# $U$-statistics of row-column exchangeable matrices. Application to ecological network analysis.

Tâm Le Minh (Université Paris-Saclay, MIA Paris-Saclay)

Stéphane Robin, Sophie Donnet, François Massol

26/07/2023

54èmes Journées de Statistique de la Société Française de Statistique

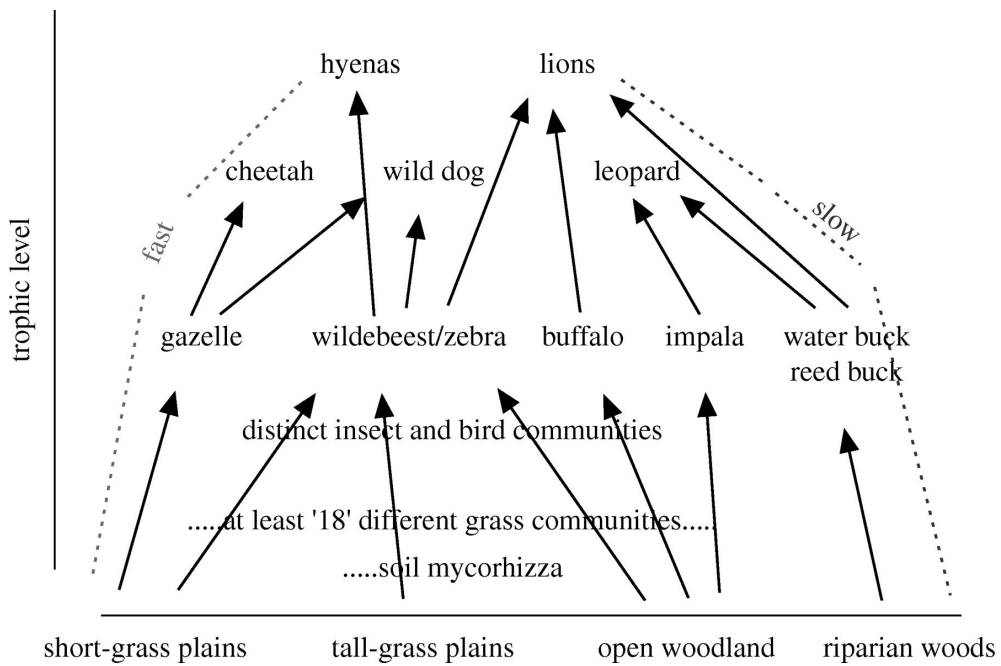
# Ecological networks

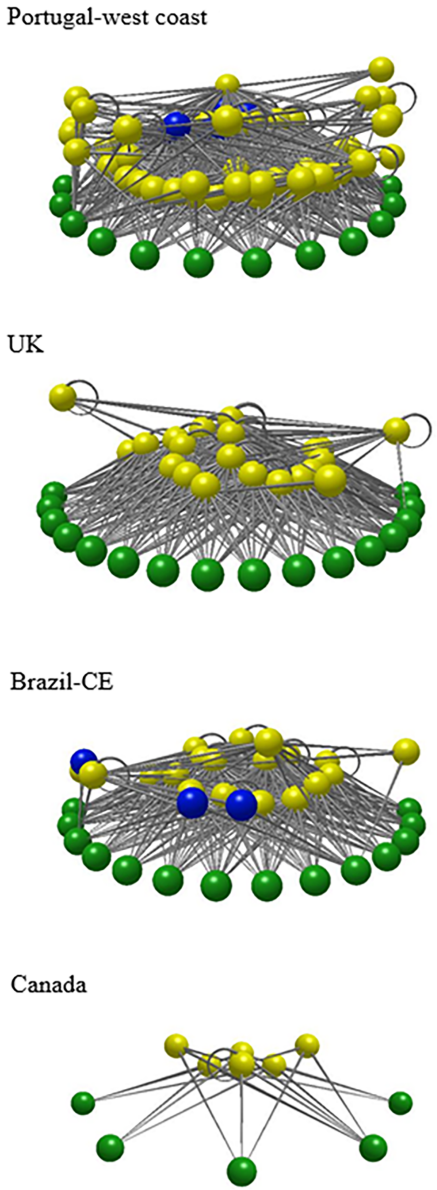
Species interactions

→ organization of an ecosystem

Network variability

→ reaction to external perturbations

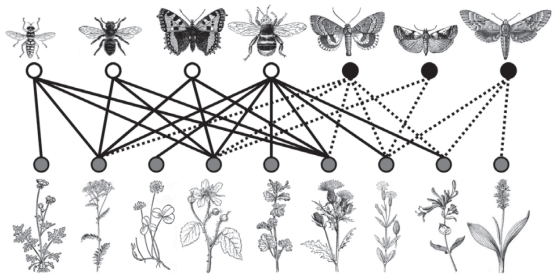


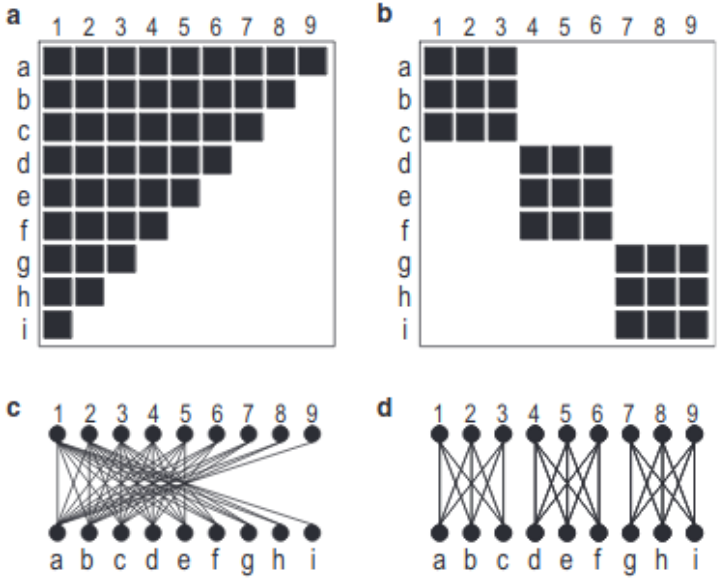



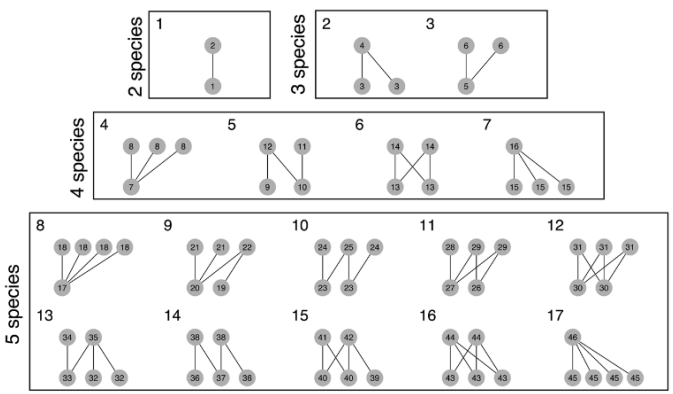
# Ecological network analysis

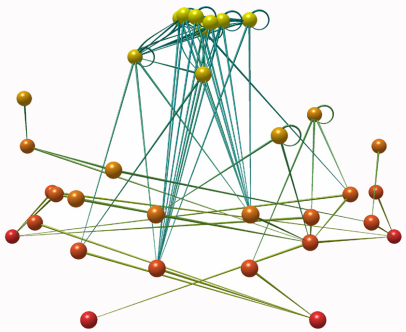
Network statistics:

- Connectance, Nestedness, Modularity,
- Subgraph densities,
- etc.

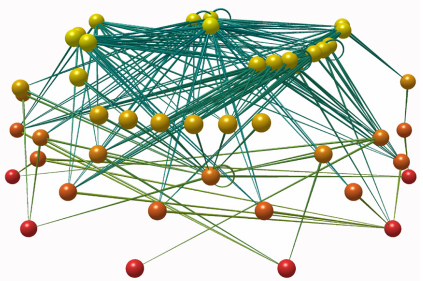




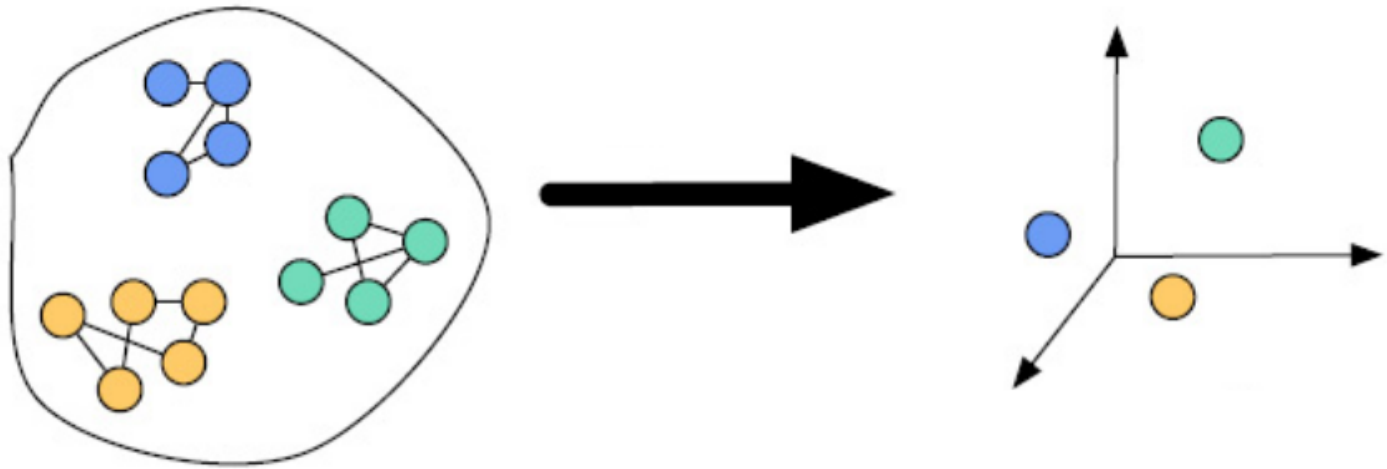
$S = 33$, $L = 99$, $C = 0.091$
$TL = 2.84$, Max$TL = 4.36$



$S = 48$, $L = 249$, $C = 0.108$
$TL = 2.72$, Max$TL = 3.78$

# General approach

Representation of a network by a vector: graph embedding



Our approach: "mapping" in a space of probabilistic network models

- Observed network = realization of a random model $\rightsquigarrow$ exchangeable networks
- Compare networks = compare models $\rightsquigarrow$ $U$-statistics

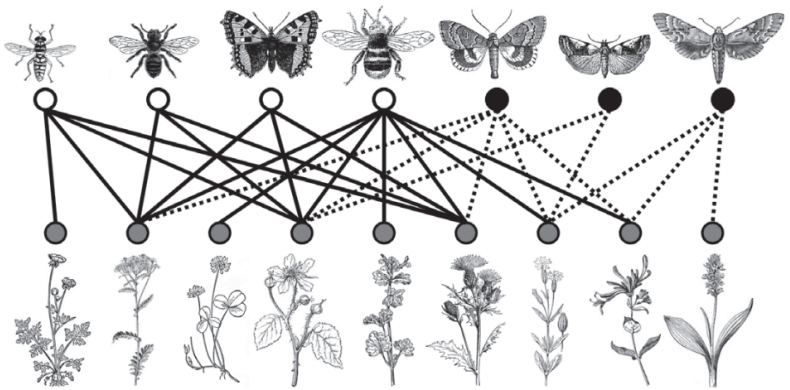
# Row-column exchangeable matrices

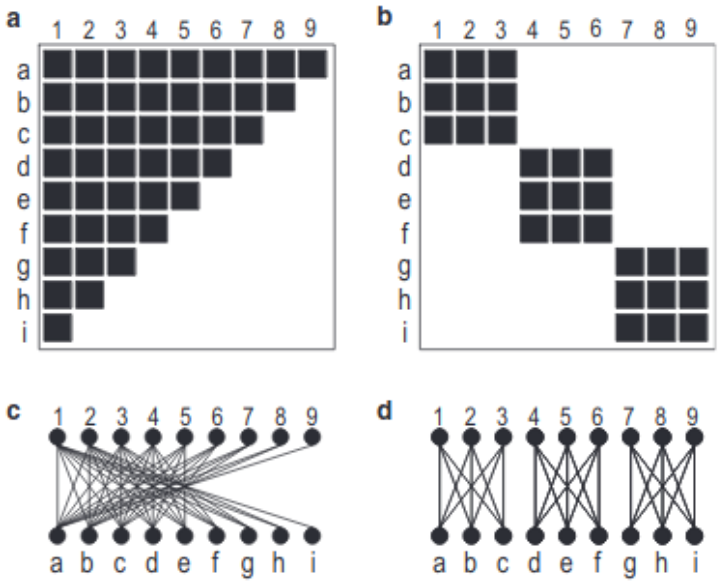
## Definition

A matrix $Y$ is row-column exchangeable (RCE) if for any permutations $\sigma_1$ and $\sigma_2$ of $\mathbb{N}$,

$$Y \overset{\mathcal{D}}{=} (Y_{\sigma_1(i),\sigma_2(j)})_{i\geq 1, j\geq 1}$$

Motivation: exchangeable bipartite networks





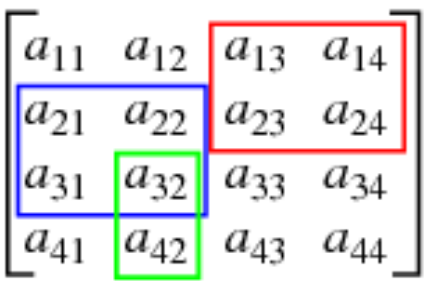
# Dissociation

## Definition

A matrix $Y$ is dissociated

$\Leftrightarrow$

For all $(m, n) \in \mathbb{N}^2$, $(Y_{ij})_{i \leq m, j \leq n}$ and $(Y_{ij})_{i > m, j > n}$ are independent.




*submatrices*

# $U$-statistics

Symmetric function of a $p \times q$ submatrix

$$h(Y_{\{i_1,\ldots,i_p;j_1,\ldots,j_q\}}) = h\left(\begin{bmatrix} Y_{i_1j_1} & Y_{i_1j_2} & \cdots & Y_{i_1j_q} \\ Y_{i_2j_1} & Y_{i_2j_2} & \cdots & Y_{i_2j_q} \\ \cdots & \cdots & \cdots & \cdots \\ Y_{i_pj_1} & Y_{i_pj_2} & \cdots & Y_{i_pj_q} \end{bmatrix}\right)$$

## $U$-statistic over a $m \times n$ network

$$U^h_{m,n} = \left[\binom{m}{p}\binom{n}{q}\right]^{-1} \sum_{\substack{\boldsymbol{i}\in\mathcal{P}_p([m]) \\ \boldsymbol{j}\in\mathcal{P}_q([n])}} h(Y_{\boldsymbol{i},\boldsymbol{j}})$$

If $Y$ is a RCE matrix, then $\mathbb{E}[U^h_{m,n}] = \mathbb{E}[h(Y_{\{1,\ldots,p;1,\ldots,q\}})]$.

# Outline

# Bipartite expected degree distribution (BEDD) models

## Poisson-BEDD model

$$\begin{array}{rcl} U_i, V_j & \overset{iid}{\sim} & \mathcal{U}[0,1] \\ Y_{ij} \mid U_i, V_j & \sim & \mathcal{P}(\lambda f(U_i) g(V_j)) \end{array}$$

where

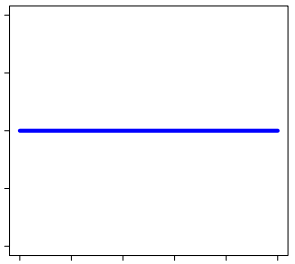
- $\lambda = \mathbb{E}[Y_{ij}]$
- $\int f = \int g = 1$.

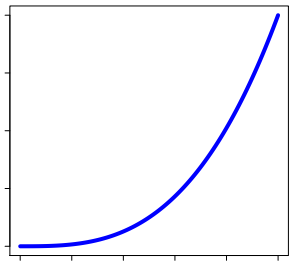
The Poisson-BEDD model is :

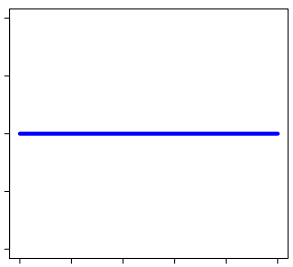
- a model with latent variables for the expected degrees of the nodes,
- RCE and dissociated,
- recoverable by a quadruplet of nodes.

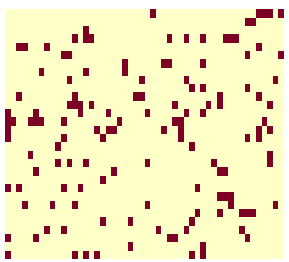
# BEDD models

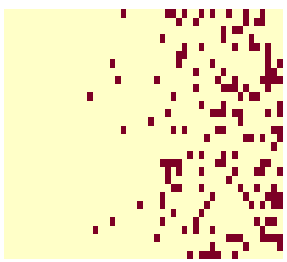
$$\begin{aligned} U_i, V_j &\overset{iid}{\sim} \mathcal{U}[0,1] \\ Y_{ij} \mid U_i, V_j &\sim \mathcal{B}(\lambda f(U_i) g(V_j)) \end{aligned}$$

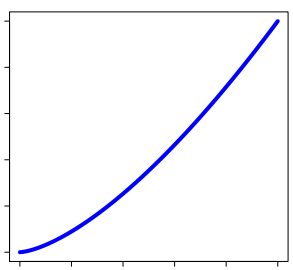
$g_0(v) =$ 

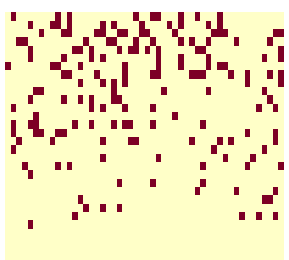
$g(v) =$ 

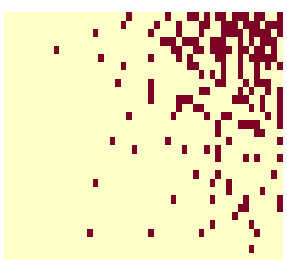
$f_0(u) =$ 





$f(u) =$ 





# Recoverability of the BEDD models

Moments of $f$ and $g$:
$\rightarrow F_k = \int f^k$, $G_k = \int g^k$

Some properties:
$\rightarrow \mathbb{E}[Y_{11}^2 - Y_{11}] = \lambda^2 F_2 G_2$
$\rightarrow \mathbb{E}[Y_{11} Y_{12}] = \lambda^2 F_2$
$\rightarrow \mathbb{E}[Y_{11} Y_{21}] = \lambda^2 G_2$

$\rightsquigarrow$ Functions on a $2 \times 2$ submatrix: $h(Y_{\{i_1,i_2;j_1,j_2\}}) = h\left(\begin{bmatrix} Y_{i_1 j_1} & Y_{i_1 j_2} \\ Y_{i_2 j_1} & Y_{i_2 j_2} \end{bmatrix}\right)$

## Recoverability theorem

The parameter $\lambda$ and the moments $F_k$ and $G_k$ of the Poisson-BEDD are recoverable by the joint distribution of a $2 \times 2$ submatrix.

# Outline

# Asymptotic distribution of $U$-statistics

$h$ is a function on a $2 \times 2$ submatrix.
$U_N^h := U_{m_N,n_N}^h$, where

- $N = m_N + n_N$,
- $m_N/N \to \rho \in ]0,1[$.

## Asymptotic normality of $U$-statistics (LM, 2023)

For RCE models, if $\mathbb{E}[h(Y_{\{1,2;1,2\}})^2] < \infty$, then

$$\sqrt{N}(U_N^h - \theta) \xrightarrow[N\to\infty]{\mathcal{D}} \mathcal{N}(0, V)$$

where

- $\theta = \mathbb{E}[h(Y_{\{1,2;1,2\}})]$,
- $V = \frac{4}{\rho}\mathrm{Cov}\big(h(Y_{\{1,2;1,2\}}), h(Y_{\{1,3;3,4\}})\big) + \frac{4}{1-\rho}\mathrm{Cov}\big(h(Y_{\{1,2;1,2\}}), h(Y_{\{3,4;1,3\}})\big)$.

# Variance estimation

Let $\widehat{V}_N$ be an estimator for $V$. If $\widehat{V}_N \xrightarrow{\mathbb{P}} V$, then

## Asymptotic normality of $U$-statistics

$$\sqrt{\frac{N}{\widehat{V}_N}}(U_N^h - \theta) \xrightarrow[N\to\infty]{\mathcal{D}} \mathcal{N}(0,1).$$

There are several consistent estimators for $V$:

- unbiased ("analytic estimator", Monte-Carlo),
- asymptotically unbiased (based on resampling, LM et al., 2023).

# Degeneracy

Analytic calculation of the variance:

$$\mathbb{V}[U_N^h] = \frac{V^{(1)}}{N} + \frac{V^{(2)}}{N^2} + \frac{V^{(3)}}{N^3} + \frac{V^{(4)}}{N^4} + o\left(\frac{1}{N^4}\right)$$

## Weak convergence in the degenerate case

If $0 = V^{(1)} = \ldots = V^{(d-1)} < V^{(d)}$, then

$$N^{\frac{d}{2}}(U_N^h - U_\infty^h) \xrightarrow[N\to\infty]{\mathcal{D}} W.$$

The limit distribution is not trivial in general, but there are cases where it is Gaussian.

# Plan

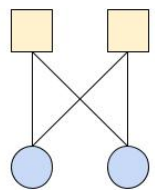
# Example 1: Subgraph densities

## Binary RCE network model

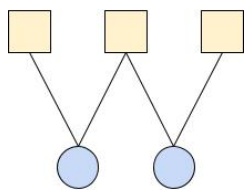
$w : \mathbb{R}^2 \to [0,1]$,

$$\begin{aligned} U_i, V_j &\overset{iid}{\sim} \mathcal{U}[0,1] \\ Y_{ij} \mid U_i, V_j &\sim \mathcal{B}(w(U_i, V_j)) \end{aligned}$$

$h_6(Y_{\{1,2;1,2\}}) = Y_{11}Y_{12}Y_{21}Y_{22}$

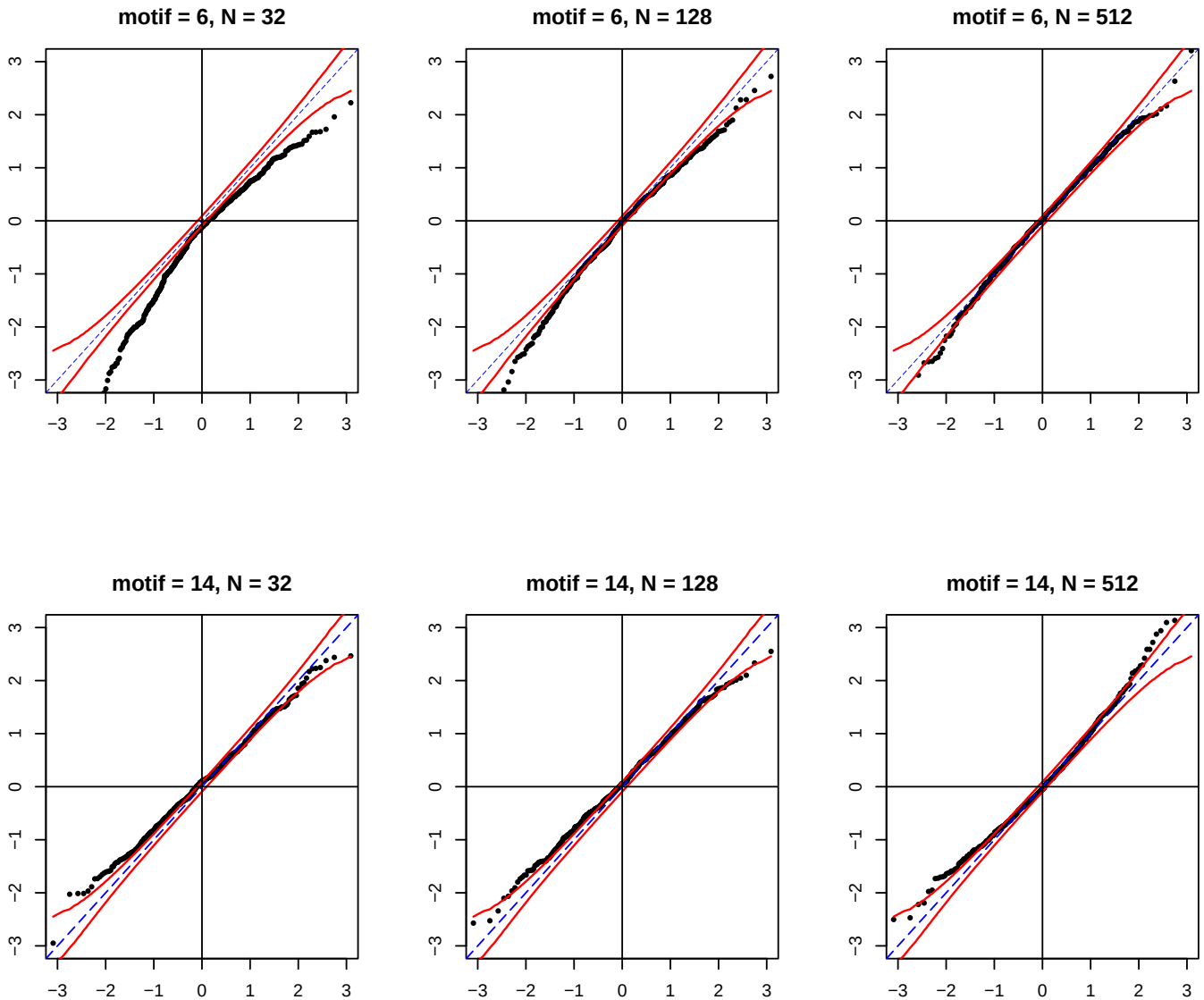


$h_{14}(Y_{\{1,2,3;1,2\}}) = Y_{11}Y_{21}Y_{22}Y_{32}(1-Y_{12})(1-Y_{31})$



The $U$-statistics $U_N^h$ are the densities of these motifs.

# Example 1: Subgraph density



# Example 2: Heterogeneity of the row degrees

## Poisson-BEDD model

$$\begin{aligned} U_i, V_j &\overset{iid}{\sim} \mathcal{U}[0,1] \\ Y_{ij} \mid U_i, V_j &\sim \mathcal{P}(\lambda f(U_i) g(V_j)) \end{aligned}$$

where

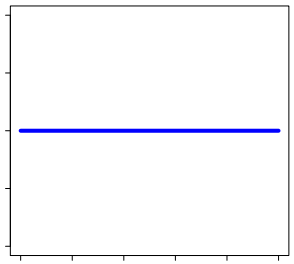
- $\lambda = \mathbb{E}[Y_{ij}]$
- $\int f = \int g = 1$, $\int f^k = F_k$, $\int g^k = G_k$.

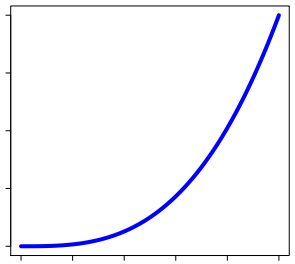
$f$ and $g$ characterize the degree distributions of the row and column nodes.

# Example 2: Heterogeneity of the row degrees

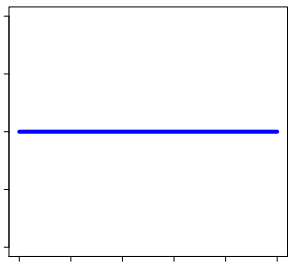
$g_0(v) =$ $g(v) =$

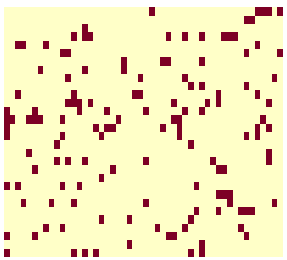
$\mathcal{H}_0 : f \equiv 1 \Leftrightarrow F_2 = 1$
$\mathcal{H}_1 : f \not\equiv 1 \Leftrightarrow F_2 > 1$

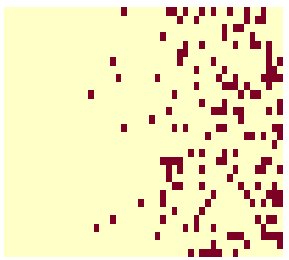




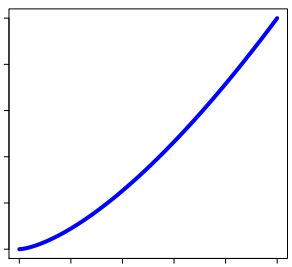
$f_0(u) =$

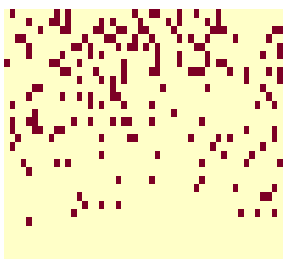


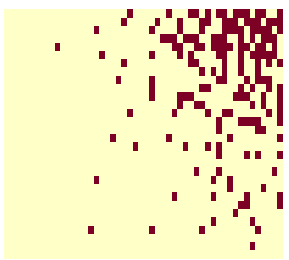




$f(u) =$







# Example 2: Heterogeneity of the row degrees

## U-statistic kernel

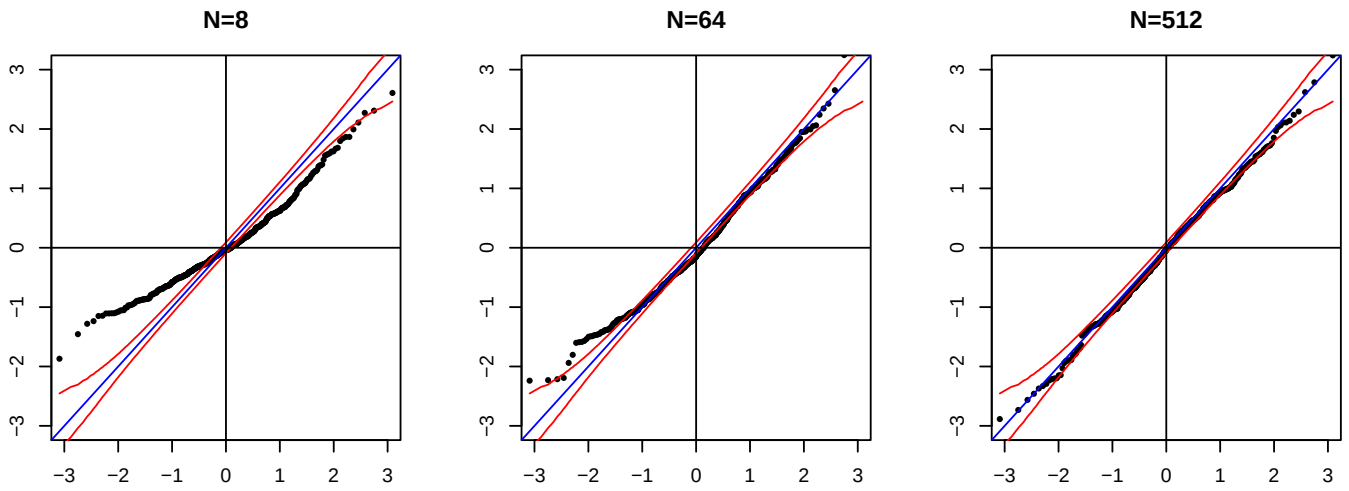
$h = h_1 - h_2$ where

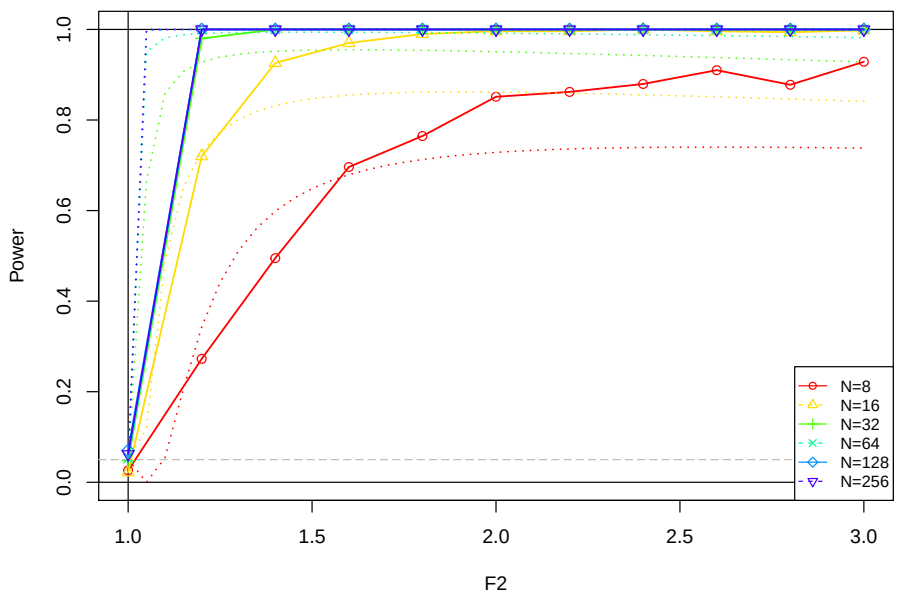
- $h_1(Y_{\{1,2;1,2\}}) = Y_{11}Y_{12}$, $\mathbb{E}h_1 = \lambda^2 F_2$
- $h_2(Y_{\{1,2;1,2\}}) = Y_{11}Y_{22}$, $\mathbb{E}h_2 = \lambda^2$

## Degeneracy under $\mathcal{H}_0$

$$\frac{N^{3/2}}{\sqrt{V}} U_N^h \xrightarrow[N\to\infty]{\mathcal{L}} \mathcal{N}(0,1)$$

# Example 2: Heterogeneity of the row degrees



# Example 3: Form of the graphon

## Graphon model

$$\begin{aligned} U_i, V_j &\overset{iid}{\sim} \mathcal{U}[0,1] \\ Y_{ij} \mid U_i, V_j &\sim \mathcal{P}\big(\lambda \tilde{w}(U_i, V_j)\big) \end{aligned}$$

where

- $\lambda = \mathbb{E}[Y_{ij}]$
- $\tilde{w} : [0,1]^2 \to [0, \frac{1}{\lambda}]$, $\iint \tilde{w} = 1$.

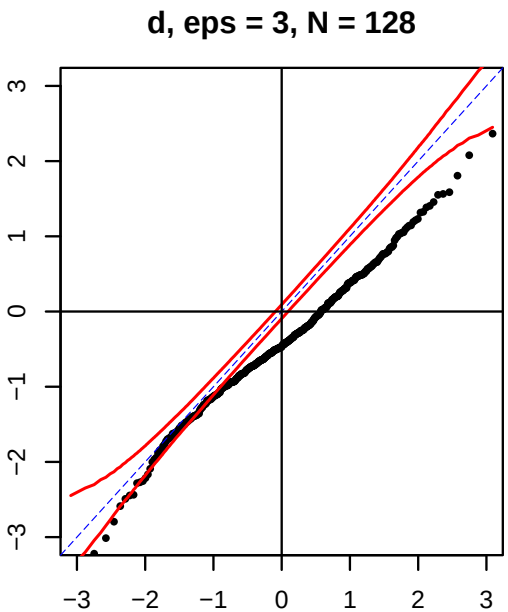
Link with the functions $f$ and $g$ of the BEDD model:

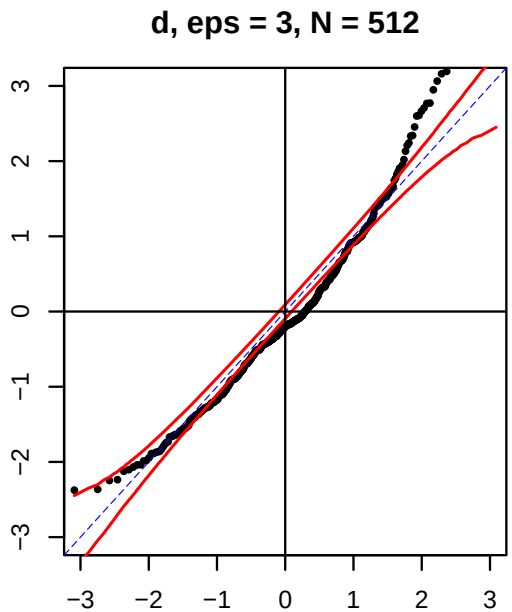
$$f(u) = \int \tilde{w}(u,v)\, dv \qquad\qquad g(v) = \int \tilde{w}(u,v)\, du$$

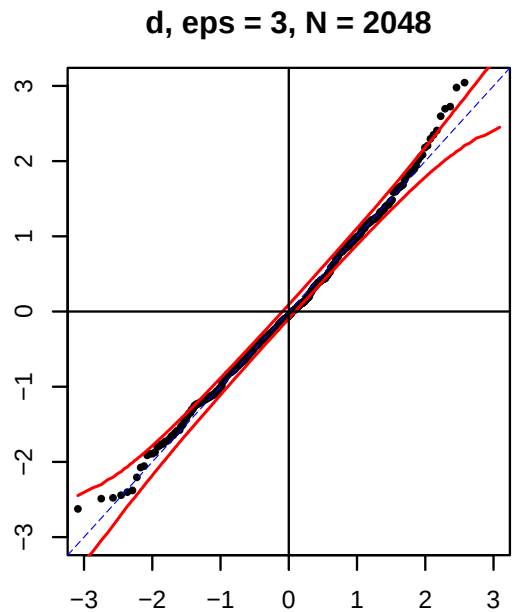
Dissimilarity measure against a BEDD:

$$d(w) = ||\tilde{w} - fg||_2^2 = \iint \left(\tilde{w}(u,v) - f(u)g(v)\right)^2 du\, dv$$

# Example 3: Form of the graphon



# Outline

# Summary

A large class of network models can be used: BEDD models, latent block models, graphon models, etc.

Diverse network questions can be investigated:

- motif densities,
- heterogeneity of the degrees,
- form of the graphon.

$U$-statistics can be used to perform statistical inference on network data with minimal assumptions:

- estimation,
- confidence intervals,
- network comparison.

# Perspectives

Bootstrapping network $U$-statistics

- especially in degenerate cases

Sparse graphs

- beyond the RCE framework, graphex models

Networks with covariates

- graph neural networks/variational autoencoders

# References

Le Minh, T. (2023). *U*-Statistics on bipartite exchangeable networks. *ESAIM: Probability and Statistics*.

Le Minh, T. (2023). Hoeffding-type decomposition for *U*-statistics on bipartite networks. *arXiv:2308.14518*.

# Thanks for your attention!