Testing independence between networks and nodal attributes via multiscale metrics

Youjin Lee*
Department of Biostatistics, Johns Hopkins School of Public Health
and
Author 2
Department of ZZZ, University of WWW

October 16, 2016

Abstract

Network dependence, which refers to the dependence between network topology and its nodal attributes, often exhibits nonlinear patterns. Unfortunately, without knowledge on metrics over network, no statistic has been proposed to test network dependence further beyond globally linear dependence. This paper introduces a multiscale dependence test statistic called Multiscale Network Test (`MNT`), which borrows the idea from diffusion maps and Multiscale Generalized Correlation (`MGC`). Our method can be applied to any exchangeable graph under some mild conditions, without further model assumption nor estimation. We prove the consistency of test statistic and demonstrate superior performance than any other model-based, global test statistics. Simulation in a variety of networks shows higher power of test, especially in stochastic block model under nonlinear dependency. The application to dMRI networks will be followed.

Keywords: distance correlation, multiscale generalized correlation, diffusion maps, exchangeable graph, stochastic block model

*The authors gratefully acknowledge *please remember to list all relevant funding sources in the unblinded version*

1 Introduction

Network, a collection of nodes and edges between them, has been a celebrated area of study over a field of sociology (Pinquart and Sörensen, 2000; Ellison et al., 2007), information theory (Gross and Acquisti, 2005), biology (Barabasi and Oltvai, 2004; Pujol et al., 2010), statistics (Raftery et al., 2012; Palla et al., 2012), economics (Banerjee et al., 2013), etc. The correlation between network relationship between nodes and their attribute values is a common interest in network analysis. According to an assumption on how they are related each other, there has been lots of efforts to manifest a network as a function of nodal attributes (Wasserman and Pattison, 1996; Howard et al., 2016) or model an outcome of nodal attribute variables through their underlying network structures (Christakis and Fowler, 2007, 2008). However, it is very obscure to determine which one should be put as a dependent variable or even whether networks are truly related to nodal attributes and how they are, if any. Moreover, a fundamental difficulty comes from the empirical fact that network often does not have a natural structure. Thus it is not easy to intuitively come up with how to represent network as a node-specific random variable. Fosdick and Hoff (2015) overcomes this issue by estimating network factors which are believed to embody each node's locations in network space. These factors are in the end used to test independence between network topology and nodal attributes by implementing standard statistical testing method. Through allowing us to choose the dimension of latent factors, they make up the constraints of parametric modeling. However their statistical modeling on networks still rely on the assumption that all the nodes in network would follow the same pattern of dependence – subject to additive and multiplicative effect. This might not be true for always. In this paper we develop a nonparametric test statistic which is also sensitive to nonlinear and local dependence pattern.

Throughout this paper, assume that we are given an unweighted and undirected, connected network, equivalently a graph $\mathbf{G}$ without self-loop, comprised of n nodes for a fixed $n \in \mathbb{N}$. Even though we assume that $\mathbf{G}$ is undirected and unweighted, we are able to extend all of the theory here to directed and even weighted network. An adjacency matrix of a given network, denoted by $\mathbf{A} = \{A_{ij} : i, j = 1, .., n\}$, is often introduced to formalize the relational data of network, where $A_{ij} = 1$ if node i and node j are adjacent each other and zero otherwise. Let us introduce a m-variate ($m \in \mathbb{N}$) variable for nodal attributes $\mathbf{X} \in \mathbb{R}^m$ which we are interested in. Investigating correlation between $\mathbf{G}$ and $\mathbf{X}$, and testing whether their distributions are independent or not is a key focus in our study. An observed network $\mathbf{G}$ can represent social network within a school and $\mathbf{X}$ is students' grades or heights, for example; or $\mathbf{G}$

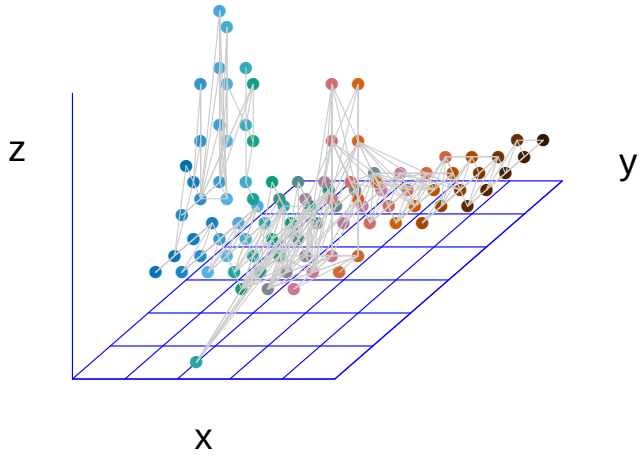

Figure 1: Physical location of one component of human brain network and its tracts that connect one vertex to another.

can be a neuronal network in human brain and $\mathbf{X}$ is a few factors of personality. For example, Figure 1 exemplifies one connected network from human brain network where dots denote nodes and tracts connected them represent edges. Location over network space, i.e. whether or how much a pair of nodes are closer than others, is different from actual spatial location, which can be measured by Euclidean distance over three-dimensional xyz space. As correlation between spatial location of subjects and its attributes has been studied, we are going to explore correlation of subjects' attributes to their network location.

The main contribution of this study is we develop multiscale test statistics which are robust to both nonlinearity and high dimensionality without modeling nor estimating networks. Having multiscale statistics is not avoidable because we regard location of or distance between nodes over network as a dynamic process. We then choose the optimal scale where distance in network space and distance in attributes maximize their correlation. To explore this time-dependent distances we define a coordinate over network space at each time in the process and test independence to attributes $\mathbf{X}$. In the next methodology section 2, we are going to define such multiscale distance and its properties. A test statistic

using multiscale distance metrics will be followed. In section 3, we demonstrate best performance of our method compared to the existing under various circumstances through numerical results. Real data example in section 4 show one of the applications among many.

2 Methodology

First step in testing network independence to attributes is figuring out a variable which configurates location of nodes over network space. We hope that our observed network be a representative realization of this variable so that our test results can be generalized to population network. To be specific we want our observations to be independent and identically-distributed ($i.i.d$) realization to guarantee representativeness and avoid redundant intra-dependence.

2.1 Exchangeable Graph

Assuming that we are given one network, equivalently one *graph* comprised of nodes and edges. If you consider each edge as $i.i.d$, then observed adjacency matrix is a random sample of a single parameter which is the probability of having edge or not. However, in this case, the resulting network model only depends on number of edges (Orbanz and Roy, 2015) and does not cover undirected or no self-loop networks, which is very common. Instead of assuming $i.i.d$ of edges right away, we are going to assume underlying distribution of edges and then consider a set of edges as $i.i.d$ conditional on random function. This idea comes from exchangeable representation of edges. The property of exchangeability is closely related to the use of $i.i.d$ random variable. A graph $\mathbf{G}$ is called exchangeable if and only if its adjacency matrix $\mathbf{A}$ is jointly exchangeable (Orbanz and Roy, 2015).

Definition 2.1 (2-array exchangeability)**.** A random 2-array (A_{ij}) is called jointly exchangeable if

$$(A_{ij}) \stackrel{d}{=} (A_{\sigma(i)\sigma(j)})$$

for every permutation σ of n, and separately exchangeable if

$$(A_{ij}) \stackrel{d}{=} (A_{\sigma(i)\sigma'(j)})$$

for every pair of permutation σ, σ' of n.

However, exchangeability itself cannot guarantee being *i.i.d.* Fortunately, thanks to the celebrated *de Finetti* A.1's representation theorem, it has been proven that there exists a random probability measure η on random variable $\mathbf{Z}$ that a sequence of $Z_1, Z_2, \ldots$ are *i.i.d* conditional η *if and only if* the sequence is exchangeable (Orbanz and Roy, 2015; Caron and Fox, 2014). *Aldous-Hoover theorem* A.2 is the representation theorem of 2-array exchangeable array, which is useful to explain jointly exchangeable adjacent matrix. Exchangeable graph is commonly called *graphon* (Lovász and Szegedy, 2006), which is defined through a random measurable functions (Chan et al., 2013).

Definition 2.2 (graphon)**.** A *graphon* with $n(\in \mathbb{N})$ nodes is defined as a function of a symmetric measurable function $g : [0,1]^2 \to [0,1]$ with input of $u_i \overset{i.i.d}{\sim} Uniform[0,1], i = 1, 2, ..., n$. Let $\mathbf{A}$ be an adjacency matrix of graphon. Then for any $i < j, i, j = 1, 2, ..., n$:

$$Pr\big(A_{ij} = 1 \big| u_i, u_j\big) = g\big(u_i, u_j\big) \tag{1}$$

By *Aldous-Hoover theorem*, we can now better represent exchangeable network through measurable function, but we are still halfway done in the case of undirected network where $A_{ij} = A_{ji}$ $(i, j = 1, 2, ..., n)$. Under undirected network without self-loop, we can represent $\{A_{ij} : i < j\}$ as a function of g of $\{u_i\}$.

$$(A_{ij}) = (A_{\sigma(i)\sigma(j)}) \Longleftrightarrow A_{ij} \overset{ind}{\sim} Bern\big(g(u_i, u_j)\big), i < j \tag{2}$$

Networks based on widely used graphical model are exchangeable. One of the most popular models is Stochastic Block Model (SBM) (Holland et al., 1983). In the simplest setting of SBM, we assume that each n nodes of $\mathbf{G}$ belongs to one of $K \in \mathbb{N}(\leq n)$ blocks or groups. Block affiliation is important in that the probability of having edges between a pair of nodes depends on which blocks they are in. Assume that latent variables corresponding to block affiliation follow $Z_1, Z_2, ..., Z_n \overset{i.i.d.}{\sim} Multinomial\big(\pi_1, \pi_2, ..., \pi_K\big)$. Then the upper triangular entries of $\mathbf{A}$ are independent and identically distributed conditional on $\{\mathbf{Z}\}$:

$$A_{ij} \overset{i.i.d.}{\sim} Bern\Big(\sum_{k,l=1}^{K} p_{kl} I\big(Z_i = k, Z_j = l\big)\Big), \forall i < j. \tag{3}$$

The above distribution can also be represented through some random function $g : [0,1]^2 \to [0,1]$. Let

$W_1, W_2, ..., W_n \overset{i.i.d.}{\sim} Unif[0,1]$ and $g(W_i, W_j) = \sum_{k,l=1}^{K} p_{kl} I\left(W_i \in [\sum_{j=0}^{k-1}\pi_j, \sum_{j=0}^{k}\pi_j], W_j \in [\sum_{j=0}^{l-1}\pi_j, \sum_{j=0}^{l}\pi_j]\right)$, where $\pi_0 = 0$ and $\sum_{j=0}^{K}\pi_j = 1$.

$$\begin{aligned} A_{ij}|g, W_i, W_j &\overset{ind}{\sim} Bern\big(g(W_i, W_j)\big), \forall i < j \\ A_{ij}|g &\overset{i.i.d}{\sim} \int\int Bern\big(g(W_i, W_j)\big) f_W(w_i) f_W(w_j) dw dw. \end{aligned} \tag{4}$$

Even though this is not the only representation of edge distribution, for any exchangeable graphs, including SBM and also Random dot product graph (RDPG), there must exist a random function g which edges are independent identically distributed conditioning on.

2.1.1 Exchangeability on point process

Graphon has been studied widely as a limit of random graphs (Lovász and Szegedy, 2006). However, despite its advantage on simple representation, it is either empty or dense. A precise definition of dense graph and sparse graph is followed by Veitch and Roy (2015).

Definition 2.3 (sparse (not dense) graph)**.** Let $\mathbf{G} = \big(V, E\big)$ be a graph and $|V|$ and $|E|$ denote the number of nodes and edges of G. Then graph G is sparse or not dense if $|E|$ is asymptotically $o(|V|^2)$, i.e.

$$\frac{\sqrt{|E|}}{|V|} \xrightarrow{p} 0 \quad \text{as} \quad n \to \infty. \tag{5}$$

Our exchangeable graph often fails to represent real network data where sparsity or scale-free distribution is fairly common. Thus, in addition to graphon, we introduce a concept of *graphex*, first proposed by Veitch and Roy (2015), which is more generalized version of graphon and also includes sparse exchangeable graphs (Caron and Fox, 2014). Caron and Fox (2014) suggested formalizing a network as point process on $\mathbb{R}_+^2$ based on *Kallengerg Representation Theorem* (Kallenberg, 1990). As we were able to represent $\{A_{ij}\}$ through random transformation of *i.i.d* uniform variables, jointly exchangeable point processing network also can be represented via a random function of *i.i.d* unit rate Poisson process and of *i.i.d* uniform variables. To be specific, undirected graph on a point process on $\mathbb{R}_+^2$ can be thought of

$$\mathbf{A} = \sum_{i,j} A_{ij} \delta_{(\theta_i, \theta_j)} \tag{6}$$

where $A_{ij} = A_{ji} \in \{0, 1\}$ with node label space $\theta \in \mathbb{R}_+$, $i, j = 1, 2, ...,$ e.g. node i is assumed to be embedded on real line, at $\theta_i \in \mathbb{R}_+$.

Definition 2.4 (graphex Kallenberg (1990))**.** Random graphs defined on exchangeable random measures are characterized by triple (I, S, g), *graphex*, where a $I \in \mathbb{R}_+$ is non-negative real, $S : \mathbb{R}_+ \to \mathbb{R}_+$ is an integrable function and $g : \mathbb{R}_+^2 \to [0, 1]$. Then conditional on g and unit rate Poisson processed $\theta \times \vartheta$, random graph $\mathbf{G}$ of graphex with node set $\{\theta\}$ is constructed as:

$$A_{\theta_i \theta_j} = g(\vartheta_i, \vartheta_j) \tag{7}$$

and exclude θ_i from $\mathbf{G}$ if θ_i is isolated and we can obtain finite subgraphs by restricting $\theta < \nu$ for some $\nu > 0$.

Then joint exchangeability applied to node itself now corresponds to joint exchangeability of a point processed node label θ, not on a node label itself:

Definition 2.5 (Joint exchangeability on point process)**.** Let $h > 0$ and $V_i = [h(i-1), hi]$ for $i \in \mathbb{N}$ then

$$\big(A(V_i \times V_j)\big) \stackrel{d}{=} \big(A(V_{\sigma(i)} \times V_{\sigma(j)}\big) \tag{8}$$

for any permutation σ of $\mathbb{N}$.

Complexities induced by Poisson process on node will require one more condition in testing in the next section. Despite its more intricate form, representation of sparse graph as exchangeable formation helps us to demonstrate the validity of the methods still applicable in real data.

2.2 Multiscale Distance Metrics

2.2.1 Diffusion maps and diffusion distance

Despite a surge of research on network representation in terms of a summarizing network factor (Hoff et al., 2002) or some meaningful coefficients, e.g.centrality (Mantzaris et al., 2013; Sporns et al., 2007), there has been no node-specific variable which provides a configuration of node over network space without losing any information. Coifman and Lafon (2006) proposed a meaningful multiscale geometries of data called *diffusion maps* while keeping every information on every local relation over

a graph. Diffusion map is constructed via Markov chain on graph. Without any model assumption on graph, an adjacency matrix $\mathbf{A}$ acts as a kernel, representing a similarity between each node in $\mathbf{G}$; thus we do not have to estimate anything in order to obtain multidimensional representation of network topology.

Let $(\mathbf{G}, \mathcal{A}, \mu)$ be a measure space. Throughout all of the arguments, assume that we have a countable node set with size of $n \in \mathbb{N}$. A network $\mathbf{G}$ is the data set of nodes and edges, and $\mathcal{A}$ is a set of pairs of nodes $\{(i,j) : v_i, v_j \in V(\mathbf{G})\}$. A measure of μ which represents a distribution of the nodes on $\mathbf{G}$, is equivalent to an adjacency matrix $\mathbf{A}$. Figure 2 illustrates one example of network $\mathbf{G}$ under SBM and its attributes which follow the provided probability matrix in the most left. However its realized edge distribution, denoted by adjacency matrix $\mathbf{A}$ and its attribute values $\mathbf{X}$ contain lots of noise. Euclidean distance of adjacency matrix $\mathbf{A}$ still possess block patterns which differentiate edge distribution depending $\mathbf{X}$ but we are going to consider different kernel for both theoretical and empirical reasons.

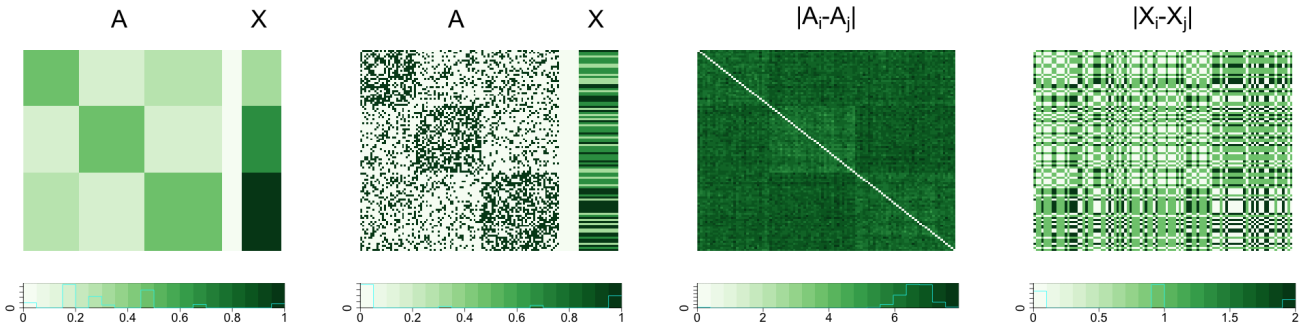

Figure 2: Population probability distribution and realized values of A and X (scaled by 1/3) in the left two panels. Euclidean distance applied to realized A and X are presented in the right two panels. Data generation follows Eq. 27.

A transition matrix P is a new kernel of a Markov chain of which element $P[i,j]$ represents the probability of travel from Node i to Node j in one time step. A transition matrix $P = \{P[i,j] : i,j = 1, ..., n\}$ in a Markov chain on $\mathbf{G}$ is defined as below:

$$P[i,j] = A_{ij} / \sum_{j=1}^{n} A_{ij}. \tag{9}$$

A corresponding probability in t step is given by the t^{th} ($t \in \mathbb{N}$) power of P. Now we assume that diffusion process occurs within a given graph with this transition probability at each time step. Distance

between a pair of nodes at each time is called *diffusion distance*. That is we have distance between every pair of nodes throughout diffusion process. How to derive diffusion distance over a directed network or weighted network is provided in Tang and Trosset (2010). Other than a transition matrix, we need a stationary probability $\pi = \{\pi(1), \pi(2), ..., \pi(n)\}$ of which $\pi(i)$ represents the probability that the diffusion process in the end is stuck in Node i regardless of the starting state. In our setting, $\pi(i)$ is assumed to be proportional to the degree of Node i, i.e. $\pi(i) = \sum_{j=1}^{n} A_{ij} / \sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij}$ $(i = 1, 2, ..., n)$.

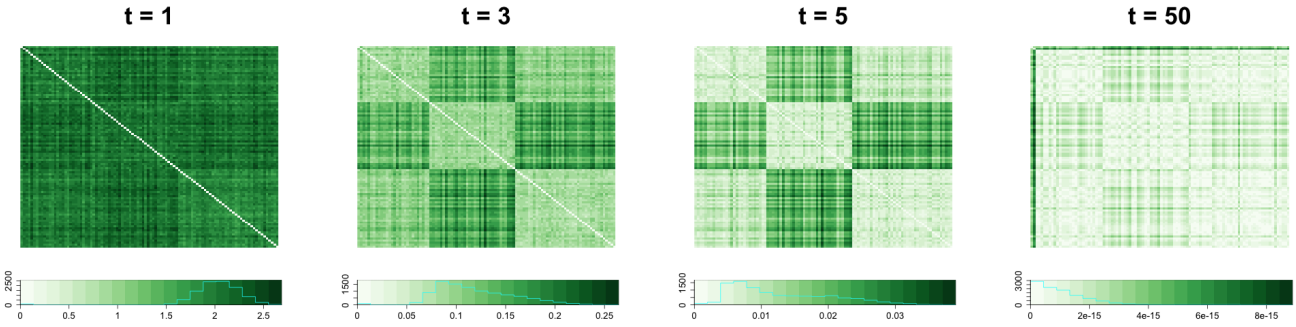

Figure 3: Diffusion distance, i.e. Euclidean distance of diffusion maps at $t = 1$, $t = 3$, $t = 5$, and $t = 50$ of sample graph G from Stochastic Block Model provided in simulation (Eq. 27).

For each time point $t \in \mathbb{N}$, we can define a diffusion distance at time t, C_t for which:

$$
\begin{aligned}
C_t^2[i,j] &= \sum_{w=1}^{n} \left(P^t[i,w] - P^t[j,w]\right)^2 \frac{1}{\pi(w)} = \sum_{w=1}^{n} \left(\frac{P^t[i,w]}{\sqrt{\pi(w)}} - \frac{P^t[j,w]}{\sqrt{\pi(w)}} \right)^2 \\
&= \| P^t[i,\cdot] - P^t[j,\cdot] \|^2_{L^2(\mathbf{G}, d\mu/\pi)} \qquad i,j = 1,2,\ldots,n
\end{aligned}
\tag{10}
$$

As diffusion time t increases, distance matrix C_t is more likely to take into account distance between two nodes which are difficult to reach each other. Basically diffusion distance at fixed time t measures the chance that we are likely to stay between Node i and Node j at t step on our journey of all other possible paths too. The higher chance is, the smaller distance between two is. Depending on the distribution of network or graph, the optimal time t when the diffusion distance is most correlated to the distance in terms of $\mathbf{X}$. If you see Figure 3, difference between blocks in distance matrix at $t = 3$ looks more distinct than that at $t = 1$. If you move to diffusion time $t = 5$, you are able to distinguish every block in upper (or lower) diagonal. However if the propagation takes enough, it becomes hard to detect the differences as nodes under the peer influence are at the end assimilated. On the other hand, since it takes into account every possible path between two nodes in contrast to adjacent relation or geodesic

distance, diffusion distance well reflects the connectivity. Simply speaking, connectivity between two nodes is higher if we need to eliminate more number of nodes to disconnect these two. It is more robust measure to the unexpected edges than geodesic distance. Often a set of nodes with higher connectivity have a higher propensity of having edges within this set and they are likely to form a cluster. This kind of cluster can be considered as a block in SBM framework.

Diffusion distance of $\mathbf{G}$ defined as above can be represented via a spectral decomposition of its transition matrix P. That is, we can derive diffusion distance using its eigenvectors and eigenvalues. The spectral analysis on diffusion distance or diffusion maps have been studied mainly for its usefulness in nonlinear dimensionality reduction (Coifman and Lafon, 2006; Lafon and Lee, 2006). Recall that diffusion distance at time t, C_t, is a functional L^2 distance, weighted by $1/\pi$ in Eq. 10. If we transform the way to represent $C_t[i,j]$ slightly, we are able to obtain an orthonormal basis of $L^2(\mathbf{G}, d\mu/\pi)$ via eigenvalues and eigenvectors. Since an adjacency matrix A does not guarantee a symmetric of P, define a symmetric kernel $Q = \Pi^{1/2} P \Pi^{-1/2}$, where Π is a $n \times n$ diagonal matrix of which ith diagonal element is $\pi(i)$. Under compactness of P, Q has a discrete set of real nonzero eigenvalues $\{\lambda_r\}_{r=\{1,2,...,q\}}$ and a set of their corresponding orthonormal eigenvectors $\{\psi_r\}_{r=\{1,2,...,q\}}$, i.e. $Q[i,j] = \sum_{r=1}^{q} \lambda_r \psi_r(i)\psi_r(j)$ $(1 \le q \le n)$. Return to transition probability between Node i and Node j,

$$\begin{aligned} P[i,j] &= \sqrt{\pi(j)/\pi(i)}\, Q[i,j] \\ &= \sum_{r=1}^{q} \lambda_r \big\{\psi_r(i)/\sqrt{\pi(i)}\big\}\big\{\psi_r(j)\sqrt{\pi(j)}\big\} \\ &:= \sum_{r=1}^{q} \lambda_r \phi_r(i)\big\{\psi_r(j)\sqrt{\pi(j)}\big\} \end{aligned} \tag{11}$$

where $\phi_r(i) := \psi_r(i)/\sqrt{\pi(i)}$. Then from $\sum_{r=1}^{q} \psi_r^2(j) = 1$ for all $j \in \{1, 2, ..., n\}$, we can represent the diffusion distance as:

$$C_t^2[i,j] = \sum_{r=1}^{n} \lambda_r^{2t} \big(\phi_r(i) - \phi_r(j)\big)^2 \tag{12}$$

That is,

$$C_t[i,j] = \| \mathbf{U}_t(i) - \mathbf{U}_t(j) \| \tag{13}$$

where

$$\mathbf{U}_t(i) = \begin{pmatrix} \lambda_1^t \phi_1(i) \\ \lambda_2^t \phi_2(i) \\ \vdots \\ \lambda_q^t \phi_q(i) \end{pmatrix} \in \mathbb{R}^q. \tag{14}$$

Now we have a family of q-variate($q \leq n$) diffusion maps $\{\mathbf{U}_t\}_{t\in\mathbb{N}}$, of which Euclidean distance is diffusion distance. Embedding each node on Euclidean metric is a novel approach in testing independence on network space; there is no estimation nor model assumption involved. However there remains a matter of dependence between observed diffusion maps.

2.2.2 Properties of diffusion maps under exchangeable graphs

We wish that diffusion maps are multivariate configuration of each node whose distance metric well reflects relative location on network space. However, due to the inter-correlated construction of $\mathbf{U}_t$, e.g. ith subject's diffusion depends on others in the given network, it is hard to say that the observed diffusion coordinates of n subjects are independent observations. As for independence of $\mathbf{U}_t$, we need a concept of exchangeable graph explained in the earlier section.

Lemma 2.1 (Exchangeability and *i.i.d* of A in graphon)**.** Assume that a connected, undirected and unweighted graph $\mathbf{G}$ is a graphon. Then 2-array of $\{A_{ij} : i = 1, 2, ..., n, i < j\}$ are *i.i.d* conditioning on some random link function $g : [0, 1]^2 \to [0, 1]$. Thus for fixed row (column) of $\mathbf{A}$, $\{A_{i1}, A_{i2}, ..., A_{in}\}\backslash\{A_{ii}\}$, $i \in \{1, 2, ..., n\}$ are conditionally i.i.d. on random link function g or equivalently, its underlying distribution.

From the above Lemma 2.1, we can also prove exchangeability and conditional *i.i.d* of diffusion maps at each time point.

Lemma 2.2 (Exchangeability and *i.i.d* of $\mathbf{U}_t$)**.** Assume that a connected, undirected and unweighted graph $\mathbf{G}$ is a graphon, i.e. any exchangeable random graph from an infinite graph. Then its transition probability so thus diffusion maps at fixed time t also exchangeable, conditional on link function of graph. Furthermore, by *de Finetti's Theorem* A.1, we can say that such diffusion maps at t are conditionally *i.i.d* given its underlying distribution, specifically given random probability measure η on U_t and random link function g.

Lemma 2.2 above provides us *i.i.d* one-parameter family of $\{\mathbf{U}_t\}_{t\in\mathbb{N}}$ conditional on a random probability measure of $\mathbf{U}_t$ and a random link function of g. At each time of t, q-variate diffusion coordinate assigns each node to the position where t step diffusion process results. Unfortunately this is a story only applied to exchangeable graph, which cannot be sparse. If we want to embed a set of nodes in sparse graphs, one more step of conditioning on point process θ is needed, explained in Def. 2.4.

Lemma 2.3 (Exchangeability and *i.i.d* of A in graphex)**.** Assume that a connected, undirected and unweighted graph $\mathbf{G}$ is a graphex. Then 2-array of $\{A_{ij} : i = 1, 2, ..., n, i < j\}$ are *i.i.d* conditioning on some random link function $g : [0, 1]^2 \to [0, 1]$ and unit-Poisson process θ. Thus for fixed row (column) of $\mathbf{A}$, $\{A_{i1}, A_{i2}, ..., A_{in}\} \setminus \{A_{ii}\}$, $i \in \{1, 2, ..., n\}$ are conditionally *i.i.d* on its underlying distribution, specifically conditioning on random link function g and θ.

Similar to Lemma 2.2, we are able to prove exchangeability of a transition matrix conditional on g and also on θ. Implicit interpretation of conditioning on node generating process θ is not very clear. However in testing independence between networks and nodal attributes, we are usually given a fixed number of nodes and network topology often implies edge structures.

2.3 Applying to Multiscale Generalized Correlation

2.3.1 Distance correlation and its multiscale version

Relationship between network and nodal attributes often exhibits local or nonlinear properties. Moreover, dimension of network spectrum, e.g. $q(\leq n)$ in case of diffusion maps, often increases as a sample size increases. Unfortunately, widely used correlation measures often fail to capture nonlinear associations especially embedded in high-dimensional data set. Székely et al. (2007) extended pairwise constructed generalized correlation coefficient and developed a novel statistic called distance correlation (`dCorr`) as a measure for all types of dependence between two random vectors in any dimension. Let us first start from a general setting that we are given $n \in \mathbb{N}$ pairs of random samples $\{(\mathbf{x}_i, \mathbf{y}_i) : \mathbf{x}_i \in \mathbb{R}^q, \mathbf{y}_i \in \mathbb{R}^m, i = 1, ..., n\}$. Define $C_{ij} = \| \mathbf{x}_i - \mathbf{x}_j \|$ and $D_{ij} = \| \mathbf{y}_i - \mathbf{y}_j \|$ for $i, j = 1, 2, \ldots, n$, where $\| \cdot \|$ denotes Euclidean distance defined on any vectors. Distance correlation (`dCorr`) is defined via distance covariance (`dCov`) $\mathcal{V}_n^2$ of $\mathbf{X}$ and $\mathbf{Y}$, which is the following:

$$\mathcal{V}_n^2(\mathbf{X}, \mathbf{Y}) = \frac{1}{n^2} \sum_{i,j=1}^{n} \tilde{C}_{ij} \tilde{D}_{ij}, \tag{15}$$

where $\tilde{C}$ and $\tilde{D}$ is a doubly-centered C and D respectively, by its column mean and row mean. Distance correlation $\mathcal{R}_n^2(\mathbf{X}, \mathbf{Y})$ is a standardized `dCov` scaled by $\mathcal{V}_n^2(\mathbf{X}, \mathbf{X})$ and $\mathcal{V}_n^2(\mathbf{Y}, \mathbf{Y})$.

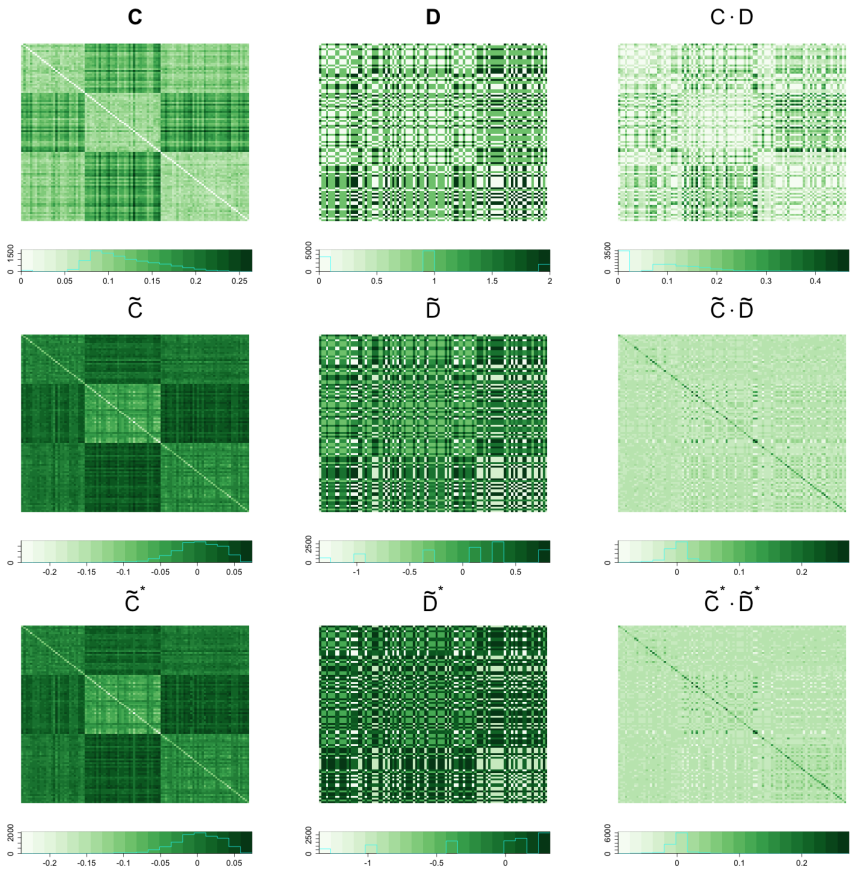

Figure 4: (a) Top : Euclidean distance of diffusion maps at time $t = 3$, C; Euclidean distance of X, D; element-wise product of C and D. (b) Middle : double-centered C, $\tilde{C}$; double-centered D, $\tilde{D}$; element-wise product of doubly centered distance matrices $\tilde{C} \cdot \tilde{D}$. (c) Bottom : truncated double-centered $\tilde{C}$ by k^*th nearest neighbor in C; truncated double-centered $\tilde{D}$ by l^*th nearest neighbor in D; element-wise product of two truncated matrices $\tilde{C}^*$ and $\tilde{D}^*$.

$$\mathcal{R}_n^2(\mathbf{X}, \mathbf{Y}) = \frac{\mathcal{V}_n^2(\mathbf{X}, \mathbf{Y})}{\sqrt{\mathcal{V}_n^2(\mathbf{X}, \mathbf{X})\mathcal{V}_n^2(\mathbf{Y}, \mathbf{Y})}} \tag{16}$$

On the other hand, a modified distance covariance (`mCov`) $\mathcal{V}_n^*$ and a modified distance correlation (`mCorr`) $\mathcal{R}_n^*$ for testing high dimensional random vectors were also proposed in Székely and Rizzo (2013a). However, `dCorr` nor even `MCorr` still perform not very well in the existence of various nonlinear dependency and under existence of outliers [Cencheng]. Out of this concern, Cencheng at al (2016) proposed Multi-scale Generalized Correlation (`MGC`) by adding local scale in a sense of nearest neighbors on correlation

coefficients. Multiscale version of distance covariance $\{\mathcal{V}^{*2}_n\}_{kl}$ is defined as following :

$$\begin{aligned}\mathcal{V}^{*2}_n(\mathbf{X},\mathbf{Y})_{kl} &= \frac{1}{n^2}\sum_{i,j=1}^{n}\tilde{C}_{ij}\tilde{D}_{ij}I\big(r(C_{ij})\leq k\big)I\big(r(D_{ij})\leq l\big)\\ &=: \frac{1}{n^2}\sum_{i,j=1}^{n}\tilde{C}^{*}_{ij}\tilde{D}^{*}_{ij}, \quad k,l=1,2,...,n\end{aligned} \tag{17}$$

where $r(C_{ij})$ ($r(D_{ij})$) denotes a rank $\mathbf{x}_i$ ($\mathbf{y}_i$) relative to $\mathbf{x}_j$ ($\mathbf{y}_j$). It basically truncates each pairwise element of distance covariance with respect to rank in terms of (Euclidean) distance. Note that if $k = l = n$, $\mathcal{V}^2_n$ and $\mathcal{V}^{*2}_n$ are equivalent. You can see panels to illustrate testing procedures and results in Figure 4. Simply speaking, given an appropriate distance matrix for $\{\mathbf{x}_i : i = 1, 2, \ldots, n\}$ and $\{\mathbf{y}_i : i = 1, 2, \ldots, n\}$ each, we take an double centering and truncate them by column ranking up to k and l respectively. Since we call all family of $\{\mathcal{R}^{*2}_n\}_{k,l=1,2,...,n}$ as `MGC`, `MGC` is more generalized version of `dCorr`. How to choose the optimal neighborhood scale of (k, l), say (k^*, l^*), is illustrated in [Cencheng] as well as its superiority and consistency. In simulation 3 we are going to show in which pattern of underlying dependency particularly `MGC` exhibits improved sensitive than the global scale. On the other hand, we often do not know the correct optimal scale of (k^*, l^*) given one single observations; while the optimal neighborhood choice can be closely estimated by simulated networks. Following the terminology made by [Cencheng], we call `MGC` when (k^*, l^*) should be estimated upon a single network as `Sample MGC` while `Oracle MGC` denotes `MGC` at near the true scale of (k^*, l^*). We can only obtain `Oracle MGC` in simulation data or when underlying network is known.

2.3.2 Choice of proper metric on network

We are still required a set of *i.i.d* node-specific coordinates of which Euclidean distance reflects a network-based distance between nodes, i.e. distance matrix C in Figure 4, in order to test independence via `MGC` (We are not always required Euclidean metric (Lyons et al., 2013) but discussion on this is out of scope for this paper). You might first propose directly using a column of an adjacency matrix so that we have a n-pair of observations $\big\{\big(\mathbf{A}_{i\cdot}, \mathbf{X}_i\big) : \mathbf{A}_{i\cdot} = (A_{i1}, ..., A_{in}) \in \mathbb{R}^n, \mathbf{X}_i \in \mathbb{R}^m, i = 1, ..., n\big\}$. In an undirected exchangeable graph, $\{\mathbf{A_{i\cdot}}\}$ cannot be independent. Even if it is in directed graph, Euclidean distance between $\{\mathbf{A}_{i\cdot} : i = 1, ..., n\}$ is not a proper metric over network space. Let us introduce a simple example. Let a given network $\mathbf{G}$ having 8 nodes be an unweighted, directed network

and possibly allowing self-loop. Let $\mathbf{A}$ be its 8×8 binary adjacency matrix. Assume `Node 1`, `Node 4` and `Node 8` have the following row entries:

$$\begin{aligned} \mathbf{A}_{1\cdot} &= \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix} \\ \mathbf{A}_{4\cdot} &= \begin{pmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \end{pmatrix} \\ \mathbf{A}_{8\cdot} &= \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \end{aligned} \tag{18}$$

which results $\| \mathbf{A}_{1\cdot} - \mathbf{A}_{4\cdot} \|^2 = 4$, $\| \mathbf{A}_{1\cdot} - \mathbf{A}_{8\cdot} \|^2 = 7$, and $\| \mathbf{A}_{4\cdot} - \mathbf{A}_{8\cdot} \|^2 = 3$. Accordingly, $\| \mathbf{A}_{4\cdot} - \mathbf{A}_{8\cdot} \| < \| \mathbf{A}_{1\cdot} - \mathbf{A}_{4\cdot} \|$. However, you can easily see that this does not make sense because `Node 4` and `Node 8` are connected each other only through `Node 1`. Therefore instead of using an adjacency matrix directly, both for theoretical and empirical reasons, we are considering embedding a vertex $v \in V(\mathbf{G})$ into its diffusion map of $\{\mathbf{U}_t\}$ and apply Euclidean distance metric. As explained before, its Euclidean distance takes into account all possible paths between every pair of nodes and measure the connectivity between them. Unlike in the other metrics in network, i.e. adjacency matrix or geodesic distance, triangle inequality holds in diffusion distance. Proof is provided in Appendix.

Corollary 2.3.1 (Triangle inequality)**.** For fixed time t, let $C_t : V(\mathbf{G})^2 \to \mathbb{R}_+$ be a diffusion distance defined on a pair of nodes in any connected and undirected graph $\mathbf{G}$. Then for any $v, w, z \in V(\mathbf{G})$,

$$C_t(v, z) \leq C_t(v, w) + C_t(w, z) \tag{19}$$

Thanks to these properties of diffusion maps, we earn better interpretation of its Euclidean distance.

2.3.3 One parameter family of test statistic

We have discussed one-parameter family of exchangeable diffusion maps and their extension to conditional *i.i.d* variable. On the other hand, all the available independence test methods, as far as we know, including distance correlation are assuming a pair of *i.i.d* sample, say $(\mathbf{X}, \mathbf{Y}) = \{(\mathbf{x}_i, \mathbf{y}_i), i = 1, 2, ..., n\}$. Thus the validity of using exchangeable samples instead of *i.i.d* ones remain to be proven. First assume that $(\mathbf{x}_i, \mathbf{y}_i)$ is identically distributed as $(\mathbf{x}, \mathbf{y})$ with finite second moment.

Lemma 2.4. Then we have

$$\mathcal{V}_n^2(\mathbf{X}, \mathbf{Y}) \quad = \|g_{\mathbf{x},\mathbf{y}}^n(t, s) - g_{\mathbf{x}}^n(t) g_{\mathbf{y}}^n(s)\|^2,$$

where $g^n_\cdot$ is the empirical characteristic function based upon $\{(\mathbf{x}_i, \mathbf{y}_i), i = 1, 2, ..., n\}$, i.e.,

$$g^n_{\mathbf{x},\mathbf{y}}(t,s) = \frac{1}{n}\sum_{j=1}^n \exp\{i\,\langle t, \mathbf{x}_j\rangle + i\,\langle s, \mathbf{y}_j\rangle\},$$
$$g^n_{\mathbf{x}}(t) = \frac{1}{n}\sum_{j=1}^n \exp\{i\,\langle t, \mathbf{x}_j\rangle\},$$
$$g^n_{\mathbf{y}}(s) = \frac{1}{n}\sum_{j=1}^n \exp\{i\,\langle s, \mathbf{y}_j\rangle\}.$$

Lemma 2.5. We have

$$\mathcal{V}^2_n(\mathbf{X},\mathbf{Y}) \longrightarrow \mathcal{V}^2(\mathbf{x},\mathbf{y}) \qquad \text{as } n \to \infty \tag{20}$$

where $\mathcal{V}^2(\mathbf{x},\mathbf{y}) := \|g_{\mathbf{x},\mathbf{y}}(t,s) - g_{\mathbf{x}}(t)g_{\mathbf{y}}(s)\|^2$, and $g_\cdot$ is a characteristic function, e.g., $g_{\mathbf{x},\mathbf{y}}(t,s) = E\{\exp\{i\,\langle t, \mathbf{x}\rangle + i\,\langle s, \mathbf{y}\rangle\}\}$. It follows that

$$\mathcal{V}^2_n(\mathbf{X},\mathbf{Y}) \to 0 \quad \text{as } n \to \infty \tag{21}$$

if and only if $g_{\mathbf{x},\mathbf{y}}(t,s) = g_{\mathbf{x}}(t)g_{\mathbf{y}}(s)$, i.e., $\mathbf{x}$ is independent of $\mathbf{y}$.

Lemma 2.4 and its following Lemma 2.5 facilitate the use of distance correlation while satisfying *Theorem 2* in Székely et al. (2007).

Theorem 2.6. Suppose that we are given n pairs of exchangeable observations $(\mathbf{X},\mathbf{Y}) = \{(\mathbf{x}_i, \mathbf{y}_i), i = 1, 2, \ldots, n\}$ having finite second moment. Assume $\mathbf{x}_i \overset{i.i.d}{\sim} \mathbf{x}$ and $\mathbf{y}_i \overset{i.i.d}{\sim} \mathbf{y}, i = 1, 2, \ldots, n$. Then

$$\mathcal{V}^2_n(\mathbf{X},\mathbf{Y}) \longrightarrow 0 \quad \text{as } n \to \infty \tag{22}$$

if and only if $\mathbf{x}$ is independent of $\mathbf{y}$. Moreover, `dCorr` and `MGC` are consistent for testing dependence between $\mathbf{x}$ and $\mathbf{y}$, i.e., the testing power converges to 1 asymptotically for any dependency of finite moments.

Note that if $\{\mathbf{x}_i : i = 1, 2, \ldots, n\}$ are *i.i.d*, they are exchangeable. thus estimated latent factors, which are assumed *i.i.d* by Fosdick and Hoff (2015) can be applied to Theorem 2.6. We already have shown that even under undirected network, diffusion maps remain exchangeable at each diffusion time point t.

Theorem 2.7. MGC is consistent in testing network independence with any exchangeable graph metric and nodal attributes, in particular testing independence between underlying distribution of diffusion maps and nodal attributes $\{(\mathbf{u}_t, \mathbf{x}) : i = 1, 2, \ldots, n\}$.

$$H_0 \quad : \quad f_{\mathbf{U}^{(t)}.\mathbf{X}} = f_{\mathbf{U}^{(\mathbf{t})}} \cdot f_{\mathbf{X}} \tag{23}$$

In particular, the consistency also holds for the estimated latent positions and adjacency matrix of directed network.

Remind that we are still testing independence of nodal attributes to conditional distribution of network topology. Underlying distribution of $\mathbf{U}_t$, $f_{\mathbf{U}_t}\big(\eta, g\big) = f_{\mathbf{U}^{(t)}}$, given a link function g and a random probability measure η_t of $\mathbf{U}_t$ has been introduced to ensure *i.i.d* from exchangeability. Since a family of diffusion maps, $\{\mathbf{U}_t\}_{t\in\mathbb{N}}$ provides a configuration of nodes in $\mathbf{G}$, the above hypothesis implies testing independence between the configuration of nodes in network space and in attribute space at each time of diffusion, but as a function of a link function g and a random function (variable) of η of $\mathbf{U}$.

Remark 1. Roughly speaking, we can say that diffusion maps are *i.i.d* function of a link function g and a random function of $\mathbf{U}$ of η. Thus testing independence between conditional U and X can be considered as testing independence between $f\big(g, \eta\big)$ and X. A link function g concerns the distribution of edges and a random function η concerns nature distribution of diffusion maps. Our testing basically examines whether how edges are constructed and how diffusions(propagation) process are correlated to nodal attributes.

We are not going to state otherwise but if you assume to be given a unit-Poisson process $\{\theta_i\}_{i=1}^n$, you can lead to the same results for sparse graphex as Theorem 2.6. Even though we are not able to present testing results in *all* alternatives, in the following section a few examples of sparse networks as well as exchangeable networks will help you understand when and why our proposing method performs better than others.

2.3.4 Each node's contribution to dependency measure

In the presence of nonlinear dependency or local (in)dependency, some nodes often exerts more reliance on their attributes than other nodes since the amount of dependence is not consistent over a set of nodes. Like other node-specific measure of its importance, e.g. centrality, each node's leverage on

dependency measure in Eq. 17 can be of interest. Here we actually measure each node's contribution to `MGC` statistic so that quantify how much its location on network space and its attributes are correlated. Let (k^*, l^*) be the optimal neighborhood choice in distance matrix (C, D). Denote the contribution of node $v \in V(G)$ to the testing statistic by $c(\cdot) : v \to \mathbb{R}$.

$$c(v) = \frac{1}{2n^2} \sum_{j=1}^{n} \left\{ \tilde{C}_{vj} \tilde{D}_{vj} I\big(r(C_{vj}) \leq k^*\big) I\big(r(D_{vj}) \leq l^*\big) + \tilde{C}_{jv} \tilde{D}_{jv} I\big(r(C_{jv}) \leq k^*\big) I\big(r(D_{jv}) \leq l^*\big) \right\} \quad (24)$$

3 Simulation Study

In simulation studies presented in this paper, we make a comparison of estimated testing power across various multivariate independence test statistics: `MGC`, `dCorr`(`mCorr`), Heller-Heller-Gorfine (`HHG`) (Heller et al., 2012), and likelihood ratio test of Fosdick and Hoff (`FH`). For computing statistical power, we used type I error $\alpha = 0.05$ and obtain p-values of each sample network via permutations. For fair comparison between these testing methods, we also present a additive model of latent factors, which is mostly targeted by `FH`. All the simulation models are illustrated by joint distribution of adjacent matrix $\mathbf{A}$, nodal attributes $\mathbf{X}$, and latent variable $\mathbf{Z}$, which explains dependence structure between $\mathbf{A}$ and $\mathbf{X}$.

$$\begin{aligned} f(\mathbf{A}, \mathbf{X}, \mathbf{Z}) &= f_{A|Z}(\mathbf{A}|\mathbf{Z}) \cdot f_{Z|X}(\mathbf{Z}|\mathbf{X}) \cdot f_X(\mathbf{X}) \\ &= f_{A|Z}(\mathbf{A}|\mathbf{Z}) \cdot f_{X|Z}(\mathbf{X}|\mathbf{Z}) \cdot f_Z(\mathbf{Z}) \end{aligned} \quad (25)$$

According to the joint model in Eq. 25, edge distribution and nodal attributes are correlated only through a node-specific latent variable $\mathbf{Z}$ no matter whether $\mathbf{X}$ is modeled via $\mathbf{Z}$ or vice versa.

For each simulated network, empirical power will be derived by comparing observed statistic to the empirical distribution under the null. Empirical distributions under both independence and dependence are shown in Figure 5 where the vertical lines indicate 95% quantiles of the null.

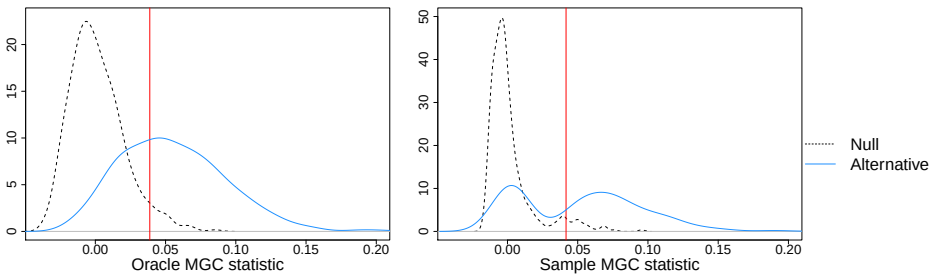

Figure 5: Statistics under null distribution and dependent distribution based on $M = 500$ independently generated SBM presented in Eq. 27

3.1 Stochastic Block Model

We mentioned in the Introduction that latent network model is very common followed by the assumption of local independence. Stochastic Block Model (SBM) is one of the most popular and also useful network generative model, especially as a tool for community detection (Karrer and Newman, 2011). We first present the simplest SBM with $K = 2$ blocks where block affiliation for each node is correlated with its attributes X. Let us introduce block affiliation (latent) variable $Z \in \{0, 1\}$. We set if $Z_i = Z_j$, $A_{ij} = 0.4$ and $A_{ij} = 0.1$ otherwise so that nodes in the same block are more likely to be adjacent than nodes having different value of Z.

- **Two Block SBM**

$$\begin{aligned} &X_i \overset{i.i.d}{\sim} f_X(x) \overset{d}{=} Bern(0.5), \quad i = 1, \dots, n \\ &Z_i | X_i \overset{i.i.d}{\sim} f_{Z|X}(z|x) \overset{d}{=} Bern(0.6)I(x = 0) + Bern(0.4)I(x = 1), \quad i = 1, \dots, n \\ &A_{ij} | Z_i, Z_j \overset{i.i.d}{\sim} f_{A|Z}(a_{ij}|z_i, z_j) \overset{d}{=} Bern(0.4)I(|z_i - z_j| = 0) + Bern(0.1)I(|z_i - z_j| > 0) \\ &\qquad i < j, i, j = 1, \dots n \end{aligned} \tag{26}$$

- **Three Block SBM 1**

$$
\begin{aligned}
&X_i \overset{i.i.d}{\sim} f_X(x) \overset{d}{=} Multi(1/3, 1/3, 1/3), i = 1, \ldots, n \\
&Z_i|X_i \overset{i.i.d}{\sim} f_{Z|X}(z|x) \overset{d}{=} Multi(0.5, 0.25, 0.25)I(x = 1) + Multi(0.25, 0.5, 0.25)I(x = 2) \\
&\quad + Multi(0.25, 0.25, 0.5)I(x = 3), \quad i = 1, \ldots, n \\
&A_{ij}|Z_i, Z_j \overset{i.i.d}{\sim} f_{A|Z}(a_{ij}|z_i, z_j) \overset{d}{=} Bern(0.5)I(|z_i - z_j| = 0) + Bern(0.2)I(|z_i - z_j| = 1) \\
&\quad + Bern(0.3)I(|z_i - z_j| = 2), \quad i < j, i, j = 1, \ldots n
\end{aligned} \tag{27}
$$

For both cases, let $A_{ij} = A_{ji}$ and $A_{ii} = 0$, $i, j = 1, \ldots n$. Figure 6 and 7 describe empirical power of `MGC`, `dCov`, and `HHG` based on diffusion distance, Euclidean distance of adjacent matrix and `FH`. In two block SBM, the performance of `MGC` is very similar to others while it is most superior in three block. This comparative advantage of `MGC` is attributed to its ability to capture non-linear dependence. Note that expectation of having edge between a pair of nodes is a monotonic function of (Euclidean) distance of their attributes in Eq. 26 but non-monotonic in Eq. 27.

$$
\begin{aligned}
&E(A_{ij}|X_i, X_j) = 0.6I(|X_i - X_j| = 0) + 0.4I(|X_i - X_i| = 1) \\
&E(A_{ij}|X_i, X_j) = 0.5I(|X_i - X_j| = 0) + 0.2I(|X_i - X_j| = 1) + 0.3I(|X_i - X_j| = 2)
\end{aligned} \tag{28}
$$

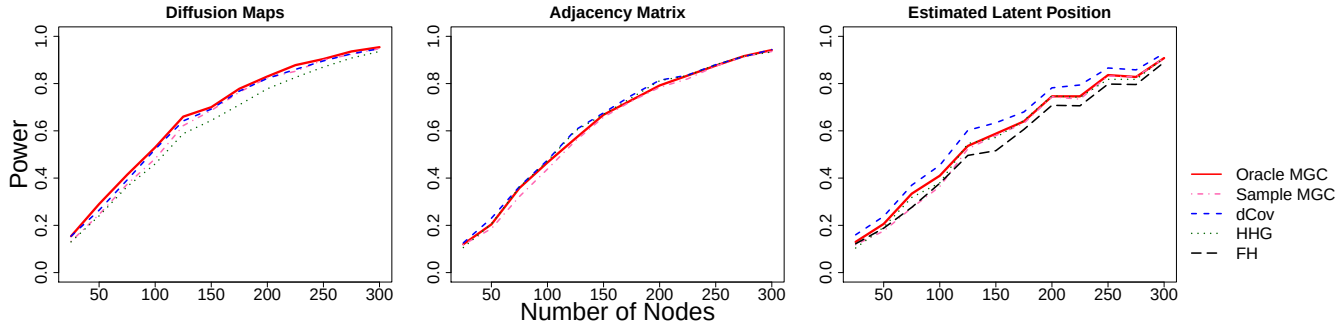

Figure 6: Empirical power based on $M = 500$ independently generated SBM presented in Eq. 26 using diffusion maps (left) and Euclidean of adjacency matrix (middle) and estimated latent position(right). The most right figure contains the results of `FH` test as well.

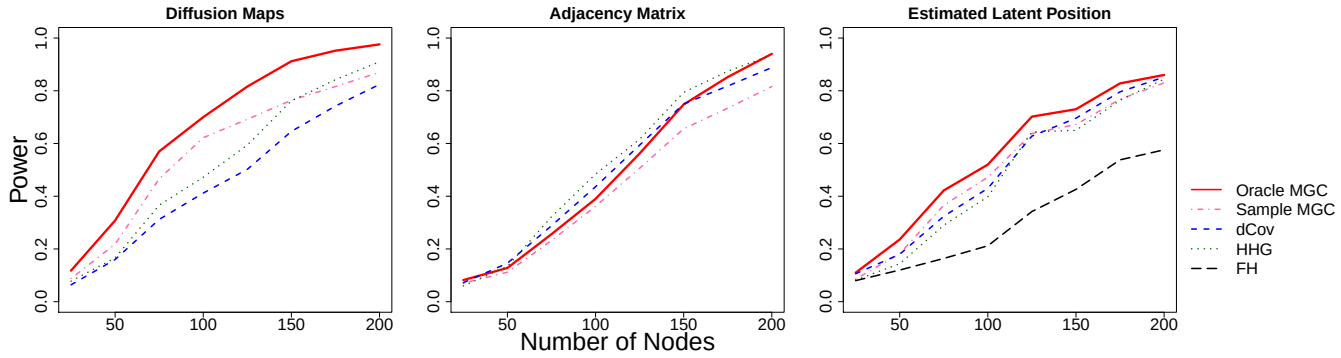

Figure 7: Empirical power based on $M = 500$ independently generated SBM presented in Eq. 27 using diffusion maps (left) and Euclidean of adjacency matrix (middle) and estimated latent position(right). The most right figure contains the results of `FH` test as well.

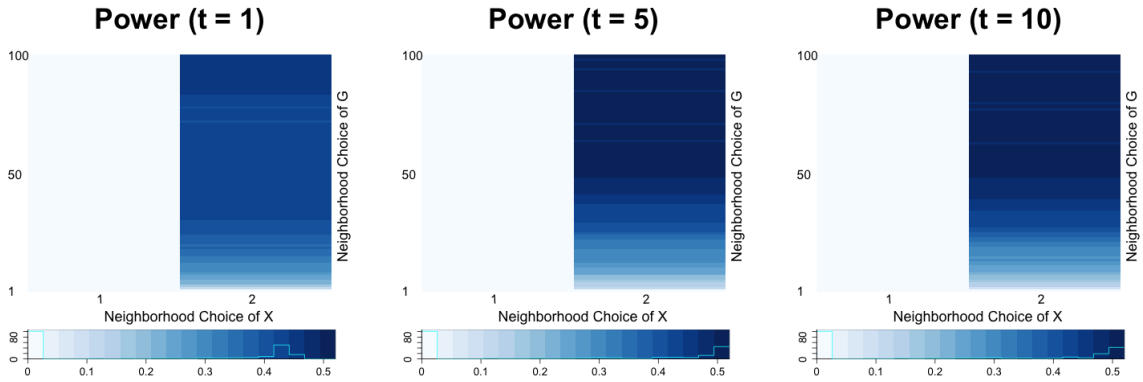

Figure 8: Empirical power heatmap at diffusion time point $t = 1$(left), $t = 5$(middle), and $t = 10$(right) based on $M = 500$ independently generated SBM presented in Eq. 26

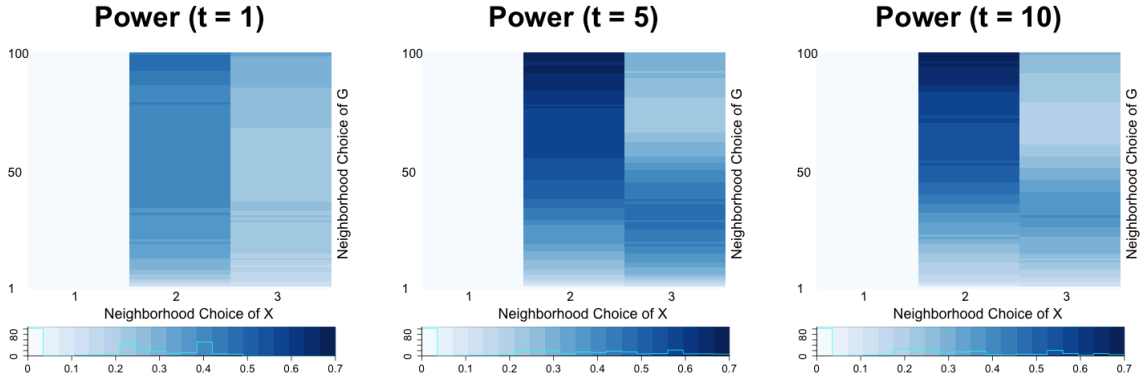

Figure 9: Empirical power heatmap at diffusion time point $t = 1$(left), $t = 5$(middle), and $t = 10$(right) based on $M = 500$ independently generated SBM presented in Eq. 27

Due to non-linear network dependence on attributes in the latter case, optimal neighborhood choice

of (k^*, l^*) results in not global scale. Figure 9 demonstrates that neighborhood choice of (k, l) in `MGC` statistic achieves its optimal in local scale. Roughly speaking, considering a pair of nodes in the nearest neighbor, in term of attribute values of X, e.g. (1,2) or (2,3), exhibits most significant dependence. When you only take into account these subsets of observations, you can actually represent $E(A_{ij}|X_i, X_j)$ as a monotonic function of $|X_i - X_j|$ (equivalently $\| \mathbf{X}_i - \mathbf{X}_j \|$ in multivariate case). On the other hand, even in the three block SBM, If you set every pair of nodes in different blocks to have the same propensity of having edges, discrepancy between local and global scale of (k, l) diminishes, which you can find in **Three Block SBM 2**, Appendix A.

To demonstrate better performance of local optimal of `MGC` over global scale of `dCorr` or `mCorr`, we control the amount of *non-linear dependency* through changing the value of $\theta \in (0, 1)$. When $\theta > 0.2$, linear dependency of edge distribution upon nodal attribute X is lost.

$$Power(\theta) = E(A_{ij}|X_i, X_j) = 0.5I(|X_i - X_j| = 0) + 0.2I(|X_i - X_j| = 1) + \theta I(|X_i - X_j| = 2) \quad (29)$$

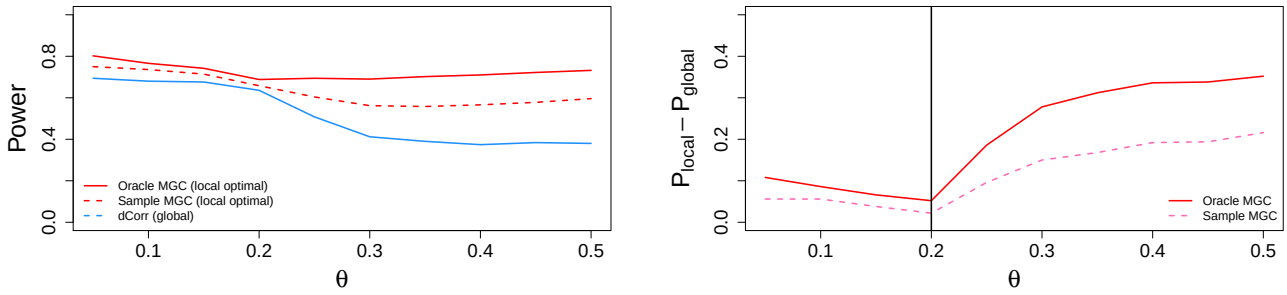

Figure 10: Change of empirical power across θ in both local and global scale of distance correlation (left). Change of difference between these two powers in Oracle and Sample `MGC`. Superiority of optimal local scale become evident from $\theta > 0.2$, when distribution of edges have non-linear dependence on X.

If you see Figure 10, power of `dCorr` starts to drop at $\theta = 0.2$ while that of `MGC` almost stays clam.

3.2 degree-corrected two block model

Under SBM, we assume that all nodes within the same block have the same expected degree. However, this block model is limited by homogeneous distribution within block and provides a poor fit to networks with hubs or highly varying node degrees within blocks or communities, which are common

in practice. On the other hand, the Degree-Corrected Stochastic Block model (DCSBM) proposed by Karrer and Newman (2011) adds an additional set of parameter, often denoted by θ, to control the node degrees. This model allows variation in node degrees within a block while preserving the overall block community structure. Consider two block SBM having same distribution of X and Z as Eq. 26 but having different adjacency matrix with more variability induced by θ :

$$\begin{aligned} &\theta_i \overset{i.i.d}{\sim} Uniform(0,2), i = 1, \ldots, n \\ &A_{ij}|\mathbf{Z}, \theta \overset{i.i.d}{\sim} f_{A|Z,\theta}(a_{ij}|z_i, z_j, \theta_i, \theta_j) \overset{d}{=} Bern(0.2 \cdot \theta_i\theta_j) I(|z_i - z_j| = 0) + Bern(0.05 \cdot \theta_i\theta_j) I(|z_i - z_j| = 1) \end{aligned} \tag{30}$$

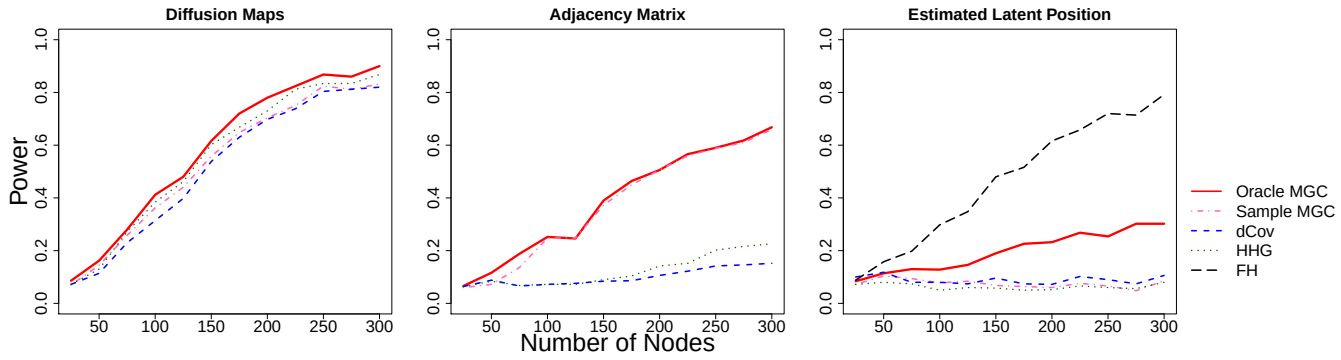

Figure 11: Empirical power based on $M = 500$ independently generated SBM presented in Eq. 30 using diffusion maps (left) and Euclidean of adjacency matrix (middle) and estimated latent position(right). The most right figure contains the results of `FH` test as well.

As in the previous two block SBM, in this simulation scheme, we also observe monotonic edge distribution function of Euclidean distance. However its variance becomes inflated due to θ (Eq. 31). If you see the testing results with Euclidean distance adjacency matrix in Figure 11, empirical power of `dCov` or `HHG` shows far less than that of `MGC`. This might imply that multiscale statistic performs better than the other global ones even in the presence of variability of degree distribution. On the other hand, node-specific variable θ is captured by latent position estimation model (Fosdick and Hoff, 2015) and

FH test looks most sensitive in this case.

$$
\begin{aligned}
E(A_{ij}|X_i, X_j) &= 0.2I(|X_i - X_j| = 0) + 0.05I(|X_i - X_i| = 1) \\
Var(A_{ij}|X_i, X_j) &= Var(Bern(0.2 \cdot \theta_i\theta_j))I(|X_i - X_j| = 0) + Var(Bern(0.05 \cdot \theta_i\theta_j))I(|X_i - X_j| = 1) \\
&> Var(Bern(0.2))I(|X_i - X_j| = 0) + Var(Bern(0.05))I(|X_i - X_j| = 1).
\end{aligned} \tag{31}
$$

3.3 Additive and multiplicative graph model

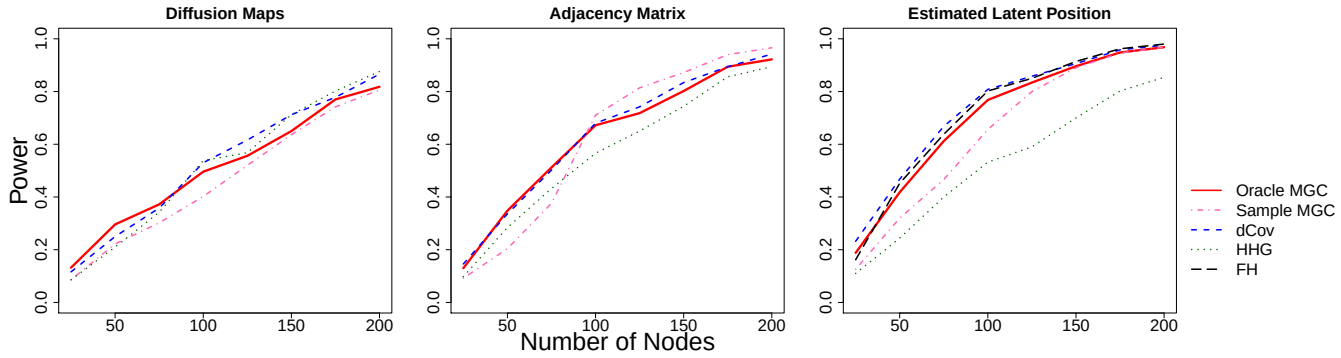

Figure 12: Empirical power based on $M = 500$ independently generated SBM presented in Eq. 32 using diffusion maps (left) and Euclidean of adjacency matrix (middle) and estimated latent position(right). The most right figure contains the results of FH test as well.

Hoff et al. (2002) proposed a approach of jointly modeling network and its attributes, where networks possess additional structure via sender-specific(or row-specific) and receiver-specific(or column-specific) latent factors. The following simulation scheme mimics the additive and multiplicative network models where distribution of A_{ij} is directly formalized through Z_i and Z_j.

$$
\begin{aligned}
&Z_i \overset{i.i.d}{\sim} f_Z(z) \overset{d}{=} Uniform[0,1]. \quad i = 1, \ldots, n \\
&X_i|Z_i \overset{i.i.d}{\sim} f_{Z|X}(z|x) \overset{d}{=} Normal(Z_i, 1), \quad i = 1, \ldots, n \\
&A_{ij}|Z_i, Z_j \overset{i.i.d}{\sim} f_{A|Z}(a_{ij}|z_i, z_j) \overset{d}{=} Bern\big((1 - z_i)^2 \times (1 - z_j)^2\big), \quad i < j, i, j = 1, \ldots, n
\end{aligned} \tag{32}
$$

A family of diffusion maps are nonparametric version of embedding nodes into multivariate variable without losing any information on adjacent relationship; while Fosdick and Hoff (2015) embedded nodes into network factors assuming that additive an multiplicative network model is *correct*. Thus in the

model explained in Eq. 32, where logit of A is an additive and multiplicative function of $\{\mathbf{w}\}$, their estimated factors would be very close to the truth, much closer than embedding made from diffusion maps. However we rarely see the network which is fitted to the model in reality. If you see that your observed network actually fits to their model, using network factors as independent observations from graph $\mathbf{G}$ and applying to `MGC` performs not very worse than `FH` statistic (Fig. 12). Simply speaking, if network really fits well to additive and multiplicative model or we have basic knowledge about network model, we is can make sure of their node-specific additive factor $\{a_i\}$ and multiplicative factor $\{\mathbf{m}_i^T : \mathbf{m}_i \in \mathbb{R}^k\}$ in testing directly. Since they assume $i.i.d$ generative model for factors of each node, $\mathbf{F}_i$, i.e. $\mathbf{F}_i = \begin{pmatrix} a_i & \mathbf{m}_i^T \end{pmatrix} \overset{i.i.d}{\sim} MVNormal$, there is nothing wrong with applying MGC using $i.i.d$ observations of $\{\mathbf{F}_i, \mathbf{X}_i\}$.

3.4 Exchangeable graph on Poisson process

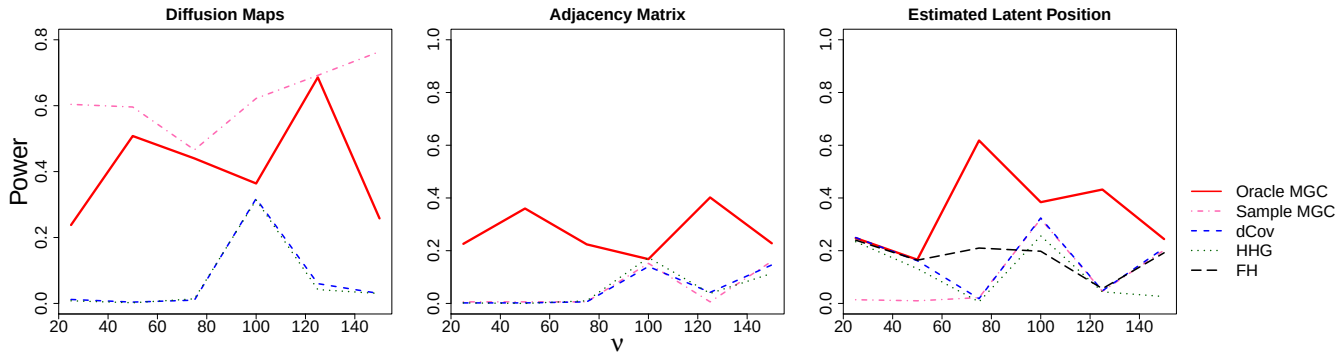

Figure 13: Empirical power based on $M = 500$ independently generated Poisson-processed exchangeable graphs presented in the model 33 using diffusion maps (left), Euclidean of adjacency matrix (middle) and estimated latent position(right). The most right figure contains the results of `FH` test as well.

We have discussed SBM and its derivative dcSBM in consideration of real network data. However an exchangeable network defined formally as Def. 2.1 is still dense or empty. Here we revisit the example of *graphex* which is provided in Veitch and Roy (2015), which has power-law degree distribution.

$$\begin{aligned} N_\nu &\overset{i.i.d}{\sim} f_N(n) \overset{d}{=} Poi(c\nu) & \vartheta_i &\overset{i.i.d}{\sim} f_\vartheta \overset{d}{=} Uniform[0,1] \\ \theta_i &\overset{i.i.d}{\sim} f_{\theta|N_\nu}(\theta) Uniform[0,\nu] & X_i &\overset{ind}{\sim} f_{X|\vartheta}(x_i) \overset{d}{=} Normal\left(\vartheta_i, 4^2\right) \\ (\theta_i,\theta_j) &\overset{ind}{\sim} f_{(\theta_i,\theta_j)|W,\vartheta}((\theta_i,\theta_j)) \overset{d}{=} Bernoulli\left(W\left(\vartheta_i,\vartheta_j\right)\right) & A_{ij} &\overset{d}{=} (\theta_i,\theta_j) \end{aligned} \tag{33}$$

where $c = 1$ and $W(\vartheta_i, \vartheta_j) = (\vartheta_i + 1)^{-2}(\vartheta_j + 1)^{-2}$ or 0 if $\vartheta_i = \vartheta_j$. Simulation results in Figure 13 does not show the linear power as ν increases since the variance of sample size (N_ν) also increases as ν increases. In most cases, performance of `Oracle MGC` looks much better than the others and `Sample MGC` succeeds in catching up `Oracle MGC` under metrics of diffusion maps.

4 Real Data Examples

4.1 MRI

needs contexts

We look into one of the connected components with $n = 95$ nodes of whole disconnected network.

$\{\mathbf{U}_t \in \mathbb{R}^{95}\}_{t\in\mathbb{N}}$ $\mathbf{X} = (x, y, z) \in \mathbb{R}^3$

It turns out that $q = 95$, i.e. full rank of transition matrix. In other words we are testing independence between 95-dimensional diffusion maps and 3-dimensional nodal attributes at each $t \in \mathbb{N}$.

test independence between brain network and its 3-dimensional locations; test independence between functional location and physical location.

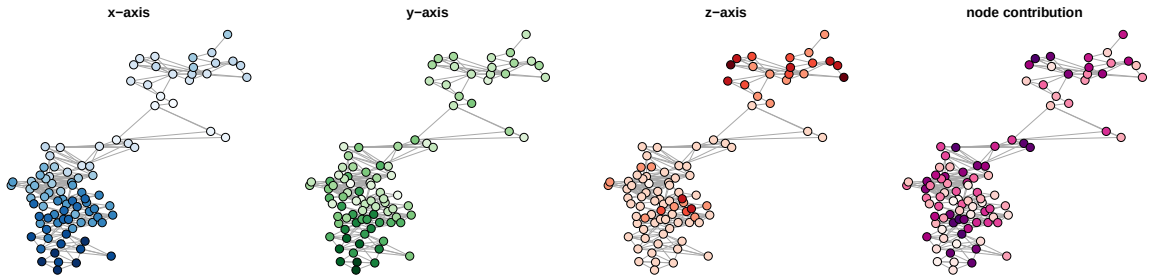

Figure 14: Subnetwork of MRI network. Darker colored nodes indicate higher positioned node in terms of x-axis(left), y-axis(middle), and z-axis(right).

5 Discussions

Throughout this study, we demonstrate that multiscale network test statistic to test network independence between network and its nodal attributes performs well in diverse settings, being supported by thorough theory on distance correlation and diffusion maps. These two tools are both specialized in detecting local and nonlinear dependence patterns, which other existing methods are lack of. However

testing independence is often the very first step in investigating relationship between network topology and nodal attributes in our interest. It is more likely that we want to know more than binary decision of rejecting or not rejecting the hypothesis. Multiscale test statistics attributed to both neighborhood choice $\{(k, l)\}$ and time spent in diffusion processes $\{t\}$ provides us a hint on latent dependence structure as well. However due to the ambiguousness of saying *optimal*, our work has some limitations; we do not suggest any theoretically supported tools to select the *optimal* time to obtain p-values of test statistics. We may further want a family of p-values as a function of t. Future research can be focused on restoring true dependence pattern or estimating *optimal* scale from a family of statistics. On the other hand, obtaining a full family of statistics are also computationally infeasible; at every Markov process, we chose the optimal region of neighborhood. As an ad hoc, we selected optimal t with highest power or lowest p-values from 1 to 10 for our simulation 3. Other than these computational issues, someone might be uncomfortable about being conditioned by a random function of g and η. However, we have to say that this is inevitable in order to argue sample properties of being *i.i.d.*, which is also very conceptual and impossible to prove. Through conditioning diffusion maps $\{\mathbf{U}_t\}$ by unknown network generative model $g(\cdot, \cdot)$ and also unknown diffusion process model $\eta(\cdot)$, we are finally able to assert that our observations are a fair sample eligible for test.

Despite a few shortcomings listed above, a range of applications of `MNT` statistics and node-specific representation of network is very diverse. Especially multiscale version of test would be very useful when indirect networking through diffusion process is not ignorable or cluster membership significantly affects attributes. Furthermore even though we specifically constraint the statistic into testing independence between network and nodal attributes, we are able to implement independence testing of two networks with same size by inputting diffusion distance at each time point from each of network in Eq. 17. This type of test will be useful when we want to show a pair of networks are topologically or structurally independent. For example, we might wonder if network on *Facebook* and network induced by club activity within class or school are independent or not; if DNA methylation network is correlated with gene expression network (Bartlett et al., 2014) so that the behavior of interest measured by DNA methylation network can be matched to that noticed by gene expression network, etc. In a broad sense, we suggest the proper metric applied to network and justify its use in measuring correlations between *network vs. attributes* or *network vs. network*. Even though we have fully presented procedures, e.g. testing on diffusion maps from $t = 1$ to $t = 10$ and obtaining a family of all local scales for each,

depending on the previous knowledge or on the results of model fitting, you may shorten the steps as well. We also presented the case where additive and multiplicative models work pretty well and how to modify the statistic in this case. Likewise the application and variation of multiscale network test is almost limitless.

A appendix

A.1 supplementary figures

Three Block SBM 2

$$
\begin{aligned}
&X_i \overset{i.i.d}{\sim} f_X(x) \overset{d}{=} Multi(1/3, 1/3, 1/3), \quad i = 1, \ldots, n \\
&Z_i|X_i \overset{i.i.d}{\sim} f_{Z|X}(z|x) \overset{d}{=} Multi(0.5, 0.25, 0.25)I(x = 1) + Multi(0.25, 0.5, 0.25)I(x = 2) \\
&\qquad + Multi(0.25, 0.25, 0.5), \quad i = 1, \ldots, n \\
&A_{ij}|Z_i, Z_j \overset{i.i.d}{\sim} f_{A|Z}(a_{ij}|z_i, z_j) \overset{d}{=} Bern(0.5)I(|z_i - z_j| = 0) \\
&\qquad + Bern(0.2)I(|z_i - z_j| > 1) \quad i < j, i, j = 1, \ldots n \\
&A_{ji} = A_{ij}; A_{ii} = 0, \quad i, j = 1, \ldots n
\end{aligned}
\tag{34}
$$

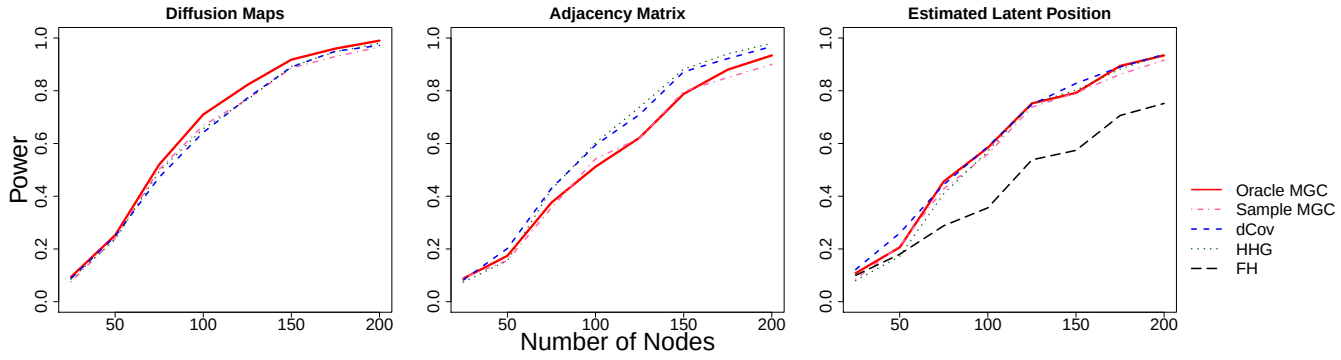

Figure 15: Empirical power based on $M = 500$ of three SBM under linear dependence using diffusion maps (left), Euclidean of adjacency matrix (middle) and estimated latent position(right). The most right figure contains the results of `FH` test as well.

A.2 Algorithms

Algorithm 1 Mutiscale representation of nodes in network

Input: Transition probability matrix of network G and time points($\in N$) of diffusion time.
Output: A list of diffusion maps at each time point.

function `DMAP` ($n \times n$ transition matrix P, time points $\{t_1, t_2, \ldots, t_K\}$)
  $\pi := \texttt{statdistr}(P)$ ▷ stationary distribution of P
  $\Pi := \texttt{Diag}(\pi)$ ▷ Diagonal matrix with diagonal element of π
  $Q := \Pi^{1/2} P \Pi^{-1/2}$
  $\lambda := \texttt{eigenvalue}(Q)$ ▷ a real-valued vector with length of $q(\leq n)$.
  $\Lambda := \texttt{Diag}(\lambda)$
  $\Psi := \texttt{eigenfunction}(Q)$ ▷ $n \times q$ real-valued matrix
  $\Phi := \Pi^{-1/2}\Psi$ ▷ $n \times q$ real-valued eigenfunction matrix of P
  for $t_i : i = 1$ **do** K
    $\texttt{Maps}[i] := \Phi\Lambda^{t_i}$
  end for
  Maps = list(Maps[1], Maps[2], ..., Maps[K])
  return Maps
end function

Algorithm 2 Multiscale Generalized Correlation (`MGC`) test statistics when diffusion maps are applied.

Input: A connected, undirected network G with its nodal attributes $\mathbf{X}$.
Output: A list of ((a) p-value of `sample MGC`, (b) estimated `sample MGC` statistic, (c) p-value map for all local correlations, (d) a set of estimated optimal neighborhood scales $\{(k^*, l^*)\}$) for each diffusion maps.

function `NetworkTest` (G, $\mathbf{X}$, $\mathbf{T}$:= (diffusion time points $\{t_1, t_2, \ldots, t_K\}$))
  $A := \texttt{get.adjacency}(G)$ ▷ obtain an adjacency matrix of network G
  $P := A$ / $\texttt{rowSums}(A)$
  $U := \texttt{dmap}(P, \mathbf{T})$ ▷ a list of diffusion maps in each time point
  for $t_i : i = 1$ **do** K
    $C := \texttt{dist}(U[i])$ ▷ distance matrix of diffusion maps at time t_i
    $D := \texttt{dist}(X)$ ▷ distance matrix of nodal attributes
    $\texttt{MGC}[i] = \texttt{MGCPermutationTest}(C, D)$
  end for
  MGC = list(MGC[1], MGC[2], ..., MGC[K])
  return MGC
end function

Algorithm 3 Node-specific contribution to detecting dependency via `MGC` statistic

Input: Distance metric of G and X for each and (one of) the estimated optimal scales $\{k^*, l^*\}$
Output: Standardized contributions of each node in network $\{c(v)\}$

function CONTRIBUTION (C, D , $\{(k^*, l^*)\}$)
  $\tilde{C} :=$ `DoubleCentering`(C)
  $\tilde{D} :=$ `DoubleCentering`(D)
  `Rank`(M_i):= (rank of node i's nearest neighbors in terms of distance matrix M)
  for $v = 1$ **do** n
    $c(v) = 0$
    for $j = 1$ **do** n
      $c(v) = c(v) + \tilde{C}_{vj}\tilde{D}_{vj}I(\texttt{Rank}(C_v) \leq k^*, \texttt{Rank}(D_v) \leq l^*)$
      $c(v) = c(v) + \tilde{C}_{jv}\tilde{D}_{jv}I(\texttt{Rank}(C_j) \geq k^*, \texttt{Rank}(D_j) \leq l^*)$
    end for
    $c(v) = c(v)/2n^2$
  end for
  `cset` $:= \{c(v) : v = 1, 2, \ldots, n\}$
  return `cset`
end function

A.3 Lemmas and Theorems

Theorem A.1 (de Finetti's Theorem)**.** 1. Let $X_1, X_2, ...$ be an infinite sequence of random variables with values in a space $\mathbf{X}$. The sequence $X_1, X_2, ...$ is exchangeable *if and only if* there is a random probability measure η on $\mathbf{X}$ such that the X_i are conditionally i.i.d. given η.

2. If the sequence is exchangeable, the empirical distributions

$$\hat{S}_n(.) := \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}(.), n \in \mathbb{N}$$

converges to η as $n \to \infty$ with probability 1.

Theorem A.2 (Aldous Hoover Theorem)**.** Let $\mathbf{A} = \{A_{ij}\}, 1 \leq i, j \leq \infty$ be a jointly exchangeable binary array if and only if there exists a random measurable function $f : [0,1]^3 \to \mathbf{A}$ such that

$$\big(A_{ij}\big) \stackrel{d}{=} \big(f\big(U_i, U_j, U_{ij}\big)\big) \tag{35}$$

where $(U_i)_{i\in\mathbb{N}}$ and $(U_{ij})_{i,j>i\in\mathbb{N}}$ with $U_{ij} = U_{ji}$ are a sequence and matrix, respectively, of i.i.d. Uniform[0,1] random variables.

Proof of Lemma 2.1 . By *Aldous-Hoover Theorem* A.2, a random array (A_{ij}) is jointly exchangeable *if and only if* it can be represented as follows :

There is a random function $g : [0,1]^2 \to [0,1]$ such that

$$(A_{ij}) \stackrel{d}{=} Bern(g(W_i, W_j)) \tag{36}$$

where $W_i \overset{i.i.d.}{\sim} Uniform(0,1)$. Thus if $\mathbf{A}$ is an adjacency matrix of an undirected, exchangeable network, for any $i < j$, $i, j = 1, ..., n$:

$$\begin{aligned} P\big(A_{ij} = a_{ij}\big) &= \int P\big(A_{ij}\big|w_i, w_j\big) Pr(W_i = w_i) Pr(W_j = w_j) dw_i dw_j \\ &= \int_0^1 \int_0^1 g(w_i, w_j)^{a_{ij}} \big(1 - g(w_i, w_j)\big)^{1-a_{ij}} dw_i dw_j \end{aligned} \tag{37}$$

Then within each row, adjacent elements are independent and also identically distributed except a diagonal element.

□

Proof of Lemma 2.2. We have shown that for fixed time t, diffusion distance is defined as an Euclidean distance of diffusion maps. Diffusion map is represented as follows :

$$\mathbf{U}_t(i) = \begin{pmatrix} \lambda_1^t\phi_1(i) & \lambda_2^t\phi_2(i) & \cdots & \lambda_q^t\phi_q(i) \end{pmatrix} \in \mathbb{R}^q. \tag{38}$$

where $\Phi = \Pi^{-1/2}\Psi$ and $Q = \Psi\Lambda\Psi^T = \Pi^{1/2}P\Pi^{-1/2}$. Thus $P\Pi^{-1/2}\Psi = \Pi^{-1/2}\Psi\Lambda$. Then for any rth row ($r \in \{1, 2, ..., q\}$, ($q \leq n$)), we can see that $P\phi_r = \lambda_r\phi_r$ where $\phi_r = \begin{pmatrix} \psi_r(1)/\sqrt{\pi(1)} & \psi_r(2)/\sqrt{\pi(2)} & \cdots & \psi_r(n)/\sqrt{\pi(n)} \end{pmatrix}$. Therefore to guarantee exchangeability (or $i.i.d$) of $\mathbf{U}_t$, it suffices to show exchangeability (or $i.i.d$) of P.

Assume joint exchangeability of $\mathbf{G}$, i.e. $(A_{ij}) \overset{d}{=} \left(A_{\sigma(i)\sigma(j)}\right)$. Since A_{ij} is binary, $A_{ij}/\sum_{ij} A_{ij} = A_{ij}/(1 + \sum_{l\neq j} A_{il})$. Moreover, A_{ij} and $(1 + \sum_{l\neq j} A_{il})$ are independent given its link function g, and $A_{\sigma(i)\sigma(j)}$ and $(1 + \sum_{l\neq j} A_{\sigma(i)\sigma(l)})$ are independent also given g. Then the following joint exchangeability of transition probability holds for $i \neq j; i, j = 1, 2, \ldots, n$:

$$(P_{ij}) = \left(\frac{A_{ij}}{1 - A_{ij} + \sum_{j=1}^{n} A_{ij}}\right) \overset{d}{=} \left(\frac{A_{\sigma(i)\sigma(j)}}{1 - A_{\sigma(i)\sigma(j)} + \sum_{\sigma(j)=1}^{n} A_{\sigma(i)\sigma(j)}}\right) = \left(P_{\sigma(i)\sigma(j)}\right) \tag{39}$$

When $i = j$, $P_{ij} = P_{\sigma(i)\sigma(j)} = 0$ for $i = 1, 2, \ldots, n$. Thus, transition probability is also exchangeable. This results exchangeable eigenfunctions $\{\Phi(1), \Phi(2), , ..., \Phi(n)\}$ where $\Phi(i) := \begin{pmatrix} \phi_1(i) & \phi_2(i) & \cdots & \phi_q(i) \end{pmatrix}^T$, $i = 1, 2, \ldots, n$. Thus diffusion maps at fixed t, $\mathbf{U}_t = \begin{pmatrix} \Lambda^t\Phi(1) & \Lambda^t\Phi(2) & \cdots & \Lambda^t\Phi(n) \end{pmatrix}$ are exchangeable. Furthermore by *de Finetti's Theorem* A.1, we can say that $\mathbf{U}(t) = \{U_1^{(t)}, U_2^{(t)}, ..., U_n^{(t)}\}$ are conditionally independent on a random probability measure η. □

Proof of Lemma 2.3. Based on Kallenberg and Exchangeable Graph (KEG) frameworks, introduced in Veitch and Roy (2015), a random array (A_{ij}) is jointly exchangeable *if and only if* it can be represented as follows : there is a random function $g : \mathbb{R}_+^2 \to [0, 1]$ such that

$$(A_{ij}) \overset{d}{=} (A_{v_i,v_j}) \overset{d}{=} Bern(g(\vartheta_i, \vartheta_j)) \tag{40}$$

where $v_i \overset{i.i.d.}{\sim} Poisson(1), \vartheta_i \overset{i.i.d.}{\sim} Poisson(1), v_i \leq \nu, i = 1, 2, ..., n$, for some pre-specified $\nu > 0$ so that finite size graphs can include vertices only if they participate in at least one edges. Thus if $\mathbf{A}$ is an adjacency matrix of such undirected, exchangeable network, for any $i < j$, $i, j = 1, ..., n$:

$$\begin{aligned} P\big(A_{ij} = a_{ij} \big| V_i, V_j\big) &= \int P\big(A_{ij} \big| v_i, v_j\big) Pr(\vartheta_i = \vartheta_i) Pr(\vartheta_j = \vartheta_j) dv_i dv_j d\vartheta_i d\vartheta_j \\ &= \int_0^\tau \int_0^\tau \int_0^\infty \int_0^\infty g(\vartheta_i, \vartheta_j)^{a_{ij}} \big(1 - g(\vartheta_i, \vartheta_j)\big)^{1-a_{ij}} \\ &\quad \times dPois_1(\vartheta_i) \times dPoi_1(\vartheta_j) d\vartheta_i d\vartheta_j. \end{aligned} \tag{41}$$

□

where $dPois_1(\cdot)$ is a probability distribution function of Poisson process with rate of 1. Thus given $\{\mathbf{V}\}$, edge probability except self-loop within each row (or column) is conditionally $i.i.d$ given a link function g and Poisson process V.

Proof of corollary 2.3.1. [Triangle inequality of diffusion distance] Let $x, y, z \in V(G)$.

$$\begin{aligned} D_t^2(x, z) &= \sum_{w \in V(G)} \big(P^t(x, w) - P^t(z, w)\big)^2 \frac{1}{\pi(w)} \\ &= \sum_{w \in V(G)} \big(P^t(x, w) - P^t(y, w) + P^t(y, w) - P^t(z, w)\big)^2 \frac{1}{\pi(w)} \\ &= \sum_{w \in V(G)} \big(P^t(x, w) - P^t(y, w)\big)^2 \frac{1}{\pi(w)} + \sum_{w \in V(G)} \big(P^t(y, w) - P^t(z, w)\big)^2 \frac{1}{\pi(w)} \\ &+ 2 \sum_{w \in V(G)} \big(P^t(x, w) - P^t(y, w)\big)\big(P^t(y, w) - P^t(z, w)\big) \frac{1}{\pi(w)} \\ &= D_t^2(x, y) + D_t^2(y, z) + 2 \sum_{w \in V(G)} \big(P^t(x, w) - P^t(y, w)\big)\big(P^t(y, w) - P^t(z, w)\big) \frac{1}{\pi(w)} \end{aligned} \tag{42}$$

Thus it suffices to show that

$$\sum_{w \in V(G)} \big(P^t(x, w) - P^t(y, w)\big)\big(P^t(y, w) - P^t(z, w)\big) \frac{1}{\pi(w)} \leq D_t(x, y) \cdot D_t(y, z). \tag{43}$$

Let $a_w = \big(P^t(x, w) - P^t(y, w)\big)\sqrt{1/\pi(w)}$ and $b_w = \big(P^t(y, w) - P^t(z, w)\big)\sqrt{1/\pi(w)}$. Then the above

inequality is equivalent to :

$$\sum_{w\in V(G)} a_w \cdot b_w \leq \sqrt{\sum_{w\in V(G)} a_w^2 \cdot \sum_{w\in V(G)} b_w^2}. \tag{44}$$

which is true by Cauchy-Schwarz inequality. □

Proof of Lemma 2.4 *Convergence of empirical characteristic function of exchangeable variables.* This follows exactly the same as *Theorem 1* in Székely et al. (2007). Note that this lemma always holds without any assumption on $\{(\mathbf{x}_j, \mathbf{y}_j), j = 1, 2, ..., n\}$, e.g., it holds without assuming exchangeability, nor identically distributed, nor finite moments.

□

Proof of Lemma 2.5 *Empirical characteristic function of exchangeable variables.* It suffices to prove the first argument 20 since the second argument 21 immediately follows from the first one by the property of characteristic functions. Proving the first one is equivalent to *Theorem 2* in Székely et al. (2007). However, they required $\{(\mathbf{x}_i, \mathbf{y}_i)\}$ to be independently identically distributed as $(\mathbf{x}, \mathbf{y})$ with finite second moments; here we have exchangeable $\{(\mathbf{x}_i, \mathbf{y}_i)\}$ instead. Followied by *de Finetti's Theorem* A.1, if and only if $\{x_i\}$ are exchangeble, there exist an underlying distribution $f_{\mathbf{x}}$ of $\mathbf{x}$ such that $\mathbf{x}_i \overset{i.i.d}{\sim} f_{\mathbf{x}}$. By the same logic, conditioning on the underlying distribution of $\mathbf{y}$, $\mathbf{y}_i \overset{i.i.d}{\sim} f_{\mathbf{y}}$. Thus under the assumption of finite second moment of the underlying, conditioned random variable $(\mathbf{x}, \mathbf{y})$, we have a strong large number for V-statistics followed by Székely et al. (2007), i.e.,

$$\int_{D(\delta)} \|g^n_{\mathbf{x},\mathbf{y}}(t,s) - g^n_{\mathbf{x}}(t)g^n_{\mathbf{y}}(s)\|^2 dw \overset{n\to\infty}{\longrightarrow} \int_{D(\delta)} \|g_{\mathbf{x},\mathbf{y}}(t,s) - g_{\mathbf{x}}(t)g_{\mathbf{y}}(s)\|^2 dw, \tag{45}$$

where $D(\delta) = \{(t,s) : \delta \leq |t|_p \leq 1/\delta, \delta \leq |s|_q \leq 1/\delta\}$, and $w(t,s)$ is the weight function chosen in Székely et al. (2007). □

Proof of Theorem 2.6 *Consistency of* `dCorr` *applied to exchangeable variables.* Under the exchangeability and finite moments assumptions of underlying distribution, it follows from Lemma 2.4 and 2.5 that $\mathcal{V}_n^2(\mathbf{X}, \mathbf{Y}) \xrightarrow{n\to\infty} 0$ if and only if underlying distribution of $\{\mathbf{x}_i\}$, $\mathbf{x}$ is independent from underlying distribution of $\{\mathbf{y}_i\}$, $\mathbf{y}$. Therefore, the `dCorr` or `mCorr` converges to 0 if and only if underlying distributions are independent; and its testing power converges to 1 under any joint distribution of finite moments. Since the multiscale generalized correlation based on any consistent

global correlation is also consistent (see in [Cencheng]), `MGC` statistic constructed by `dCorr` or `mCorr` is also consistent in testing dependence. □

Proof of Theorem 2.7 *Consistency of `MGC` applied to exchangeable variables.* Suppose that we have undirected, connected network $\mathbf{G}$ with a family of diffusion maps $\{\mathbf{u}_t\}$ and with nodal attributes $\{\mathbf{x}\}$. We have shown in the Lemma 2.2 that $\{\mathbf{u}_t\}$ are exchangeable for each $t\mathbb{N}$. Thus there exists an underlying distribution of $\mathbf{u}_t$ such that $\mathbf{u} \overset{i.i.d}{\sim} f_{\mathbf{u}^{(t)}}$, $t = 1, 2, \ldots$; and we have $\mathbf{x}_i \overset{i.i.d}{\sim} f_X$. Under the assumption of finite second moment of $\mathbf{u}^{(t)}$ and $\mathbf{x}$, `MGC` statistics constructed by $\{(\mathbf{u}_{ti}, x_i) : i = 1, 2, \ldots, n\}$ yield consistent testing which determines the independence between underlying distributions of $\mathbf{u}^{(t)}$ and $\mathbf{x}$, followed by Lemma 2.5. From the same setting of network $\mathbf{G}$, we have estimated *i.i.d* node-specific network factors $\{\mathbf{F}_i\}$, we have n-pair of *i.i.d* $\{(\mathbf{F}_i, \mathbf{x}_i)\}$ and they can be applied it to *MGC* without assuming conditioning underlying distribution. In case of using adjacency matrix directly into test, we must assume the adjacency matrix comes from directed network $\mathbf{G}$, i.e. $A_{ij} \overset{i.i.d}{\sim} f_A$ for all $i, j = 1, 2, \ldots, n$; otherwise, each column is dependent on one another. □

References

Banerjee, A., A. G. Chandrasekhar, E. Duflo, and M. O. Jackson (2013). The diffusion of microfinance. *Science 341*(6144), 1236498.

Barabasi, A.-L. and Z. N. Oltvai (2004). Network biology: understanding the cell's functional organization. *Nature reviews genetics 5*(2), 101–113.

Bartlett, T. E., S. C. Olhede, and A. Zaikin (2014). A dna methylation network interaction measure, and detection of network oncomarkers. *PloS one 9*(1), e84573.

Caron, F. and E. B. Fox (2014). Sparse graphs using exchangeable random measures. *arXiv preprint arXiv:1401.1137*.

Chan, S. H., T. B. Costa, and E. M. Airoldi (2013). Estimation of exchangeable graph models by stochastic blockmodel approximation. In *Global Conference on Signal and Information Processing (GlobalSIP), 2013 IEEE*, pp. 293–296. IEEE.

Christakis, N. A. and J. H. Fowler (2007). The spread of obesity in a large social network over 32 years. *New England journal of medicine 357*(4), 370–379.

Christakis, N. A. and J. H. Fowler (2008). The collective dynamics of smoking in a large social network. *New England journal of medicine 358*(21), 2249–2258.

Coifman, R. R. and S. Lafon (2006). Diffusion maps. *Applied and computational harmonic analysis 21*(1), 5–30.

Ellison, N. B., C. Steinfield, and C. Lampe (2007). The benefits of facebook friends: social capital and college students use of online social network sites. *Journal of Computer-Mediated Communication 12*(4), 1143–1168.

Fosdick, B. K. and P. D. Hoff (2015). Testing and modeling dependencies between a network and nodal attributes. *Journal of the American Statistical Association 110*(511), 1047–1056.

Gross, R. and A. Acquisti (2005). Information revelation and privacy in online social networks. In *Proceedings of the 2005 ACM workshop on Privacy in the electronic society*, pp. 71–80. ACM.

Heller, R., Y. Heller, and M. Gorfine (2012). A consistent multivariate test of association based on ranks of distances. *Biometrika*, ass070.

Hoff, P. D., A. E. Raftery, and M. S. Handcock (2002). Latent space approaches to social network analysis. *Journal of the american Statistical association 97*(460), 1090–1098.

Holland, P. W., K. B. Laskey, and S. Leinhardt (1983). Stochastic blockmodels: First steps. *Social networks 5*(2), 109–137.

Howard, M., E. Cox Pahnke, W. Boeker, et al. (2016). Understanding network formation in strategy research: Exponential random graph models. *Strategic Management Journal 37*(1), 22–44.

Kallenberg, O. (1990). Exchangeable random measures in the plane. *Journal of Theoretical Probability 3*(1), 81–136.

Karrer, B. and M. E. Newman (2011). Stochastic blockmodels and community structure in networks. *Physical Review E 83*(1), 016107.

Lafon, S. and A. B. Lee (2006). Diffusion maps and coarse-graining: A unified framework for dimensionality reduction, graph partitioning, and data set parameterization. *IEEE transactions on pattern analysis and machine intelligence 28*(9), 1393–1403.

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

Lyons, R. et al. (2013). Distance covariance in metric spaces. *The Annals of Probability 41*(5), 3284–3305.

Mantzaris, A. V., D. S. Bassett, N. F. Wymbs, E. Estrada, M. A. Porter, P. J. Mucha, S. T. Grafton, and D. J. Higham (2013). Dynamic network centrality summarizes learning in the human brain. *Journal of Complex Networks 1*(1), 83–92.

Orbanz, P. and D. M. Roy (2015). Bayesian models of graphs, arrays and other exchangeable random structures. *IEEE transactions on pattern analysis and machine intelligence 37*(2), 437–461.

Palla, K., D. Knowles, and Z. Ghahramani (2012). An infinite latent attribute model for network data. *arXiv preprint arXiv:1206.6416*.

Pinquart, M. and S. Sörensen (2000). Influences of socioeconomic status, social network, and competence on subjective well-being in later life: a meta-analysis. *Psychology and aging 15*(2), 187.

Pujol, A., R. Mosca, J. Farrés, and P. Aloy (2010). Unveiling the role of network and systems biology in drug discovery. *Trends in pharmacological sciences 31*(3), 115–123.

Raftery, A. E., X. Niu, P. D. Hoff, and K. Y. Yeung (2012). Fast inference for the latent space network model using a case-control approximate likelihood. *Journal of Computational and Graphical Statistics 21*(4), 901–919.

Sporns, O., C. J. Honey, and R. Kötter (2007). Identification and classification of hubs in brain networks. *PloS one 2*(10), e1049.

Székely, G. J. and M. L. Rizzo (2013a). The distance correlation t-test of independence in high dimension. *Journal of Multivariate Analysis 117*, 193–213.

Székely, G. J. and M. L. Rizzo (2013b). Energy statistics: A class of statistics based on distances. *Journal of statistical planning and inference 143*(8), 1249–1272.

Székely, G. J., M. L. Rizzo, N. K. Bakirov, et al. (2007). Measuring and testing dependence by correlation of distances. *The Annals of Statistics 35*(6), 2769–2794.

Tang, M. and M. Trosset (2010). Graph metrics and dimension reduction. *Indiana University, Indianapolis, IN*.

Veitch, V. and D. M. Roy (2015). The class of random graphs arising from exchangeable random measures. *arXiv preprint arXiv:1512.03099*.

Wasserman, S. and P. Pattison (1996). Logit models and logistic regressions for social networks: I. an introduction to markov graphs andp. *Psychometrika 61*(3), 401–425.

SUPPLEMENTARY MATERIAL

All of the `R` functions and simulation data in `RData` format are provided in https://github.com/neurodata/Multiscale-Network-Test.