Nonparametric Network Dependence Testing by Diffusion Maps

Abstract

Deciphering the association of network structures with their nodal attributes of interest is a core problem in network science. As the network topology is structured and often high-dimensional, many traditional nonparametric tests are no longer applicable and instead parametric approaches are dominant in network inferences. We propose a new procedure for testing the dependence between network topology and nodal attributes. To deal with the structured data of the network, we introduce a family of random vectors, called diffusion maps, which embed each node into the Euclidean space. The diffusion maps then provide network metrics on which we apply nonparametric distance-based correlation tests. We demonstrate that our testing method, local optimal distance-based correlation test combined with proper network metrics, not only yields consistent test statistics under common network models, but also significantly surpasses the testing power of existing benchmarks under various circumstances. In particular when the amount of dependency differs between the nodes, the statistic not only efficiently detects the dependency but also measures each node's contribution to testing dependence.

Keywords: testing independence, exchangeable graph, diffusion distance, distance correlation, multiscale generalized correlation

1 Introduction

Propelled by increasing demand and supply of graph data from various disciplines, the ubiquitous influence of network inferences has motivated numerous recent advances and applications in statistics, physics, computer science, biology, social science, etc., which further poses many new challenges to data scientists. One of the most fundamental statistical questions is to determine and characterize the relationship among multiple modalities of a given data set, for which the first step is to test the existence of any dependency between the data sets. However, the lack of a principal notion of correlation in the graph domain has not only hindered the progress of nonparametric dependency testing methods, but also deterred a rich literature of statistical techniques in other inferences (e.g., regression, feature screening, two-sample test) from being directly applied to graphs.

1.1 Data Structure of Graph with Nodal Attributes

A graph, or equivalently a network, can be formally defined as collection of nodes and edges. Mathematically, a graph $\mathbf{G} = (V, E)$ can be formally defined as collection of a set V of nodes (or vertices) and a set E of edges, which is often represented via an adjacency matrix $\mathbf{A} = \{A_{ij} : i, j =$

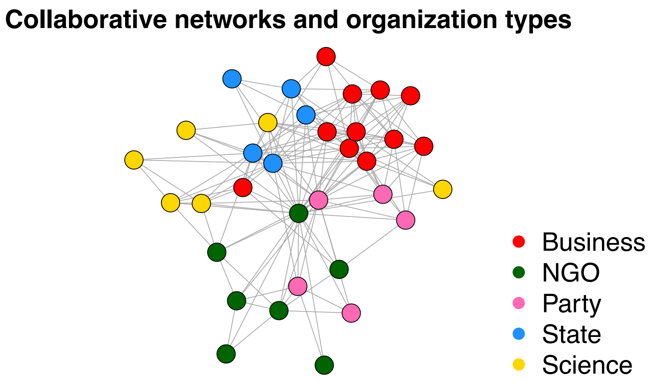

Figure 1: First figure needs to be replaced by brain network.. after the real data example in section 5 is affirmed.You may conjecture that organizations with the same type are more likely to collaborate each other at first glance; but there has been a lack of statistical method to test if there exists any significant relationship between network topology and node-specific attributes and if any, which node exerts the most dependency on network.

$1, .., n = |V|\}$, e.g. for an unweighted and undirected network $A_{ij} = 1$ if node i and node j are connected by an edge, and zero otherwise. Let $\mathbf{X} = \{\mathbf{x}_i \in \mathbb{R}^{q_x} : i = 1, \ldots, n\}$ be nodal attributes, a random variable or random vector associated with each node. Assume that we are given a $n \times n$ adjacency matrix $\mathbf{A}$ and nodal attributes $\mathbf{X}$ for each of n nodes. Since $\mathbf{A}$ is a symmetric square matrix when $\mathbf{G}$ is undirected, it does not satisfy traditional data assumptions, e.g., each observation can be assumed independently and identically distributed, the sample size increases faster than the feature dimension, etc.. These are the notable obstacles in directly applying conventional statistical methods. Therefore, graph inferences have long relied on specifying a particular statistical model for graphs, such as the Erdos-Renyi model [1, 2], stochastic block model [3–6] and its degree-corrected version [7, 8], the latent position model [9, 10], the random dot product model [11, 12], etc. However, model-based statistical methods often have limited applicability, e.g., connected-ness, unweighted-ness, and undirected-ness are the most common assumptions underlying statistical network models, which only represent a subset of real networks. Even under the model assumptions, how to select the model parameter can be expensive and unwarranted in practice, e.g., how to choose the dimension of latent factors for a given graph when a latent position model is assumed. Moreover, model mis-specification can largely affect the inference performance on networks. It is thus desirable to develop robust graph analysis approaches that are less dependent on models and parameters [13].

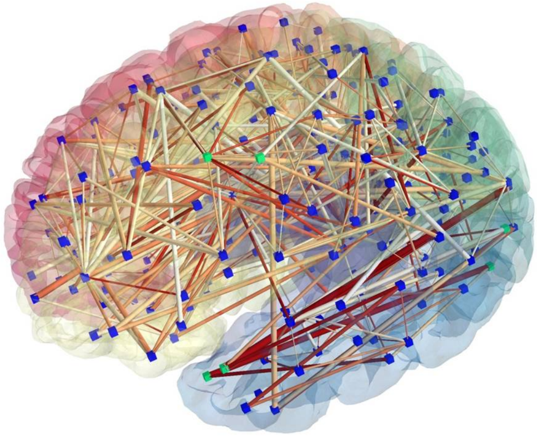

Figure 2: First figure needs to be replaced by brain network.. after the real data example in section 5 is affirmed.

1.2 Testing Network Dependence

When it comes to investigating the relationships among network data, one of the core problems is to detect dependency between network topology and nodal attributes, i.e., certain properties of interest defined on the nodes. For example, each person on Facebook not only has a number of distinct attributes (e.g., occupations, sex, personal behaviors), but also interacts with other persons via the social network; in neuro-science, each brain region has its own distinct functionality, and is also connected with other regions in the brain map. Figure 1, as an example, illustrates the collaborative networks between the political organizations which have different properties. Identifying dependency between network and nodal attributes, e.g. between collaborative relationships and the properties of each organization, however, has primarily focused on their relationship explained only by network model under the boundary of model assumption [10, 14, 15], thus suffers from the same problems all other model-based methods face. For example, the parametric network test proposed by Fosdick and Hoff [10] assumes a multivariate normal distribution of the latent factors as the generative model, estimates the latent factor of each node (which requires estimating its dimension q), then proceeds to test network dependence on the covariance by the standard likelihood ratio test. To our best knowledge so far, there is no principled method to compute a correlation measure on graphs which is consistent and model-free while overcoming all existing restraints on network analysis.

1.3 Testing Dependence in General Settings

On the other hand, the general problem of dependence testing between two random vectors has seen notable progress in recent years. The Pearson's correlation [16] is the most classical approach, which determines the existence of linear relationship via a correlation coefficient in the range of $[-1, 1]$, with 0 indicating no linear association while ±1 indicating perfect linear association. To capture all types of dependencies not limited to linear relationship, new correlation measures and nonparametric statistics have been suggested recently, such as the Mantel coefficient [17], RV coefficient [18], distance correlation (`dCorr`) and energy statistic [19–21], kernel-based independence test [22], Heller-Heller-Gorfine (`HHG`) test [23, 24], and multiscale generalized correlation (`MGC`) [25]. In particular, the distance correlation by Szekely et al. [19] is the first correlation measure that is consistent against all possible dependencies (with finite moments). The multiscale generalized correlation statistic (`MGC`) by Shen et al. [25] inherits the same consistency of distance correlation with remarkably better finite-sample testing powers under high-dimensional and nonlinear dependencies. The `MGC` defines a family of distance-based local correlations at every local scale and efficiently searches the optimal correlation in testing. Since all the above methods do not depend on particular models and also do not require explicit model parameter tuning, the network dependency testing may be significantly improved if some of them can be employed on graphs.

1.4 Outline of the Article

In the following Section 2, we introduce the background for network metrics and distance-based tests, which will be the ingredients for network dependence test. In Section 3 we demonstrate that our proposed test is theoretically sound under very mild condition, which includes almost all existing generative graph models while overcoming the theoretical barricades by the distinct structure of network data and relaxing the limitations of model-based method for network testing. Moreover, You can find in Section 4 that our proposed statistic offers major power improvement under various scenarios in finite-sample testing. The combined advantages of the network metrics and the testing method over the existing benchmarks are illustrated via comprehensive simulations under popular network models.

2 Method

2.1 Diffusion Maps and Diffusion Distances

In this section, we introduce the diffusion maps as a family of network geometries for a graph [26]. Coifman and Lafon [26, 27] proposed multiscale geometries of data called diffusion maps, which are constructed by iterating the transition matrix that determines the probability of moving forward from one node to the others during the random walk. The transition matrix here can be based on any reasonable kernels that represent the similarity between the node while satisfying the assumptions [26, 28]. We are going to define such transition matrix $\mathbf{P}$ via an adjacency matrix as a kernel function. For example, given an undirected non-empty graph $\mathbf{G}$, we have $P_{ij} = A_{ij} / \sum_{j=1}^{n} A_{ij}$ if $\sum_{j=1}^{n} A_{ij} > 0$ and $P_{ij} = 0$ otherwise $(i, j = 1, \ldots, n)$. When this kernel satisfies the properties of symmetry, positivity, and positive semi-definiteness, all of which the transition matrix $\mathbf{P}$ based on a symmetric adjacency matrix obeys, the diffusion map $\mathbf{u}(i)$ corresponding to node i at random walk iteration t is computed as follows :

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

where $\{\lambda_j\}$ and $\{\phi_j\}$ are the non-zero eigenvalues and corresponding eigenvectors of the transition matrix $\mathbf{P}$; q is the number of non-zero eigenvalues; λ_j^t is the t^{th} power of the eigenvalue; and $(\cdot)^T$ is the matrix transpose. Then diffusion maps locate each node's position at every diffusion time t and provide node-wise multivariate coordinates through $\{\mathbf{u}_t(i) : i = 1, \ldots, n; t \in \mathbb{N}\}$. It is not necessary to utilize whole set of non-zero eigenvalues in constructing diffusion maps but it is common to use a part of them having the largest absolute values considering the efficiency in dimensional reduction [26]. However since we are going to borrow the properties of transition matrix $\mathbf{P}$ to show that diffusion maps are independent sample under some conditions, deriving diffusion maps with a full set of $\{\lambda_j\}$ and $\{\phi_j\}$ is required. A family of diffusion maps as a function of the eigenvalues and eigenvectors of $\mathbf{P}$ can always be obtained when a symmetric kernel is given. When a given non-empty graph $\mathbf{G}$ is directed, i.e. when the probability for a random walk from node $x(\in V)$ to $y(\in V)$ differs from that from y to x, we are not able to represent diffusion maps via spectral properties of $\mathbf{P}$ based on an asymmetric kernel [29] (Appendix 7.2). In that case we might set a new symmetric weight between node i and node j, for example, $\tilde{w}_{ij}$, proportional to the average of both weights assigned to each direction, e.g. $\tilde{w}_{ij} := (w_{ij} + w_{ji})/2$. From now on we are going to restrict our arguments to an undirected and unweighted graph for simplicity.

The *diffusion distance* between node i and node j, $C_t(i,j)$, considering the propagation of information through Markov chains at diffusion time t, is formally defined as follows:

$$C_t^2(i,j) := \sum_{u \in V} \left(P_{iu}^t - P_{ju}^t \right)^2 / \pi(u), \tag{2}$$

where $\pi(u)$ is a stationary probability of node $u (\in V)$. It has shown that the above diffusion distance is exactly equivalent to the Euclidean distance of the diffusion maps.

$$C_t^2(i,j) = \parallel \mathbf{u}_t(i) - \mathbf{u}_t(j) \parallel \quad i,j = 1,2,\ldots,n. \tag{3}$$

As the diffusion time t increases, the corresponding diffusion distance C_t reveals the geometric structure of the network topology in a larger and larger scale, and is thus more likely to take into account of two nodes which are relatively difficult to reach each other when differentiating the distances of each pair. Figure 3 shows how well diffusion distance notices the community structure in a graph (generated by the stochastic block model by Equation 9) while differentiating distances across blocks, when a reasonable t is chosen in the family of diffusion distances $\{C_t : t \in \mathbb{N}\}$. Compared to adjacent relation or geodesic distance, where the set of distances are differentiated as little as possible or as much as possible respectively, diffusion distance better reflects the connectivity since it considers every possible path between the two nodes in its computation.

Although the parameter t may seem like another model parameter to tune, in practice $t \in [3,10]$ usually yields similar inference results. Therefore, throughout the paper we always take $t = 5$ in the simulations, and drop the subscript t in the diffusion maps $\mathbf{U}$ from now on.

2.2 Dependence Testing via MGC

The results in Section 2.1 allow us to cast the network dependency test into the following framework: given sample data $(\mathbf{U}, \mathbf{X}) = \{(\mathbf{u}_i, \mathbf{x}_i); i = 1,2,\ldots,n\}$ defined as Equation 1 that are

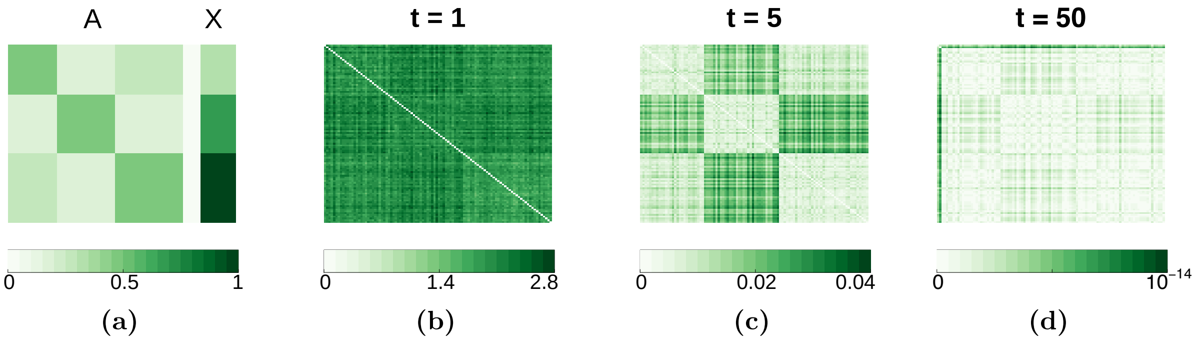

Figure 3: Panel (a) shows the adjacency matrix $\mathbf{A}$ and nodal attributes $\mathbf{X}$ generated by Equation 9. Panel (b), (c), and (d) shows the diffusion distances of the graph, as a proposed network metric to provide a one-parameter family of network-based distances. As t increases, there is a slight change in pattern, and the diffusion distance at $t = 5$ illustrates a very distinct block structures and thus has a very clear dependency to the attributes $\mathbf{X}$.

identically distributed as $(\mathbf{u}, \mathbf{x}) \in \mathbb{R}^{q \times q_x}$ (q and q_x are the respective feature dimension), we are looking to test whether their joint distribution equals the product of the marginals, i.e.,

$$H_0 : f_{\mathbf{ux}} = f_{\mathbf{u}} f_{\mathbf{x}},$$

$$H_A : f_{\mathbf{ux}} \neq f_{\mathbf{u}} f_{\mathbf{x}}.$$

If a pair of data $(\mathbf{u}_i, \mathbf{x}_i)$ can be further assumed independently distributed for each i, we can directly use a wide range of consistent test statistics, including the distance correlation, the `HHG` test, and `MGC`. Take the distance correlation for example: denote $C_{ij} =\| \mathbf{u}_i - \mathbf{u}_j \|$ and $D_{ij} =\| \mathbf{x}_i - \mathbf{x}_j \|$ for $i, j = 1, 2, \ldots, n$, where $\| \cdot \|$ is the Euclidean distance. The sample distance covariance is defined as

$$\texttt{dCov}(\mathbf{U}, \mathbf{X}) = \frac{1}{n^2} \sum_{i,j=1}^{n} \tilde{C}_{ij} \tilde{D}_{ij}, \tag{4}$$

where $\tilde{C}$ and $\tilde{D}$ is doubly-centered C and D by its column mean and row mean respectively, i.e., $\tilde{C} = HCH$, where $H = I_n - \frac{J_n}{n}$ (the double centering matrix), I_n is the $n \times n$ identity matrix (ones on the diagonal, zeros elsewhere), and J_n is the $n \times n$ matrix of all ones. The distance correlation (`dCorr`) follows by normalizing the distance covariance and is in the range of $[0, 1]$. The best property of distance correlation is its consistency against almost all alternatives, i.e., $\texttt{dCorr}(\mathbf{U}, \mathbf{X})$ has testing power 1 for sufficiently large n, for any joint distributions of finite second moment. In addition, a modified distance correlation (`mCorr`) was also proposed by Szekely and Rizzo [30] especially for testing high dimensional random vectors but unfortunately, the mCorr often fails to capture nonlinear associations especially embedded in high-dimensional data set [25, 31].

The MGC test inherits the consistency of distance correlation and significantly improves the finite-sample testing power via utilizing the correlation from a subset of data points. To be specific, we first compute all local covariances c_n^{kl} still based on the distance matrices of $\mathbf{U}$ and $\mathbf{X}$ but only including up to k-nearest points and up to l-nearest points for each data set.

$$c_n^{kl} = \frac{1}{n^2} \sum_{i,j=1}^{n} \tilde{C}_{ij} \tilde{D}_{ij} I\big(r(C_{ij}) \leq k\big) I\big(r(D_{ij}) \leq l\big), \quad k = 1, \ldots, K_{\mathbf{U}}; l = 1, \ldots, L_{\mathbf{L}}, \tag{5}$$

where $r(C_{ij})$ denotes a rank function of data $\mathbf{U}$ indicating the rank of $\mathbf{u}_i$ with respect to $\mathbf{u}_j$, and the same definition for $r(D_{ij})$ in data $\mathbf{X}$; $K_{\mathbf{U}}(\leq n)$ and $L_{\mathbf{X}}(\leq n)$ is the number of distinct values in data $\mathbf{U}$ and $\mathbf{X}$ respectively. Then the local correlations are the normalizations of the local covariance into $[-1, 1]$, and the MGC statistic is denoted by $c_n^{k^* l^*}$ via locating the optimal choice of neighborhood (k^*, l^*) among all possible neighborhood choices. Finding the optimal neighborhood, however, is not plagued by inflated false positive problem since MGC finds a significant region among the p-values of all scales thus it filters out accidentally low p-value. Moreover this p-value map of all scales, so-called multiscale map, provides a glimpse of nature of the dependence, i.e. existence or pattern of dependency between the two data sets.

cs: I rephrased the above MGC definition a little. Still need a few tweaks later. Better explain a bit on computation advantage, and what it means to be optimal here. last two sentences are added.

3 Theoretical Properties

However, as the *i.i.d.* assumption on $\mathbf{U}$ is not guaranteed by nature under network topology, the consistency of distance correlation no longer holds when applied to the arbitrary graphs. Moreover, neither the Euclidean distance of the adjacency vector nor the shortest-path distance can work together with distance correlation without breaking its consistency proof. We are going to introduce one of the statistical network models under which diffusion maps provide *i.i.d.* network observations.

A graph $\mathbf{G}$ is called exchangeable if and only if its adjacency matrix $\mathbf{A}$ is jointly exchangeable [32], i.e. for every permutation σ of n elements, $(A_{ij}) \overset{d}{=} (A_{\sigma(i)\sigma(j)})$. Exchangeability is a mild condition that most generative statistical network models satisfy, including all aforementioned models such as the stochastic block model and latent position model [4, 12, 33]. Lemma 3.1 proves that the diffusion map as node-wise multivariate coordinates $\mathbf{U}_t = \{(\mathbf{u}_t(1), \mathbf{u}_t(2), \cdots, \mathbf{u}_t(n)) : t \in \mathbb{N}\}$ can furnish conditional *i.i.d.* samples for nodes in an exchangeable graph, with the proof supplied in the Appendix.

Lemma 3.1 (Conditional *i.i.d.* of diffusion map $\mathbf{U}_t$)**.** Assume that $\mathbf{G}$ is an exchangeable random graph. Then as $n \to \infty$, the diffusion map $\{\mathbf{u}_t(i) : i = 1, \dots, n\}$ are conditionally *i.i.d.* given its underlying distribution.

Now assume that $\mathbf{G}$ is an exchangeable random graph and its diffusion maps are $\mathbf{U}$ with finite moment; and the nodal attributes $\mathbf{X} = \{\mathbf{x}_i : i = 1, 2, \dots, n\}$ are *i.i.d.* as a random vector $\mathbf{x}$ of finite moment. Lemma 3.1 shows that diffusion map at any fixed time t, $\mathbf{U} = \{\mathbf{u}_i : i = 1, \dots, n\}$ are conditional *i.i.d.* for an exchangeable graph, i.e., there exists an underlying distribution random variable $\mathbf{u}$ such that $\mathbf{u}_i | \mathbf{u}$ are *i.i.d.* as $n \to \infty$. The following two Lemmas serve as the foundation for using exchangeable observations in distance-based independence testing.

Lemma 3.2. Let $\mathcal{V}_n^2(\mathbf{U}, \mathbf{X})$ be the distance covariance (`dCov`) of $(\mathbf{U}, \mathbf{X}) = \{(\mathbf{u}_i, \mathbf{x}_i) : i = 1, \dots, n\}$ defined as Equation 4. Then we have

$$\mathcal{V}_n^2(\mathbf{U}, \mathbf{X}) \quad = \|g^n_{\mathbf{u},\mathbf{x}}(t,s) - g^n_{\mathbf{u}}(t) g^n_{\mathbf{x}}(s)\|^2, \tag{6}$$

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

Lemma 3.3. Then under the conditions above on $(\mathbf{U}, \mathbf{X})$, we have

$$\mathcal{V}_n^2(\mathbf{U}, \mathbf{X}) \quad \longrightarrow \mathcal{V}^2(\mathbf{u}, \mathbf{x}) \qquad \text{as } n \to \infty \tag{7}$$

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

$$\mathcal{V}_n^2(\mathbf{U}, \mathbf{X}) \quad \to 0 \qquad \text{as } n \to \infty \tag{8}$$

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

Followed by Lemma 3.1, Lemma 3.2 and Lemma 3.3, our next result shows that both the `dCorr` and `MGC` defined on the diffusion distance can have the same consistency when extended to network dependency test of exchangeable graphs.

Theorem 3.4 (`MGC` Consistency via Diffusion Distance)**.** Under the conditions above,

$$\texttt{dCorr}(\mathbf{U},\mathbf{X}) \longrightarrow 0 \text{ as } n \to \infty$$

if and only if $\mathbf{U}$ is independent of $\mathbf{X}$. And both `MGC` and `dCorr` are consistent for testing independence between any $\mathbf{U}$ and $\mathbf{X}$ satisfying the above condition.

Therefore, our approach not only yields an easy-to-use methodology in network dependence testing, but also enjoys solid theoretical property and thus offers a principal approach to study correlation on network data. You can find the proof of this theorem in Appendix 7.1.

cs: I think, maybe we should take certain contents out of Section 2, and make a separate review section on diffusion maps and MGC. Then in the results section, we can be more concentrated and clearer about our contribution and theorems. Also add another subsection describing the full network testing procedure. What do you think? I am very certain what it does mean but I agree that we now have heavy section 2. Let us discuss this next week.

cs: For the directed case, if it does not work we simply exclude; otherwise we can either add it here or in the appendix (so as not to further complicate the main content) As long as I found, spectral decomposition diffusion maps of asymmetric weight matrix is not tractable; so it is not unusual to transform asymmetric kernel to the symmetric. Why don't we just mention one sentence suggesting such alternatives?

4 Simulation Study

Next we investigate our approach via simulated models and empirical performances. In the simulation studies, we compare the empirical testing powers of four test statistics: `MGC`, `mCorr`, `HHG`, and the likelihood ratio test proposed by Fosdick and Hoff (`FH`) [10]. For the first three statistics, we further consider three different metrics of the network topology: the Euclidean distances of the diffusion maps (DM), of each column of adjacency matrix (AM), and of the latent factors (LF), which is based on singular value decomposition of the adjacency matrix. The FH likelihood ratio test must always be based on the latent factors.

Note that the latent-factor-based FH test requires a selection of a dimension parameter q, which we vary $q \in [1, 10]$ and take the optimal power within the range (e.g., as a benchmark, the FH test actually has its power maximized over the parameter range). While for the diffusion maps, it suffices to fix $t = 5$ as discussed in Section 2.1.

For each simulation model and each test, we repeatedly generate sample graph and attributes for 500 times, carry out the permutation test, and reject the null if the resulting p-value is less than $\alpha = 0.05$. The testing power of each method equals the percentage of correct rejection.

4.1 Stochastic Block Model

Let us first consider SBM with 3 blocks, i.e., partition the nodes into 3 communities, and generate the edges by a Bernoulli random variable whose probability is determined by the communities of the connecting nodes. Assume $n = 100$ nodes whose attribute values $\mathbf{x}_i$ takes values of $0, 1,$ or 2 in ordinal scale equally likely. The edge probability is designed as

$$E(A_{ij}|X_i, X_j) = 0.5I(|X_i - X_j| = 0) + 0.2I(|X_i - X_j| = 1) + 0.3I(|X_i - X_j| = 2), \quad i, j = 1, \ldots, n = 100. \tag{9}$$

Namely, within-block edge probability is 0.5, between-block edge probability is 0.2 or 0.3 depending on the communities defined by (dis)similarity in nodal attributes value of X. This 3-block model describes a nonlinear dependency, where MGC has been shown to work better than the dCorr or HHG given a pair of random vectors [25]. We now want to look at the performance of MGC given a graph object and a random vector of nodal attributes. A visualization of the statistics from one sample graph is offered in Figure 3. After repeatedly generating the data set and implementing independence testing by all the methods mentioned, the powers are computed and shown in Figure 4, for which MGC combined with diffusion maps indeed yields the most superior power comparing to all other benchmarks.

To further understand the advantage of MGC, we fix the diffusion distance as the network metric, and compare different test statistics. Based on the same three-block model, the edge probability is

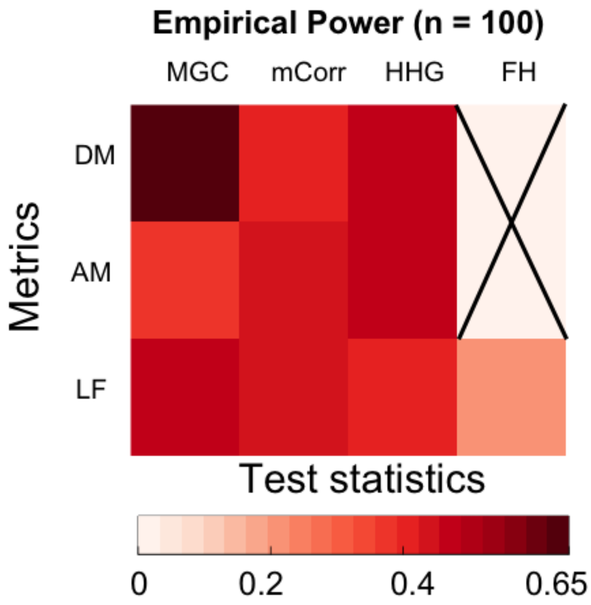

Figure 4: The power heatmap under the SBM with three blocks (Equation 9) demonstrates that for all possible combinations of test statistics with distance metrics the MGC with the diffusion maps yields the best power comparing to all other methods.

now generated as follows,by controlling the amount of *nonlinear dependency* through $\theta \in (0,1)$:

$$E(A_{ij}|X_i,X_j) = 0.5I(|X_i - X_j| = 0) + 0.2I(|X_i - X_j| = 1) + \theta I(|X_i - X_j| = 2), \quad i,j = 1,\ldots,n = 100. \tag{10}$$

When $\theta > 0.2$, the network dependency changes from a close to linear relationship to strongly nonlinear. Figure 5 shows the testing power with respect to increasing θ, and there is a clear trend that both the mCorr and FH tests have deteriorating power while MGC has a very stable performance against varying θ. The same phenomenon holds by varying other edge probabilities. These observations support the argument that the MGC can better capture the nonlinear dependencies for network dependence testing, and is the best method to couple with the diffusion distance.

4.2 Degree-corrected Stochastic Block Model

Our next simulation shifts to the degree-corrected stochastic block model (DC-SBM) with two blocks. The DC-SBM adds another random variable V_i associated with each node to vary the node degrees, which is a generalization of the stochastic block model and provides a better fit to real networks. Setting $n = 250$, suppose that the nodal attributes X_i takes binary values in 0 and 1

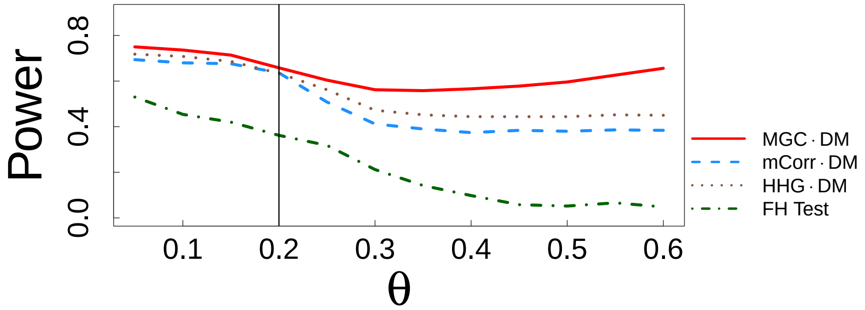

Figure 5: The power curve with respect to increasing θ in the SBM with three blocks, for `MGC`, `mCorr`, and `FH`. Larger θ implies stronger nonlinear dependency, while $\theta < 0.2$ has close-to-linear dependency. The `MGC` is the best performing method throughout all possible θ.

equally likely, and the edge probabilities are specified by

$$E(A_{ij}|\mathbf{X},\mathbf{V}) = 0.2V_iV_j \cdot I(|X_i - X_j| = 0) + 0.05V_iV_j \cdot I(|X_i - X_j| = 1), \qquad (11)$$

where $V_i \overset{i.i.d}{\sim} Uniform(1-\tau, 1+\tau)$ for $i = 1, \ldots, n$, and τ $(\in [0,1])$ is a parameter to control the amount of variability in the edge distribution. Again, the `MGC` coupled with diffusion maps, i.e. `MGC` $\circ$ `DM`, is the best method in power throughout τ; and in Figure 6 (a) we show the testing power restricted to `MGC` but varying the distance metrics, which shows the diffusion distance is indeed the best distance metric for network dependence testing.

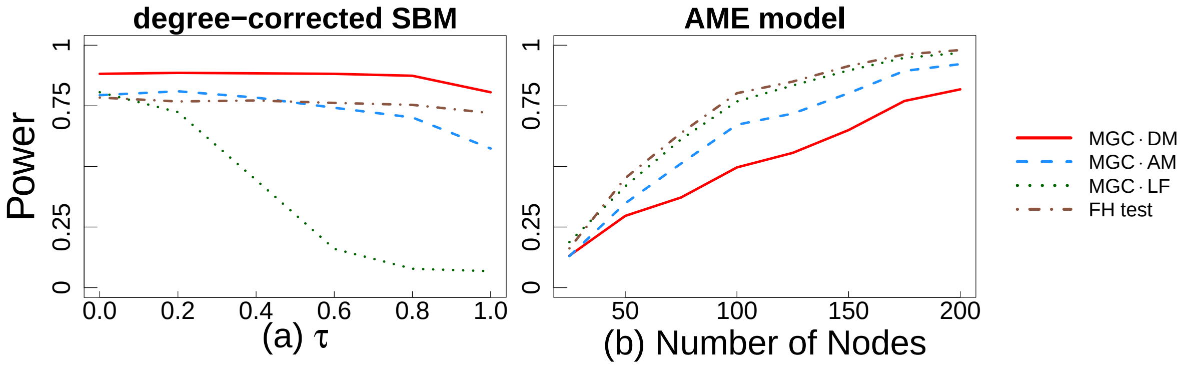

Figure 6: (a) In DC-SBM where the variability in degree distribution increases as τ increases, testing power of diffusion maps are more likely to be robust against increasing variability compared to other network metrics, e.g. adjacency matrix or latent positions. The `FH` test statistics allowing different dimensions of network factors perform consistently well but still have less power than the `MGC`. (b) The `MGC` utilizing diffusion distances loses some power under additive and multiplicative model which favors estimated latent position metrics, but `MGC` does as good as `FH` tests under latent factor metrics which closes to the truth. This reveals the flexibility in selecting distance matrix used in `MGC` statistics, which can be chosen depending on model fit or preliminary knowledge.

4.3 Additive and Multiplicative Graph Model

Fosdick and Hoff [10] proposed an approach of modeling network as an additive and multiplicative effect (AME) of node-specific latent factors. Whereas AME model embeds the nodes into the latent factors assuming that their network model is *correct*, a family of diffusion maps configure each node as a multivariate variable without losing any information on the adjacent matrix or weight matrix. Thus in the following model 12, where logit of A obeys the presumed, additive and multiplicative model of latent factors of Z, the estimated latent factors would be very close to the truth.

$$\begin{aligned} & Z_i \overset{i.i.d}{\sim} f_Z(z) \overset{d}{=} Uniform[0,1]. \quad i = 1, \ldots, n \\ & X_i|Z_i \overset{i.i.d}{\sim} f_{X|Z}(x_i|z_i) \overset{d}{=} Normal(z_i, 1), \quad i = 1, \ldots, n \\ & A_{ij}|Z_i, Z_j \overset{i.i.d}{\sim} f_{A|Z}(a_{ij}|z_i, z_j) \overset{d}{=} Bern\big((1-z_i)^2 \times (1-z_j)^2\big), \quad i, j = 1, \ldots, n; i < j. \end{aligned} \tag{12}$$

Even though we rarely see the network nearly follows the model in reality, if so, using the estimated network factors as independent observations from graph $\mathbf{G}$ and applying them to `MGC` performs not very worse than `FH` statistic (Figure 6). In other words, if the network really fits well to the network model with node-specific latent factors as covariates, then it would be safe to use those factors in the `MGC` statistic directly. Since they assume $i.i.d$ generative model for the factors, it is still valid to apply `MGC` using these $i.i.d$ observations of estimated factors.

5 Real Data Example

- We have a connected and undirected brain network with 95 nodes and 337 edges with three dimensional nodal attributes which specify $x - y - z$ physical locations of each node.
- All the benchmarks, including `FH` test result in very significant network dependence to physical location.
- Among x, y and z, it looks like x contributes to the dependency a lot.
- I deliberately divided the nodes into three groups by x values.
- As contamination rate $(10\%, 20\%, \cdots, 100\%)$ increases, a large portion of nodes in each group

are randomly selected and edges between them are also randomly generated by pre-specified probability scheme.

- This random mechanism depends on group and imitates nonlinear dependence.
- In the figures, empirical power changes of different testing methods (MGC, mCorr, HHG, FH) with different metrics (DM, AM, LF) will be presented according to different level of contamination.

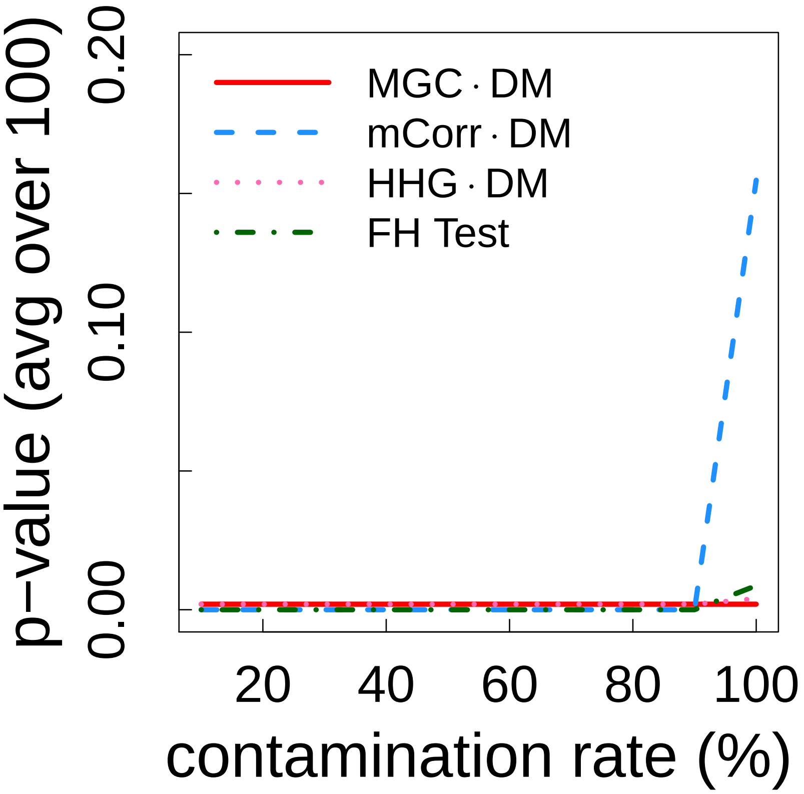

Figure 7: Example plot

cs: can we expand the real data section? It looks too short for now, although it is probably ok for stat journal. I am considering manipulation MRI data in order to show the case where MGC ¿ others...

6 Discussion

In this study, we combined recent progress in dependency testing and metric learning into the graph domain. We showed that MGC on the diffusion distance offers an elegant and powerful solution to the network dependency problem, which overcomes many challenges and restraints in the domain of network analysis. We proved that our method is consistent under a mild condition inclusive of almost all popular graph models; and empirically demonstrated that it has superior power over all benchmarks, with MGC and the diffusion distance being the core elements behind the success. Moreover our proposed method not only detects the dependence between network topology and nodal attributes but also helps us to reveal possibly diverse dependence patterns through multiscale correlation maps or multiscale p-values as a function of diffusion time.

However, obtaining a full family of statistics according to all diffusion times are computationally infeasible. As an ad hoc, we selected an *optimal* diffusion time t as $t = 5$ based on empirical results in simulation. How to choose a better diffusion time t, or find a t with provable finite-sample performance, may establish more solid foundation of this approach.

On the other hand, the network dependence testing here is actually equivalent to the two-sample test, i.e., whether two graphs come from the same distribution; thus our approach readily offers a new nonparametric two-sample test on networks, for which more investigation will bring a valuable addition to the graph analysis. Furthermore, with a few alterations, the new correlation measure on graph may be utilized for other tasks, such as feature screening, outlier detections, clustering, and classification, etc. Fourth, as a next step of this paper, we will utilize this method to a wide range of graphs available in social network and brain analysis, to answer domain specific practical questions.

References

[1] P Erdos and A Renyi. On random graphs. i. *Publicationes Mathematicae*, 6:290–297, 1959.

[2] E.N. Gilbert. Random graphs. *Annals of Mathematical Statistics*, 30(4):1141–1144, 1959.

[3] P. Holland, K. Laskey, and S. Leinhardt. Stochastic blockmodels: First steps. *Social Networks*, 5(2):109–137, 1983.

[4] Karl Rohe, Sourav Chatterjee, and Bin Yu. Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics*, pages 1878–1915, 2011.

[5] D. Sussman, M. Tang, D. Fishkind, and C. Priebe. A consistent adjacency spectral embedding for stochastic blockmodel graphs. *Journal of the American Statistical Association*, 107(499):1119–1128, 2012.

[6] Jing Lei and Alessandro Rinaldo. Consistency of spectral clustering in stochastic block models. *The Annals of Statistics*, 43(1):215–237, 2015.

[7] Brian Karrer and Mark EJ Newman. Stochastic blockmodels and community structure in networks. *Physical Review E*, 83(1):016107, 2011.

[8] Y. Zhao, E. Levina, and J. Zhu. Consistency of community detection in networks under degree-corrected stochastic block models. *Annals of Statistics*, 40(4):2266–2292, 2012.

[9] M. Tang, D. L. Sussman, and C. E. Priebe. Universally consistent vertex classification for latent positions graphs. *Annals of Statistics*, 41(3):1406–1430, 2013.

[10] Bailey K Fosdick and Peter D Hoff. Testing and modeling dependencies between a network and nodal attributes. *Journal of the American Statistical Association*, 110(511):1047–1056, 2015.

[11] S. Young and E. Scheinerman. Random dot product graph models for social networks. In *Algorithms and Models for the Web-Graph*, pages 138–149. Springer Berlin Heidelberg, 2007.

[12] Daniel L Sussman, Minh Tang, and Carey E Priebe. Consistent latent position estimation and vertex classification for random dot product graphs. *IEEE transactions on pattern analysis and machine intelligence*, 36(1):48–57, 2014.

[13] L. Chen, C. Shen, J. T. Vogelstein, and C. E. Priebe. Robust vertex classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(3):578–590, 2016.

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

[15] Michael Howard, Emily Cox Pahnke, Warren Boeker, et al. Understanding network formation in strategy research: Exponential random graph models. *Strategic Management Journal*, 37(1):22–44, 2016.

[16] K. Pearson. Notes on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58:240–242, 1895.

[17] N. Mantel. The detection of disease clustering and a generalized regression approach. *Cancer Research*, 27(2):209–220, 1967.

[18] P. Robert and Y. Escoufier. A unifying tool for linear multivariate statistical methods: The rv - coefficient. *Journal of the Royal Statistical Society. Series C*, 25(3):257–265, 1976.

[19] Gábor J Székely, Maria L Rizzo, Nail K Bakirov, et al. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6):2769–2794, 2007.

[20] G. Szekely and M. Rizzo. The distance correlation t-test of independence in high dimension. *Journal of Multivariate Analysis*, 117:193–213, 2013.

[21] M. Rizzo and G. Szekely. Energy distance. *Wiley Interdisciplinary Reviews: Computational Statistics*, 8(1):27–38, 2016.

[22] A. Gretton and L. Gyorfi. Consistent nonparametric tests of independence. *Journal of Machine Learning Research*, 11:1391–1423, 2010.

[23] R. Heller, Y. Heller, and M. Gorfine. A consistent multivariate test of association based on ranks of distances. *Biometrika*, 100(2):503–510, 2013.

[24] Ruth Heller, Yair Heller, Shachar Kaufman, Barak Brill, and Malka Gorfine. Consistent distribution-free k-sample and independence tests for univariate random variables. *Journal of Machine Learning Research*, 17(29):1–54, 2016.

[25] Cencheng Shen, Carey E Priebe, Mauro Maggioni, and Joshua T Vogelstein. Discovering relationships across disparate data modalities. *arXiv preprint arXiv:1609.05148*, 2016.

[26] Ronald R Coifman and Stéphane Lafon. Diffusion maps. *Applied and computational harmonic analysis*, 21(1):5–30, 2006.

[27] Stephane Lafon and Ann B Lee. 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, 2006.

[28] Ronald R Coifman, Stephane Lafon, Ann B Lee, Mauro Maggioni, Boaz Nadler, Frederick Warner, and Steven W Zucker. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *Proceedings of the National Academy of Sciences of the United States of America*, 102(21):7426–7431, 2005.

[29] Minh Tang and Michael Trosset. Graph metrics and dimension reduction. *Indiana University, Indianapolis, IN*, 2010.

[30] Gábor J Székely and Maria L Rizzo. The distance correlation t-test of independence in high dimension. *Journal of Multivariate Analysis*, 117:193–213, 2013.

[31] Ruth Heller, Yair Heller, and Malka Gorfine. A consistent multivariate test of association based on ranks of distances. *Biometrika*, page ass070, 2012.

[32] Peter Orbanz and Daniel M Roy. Bayesian models of graphs, arrays and other exchangeable random structures. *IEEE transactions on pattern analysis and machine intelligence*, 37(2):437–461, 2015.

[33] Adrien Todeschini and François Caron. Exchangeable random measures for sparse and modular graphs with overlapping communities. *arXiv preprint arXiv:1602.02114*, 2016.

[34] Persi Diaconis and David Freedman. Finite exchangeable sequences. *The Annals of Probability*, pages 745–764, 1980.

7 Appendix

7.1 Proofs

Proof of Lemma 3.1. To prove conditional *i.i.d.* of the diffusion map $\mathbf{U} = \{\mathbf{u}(i) : i = 1, \dots, n\}$ as $n \to \infty$ given its underlying distribution, by the celebrated *de Finetti's Theorem* [34], it suffices to prove that $\mathbf{u}(i)$ for $i = 1, \dots, n$ are exchangeable, i.e., for any permutation σ, the permuted sequence $\mathbf{u}(\sigma(1)), \mathbf{u}(\sigma(2)), \dots, \mathbf{u}(\sigma(n))$ distributes the same as the original sequence $\mathbf{u}(1), \mathbf{u}(2), \dots, \mathbf{u}(n)$. Let$\mathbf{V}$ be a $q \times n$ matrix having $\mathbf{u}(i)$ as a i^{th} column $(i = 1, \dots, n)$. Denoting the permutation matrix as M, it is to show that $\mathbf{V}$ always distributes the same as $\mathbf{V}M^T$ in matrix notation.

Recall that the diffusion map at time t, $\mathbf{u}(i)$ is represented as follows :

$$\mathbf{u}(i) = \begin{pmatrix} \lambda_1^t \phi_1(i) & \lambda_2^t \phi_2(i) & \cdots & \lambda_q^t \phi_q(i) \end{pmatrix} \in \mathbb{R}^q; \quad i = 1, \dots, n. \tag{13}$$

Thus when $\Lambda = diag\{\lambda_1, \lambda_2, \dots, \lambda_q\}$ denotes the diagonal matrix of non-zero eigenvalues and $\Phi = [\phi_1, \phi_2, \cdots, \phi_q]$ is the matrix having column vectors as the corresponding eigenvectors of the transition matrix $\mathbf{P}$, $\mathbf{P} = \Phi \Lambda \Phi^T$. Then given the graph $\mathbf{G}$ is exchangeable, i.e., $A_{\sigma(i)\sigma(j)} \overset{d}{=} A_{ij}$, we have

$$\begin{aligned} \mathbf{P}_{\sigma(i)\sigma(j)} &= A_{\sigma(i)\sigma(j)} / \sum_j A_{\sigma(i)\sigma(j)} \\ &\overset{d}{=} A_{ij} / \sum_j A_{ij} \\ &= \mathbf{P}_{ij}, \end{aligned}$$

from which it follows that

$$\begin{gathered} M\mathbf{P}M^T \overset{d}{=} \mathbf{P} \\ \Rightarrow (M\Phi)\Lambda(M\Phi)^T \overset{d}{=} \Phi\Lambda\Phi^T \\ \Rightarrow \mathbf{V}M^T = \Lambda(M\Phi)^T \overset{d}{=} \Lambda\Phi^T = \mathbf{V} \end{gathered}$$

Therefore, the diffusion maps are exchangeable, and also conditional *i.i.d.* asymptotically by *Finet-*

tis Theorem [32, 34]. Note that with a part of eigenvalues, e.g. up to s largest in absolute values [26], we are not able to have exchangeability of diffusion map. □

Proof of Lemma 3.2 *Convergence of empirical characteristic functions of exchangeable variables.* This follows exactly the same as *Theorem 1* in [19]. Note that this Lemma always holds without any assumption on $(\mathbf{U}, \mathbf{X}) = \{(\mathbf{u}_i, \mathbf{x}_j) : i = 1, 2, ..., n\}$, e.g., it holds without assuming exchangeability, nor identically distributed, nor finite second moments. □

Proof of Lemma 3.3 *Empirical characteristic function of exchangeable variables.* It suffices to prove the first argument 7 since the second argument 8 immediately follows from the first one by the property of characteristic functions. Proving the first one is equivalent to *Theorem 2* in [19]. However, they required a given pair of data $\{(\mathbf{u}_i, \mathbf{x}_i) : i = 1, \ldots, n\}$ to be independently identically distributed as $(\mathbf{u}, \mathbf{x})$ with finite second moments; here we have exchangeable data $\{\mathbf{u}_i : i = 1, \ldots, n\}$ instead.

Followed by *de Finetti's Theorem* [34], if and only if $\{\mathbf{u}_i : i = 1, \ldots, n\}$ are (infinitely) exchangeable, there exists an underlying distribution $f_{\mathbf{u}}$ of $\mathbf{u}$ such that $\mathbf{u}_i \overset{i.i.d}{\sim} f_{\mathbf{u}}$. Then we can also have joint distribution of the data set. Let $(\mathbf{u}_i, \mathbf{x}_i) \overset{i.i.d}{\sim} f_{\mathbf{u},\mathbf{x}}$. Then under the assumption of finite second moment of the underlying distributions and measurable, conditioned random functions, we have a strong large number for V-statistics followed by [19], i.e.,

$$\int_{D(\delta)} \|g^n_{\mathbf{u},\mathbf{x}}(t,s) - g^n_{\mathbf{u}}(t) g^n_{\mathbf{x}}(s)\|^2 dh \overset{n\to\infty}{\longrightarrow} \int_{D(\delta)} \|g_{\mathbf{u},\mathbf{x}}(t,s) - g_{\mathbf{u}}(t) g_{\mathbf{x}}(s)\|^2 dh, \tag{14}$$

where $D(\delta) = \{(t, s) : \delta \leq |t|_p \leq 1/\delta, \delta \leq |s|_q \leq 1/\delta\}$, and $h(t, s)$ is the weight function chosen in [19]. □

Proof of Theorem 3.4 `MGC` *Consistency via Diffusion Distance.* Combining equation 7 and equation 14 yields

$$\texttt{dCov}(\mathbf{U}, \mathbf{X}) \longrightarrow \int \|g_{\mathbf{u},\mathbf{x}}(t,s) - g_{\mathbf{u}}(t) g_{\mathbf{x}}(s)\|^2 dw, \tag{15}$$

which clearly equals 0 if and only if independence holds. As distance correlation is just a normalized version of distance covariance, we also have

$$\texttt{dCorr}(\mathbf{U}, \mathbf{X}) \quad \longrightarrow 0, \tag{16}$$

if and only if the diffusion maps $\mathbf{U}$ is independent of the nodal attributes $\mathbf{X}$.

By [25], Equation 16 holds under the same condition, when `dCorr` is replaced by `MGC`. Therefore, both `MGC` and `dCorr` are consistent in network dependence testing between the diffusion maps $\mathbf{U}$ and the nodal attributes $\mathbf{X}$. □

7.2 Diffusion Distance in Directed Graphs

When a graph is directed, a kernel of an adjacency matrix is not symmetric. Thus, keeping the formal definition of diffusion distance as Equation 2, we have a slightly different representation from Equation 3 mainly due to $\pi(i)P_{ij} \neq \pi(j)P_{ji}$. Let $\tilde{\mathbf{P}}$ be a time-reversal or transpose matrix of $\mathbf{P}$. Then the squared diffusion distance between node i and node j can be derived through:

$$\begin{aligned} C_t^2(i,j) = &\big(\mathbf{P}^t\tilde{\mathbf{P}}^t\Pi^{-1}\big)(i,i) - \big(\mathbf{P}^t\tilde{\mathbf{P}}^t\Pi^{-1}\big)(j,i) \\ &- \big(\mathbf{P}^t\tilde{\mathbf{P}}^t\Pi^{-1}\big)(j,j) + \big(\mathbf{P}^t\tilde{\mathbf{P}}^t\Pi^{-1}\big)(i,j), \end{aligned} \tag{17}$$

where a Π is a diagonal matrix with diagonal terms of $\{\pi(u) : u \in V\}$. Different from the former undirected case, $\Pi^{1/2}\mathbf{P}\Pi^{-1/2}$ does not yield a symmetric matrix which makes concise representation of $\mathbf{P}$ possible. Tang [29] claims that embedding of C_t using the eigenvalues and eigenvectors of $\mathbf{P}$ is not possible, even though it is shown that the diffusion distance in directed graph is still Type-2 Euclidean distance matrix (EDM-2) when $\mathbf{P}$ is irreducible.

7.3 Algorithms

Algorithm 1 Mutiscale representation of nodes in network

```
Require: Transition probability matrix $P$ of network $G$ and a set of time points $\{t_i : t_i \in \mathbb{N}\}$ of diffusion time.
Ensure: 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 $\pi$
    $Q := \Pi^{1/2} P \Pi^{-1/2}$
    $\lambda := \texttt{eigenvalue}(Q)$                     ▷ 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
    $\texttt{Maps}$ = list( $\texttt{Maps}[1]$, $\texttt{Maps}[2]$, ..., $\texttt{Maps}[K]$ )
     return Maps
end function
```

Algorithm 2 Multiscale Generalized Correlation (MGC) test statistics with diffusion maps as a network-based distance.

```
Require: A connected, undirected network $G$ with its nodal attributes $\mathbf{X}$.
Ensure: 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 $\{t_1, t_2, \ldots, t_K\}$) )
    $A$ := get.adjacency($G$)                          ▷ obtain an adjacency matrix of network $G$
    $P$ := $A$ / rowSums($A$)
    $U$ := dmap($P$, $\mathbf{T}$)                     ▷ a list of diffusion maps in each time point
    for $t_i$ : $i = 1$ do $K$
        $C$ := dist($U[i]$)                            ▷ distance matrix of diffusion maps at time $t_i$
        $D$ := dist($X$)                               ▷ distance matrix of nodal attributes
        MGC[$i$] = 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

```
Require: Distance metric of graph $G$, $C$, and attributes $X$, $D$, and (one of) the estimated optimal
    scales $\{k^*, l^*\}$
Ensure: unstandardized contributions of each node in network $\{c(v)\}$
function Contribution ( C, D , $\{(k^*, l^*)\}$ )
    $\tilde{C}$ := DoubleCentering($C$)
    $\tilde{D}$ := DoubleCentering($D$)
    Rank($M_{ij}$):= (rank of node $j$ with respect to node $i$)
    for $v = 1$ do $|V(G)|$                            ▷ iterate over every each node
        $c(v) = 0$
        for $j = 1$ do $n$
            $c(v) = c(v) + \tilde{C}_{vj}\tilde{D}_{vj} I(\mathtt{Rank}(C_{vj}) \leq k^*, \mathtt{Rank}(D_{vj}) \leq l^*)$
        end for
    end for
    cset := $\{c(v) : v = 1, 2, \ldots, |V(G)|\}$
     return cset
end function
```

SUPPLEMENTARY MATERIAL

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