Asymptotic behavior of U-statistics on row-column exchangeable matrices

Application to bipartite network data analysis

Tâm Le Minh

Inria-Université Grenoble Alpes, France

August 16th, 2024

Networks

Network data

→ relational data (*links*) between entities (*nodes*)

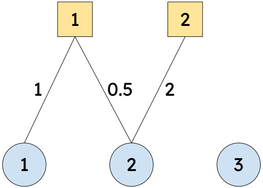
Bipartite networks

→ two types of nodes

→ links between nodes of different types

→ represented by a *bipartite graph* or an *adjacency matrix*

→ *binary* if only 0-1s, *weighted* otherwise



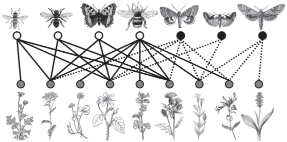
$$\begin{pmatrix} 1 & 0.5 & 0 \\ 0 & 2 & 0 \end{pmatrix}$$

Matrix representation of a bipartite network

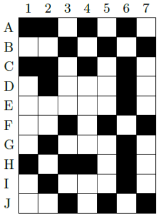
Examples of bipartite network data:

- recommender systems (users vs. items)
- scientific authorship (authors vs. articles)
- ecological networks (plant vs. animal species)

⇝ Bipartite network = rectangular matrix Y of size $m \times n$



Source: Macgregor et al. (2015)

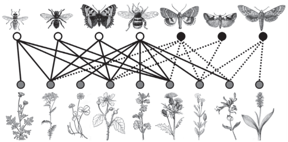


Matrix representation of a bipartite network

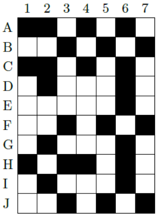
Examples of bipartite network data:

- recommender systems (users vs. items)
- scientific authorship (authors vs. articles)
- ecological networks (plant vs. animal species)

⇝ Bipartite network = rectangular matrix Y of size $m \times n$



Source: Macgregor et al. (2015)



<u>Objective:</u> characterize the network topology and perform statistical inference

Outline

1. Introduction
   - (Network) *U*-statistics
   - Exchangeable network models
2. Asymptotic behavior of network *U*-statistics
   - Convergence in distribution
   - Hoeffding decomposition
   - Degenerate case
3. Conclusion

Classic U-statistics

U-statistic on a vector of variables $(X_1, ..., X_n)$

$h : \mathbb{R}^k \to \mathbb{R}$ symmetric function, $[n] := \{1, ..., n\}$,

$$U_n = \binom{n}{k}^{-1} \sum_{\boldsymbol{I} \in \mathcal{P}_k([n])} h(X_{i_1}, ..., X_{i_k}).$$

Network U-statistics

Symmetric function h of a $p \times q$ matrix

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

Network U-statistics

Symmetric function h of a $p \times q$ matrix

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

U-statistic over a network of size $m \times n$

$$U_{m,n} = \binom{m}{p}^{-1} \binom{n}{q}^{-1} \sum_{\substack{\boldsymbol{I} \in \mathcal{P}_p([m]) \\ \boldsymbol{J} \in \mathcal{P}_q([n])}} h(Y_{\boldsymbol{I},\boldsymbol{J}})$$

Network U-statistics

Symmetric function h of a $p \times q$ matrix

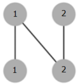
$$h(Y_{\{i_1,\dots,i_p;j_1,\dots,j_q\}}) = h\left(\begin{bmatrix} Y_{i_1 j_1} & Y_{i_1 j_2} & \dots & Y_{i_1 j_q} \\ Y_{i_2 j_1} & Y_{i_2 j_2} & \dots & Y_{i_2 j_q} \\ \dots & \dots & \dots & \dots \\ Y_{i_p j_1} & Y_{i_p j_2} & \dots & Y_{i_p j_q} \end{bmatrix}\right)$$

U-statistic over a network of size $m \times n$

$$U_{m,n} = \binom{m}{p}^{-1} \binom{n}{q}^{-1} \sum_{\substack{\boldsymbol{I} \in \mathcal{P}_p([m]) \\ \boldsymbol{J} \in \mathcal{P}_q([n])}} h(Y_{\boldsymbol{I},\boldsymbol{J}})$$

Example: subgraph density

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



Outline

RCE models

Row-column exchangeability (RCE, Aldous, 1981)

For all permutations σ_1 and σ_2,

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

RCE models

Row-column exchangeability (RCE, Aldous, 1981)

For all permutations σ_1 and σ_2,

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

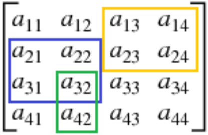
Meaning: Omission of the node labels

- No node labels, shuffling the rows or columns = same network.
- OK if interested in the general topology of the network.

Dissociation

Dissociated matrix (Silverman, 1976)

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.

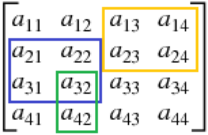


submatrices

Dissociation

Dissociated matrix (Silverman, 1976)

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

Graphon form of RCE dissociated models (Diaconis and Janson, 2008)

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

where $w : [0,1]^2 \to \mathbb{R}$ (graphon, W-graph model).

Outline

Asymptotic normality

Asymptotic framework

- $N := m_N + n_N$, $U_N := U_{m_N,n_N}$,
- $m_N/N \xrightarrow[N\to\infty]{} \rho \in]0,1[$, $n_N/N \xrightarrow[N\to\infty]{} 1-\rho$.

Asymptotic normality

Asymptotic framework

- $N := m_N + n_N$, $U_N := U_{m_N, n_N}$,
- $m_N/N \xrightarrow[N\to\infty]{} \rho \in]0,1[$, $n_N/N \xrightarrow[N\to\infty]{} 1-\rho$.

Asymptotic normality of U-statistics of RCE matrices

If Y is row-column exchangeable and dissociated and $\mathbb{E}[h(Y_{\{1,\dots,p;1,\dots,q\}})^2] < \infty$, then

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

with

$$\theta = \mathbb{E}[h(Y_{\{1,\dots,p;1,\dots,q\}})] \qquad \text{and} \qquad V = \frac{p^2}{\rho} v^{1,0} + \frac{q^2}{1-\rho} v^{0,1},$$

where $v^{r,c} = \mathrm{Cov}\big(h(Y_{\boldsymbol{I},\boldsymbol{J}}), h(Y_{\boldsymbol{I}',\boldsymbol{J}'})\big)$, $\boldsymbol{I}$ and $\boldsymbol{I}'$ share r elements, $\boldsymbol{J}$ and $\boldsymbol{J}'$ share c elements.

Outline

A gentle introduction to the Hoeffding decomposition

Let h be a symmetric function and $(X_1, X_2, ...)$ i.i.d. random variables.

$$U_n = \binom{n}{k}^{-1} \sum_{\boldsymbol{I} \in \mathcal{P}_k([n])} h(X_{\boldsymbol{I}}),$$

where $X_{\boldsymbol{I}} = \{X_{i_1}, ..., X_{i_k}\}$.

Principle: An orthogonal decomposition of U_n based on spaces generated by observations.

Hoeffding decomposition: the i.i.d. case

Let $\langle X_1, X_2 \rangle = \mathbb{E}[X_1 X_2]$ and

$$L_2\big(\{X_1, ..., X_k\}\big) = \big\{f(X_1, ..., X_k) : \mathbb{E}[f(X_1, ..., X_k)^2] < \infty\big\}.$$

Typically, we have

$$h(X_{\boldsymbol{I}}) \in L_2(X_{\boldsymbol{I}}).$$

Hoeffding decomposition: the i.i.d. case

Let $\langle X_1, X_2 \rangle = \mathbb{E}[X_1 X_2]$ and

$$L_2\big(\{X_1, ..., X_k\}\big) = \big\{f(X_1, ..., X_k) : \mathbb{E}[f(X_1, ..., X_k)^2] < \infty\big\}\,.$$

Typically, we have

$$h(X_{\boldsymbol{I}}) \in L_2(X_{\boldsymbol{I}}).$$

Projection subspaces L_2^*

$$L_2^*(X_{\boldsymbol{I}}) = L_2(X_{\boldsymbol{I}}) \cap \left(\cup_{\boldsymbol{I}' \subset \boldsymbol{I}} L_2(X_{\boldsymbol{I}'})^{\perp}\right).$$

Hoeffding decomposition: the i.i.d. case

Examples:

- $L_2^*(\emptyset) = L_2(\emptyset) = \mathbb{R}$,
- $L_2^*(\{X_1\}) = L_2(\{X_1\}) \cap L_2(\emptyset)^\perp$,
- $L_2^*(\{X_1, X_2\}) = L_2(\{X_1, X_2\}) \cap \left(L_2(\{X_1\})^\perp \cup L_2(\{X_2\})^\perp \cup L_2(\emptyset)^\perp\right)$.

Hoeffding decomposition: the i.i.d. case

Examples:

- $L_2^*(\emptyset) = L_2(\emptyset) = \mathbb{R}$,
- $L_2^*(\{X_1\}) = L_2(\{X_1\}) \cap L_2(\emptyset)^\perp$,
- $L_2^*(\{X_1, X_2\}) = L_2(\{X_1, X_2\}) \cap \left(L_2(\{X_1\})^\perp \cup L_2(\{X_2\})^\perp \cup L_2(\emptyset)^\perp\right)$.

Orthogonal decomposition of $L_2(X_{\boldsymbol{I}})$

$$L_2(X_{\boldsymbol{I}}) = \overset{\perp}{\bigoplus_{\boldsymbol{I}' \subseteq \boldsymbol{I}}} L_2^*(X_{\boldsymbol{I}'}).$$

Hoeffding decomposition: the i.i.d. case

Projections of $h(X_{\boldsymbol{I}}) \in L_2(X_{\boldsymbol{I}})$ on the L_2^* spaces

Defined by recursion:

- $p(\emptyset) = \mathbb{E}[h(X_{\boldsymbol{I}})] \in L_2^*(\emptyset)$,
- For $\boldsymbol{I}' \subseteq \boldsymbol{I}$,

$$p(X_{\boldsymbol{I}'}) = \mathbb{E}[h(X_{\boldsymbol{I}})|X_{\boldsymbol{I}'}] - \sum_{\boldsymbol{I}'' \subset \boldsymbol{I}'} p(X_{\boldsymbol{I}''}) \in L_2^*(X_{\boldsymbol{I}'}).$$

Hoeffding decomposition: the i.i.d. case

Projections of $h(X_{\boldsymbol{I}}) \in L_2(X_{\boldsymbol{I}})$ on the L_2^* spaces

Defined by recursion:

- $p(\emptyset) = \mathbb{E}[h(X_{\boldsymbol{I}})] \in L_2^*(\emptyset)$,
- For $\boldsymbol{I}' \subseteq \boldsymbol{I}$,

$$p(X_{\boldsymbol{I}'}) = \mathbb{E}[h(X_{\boldsymbol{I}})|X_{\boldsymbol{I}'}] - \sum_{\boldsymbol{I}'' \subset \boldsymbol{I}'} p(X_{\boldsymbol{I}''}) \in L_2^*(X_{\boldsymbol{I}'}).$$

Decomposition of $h(X_{\boldsymbol{I}})$

$$h(X_{\boldsymbol{I}}) = \sum_{\boldsymbol{I}' \subseteq \boldsymbol{I}} p(X_{\boldsymbol{I}'}) = \sum_{0 \le d \le k} \sum_{\boldsymbol{I}' \in \mathcal{P}_d(\boldsymbol{I})} p^{(d)}(X_{\boldsymbol{I}'}).$$

Application: CLT for U-statistics

Decomposition of U_n

$$U_n = \sum_{0 \leq d \leq k} \binom{k}{d} \binom{n}{d}^{-1} \sum_{\boldsymbol{I} \in \mathcal{P}_d([n])} p^{(d)}(X_{\boldsymbol{I}}).$$

Application: CLT for U-statistics

Decomposition of U_n

$$U_n = \sum_{0 \leq d \leq k} \binom{k}{d} \binom{n}{d}^{-1} \sum_{\boldsymbol{I} \in \mathcal{P}_d([n])} p^{(d)}(X_{\boldsymbol{I}}).$$

Asymptotic normality of U_n

$$U_n - \theta = \frac{k}{n} \sum_{1 \leq i \leq n} p^{(1)}(X_i) + o(n^{-1}).$$

$\rightsquigarrow$ Central Limit Theorem.

Case of network U-statistics

U-statistics on a RCE (row-column exchangeable) matrix:

$$U_{m,n} = \binom{m}{p}^{-1} \binom{n}{q}^{-1} \sum_{\substack{\boldsymbol{I} \in \mathcal{P}_p([m]) \\ \boldsymbol{J} \in \mathcal{P}_q([n])}} h(Y_{\boldsymbol{I},\boldsymbol{J}}),$$

... but the entries of $Y_{\boldsymbol{I},\boldsymbol{J}}$ are not independent.

Case of network U-statistics

U-statistics on a RCE (row-column exchangeable) matrix:

$$U_{m,n} = \binom{m}{p}^{-1} \binom{n}{q}^{-1} \sum_{\substack{\boldsymbol{I} \in \mathcal{P}_p([m]) \\ \boldsymbol{J} \in \mathcal{P}_q([n])}} h(Y_{\boldsymbol{I},\boldsymbol{J}}),$$

... but the entries of $Y_{\boldsymbol{I},\boldsymbol{J}}$ are not independent.

Aldous-Hoover-Kallenberg (AHK) representation (Kallenberg, 2005)

If Y is RCE dissociated, then there are

- $(\xi_i)_{1 \leq i < \infty}$, $(\eta_j)_{1 \leq j < \infty}$ and $(\zeta_{ij})_{1 \leq i < \infty, 1 \leq j < \infty}$ uniform i.i.d. random variables,
- a function φ,

such that

$$Y_{ij} \overset{a.s.}{=} \varphi(\xi_i, \eta_j, \zeta_{ij}).$$

Remark: The AHK variables are unobserved!

Graph sets of AHK variables

Representation of some AHK sets by **bipartite graphs** where:

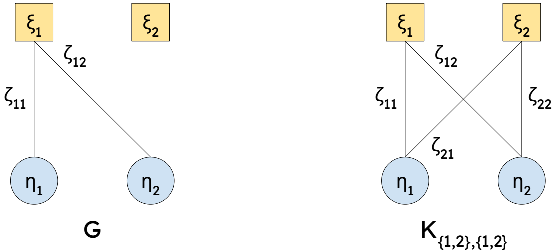
- sets of $(\xi_i)_i$ and $(\eta_j)_j$ = two types of vertices,
- set of $(\zeta_{ij})_{(i,j)}$ = edges.

Graph sets of AHK variables

Representation of some AHK sets by **bipartite graphs** where:

- sets of $(\xi_i)_i$ and $(\eta_j)_j$ = two types of vertices,
- set of $(\zeta_{ij})_{(i,j)}$ = edges.

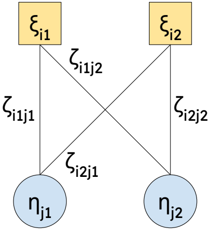
Examples:



Notation: $K_{\boldsymbol{I},\boldsymbol{J}}$ = fully connected graph with sets of vertices $\boldsymbol{I}$ and $\boldsymbol{J}$.

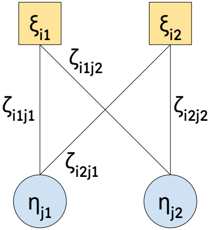
Decomposition of the probability space

$$Y_{\{i_1,i_2;j_1,j_2\}} = \begin{bmatrix} \varphi(\xi_{i_1},\eta_{j_1},\zeta_{i_1 j_1}) & \varphi(\xi_{i_1},\eta_{j_2},\zeta_{i_1 j_2}) \\ \varphi(\xi_{i_2},\eta_{j_1},\zeta_{i_2 j_1}) & \varphi(\xi_{i_2},\eta_{j_2},\zeta_{i_2 j_2}) \end{bmatrix} \in L_2(K_{\{i_1,i_2\},\{j_1,j_2\}})$$



Decomposition of the probability space

$$Y_{\{i_1,i_2;j_1,j_2\}} = \begin{bmatrix} \varphi(\xi_{i_1}, \eta_{j_1}, \zeta_{i_1 j_1}) & \varphi(\xi_{i_1}, \eta_{j_2}, \zeta_{i_1 j_2}) \\ \varphi(\xi_{i_2}, \eta_{j_1}, \zeta_{i_2 j_1}) & \varphi(\xi_{i_2}, \eta_{j_2}, \zeta_{i_2 j_2}) \end{bmatrix} \in L_2(K_{\{i_1,i_2\},\{j_1,j_2\}})$$



Decomposition of $L_2(G)$

Defined by recursion:

- $L_2^*(\emptyset) = \mathbb{R}$
- $L_2^*(G) = L_2(G) \cap \left(\cup_{G' \subset G} L_2(G')^\perp\right)$

We have

$$L_2(G) = \bigoplus_{G' \subseteq G}^{\perp} L_2^*(G').$$

Decomposition for network U-statistics

Projections of $h(Y_{\boldsymbol{I},\boldsymbol{J}})$

- $p(\emptyset) = \mathbb{E}[h(Y_{\boldsymbol{I},\boldsymbol{J}})] = \theta \in L_2^*(\emptyset)$,
- For some graph $G \subseteq K_{\boldsymbol{I},\boldsymbol{J}}$,

$$p(G) = \mathbb{E}[h(Y_{\boldsymbol{I},\boldsymbol{J}})|G] - \sum_{G' \subset G} p(G') \in L_2^*(G).$$

Decomposition for network U-statistics

Projections of $h(Y_{\boldsymbol{I},\boldsymbol{J}})$

- $p(\emptyset) = \mathbb{E}[h(Y_{\boldsymbol{I},\boldsymbol{J}})] = \theta \in L_2^*(\emptyset)$,
- For some graph $G \subseteq K_{\boldsymbol{I},\boldsymbol{J}}$,

$$p(G) = \mathbb{E}[h(Y_{\boldsymbol{I},\boldsymbol{J}}) | G] - \sum_{G' \subset G} p(G') \in L_2^*(G).$$

Hoeffding decomposition

Let $\Gamma_{\boldsymbol{I},\boldsymbol{J}}^{r,c}$ the set of subgraphs of $K_{\boldsymbol{I},\boldsymbol{J}}$ with r row vertices and c column vertices.

$$U_{m,n} = \sum_{\substack{0 \leq r \leq p \\ 0 \leq c \leq q}} \binom{p}{r} \binom{q}{c} \binom{m}{r}^{-1} \binom{n}{c}^{-1} \sum_{\substack{\boldsymbol{I} \in \mathcal{P}_r([m]) \\ \boldsymbol{J} \in \mathcal{P}_c([n])}} \sum_{G \in \Gamma_{\boldsymbol{I},\boldsymbol{J}}^{r,c}} p^{(r,c)}(G).$$

CLT for network U-statistics

Reminder: Asymptotic framework

- $N := m_N + n_N$, $U_N := U_{m_N,n_N}$,
- $m_N/N \xrightarrow[N\to\infty]{} \rho \in]0,1[$, $n_N/N \xrightarrow[N\to\infty]{} 1-\rho$.

$$U_N - \theta = \frac{p}{m_N}\sum_{i=1}^{m_N} p^{(1,0)}(\xi_i) + \frac{q}{n_N}\sum_{j=1}^{n_N} p^{(0,1)}(\eta_j) + o(N^{-1}).$$

CLT for network U-statistics

Reminder: Asymptotic framework

- $N := m_N + n_N$, $U_N := U_{m_N,n_N}$,
- $m_N/N \xrightarrow[N\to\infty]{} \rho \in]0,1[$, $n_N/N \xrightarrow[N\to\infty]{} 1-\rho$.

$$U_N - \theta = \frac{p}{m_N}\sum_{i=1}^{m_N} p^{(1,0)}(\xi_i) + \frac{q}{n_N}\sum_{j=1}^{n_N} p^{(0,1)}(\eta_j) + o(N^{-1}).$$

Central Limit Theorem:

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

with

$$V = \frac{p^2}{\rho}\mathbb{V}[p^{(1,0)}(\xi_1)] + \frac{q^2}{1-\rho}\mathbb{V}[p^{(0,1)}(\eta_1)].$$

Outline

1. Introduction
   - (Network) U-statistics
   - Exchangeable network models
2. Asymptotic behavior of network U-statistics
   - Convergence in distribution
   - Hoeffding decomposition
   - Degenerate case
3. Conclusion

Principal support graphs

$$U_N = \sum_{\substack{0 \le r \le p \\ 0 \le c \le q}} \binom{p}{r} \binom{q}{c} \binom{m_N}{r}^{-1} \binom{n_N}{c}^{-1} \sum_{\substack{\boldsymbol{I} \in \mathcal{P}_r([m_N]) \\ \boldsymbol{J} \in \mathcal{P}_c([n_N])}} \sum_{G \in \Gamma^{r,c}_{\boldsymbol{I},\boldsymbol{J}}} p^{(r,c)}(G).$$

Principal support graphs (Janson and Nowicki, 1991)

- The **principal support graphs** of U_N are the graphs G such that $p(G) \neq 0$ with the smallest number of nodes $d = r + c$,
- d is called the **principal degree** of U_N.

Principal support graphs

$$U_N = \sum_{\substack{0 \leq r \leq p \\ 0 \leq c \leq q}} \binom{p}{r} \binom{q}{c} \binom{m_N}{r}^{-1} \binom{n_N}{c}^{-1} \sum_{\substack{\mathbf{I} \in \mathcal{P}_r([m_N]) \\ \mathbf{J} \in \mathcal{P}_c([n_N])}} \sum_{G \in \Gamma^{r,c}_{\mathbf{I},\mathbf{J}}} p^{(r,c)}(G).$$

Principal support graphs (Janson and Nowicki, 1991)

- The **principal support graphs** of U_N are the graphs G such that $p(G) \neq 0$ with the smallest number of nodes $d = r + c$,
- d is called the **principal degree** of U_N.

$$U_N = \binom{N}{d}^{-1} \sum_{r+c=d} \frac{p!}{(p-r)!\rho^r} \frac{q!}{(q-c)!(1-\rho)^c} \sum_{\substack{\mathbf{I} \in \mathcal{P}_r([m_N]) \\ \mathbf{J} \in \mathcal{P}_c([n_N])}} \sum_{G \in \Gamma^{r,c}_{\mathbf{I},\mathbf{J}}} p^{(r,c)}(G) + o(N^{-d}).$$

Convergence of degenerate U-statistics

Convergence of degenerate U-statistics

$$N^{d/2}(U_N - \theta) \xrightarrow[N\to\infty]{\mathcal{D}} W,$$

where W is a random variable with finite variance $V^{(d)}$, only depending on the principal support graphs.

Convergence of degenerate U-statistics

Convergence of degenerate U-statistics

$$N^{d/2}(U_N - \theta) \xrightarrow[N\to\infty]{\mathcal{D}} W,$$

where W is a random variable with finite variance $V^{(d)}$, only depending on the principal support graphs.

Gaussian case

If all the principal support graphs are connected, then

$$N^{d/2}(U_N - \theta) \xrightarrow[N\to\infty]{\mathcal{D}} \mathcal{N}(0, V^{(d)}).$$

Outline

Summary

A Hoeffding decomposition for U-statistics on RCE matrices can be defined using **graph sets of Aldous-Hoover-Kallenberg variables**.

The asymptotic behavior of these U-statistics is characterized by the **topology of the principal support graphs**.

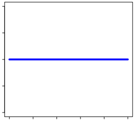
These results can be used to perform statistical inference on bipartite network data with symmetry assumptions (exchangeable nodes = graphon models).

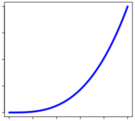
Preprint:

Le Minh, T. (2024). Characterization of the asymptotic behavior of U-statistics on row-column exchangeable matrices. *arXiv:2401.07876*.

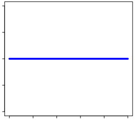
Outline

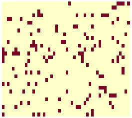
Poisson product graphon

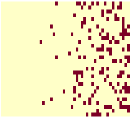
$g(v) =$ 

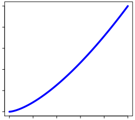
$g(v) =$ 

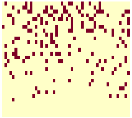
$$\begin{aligned} \xi_i, \eta_j &\overset{iid}{\sim} \mathcal{U}[0,1] \\ Y_{ij} \mid \xi_i, \eta_j &\sim \mathcal{P}(\lambda f(\xi_i) g(\eta_j)) \end{aligned}$$

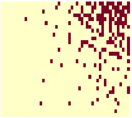
$f(u) =$ 





$f(u) =$ 





$$\boxed{\int f = 1, F_2 := \int f^2}$$

Example 1: Test of $\mathcal{H}_0 : F_2 = 1$

Estimators

- $h_1(Y_{\{1,2;1,2\}}) = Y_{11}Y_{12}$, $\mathbb{E}h_1 = \lambda^2 F_2$
- $h_2(Y_{\{1,2;1,2\}}) = Y_{11}Y_{22}$, $\mathbb{E}h_2 = \lambda^2$
- $h = h_1 - h_2$, $U_N^h := U_N^{h_1} - U_N^{h_2}$, $\mathbb{E}[U_N^h] = \lambda^2(F_2 - 1)$.

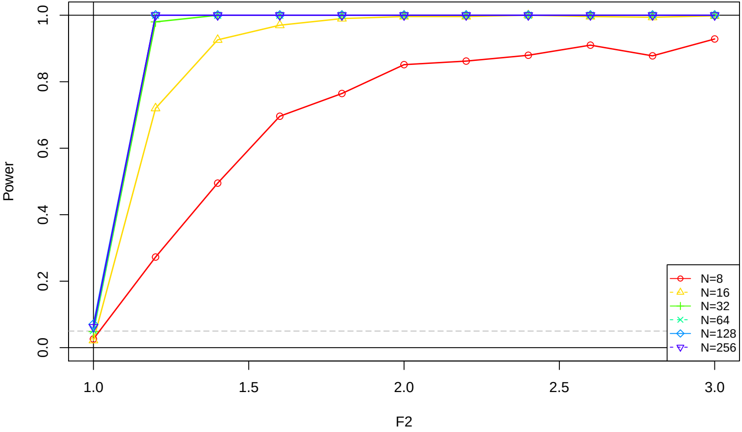
Limit distribution

If $F_2 = 1$, then U_N^h is degenerate, with principal degree $= 3$,

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

where $V = \frac{2\lambda^2}{\rho(1-\rho)^2}$.

Example 1: Test on $F_2 = 1$



Null hypothesis $\mathcal{H}_0 : F_2 = 1$

Example 2: Estimation of F_2, network comparison

Estimators

- $h_1(Y_{\{1,2;1,2\}}) = Y_{11}Y_{12}$, $\mathbb{E}h_1 = \lambda^2 F_2$
- $h_2(Y_{\{1,2;1,2\}}) = Y_{11}Y_{22}$, $\mathbb{E}h_2 = \lambda^2$
- $\widehat{\theta}_N := U_N^{h_1} / U_N^{h_2}$

Limit distribution

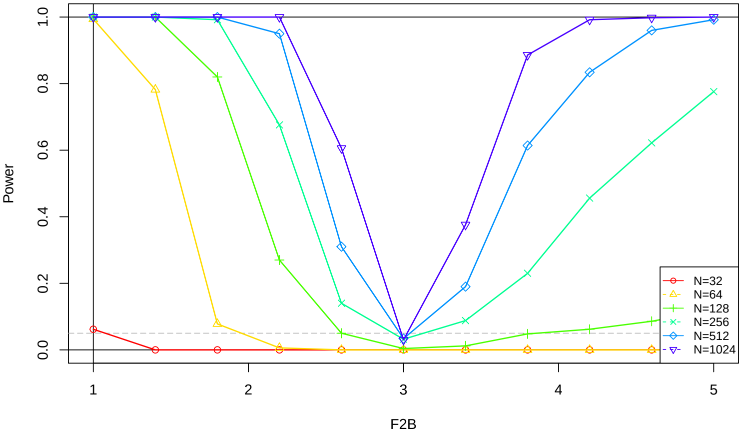
$$\sqrt{\frac{N}{V^\delta}}\left(\widehat{\theta}_N - F_2\right) \xrightarrow[N\to\infty]{\mathcal{D}} \mathcal{N}(0,1)$$

where $V^\delta = \frac{1}{\lambda^4}V^{h_1} - \frac{2F_2}{\lambda^4}C^{h_1,h_2} + \frac{F_2^2}{\lambda^4}V^{h_2}$ (delta-method).

Compare the value of F_2 for 2 independent networks Y^A and Y^B:

$$Z_N = \widehat{\theta}_{N_A}(Y^A) - \widehat{\theta}_{N_B}(Y^B).$$

Example 2: Estimation of F_2, network comparison



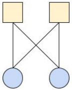
Null hypothesis $\mathcal{H}_0 : F_2^A = F_2^B$, simulations with fixed $F_2^A = 3$

Example 1: Motif counts

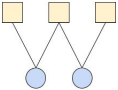
Exchangeable binary network model

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

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

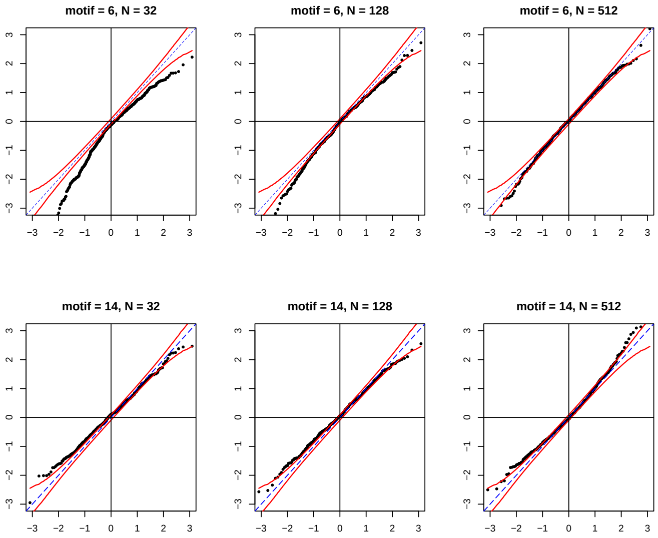


$$h_{14}(Y_{\{1,2,3;1,2\}}) = Y_{11} Y_{21} Y_{22} Y_{32} (1 - Y_{12})(1 - Y_{31})$$



The respective U_N^h are the densities of these motifs in the observed network.

Example 1: Motif counts



Example 2: Form of a 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

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

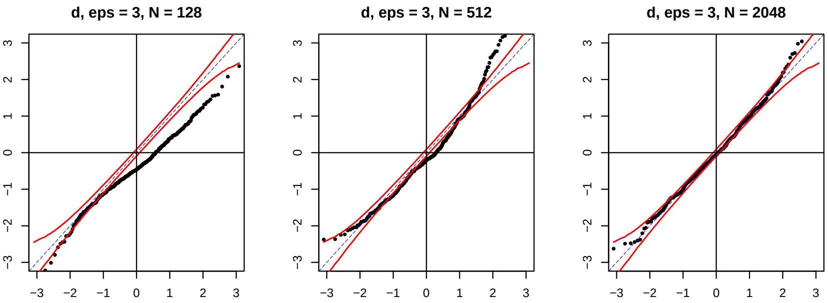
Link with the f and g functions of the BEDD model:

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

Dissimilarity measure to the BEDD model:

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

Example 2: Form of a graphon



Example 3: Assessing overdispersion

Overdispersed Poisson-BEDD model

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

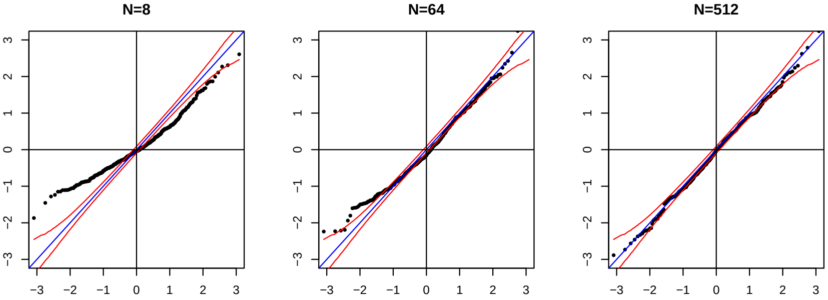
where

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

The model becomes a simple Poisson-BEDD if $W_{ij} = 1$, i.e. $W_2 := \mathbb{E}[W_{ij}^2] = 1$.

W_2 is a property of the model, not a parameter of the BEDD model.

Example 4: Test on $F_2 = 1$



References

Aldous, D. J. (1981). Representations for partially exchangeable arrays of random variables. *Journal of Multivariate Analysis*.

Diaconis, P. and Janson, S. (2008). Graph limits and exchangeable random graphs. *Rendiconti di Matematica e delle sue Applicazioni*.

Janson, S. and Nowicki, K. (1991). The asymptotic distributions of generalized U-statistics with applications to random graphs. *Probability Theory and Related Fields*.

Kallenberg, O. (2005). *Probabilistic symmetries and invariance principles* (Vol. 9).

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*.

Le Minh, T. (2024). Characterization of the asymptotic behavior of U-statistics on row-column exchangeable matrices. *arXiv:2401.07876*.

Silverman, B. W. (1976). Limit theorems for dissociated random variables. *Advances in Applied Probability*.