Network Dependence Testing via Diffusion Maps and Distance-Based Correlations

Nonparametric Statistics Student Paper Competition

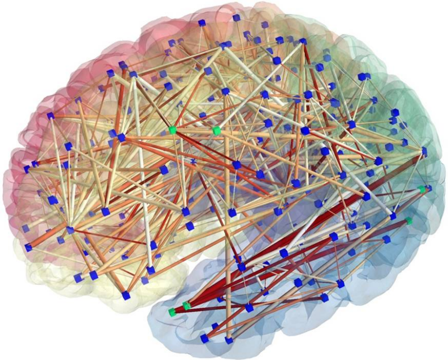
Youjin Lee, Cencheng Shen, and Joshua T. Vogelstein

Johns Hopkins University

Social Network

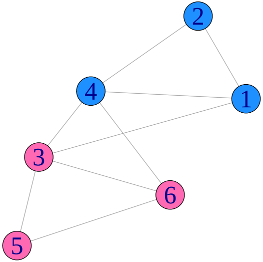
Personal characteristics may be dependent on social connections.

Networks in Science



Scientists might be interested in the correlation between the characteristics of each voxel and their functional connectivity.

Network as a Graph



- Graph $\mathbf{G} = \big(V(\mathbf{G}), E(\mathbf{G})\big)$
- Node set : V($\mathbf{G}$) = $\{1, 2, 3, 4, 5, 6\}$
- Edges set : E($\mathbf{G}$) = $\{(1, 2), (1, 3), \ldots, (5, 6)\}$ $\Rightarrow$ 6 $\times$ 6 matrix $\mathbf{A}$
- $\mathbf{X} = \{B, B, P, B, P, P\}$: Nodal attributes, e.g., weight of each subject or composition of each voxel.

Networks with Nodal Attributes $(\mathbf{A}, \mathbf{X})$

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 1 & 0 \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} B \\ B \\ P \\ B \\ P \\ P \end{bmatrix}$$

*Are **B**lue or **P**ink nodes more likely to connect within the same color than between **B**lue and **P**ink ?*

$\Rightarrow$ **Independence test** with $H_0 : f_{Net \cdot \mathbf{X}} = f_{Net} \circ f_{\mathbf{X}}$

Testing Independence between Network Topology and Attributes

What can we do with a pair of $(\mathbf{A}, \mathbf{X}) = \{(\mathbf{a}_{ij}, \mathbf{x}_i) : i, j = 1, 2, \ldots, n\}$?

- Model-based representation of network as a function of node-specific network factor $\mathbf{u}_i$.
- Testing independence with high dimensional independent and identically distributed (i.i.d.) network representation $\{\mathbf{u}_i\}$ and nodal attributes $\{\mathbf{x}_i\}$.

Challenges

1. Efficient multivariate independence test statistics.
2. Node-wise i.i.d. representation of network topology without network model specification.

Testing Independence in General Setting

Given a pair of sample data $(\mathbf{U}, \mathbf{X}) = \{(\mathbf{u}_i, \mathbf{x}_i) : i = 1, 2, \ldots, n\}$ i.i.d. as $(\mathbf{u}, \mathbf{x}) \in \mathbb{R}^{q \times q_x}$,

- **Pearson's correlation** : measures linear correlation between two random variables.
- **Mantel coefficient** : Pearson's correlation coefficients on distance matrices $C_{ij} := dist_{\mathbf{U}}(\mathbf{u}_i, \mathbf{u}_j)$ and $D_{ij} := dist_{\mathbf{X}}(\mathbf{x}_i, \mathbf{x}_j)$.
- **Heller-Heller-Gorfine** (`HHG`) test: uses ranks of C_{ij} and D_{ij}.
- **Distance Correlation** (`dCorr`) test: uses properly centered C_{ij} and D_{ij}.

Distance-Based Correlations for Multivariate Independence Test

- Define pairwise distances within each data set using the Euclidean distance - $\mathbf{C}_{ij} = \| \mathbf{u}_i - \mathbf{u}_j \|$ and $\mathbf{D}_{ij} = \| \mathbf{x}_i - \mathbf{x}_j \|$
- Double-center each of distance matrices : $\tilde{\mathbf{C}}_{ij} := c_{ij} - \bar{c}_{i\cdot} - \bar{c}_{\cdot j} + \bar{c}_{\cdot\cdot}$

$$\texttt{dCov}(C, D) = \frac{1}{n^2} \sum_{i,j=1}^{n} \tilde{\mathbf{C}}_{ij} \tilde{\mathbf{D}}_{ij} \tag{1}$$

- Standardized $\texttt{dCorr}(C, D)$ is proven to be consistent test statistic against **all** possible dependencies in multivariate **U** and **X** under the finite first moment.

Multiscale Generalized Correlations

- Under high-dimensional and nonlinear dependencies, performance of `dCorr` and its unbiased version for multivariate test, `mCorr` work less efficiently.
- Multiscale Generalized Correlations (`MGC`) is a local version of distance correlation (Shen et al., 2017).
- `MGC` $\rho^* = \texttt{dCorr}_n^{kl^*}$ for optimal scale of neighborhood choice $(k, l)^*$ $\rightarrow$ considers only k^*-nearest neighbors in each point of $\mathbf{U}$ and l^*-nearest neighbors in each point of $\mathbf{X}$.
- `MGC` achieves higher power under **nonlinear dependencies** than `mCorr`.

Network Metrics

- Ingredient for $\mathtt{MGC}$: $\mathbf{C}_{ij} = \| \mathbf{u}_i - \mathbf{u}_j \|$ & $\mathbf{D}_{ij} = \| \mathbf{x}_i - \mathbf{x}_j \|$
- Cannot directly use an adjacency matrix $\mathbf{A}$ for $\mathbf{u}$.

Requirements

1. IID : $\mathbf{u}_i \overset{i.i.d.}{\sim} f_{\mathbf{u}}$
2. Finite moment and finite dimension.
3. Represent node-wise position over network in a robust way.

Diffusion Maps as Network Representation

Diffusion Maps : Network representation satisfying above three requirements under an **exchangeable graph**.

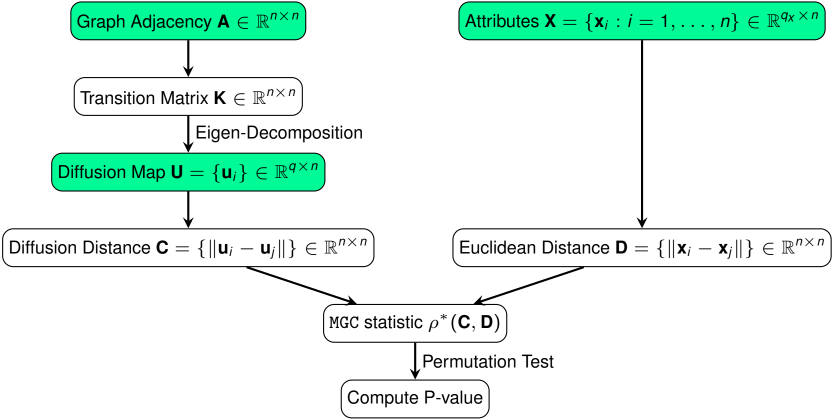
- Assume a $n \times n$ symmetric and positive kernel matrix $\mathbf{K}$.
- Transform it into a transition matrix $\tilde{\mathbf{K}}_{ij} = \mathbf{K}_{ij} / \sum_{j=1}^{n} \mathbf{K}_{ij}$.

  : Probability of traveling from node i to node j.
- Derive eigenvalues and eigenvectors $\{\lambda_j\}$ and $\{\phi_j\}$ for diffusion maps $\{\mathbf{u}_j\}$.

$$\mathbf{u}_i = \begin{pmatrix} \lambda_1^t \phi_1(i) & \lambda_2^t \phi_2(i) & \cdots & \lambda_q^t \phi_q(i) \end{pmatrix} \in \mathbb{R}^q; \quad i = 1, \ldots, n, \quad (2)$$

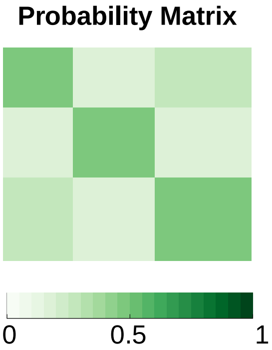
- Diffusion distance ($\mathbf{C}_{ij} = \| \mathbf{u}_i - \mathbf{u}_j \|$) takes into account every possible path from node i to node j $\rightarrow$ robust against perturbation.

Network Dependence Testing



Flowchart for network dependence testing via diffusion maps and MGC. The above procedure provides a consistent test under popular network models, like stochastic block model (SBM) and random-dot-product graph (RDPG).

Simulation - Stochastic Block Model



(a) Data generating matrix.

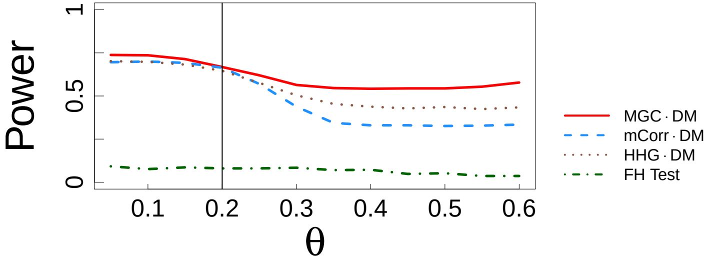
(b) Adjacency matrix.

In Stochastic Block Model (SBM), edge probability depends on block membership of each node.

Simulation - Stochastic Block Model

Z : block membership (1,2,3) X : nodal attributes ($\approx Z$)

$$E(A_{ij}|z_i, z_j) = 0.5\mathbf{I}(|z_i - z_j| = 0) + 0.2\mathbf{I}(|z_i - z_j| = 1) + \theta\mathbf{I}(|z_i - z_j| = 2)$$



When $\{\theta : \theta > 0.2\}$, such SBM generates **nonlinear dependency** and it becomes strongly nonlinear as θ gets further away from 0.2.

Simulation - Two Graph Test

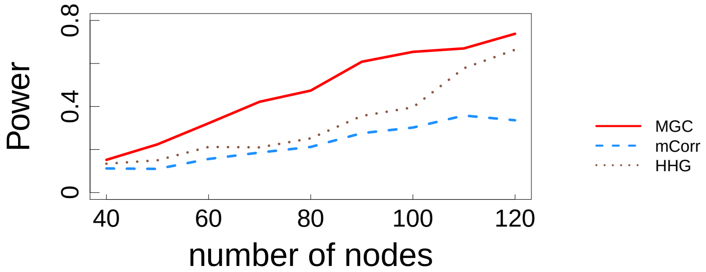
Random Dot Product Graph (RDPG) : $P(A_{ij} = 1|\mathbf{W}) = \langle \mathbf{W}_i, \mathbf{W}_j \rangle$.

Latent factor 1: $y_{ki} \overset{i.i.d.}{\sim} Unif(0,1), \quad i = 1, 2, \ldots, n; \; k = 1, 2, \ldots, 5$

Latent factor 2: $w_i := (1 - y_{i1})^2, \quad i = 1, 2, \ldots, n$

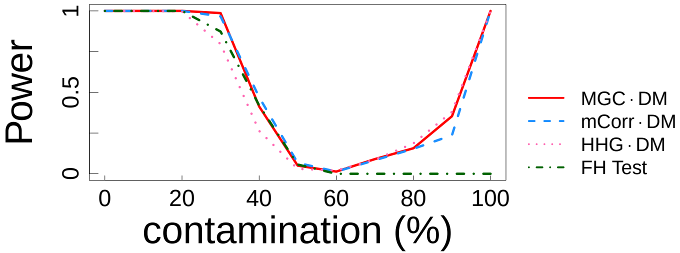
$$\mathbf{G}_1 : A_{ij}^{(1)} \big| \mathbf{y}_i, \mathbf{y}_j \sim Bernoulli\big(\langle \mathbf{y}_i/5, \mathbf{y}_j/5 \rangle\big), \quad \mathbf{y}_j \in \mathbb{R}^5$$

$$\mathbf{G}_2 : A_{ij}^{(2)} \big| w_i, w_j \sim Bernoulli\big(\langle w_i, w_j \rangle\big), \quad \forall i < j; \; i, j = 1, 2, \ldots, n.$$



Experiment on Brain Network

- **G** : node - voxel in the brain, $|V(\mathbf{G})| = 95$.
  edge - brain fibers connecting each region, $|E(\mathbf{G})| = 337$.
- **X** : 3-dimensional voxel-wise coordinates.
- At $c\%$ of contamination, $c\%$ of edges are randomly selected to be flipped, e.g., connected $\rightarrow$ dis-connected



Powers are obtained through 300 random contaminations for each of contamination level $c\% \in \{10\%, 20\%, \ldots, 90\%\}$.

Discussion

- Testing independence between network topology and nodal attributes.
- **Diffusion distance** (network metrics) + **MGC** (test statistics) $\rightarrow$ successful in nonlinear, high-dimensional, and noisy dependencies without worrying about network model mis-specifications.
- The method can be extended to independence test between two graphs.

Thank You!

Contact : ylee160@jhu.edu
Draft : `arXiv:1703.10136`