_s)}_{\text{random fluctuations}} + \underbrace{(\phi_s - \overline{\phi}_s)}_{\text{null under BEDD}} + \underbrace{(\overline{\phi}_s - \overline{F}_s)}_{\text{estimation error }\to 0},$$

$+$ construct a counting martingale for $F_s - \phi_s$ [GL17].

Distribution of the count

Asymptotic normality for non-star motifs. Under BEDD (and sparsity conditions):

$$(F_s - \overline{F}_s) \Big/ \sqrt{\widehat{\mathbb{V}}(F_s)} \quad \xrightarrow{m,n\to\infty} \quad \mathcal{N}(0,1)$$

Proof [OLR22]:

$$\text{decompose} \qquad F_s - \overline{F}_s = \underbrace{(F_s - \phi_s)}_{\text{random fluctuations}} + \underbrace{(\phi_s - \overline{\phi}_s)}_{\text{null under BEDD}} + \underbrace{(\overline{\phi}_s - \overline{F}_s)}_{\text{estimation error} \to 0},$$

$+$ construct a counting martingale for $F_s - \phi_s$ [GL17].

Consequence. We know the expected behavior of motif counts under the BEDD model ($=$ 'null model').

Outline

Goodness-of-fit (GOF)

Aim. Test if the observed data arise from a given model.

More cautious: 'if the model fits the data reasonably well'

Goodness-of-fit (GOF)

Aim. Test if the observed data arise from a given model.

More cautious: 'if the model fits the data reasonably well'

Typical approach.

1. Define some statistic (= function of the data) T,
2. Establish the distribution of T under the model,
3. Compare the observed value of T with its distribution under the model.

Goodness-of-fit (GOF)

Aim. Test if the observed data arise from a given model.

More cautious: 'if the model fits the data reasonably well'

Typical approach.

1. Define some statistic (= function of the data) T,
2. Establish the distribution of T under the model,
3. Compare the observed value of T with its distribution under the model.

Example.

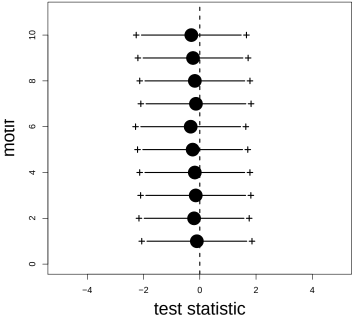
1. Data = observed plant-pollinator network
2. Statistic T = motif count N_s
3. Model = BEDD

Goodness-of-fit (GOF) of the BEDD model

Zackenberg network.

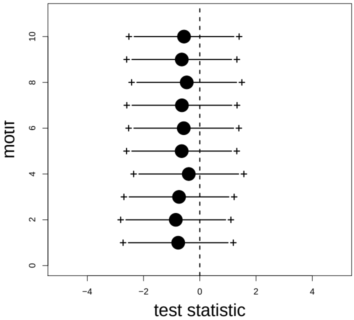
Raw statistic:

$$T_s = \frac{N_s - \widehat{\mathbb{E}} N_s}{\sqrt{\widehat{\mathbb{V}} N_s}}$$



Goodness-of-fit (GOF) of the BEDD model

Zackenberg network.



Raw statistic:

$$T_s = \frac{N_s - \widehat{\mathbb{E}} N_s}{\sqrt{\widehat{\mathbb{V}} N_s}}$$

Corrected stat.: accounts for the estimation error in $\widehat{\mathbb{E}} N$

$$T_s' = \frac{N_s - (\widehat{\mathbb{E}} N_s - \widehat{\mathbb{B}}(\widehat{\mathbb{E}} N_s))}{\sqrt{\widehat{\mathbb{V}}(N_s - \widehat{\mathbb{E}} N_s)}}$$

Testing degree imbalance

Question. Is there some degree imbalance between plants?

Testing degree imbalance

Question. Is there some degree imbalance between plants?

Statistical test.

- Assume $A \sim BEDD(\rho, g, h)$,
$$H_0 = \{h = 1\}$$
- For motif s, evaluate $\widehat{\mathbb{E}}_0(N_s)$ and $\widehat{\mathbb{E}}_0(N_s)$ and compare
$$W_s = (N_s - \widehat{\mathbb{E}}_0(N_s))/\sqrt{\widehat{\mathbb{V}}_0(N_s)}$$
with $\mathcal{N}(0, 1)$

Testing degree imbalance

Question. Is there some degree imbalance between plants?

Statistical test.

- Assume $A \sim BEDD(\rho, g, h)$,
$$H_0 = \{h = 1\}$$
- For motif s, evaluate $\widehat{\mathbb{E}}_0(N_s)$ and $\widehat{\mathbb{E}}_0(N_s)$ and compare
$$W_s = (N_s - \widehat{\mathbb{E}}_0(N_s))/\sqrt{\widehat{\mathbb{V}}_0(N_s)}$$
with $\mathcal{N}(0, 1)$

Example. (only one significant difference)

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}

Comparing network imbalances

Question. Do network A and B share the same imbalance for insects?

Comparing network imbalances

Question. Do network A and B share the same imbalance for insects?

Statistical test.

- Assume $A \sim BEDD(\rho^A, g^A, h^A)$ and $B \sim BEDD(\rho^B, g^B, h^B)$
$$H_0 = \{g^A = g^B\}$$
- For motif s, evaluate $\widehat{\mathbb{E}}_{\widehat{\rho}^A, \widehat{g}^B, \widehat{g}^A}(N_s^A)$ and $\widehat{\mathbb{E}}_{\widehat{\rho}^B, \widehat{g}^A, \widehat{g}^B}(N_s^B)$ and compare
$$W_s = \frac{(N_s^A - \widehat{\mathbb{E}}_0(N_s^A)) - (N_s^B - \widehat{\mathbb{E}}_0(N_s^B))}{\sqrt{\widehat{\mathbb{V}}_0(N_s^A) + \widehat{\mathbb{V}}_0(N_s^B)}}$$

with $\mathcal{N}(0, 1)$

Comparing network imbalances

Question. Do network A and B share the same imbalance for insects?

Statistical test.

- Assume $A \sim BEDD(\rho^A, g^A, h^A)$ and $B \sim BEDD(\rho^B, g^B, h^B)$

$$H_0 = \{g^A = g^B\}$$

- For motif s, evaluate $\widehat{\mathbb{E}}_{\widehat{\rho}^A, \widehat{g}^B, \widehat{g}^A}(N_s^A)$ and $\widehat{\mathbb{E}}_{\widehat{\rho}^B, \widehat{g}^A, \widehat{g}^B}(N_s^B)$ and compare

$$W_s = \frac{(N_s^A - \widehat{\mathbb{E}}_0(N_s^A)) - (N_s^B - \widehat{\mathbb{E}}_0(N_s^B))}{\sqrt{\widehat{\mathbb{V}}_0(N_s^A) + \widehat{\mathbb{V}}_0(N_s^B)}}$$

with $\mathcal{N}(0,1)$

Example. (no significant difference)

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

Outline

Network embedding: Multivariate analysis

Analysing multiple networks. Principle

- 'Embed' each network into a convenient space (e.g. $\mathbb{R}^d$)
- Use standard multivariate analysis (clustering, PCA, MDS, ...)

Network embedding: Multivariate analysis

Analysing multiple networks. Principle

- 'Embed' each network into a convenient space (e.g. $\mathbb{R}^d$)
- Use standard multivariate analysis (clustering, PCA, MDS, ...)

Using motifs. K networks

$$(\text{Network})_k \quad \rightarrow \quad (N_1^k, \dots, N_S^k) \in \mathbb{R}^S$$

but need to correct for: network sizes, correlation between motif frequencies, etc...

Network embedding: Multivariate analysis

Analysing multiple networks. Principle

- 'Embed' each network into a convenient space (e.g. $\mathbb{R}^d$)
- Use standard multivariate analysis (clustering, PCA, MDS, ...)

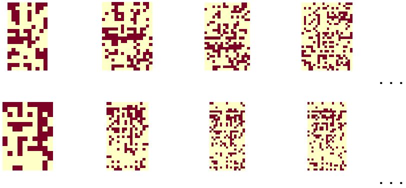
Using motifs. K networks

$$(\text{Network})_k \quad \to \quad (N_1^k, \dots, N_S^k) \in \mathbb{R}^S$$

but need to correct for: network sizes, correlation between motif frequencies, etc...

Zackenberg dataset. $K = 46$ networks

- 2 years
- 1 network observed every few days

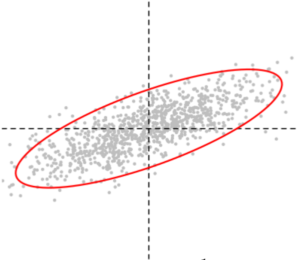


Choleski transform

Aim: 'Remove' correlation and variance heterogeneity

Covariance matrix (X_1, X_2):

$$\Sigma(X_1, X_2) = \left[\begin{array}{cc} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{array} \right]$$



Diagonalization: $\Sigma = P \Lambda P^{-1}$

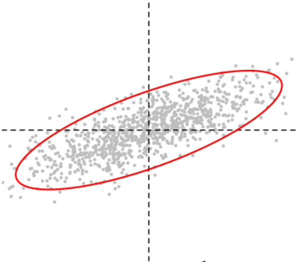
Choleski matrix: $\Sigma^{-1/2} = P \Lambda^{-1/2} P^{-1}$

Choleski transform

Aim: 'Remove' correlation and variance heterogeneity

Covariance matrix (X_1, X_2):

$$\Sigma(X_1, X_2) = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}$$

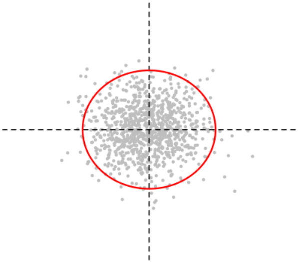


Diagonalization: $\Sigma = P \Lambda P^{-1}$

Choleski matrix: $\Sigma^{-1/2} = P \Lambda^{-1/2} P^{-1}$

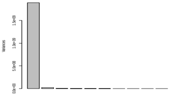
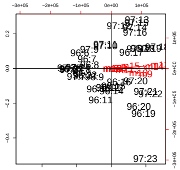
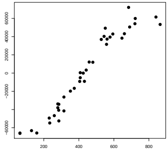
Choleski transform:

$$\begin{bmatrix} X_1' \\ X_2' \end{bmatrix} = \Sigma^{-1/2} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$$

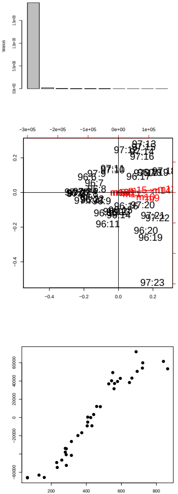
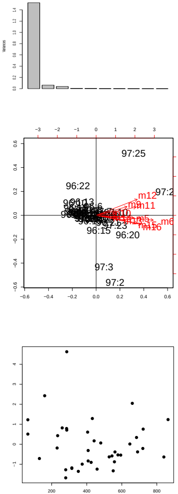


$$\Sigma(X_1', X_2') = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

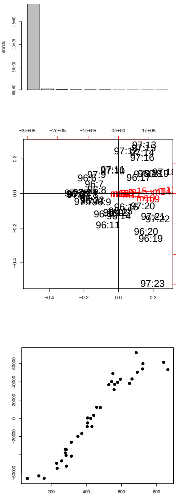
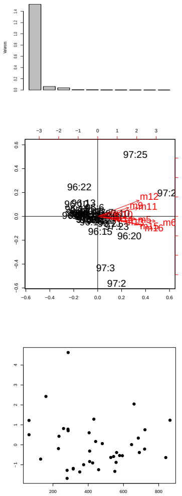
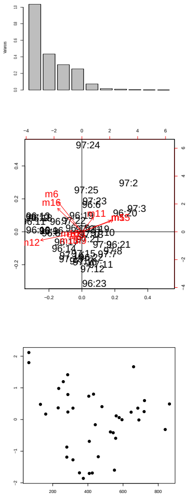
Network embedding: Zackenberg's data [SROB16]

Raw counts			
			
			
			

Network embedding: Zackenberg's data [SROB16]

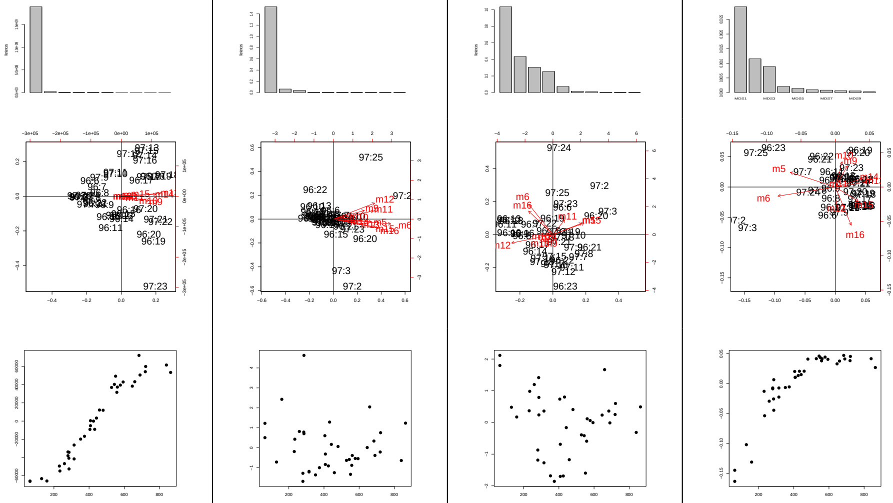
Raw counts	Corrected stat.		
			

Network embedding: Zackenberg's data [SROB16]

Raw counts	Corrected stat.	Choleski
		

Network embedding: Zackenberg's data [SROB16]

Raw counts	Corrected stat.	Choleski	Bray-Curtis MDS



References I

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

C. Gao and J. Lafferty. Testing for global network structure using small subgraph statistics. Technical Report 1704.06742, arXiv, 2017.

S. Ouadah, P. Latouche, and S. Robin. Motif-based tests for bipartite networks. *Electronic Journal of Statistics*, 16(1):293 – 330, 2022.

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.

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, 2019.

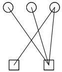
S. Saavedra, R. P Rohr, J. M Olesen, and J. Bascompte. Nested species interactions promote feasibility over stability during the assembly of a pollinator community. *Ecology and evolution*, 6(4):997–1007, 2016.

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, 2019.

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

Super-motifs

Motif:



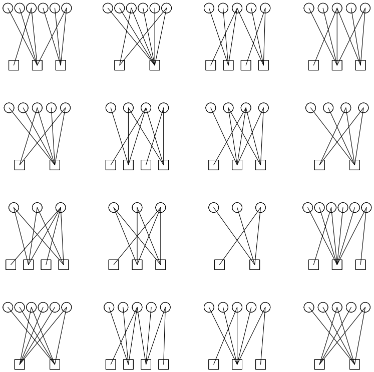
Variance:

$$N_s^2 = \left(\sum_\alpha Y_{s\alpha}\right)^2$$

$$= \sum_{\alpha,\beta:\alpha\cap\beta=\emptyset} Y_{s\alpha} Y_{s\beta}$$

$$+ \sum_{\alpha,\beta:\alpha\cap\beta\neq\emptyset} \underbrace{Y_{s\alpha} Y_{s\beta}}_{\text{occurrence of a super-motif}}$$

Some super-motifs:



... 396 super-motifs

Covariance: same game, for $Y_{s\alpha} Y_{s'\beta}$ with $s \neq s'$

Asymptotic distribution of the count

Estimated probability.

$$\overline{\phi}_s := \phi_2 \phi_4 \,/ \phi_1 \qquad \to \qquad \overline{F}_s := F_2 F_4 \,/ F_1$$

where F_1, F_2, F_4 = observed frequencies of top stars, bottom stars and edges.

Asymptotic distribution of the count

Estimated probability.

$$\overline{\phi}_s := \phi_2\phi_4 \,/\phi_1 \qquad \to \qquad \overline{F}_s := F_2F_4 \,/F_1$$

where F_1, F_2, F_4 = observed frequencies of top stars, bottom stars and edges.

Asymptotic normality for non-star motifs. Under BEDD (and sparsity conditions):

$$(F_s - \overline{F}_s) \Big/ \sqrt{\widehat{\mathbb{V}}(F_s)} \quad \xrightarrow{m,n\to\infty} \quad \mathcal{N}(0,1)$$

Proof:

- decompose

$$F_s - \overline{F}_s = \underbrace{(F_s - \phi_s)}_{\text{random fluctuations}} + \underbrace{(\phi_s - \overline{\phi}_s)}_{\text{null under BEDD}} + \underbrace{(\overline{\phi}_s - \overline{F}_s)}_{\text{estimation error }\to 0},$$

- construct a counting martingale [GL17] for $F_s - \phi_s$

Asymptotic distribution of the count

Estimated probability.

$$\overline{\phi}_s := \phi_2\phi_4 / \phi_1 \qquad \to \qquad \overline{F}_s := F_2 F_4 / F_1$$

where F_1, F_2, F_4 = observed frequencies of top stars, bottom stars and edges.

Asymptotic normality for non-star motifs. Under BEDD (and sparsity conditions):

$$(F_s - \overline{F}_s) \Big/ \sqrt{\widehat{\mathbb{V}}(F_s)} \quad \xrightarrow{m,n\to\infty} \quad \mathcal{N}(0,1)$$

Proof:

- decompose

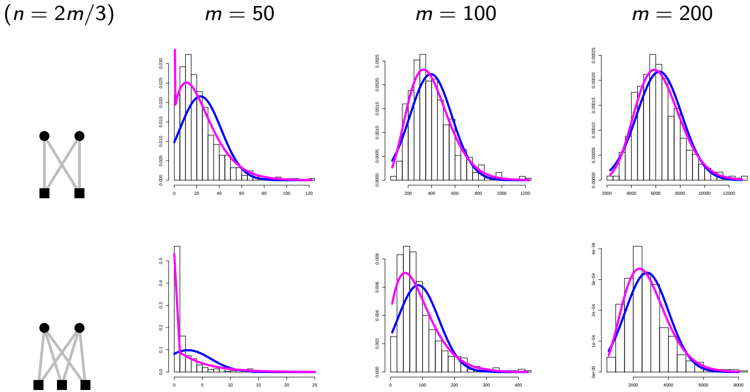
$$F_s - \overline{F}_s = \underbrace{(F_s - \phi_s)}_{\text{random fluctuations}} + \underbrace{(\phi_s - \overline{\phi}_s)}_{\text{null under BEDD}} + \underbrace{(\overline{\phi}_s - \overline{F}_s)}_{\text{estimation error } \to 0},$$

- construct a counting martingale [GL17] for $F_s - \phi_s$

Test statistic. Under BEDD:

$$N_s \approx \mathcal{N}\left(\widehat{\mathbb{E}}(N_s), \widehat{\mathbb{V}}(N_s)\right) \qquad \Leftrightarrow \qquad \left(N_s - \widehat{\mathbb{E}}(N_s)\right) \Big/ \sqrt{\widehat{\mathbb{V}}(N_s)} \approx \mathcal{N}(0,1)$$

In practice: Asymptotic normality

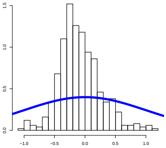


Normal distribution, Poisson-geometric distribution with same mean and variance [Sta01,PDK+08]

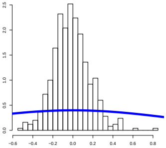
In practice: Test statistic

Need to account for the estimation error of $\widehat{\mathbb{E}}N$

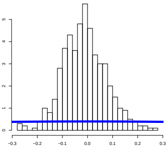
$m = 50$



$m = 100$



$m = 200$



Regular stat.:

$$\frac{N - \widehat{\mathbb{E}}N}{\sqrt{\widehat{\mathbb{V}}N}}$$

In practice: Test statistic

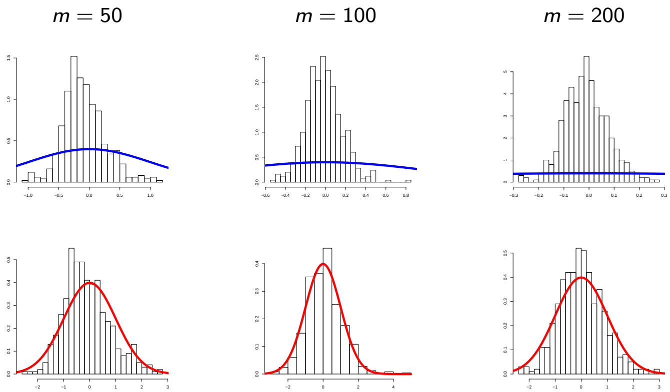
Need to account for the estimation error of $\widehat{\mathbb{E}}N$

Regular stat.:

$$\frac{N - \widehat{\mathbb{E}}N}{\sqrt{\widehat{\mathbb{V}}N}}$$

Correction:

$$\frac{N - (\widehat{\mathbb{E}}N - \widehat{\mathbb{B}}(\widehat{\mathbb{E}}N))}{\sqrt{\widehat{\mathbb{V}}(N - \widehat{\mathbb{E}}N)}}$$



- Need to evaluate $\mathbb{V}(N - \widehat{\mathbb{E}}(N))$ and $\mathbb{B}(\widehat{\mathbb{E}}N)$: resort to Taylor expansion ($\Delta-$method)