Network Dependence Testing via Diffusion Maps and Distance-Based Correlations

Youjin Lee[1], Cencheng Shen[2], and Joshua T. Vogelstein[2,3,4]

[1]Department of Biostatistics, Johns Hopkins University
[2]Center for Imaging Science, Johns Hopkins University
[3]Department of Biomedical Engineering and Institute for Computational Medicine, Johns Hopkins University
[4]Institute for Data-Intensive Engineering & Science, Johns Hopkins University

Abstract

Deciphering potential associations between network structures and the corresponding nodal attributes of interest is a core problem in network science. As the network topology is structured and often high-dimensional, many nonparametric statistical tests are not directly applicable, whereas model-based approaches are dominant in network inference. In this paper, we propose a model-free approach to test independence between network topology and nodal attributes, via diffusion maps and distance-based correlations. We prove in theory that the diffusion maps based on the adjacency matrix from an infinitely exchangeable graph can provide a set of conditionally independent coordinates for each node in graph, which yields a consistent test statistic for network dependence testing with distance-based correlations combined. The new approach excels in capturing nonlinear and high-dimensional network dependencies, and is robust against parameter choices and noise, as demonstrated by superior testing powers throughout various popular network models. An application on brain data is provided to illustrate its advantage and utility.

Keywords: kernel matrix, diffusion distance, infinitely exchangeable graph, multiscale generalized correlation

1 Introduction

Propelled by increasing demand and supply of network data, the ubiquitous influence of network inferences has motivated numerous recent advances and applications in statistics, physics, computer science, biology, social science, etc., which poses many new challenges to data scientists and statisticians due to its distinct structure. A graph, or equivalently a network, is formally defined as an ordered pair $\mathbf{G} = (V, E)$, where V represents the set of nodes (or vertices) and E is the set of edges. The graph can be conveniently represented by the adjacency matrix $\mathbf{A} = \{\mathbf{A}_{ij} : i, j = 1, .., n = |V|\}$, with $\mathbf{A}_{ij}$ being the edge weight between node i and node j, e.g., for an unweighted and undirected network, $\mathbf{A}_{ij} = \mathbf{A}_{ji} = 1$ if and only if node i and node j are connected by an edge, and zero otherwise.

Due to the distinct structure of graphs from traditional data analysis, statistical studies of network sciences are often model-based, e.g., the Erdos-Renyi model (Erdos and Renyi, 1959; Gilbert, 1959), stochastic block model (Holland et al., 1983; Rohe et al., 2011; Sussman et al., 2012; Lei and Rinaldo, 2015) and its degree-corrected version (Karrer and Newman, 2011; Zhao et al., 2012), the latent position model (Tang et al., 2013; Fosdick and Hoff, 2015), the random dot product model (Young and Scheinerman, 2007; Sussman et al., 2014) have all played significant roles in a range of tasks such as hypothesis testing, community detection, dimension reduction, etc. Despite their success so far, model-based statistical methods have limited applicability on real networks, e.g., statistical models and relevant inferences usually assume a connected, unweighted, and undirected graph, which only represents a subset of real networks; whether the given graph actually satisfies the presumed model can be challenging and expensive to verify; parameter selection under a given model is often difficult and un-guaranteed in practice, etc. Indeed, model misspecification can largely affect the inference performance on networks (Chen et al., 2016);

and it is becoming more desirable to develop robust and model-free approaches on graph analysis, so as to better handle a large variety of disparate graphs.

In this paper, we focus on one of the most fundamental statistical inference tasks on networks– testing independence between the nodal attributes and network topology; that is, given an adjacency matrix $\mathbf{A}$ and corresponding nodal attributes $\mathbf{X} = \{\mathbf{x}_i \in \mathbb{R}^{q_x} : i = 1, \ldots, n\}$ associated with each node, we would like to test whether the graph topology constructed via $\mathbf{A}$ has any relationship with the nodal attributes $\mathbf{X}$. For example, each person on Facebook has a number of distinct attributes (e.g., occupations, sex, personal behaviors), and also has connections with other persons and entities via the social network; in neuroscience, each brain region has its own distinct functionality, and is also linked with other regions in the brain map. Therefore, determining the existence of relationship between network connectivity and certain properties of the nodes is often a crucial first step in exploring the network data.

A notable obstacle in the network dependence testing problem is that the adjacency matrix $\mathbf{A}$ is structured (e.g., $\mathbf{A}$ is a symmetric matrix under un-directed graph), which prevents many well-established methodology from being directly used, e.g., unlike in traditional data analysis where one can comfortably assume independently identically distributed (i.i.d.) of each observation, there is no correlation measure straightly applicable between $\mathbf{A}$ and $\mathbf{X}$. Therefore, the primary focus so far has been on developing parametric tests on networks (Wasserman and Pattison, 1996; Fosdick and Hoff, 2015; Howard et al., 2016), which are limited by the boundary of model assumption. For example, the network dependence test proposed by Fosdick and Hoff (Fosdick and Hoff, 2015) assumes that the adjacency matrix is generated from a multivariate normal distribution of the latent factors, then proceeds to estimate the latent factor associated with each node from $\mathbf{A}$ (requiring dimension selection of the latent factor), followed by applying the standard likelihood ratio test on the normal distribution.

To tackle the challenges from parametric tests, we propose a new model-free method to test network dependency, via diffusion maps and distance-based correlations. We first compute the diffusion maps (Coifman et al., 2005; Coifman and Lafon, 2006; Lafon and Lee, 2006) from the adjacency matrix to obtain a node-wise embedding, then proceed to test independence between the diffusion distance of the adjacency matrix and the Euclidean distance of the nodal attributes. Assuming an infinitely exchangeable graph and mild regularity conditions, the new approach provides a set of asymptotic conditional i.i.d. embedding for the nodes, which results in consistent test statistics for network dependence testing by taking advantages of the recent progress in the distance-based correlation framework (Székely et al., 2007; Székely and Rizzo, 2013a; Shen et al., 2017). Moreover, our method is computationally inexpensive and robust against parameter mis-specifications, very efficient in capturing a wide variety of nonlinear and high-dimensional relationships, and readily extend-able to testing independence between two graphs.

The remaining of the paper is organized as follows: In Section 2 we briefly review diffusion maps and distance-based correlations, which are the main ingredients for the proposed test. In Section 3 we formally propose the testing method and prove the relevant theoretical properties under mild distributional assumptions, followed by discussions on implementation details and immediate extensions. Section 4 is dedicated to numerical experiments, where we illustrate the advantages of the proposed method under various scenarios. Proofs are put into the supplementary material, and the R codes and accompanying data are available online [1].

[1] https://github.com/neurodata/Multiscale-Network-Test

2 Preliminaries

2.1 Diffusion Maps

The diffusion map is introduced as a feature extraction algorithm by Coifman and Lafon (Coifman et al., 2005; Coifman and Lafon, 2006; Lafon and Lee, 2006), which computes a family of embeddings in the Euclidean space by eigen-decomposition on a diffusion operator of the given data.

To derive the diffusion maps for sample observations of size n, the first step is to find an $n \times n$ kernel matrix $\mathbf{K}$ that represents the similarity within the sample data, under the restriction that the kernel is symmetric and positive preserving, i.e., $\mathbf{K}_{ij} = \mathbf{K}_{ji}$, and $\mathbf{K}_{ij} \geq 0$ for all $i, j \in \{1, 2, \ldots, n\}$. Next we normalize the kernel matrix into a transition matrix $\widetilde{\mathbf{K}}$, where $\widetilde{\mathbf{K}}_{ij} = \mathbf{K}_{ij} / \sum_{j=1}^{n} \mathbf{K}_{ij}$ when $\sum_{j=1}^{n} \mathbf{K}_{ij} \neq 0$, and $\widetilde{\mathbf{K}}_{ij} = 0$ otherwise.

Fixing q as the embedding dimension and t as the iteration time step, the diffusion map $\mathbf{U} = \{\mathbf{u}_i, i = 1, \ldots, n\}$ is computed as

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

where $\{\lambda_j\}$ and $\{\phi_j\}$ denote the q largest eigenvalues and corresponding eigenvectors of $\widetilde{\mathbf{K}}$, and λ_j^t denotes the t^{th} power of the j^{th} eigenvalue.

The diffusion distance between the i^{th} observation and the j^{th} observation is defined as the similarity of the two points in the observation space with respect to the connectivity between them, which turns out to equal the Euclidean distance in the diffusion coordinate, i.e,

$$\mathbf{C}(i, j) = \|\mathbf{u}_i - \mathbf{u}_j\|; \quad i, j = 1, 2, \ldots, n. \tag{2}$$

Alternatively, the diffusion maps can be viewed as a nonlinear embedding method that integrates local similarities at different time steps. Compared to other nonlinear dimension reduction or feature extraction methods, the algorithm of diffusion maps is robust to noise perturbation and computationally inexpensive (i.e., the eigenvalue decomposition is the only time-consuming step).

Note that the diffusion maps require two parameters to be specified, namely the time step $t \in [1, \infty]$ and the number of eigenvalues $q \in [1, n]$. The parameters are always selected and fixed prior to applying the diffusion maps, which will be further elaborated in Section 3.

2.2 Distance-Based Correlations

The general problem of dependence testing between two random variables has seen notable progress in recent years. The Pearson's correlation (Pearson, 1895) 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 and ± 1 indicating perfect linear association. To better capture the dependencies not limited to linear relationship, a variety of distance-based correlation measures have been suggested recently, such as the Mantel coefficient (Mantel, 1967), distance correlation (`dCorr`) and energy statistic (Székely et al., 2007; Székely and Rizzo, 2013a; Rizzo and Székely, 2016), kernel-based independence test (Gretton and Gyorfi, 2010), Heller-Heller-Gorfine (`HHG`) test (Heller et al., 2013, 2016), and multiscale generalized correlation (`MGC`) (Shen et al., 2017), among others. In particular, the distance correlation by Székely et al. (2007) is the first correlation measure that is consistent against all possible dependencies with finite second moment. The multiscale generalized correlation statistic (`MGC`) by Shen et al. (2017) inherits the same consistency of distance correlation with remarkably better finite-sample testing powers under high-dimensional and nonlinear dependencies, by defining a family of distance-based local correlations and efficiently searching for the optimal correlation in testing. Here we briefly introduce `dCorr` and `MGC`.

Suppose a pair of sample data $(\mathbf{U}, \mathbf{X}) = \{(\mathbf{u}_i, \mathbf{x}_i);\ i = 1, 2, \ldots, n\}$ are independently and identically distributed as $(\mathbf{u}, \mathbf{x}) \in \mathbb{R}^{q \times q_x}$ (q and q_x are the respective feature dimension). Denote the pairwise distances within $\mathbf{U} = \{\mathbf{u}_i\}_{i=1}^n$ and $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^n$ as $\mathbf{C}_{ij} = \|\mathbf{u}_i - \mathbf{u}_j\|$ and $\mathbf{D}_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|$ for $i, j = 1, 2, \ldots, n$, where $\| \cdot \|$ is the Euclidean distance.

Then the distance covariance of the sample data is defined as

$$\texttt{dCov}(\mathbf{C}, \mathbf{D}) = \frac{1}{n^2} \sum_{i,j=1}^{n} \tilde{\mathbf{C}}_{ij} \tilde{\mathbf{D}}_{ij}, \tag{3}$$

where $\tilde{\mathbf{C}}$ and $\tilde{\mathbf{D}}$ doubly-center $\mathbf{C}$ and $\mathbf{D}$ by its column mean and row mean respectively, i.e., $\tilde{\mathbf{C}} = \mathbf{HCH}$, where $\mathbf{H} = I_n - J_n/n$ (the double centering operation 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 above distance covariance, and lies in the range of $[0, 1]$. In Székely et al. (2007), the `dCorr` has been shown to be asymptotically 0 if and only if $\mathbf{u}$ is independent of $\mathbf{x}$, resulting in a consistent statistic for independence testing. In addition, a modified version of distance correlation (`mCorr`) is proposed to eliminate the sample bias of `dCorr` for improved testing performance (Székely and Rizzo, 2013b, 2014).

The `MGC` statistic is an optimal local version of a given distance-based global correlation, aiming for improved finite-sample testing power. Taking the distance correlation for example, it first derives all local distance covariances $\texttt{dCov}_n^{kl}$ as

$$\texttt{dCov}_n^{kl} = \frac{1}{n^2} \sum_{i,j=1}^{n} \tilde{\mathbf{C}}_{ij} \tilde{\mathbf{D}}_{ij} I\big(r(\mathbf{C}_{ij}) \leq k\big) I\big(r(\mathbf{D}_{ij}) \leq l\big); \quad k, l = 1, \ldots, n, \tag{4}$$

where $r(\mathbf{C}_{ij})$ is the rank function of $\mathbf{u}_i$ relative to $\mathbf{u}_j$, i.e., $r(\mathbf{C}_{ij}) = k$ if $\mathbf{u}_i$ is the k^{th} nearest neighbor of $\mathbf{u}_j$, and define equivalently $r(\mathbf{D}_{ij})$ for $\mathbf{X}$. Then the local distance correlations $\texttt{dCorr}_n^{kl}$ are the normalizations of the local distance covariances into $[-1, 1]$. Among all possible neighborhood choices, the `MGC` statistic ρ^* equals the largest local distance correlation up-to a smoothing operation $S(\cdot)$, i.e.,

$$\rho^* = \texttt{dCorr}_n^{(kl)^*}, \text{ where } (kl)^* = \arg\max_{(kl)} S(\texttt{dCorr}_n^{kl}). \tag{5}$$

The `MGC` statistic is also applicable to `mCorr` as well, which is the default implementation for the proposed approach in Section 3. More details of `mCorr` implementation and the smoothing function are available in Shen et al. (2017).

It is worthwhile to note that `MGC` is computationally efficient, despite of searching over all possible neighborhoods for the optimal local correlation. Assuming sample size is n and the maximal feature dimension of $\mathbf{U}$ and $\mathbf{X}$ is q, `MGC` runs in $O(n^2 \max\{\log n, q\})$. In comparison, `dCorr` and `mCorr` run in $O(n^2 q)$, while the `HHG` statistic (another distance-based method excellent for testing nonlinear dependency) has the same complexity as `MGC`. The space requirement for any above-mentioned method is $O(n^2)$.

3 Network Dependence Testing

In this section we propose our approach built on diffusion maps and `MGC`. Theoretical properties are presented in Section 3.2, with discussions followed in Section 3.3.

3.1 A Nonparametric Approach

Input: The adjacency matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ for the network topology, and the nodal attributes $\mathbf{X} = \{\mathbf{x}_i \in \mathbb{R}^{q_x}, i = 1, \ldots, n\}$. Note that the input graph can be of any structure, e.g., directed or undirected, weighted or unweighted, connected or disconnected, has self-loop or not.

Step 1: Transform the adjacency matrix into a proper Euclidean embedding via the diffusion maps. We set the symmetric kernel matrix as $\mathbf{K}_{ij} = (\mathbf{A}_{ij} + \mathbf{A}_{ji})/2$, then the diffusion map $\mathbf{U} = \{\mathbf{u}_i, i = 1, \ldots, n\}$ follows by applying the eigen-decomposition to the normalized kernel matrix $\widetilde{\mathbf{K}}$ as in Equation 1. For the parameters, we fix $t = 3$, and select q by the third elbow of the eigenvalue scree plot using the profile likelihood method in Zhu and Ghodsi (2006).

Step 2: Calculate the `MGC` statistic $\rho^*(\mathbf{C}, \mathbf{D})$ from the pairwise distances. $\mathbf{C}$ is computed as the diffusion distance matrix from $\mathbf{U}$, and $\mathbf{D}$ is the Euclidean distance matrix of $\mathbf{X}$, followed by applying Equation 4 and 5.

Step 3: Compute the p-value of the statistic via permutation test. Given any permutation σ of size n, we randomly permute the sample indices of the nodal attributes to $\mathbf{X}_\sigma = \{\mathbf{x}_{\sigma(i)}\}$, and compare $\rho^*(\mathbf{C}, \mathbf{D})$ to $\rho^*(\mathbf{C}, \mathbf{D}_\sigma)$. Using r random permutations, the p-value equals the proportion of times that the permuted `MGC` statistic is no smaller than the original `MGC` statistic.

Output: The p-value of the test, which determines the rejection of the independence hypothesis based on a given critical level α.

A flowchart of the above procedure is provided in Figure 1.

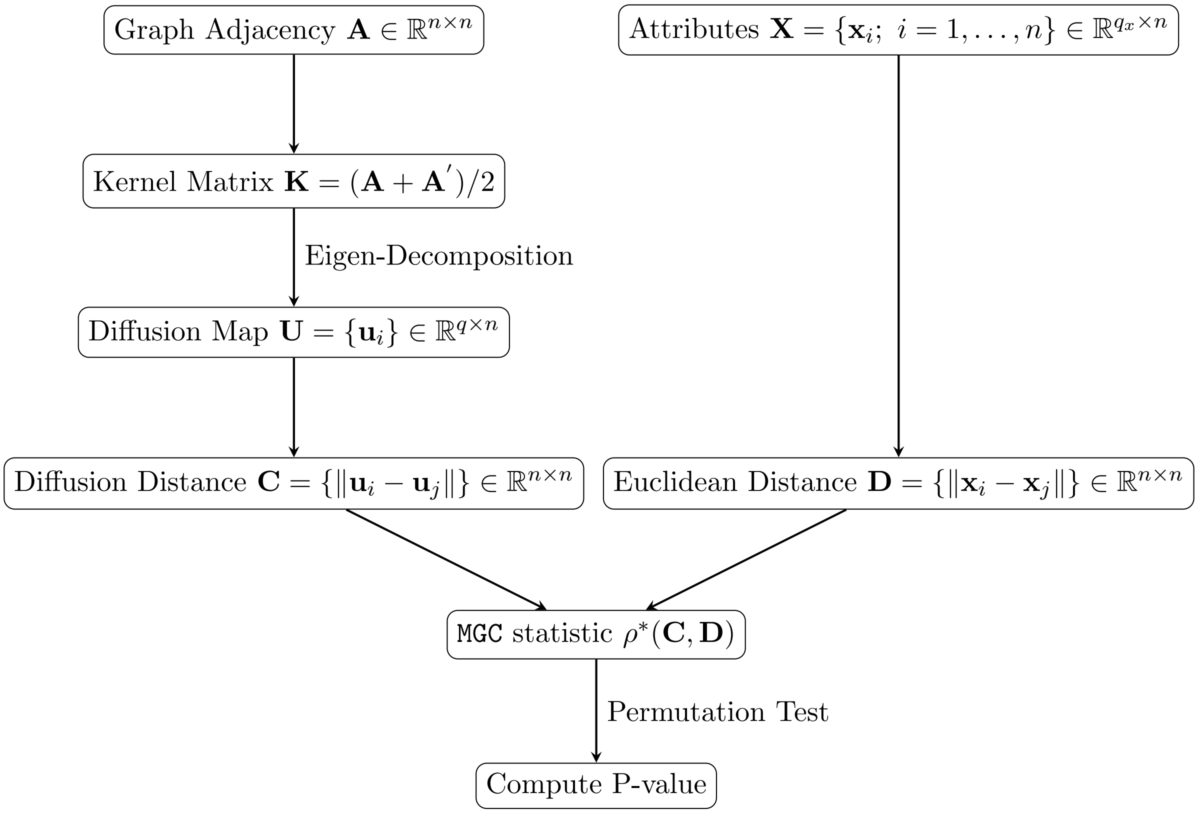

Graph Adjacency $\mathbf{A} \in \mathbb{R}^{n \times n}$
Attributes $\mathbf{X} = \{\mathbf{x}_i;\ i = 1, \ldots, n\} \in \mathbb{R}^{q_x \times n}$
Kernel Matrix $\mathbf{K} = (\mathbf{A} + \mathbf{A}^{'})/2$
Eigen-Decomposition
Diffusion Map $\mathbf{U} = \{\mathbf{u}_i\} \in \mathbb{R}^{q \times n}$
Diffusion Distance $\mathbf{C} = \{\|\mathbf{u}_i - \mathbf{u}_j\|\} \in \mathbb{R}^{n \times n}$
Euclidean Distance $\mathbf{D} = \{\|\mathbf{x}_i - \mathbf{x}_j\|\} \in \mathbb{R}^{n \times n}$
MGC statistic $\rho^*(\mathbf{C}, \mathbf{D})$
Permutation Test
Compute P-value

Figure 1: Flowchart for Network Dependence Testing via Diffusion Maps and MGC

3.2 Theoretical Properties

Throughout this section, we assume that the parameters t and q, and the dimension of the nodal attributes q_x are fixed and finite. Moreover, we require the following distributional conditions on the graph $\mathbf{G}$ and the nodal attributes $\mathbf{X}$:

(C1) Graph $\mathbf{G}$ is from an infinitely exchangeable graph, i.e., the adjacency matrix $\mathbf{A}$ is generated by an $n \times n$ random matrix $\mathcal{A}$ such that

$$\mathcal{A}_{ij} \overset{d}{=} \mathcal{A}_{\sigma(i)\sigma(j)}$$

for any $i, j = 1, \ldots, n$ and any permutation σ of size n, where n can be finite or infinite. The notation $\overset{d}{=}$ stands for equality in distribution.

(C2) The underlying distribution of the nodal attributes are of finite second moment, i.e., we assume $\{\mathbf{x}_i;\ i = 1, \ldots, n\}$ are independently and identically distributed as x, whose second moment is finite.

Condition (C1) is straightforward to verify and most popular statistical networks models satisfy this condition under regularity conditions. For example, both the stochastic block model and latent position model can be thought of as a subgraph of an infinitely exchangeable random graph, when the block-membership or latent positions are assumed independently and identically distributed in each model. From condition (C1), we prove that the diffusion map $\mathbf{U}$ can furnish asymptotic conditional i.i.d. embedding for the set of nodes V.

Theorem 1. Assume $\mathbf{G}$ satisfies (C1). Then the derived diffusion map $\{\mathbf{u}_i\}$ are conditionally i.i.d. as $n \to \infty$, i.e., there exists a latent variable u such that $\mathbf{u}_i | u$ are i.i.d. for $i = 1, \ldots, n$ as $n \to \infty$.

The conditional i.i.d. property of the diffusion map can lead to the consistency of distance-based correlations for the network dependence test, as long as the underlying distributions of the diffusion map and nodal attributes are of finite second moment. Therefore, condition (C2) is merely a regularity condition on the distribution of nodal attributes; while the next lemma states that the diffusion map always satisfies the finite-moment assumption.

Lemma 1. Under the same assumptions of Theorem 1, the latent variable u of the diffusion map is of finite second moment.

The consistency of the proposed test method is established in the next theorem.

Theorem 2. Assume the graph $\mathbf{G}$ and the attributes $\mathbf{X}$ satisfy condition (C1) and (C2). Then the `MGC` statistic between the derived diffusion map $\mathbf{U}$ and the nodal attributes $\mathbf{X}$ satisfies:

$$\rho^*(\mathbf{C}, \mathbf{D}) \longrightarrow 0 \text{ as } n \to \infty \tag{6}$$

if and only if u is independent of x. Namely, it is a consistent statistic for testing independence between the network topology and nodal attributes.

The following corollary summarizes some immediate extensions of the main testing procedure.

Corollary 1. Theorem 2 still holds, when any of the following changes are applied to the testing procedure described in Section 3.1:

(1) Instead of $t = 3$ and q as the third elbow, t and q are chosen as any other finite positive integers;

(2) The nodal attributes in the input are replaced by a second graph of the same node set, which satisfies condition (C1) and has its' distance matrix $\mathbf{D}$ computed by the same diffusion map process;

(3) The `MGC` statistic in step 2 is replaced by either `dCorr` or `mCorr`.

Namely, the theoretical consistency always holds regardless of the choice on t and q; the same approach can be applied to test dependency between two graphs of arbitrary structures, e.g., testing between a weighted directed graph and an unweighted undirected graph, by deriving the diffusion distance for each graph individually; that is, once the appropriate distance is determined, distance-based measures other than MGC can also be consistent, such as dCorr and mCorr.

To that end, we offer more discussions on the parameter selection in the next section, consider a two-graph testing scenario in Section 4.4, and use mCorr and HHG in the simulations to make meaningful comparisons with the main approach.

All proofs are supplied in Appendix A.

3.3 Discussions

In this section, we discuss the parameter selection and the advantages of diffusion maps over other embeddings.

There are two parameters to choose for the diffusion map, i.e., the time step t, and the dimension of graph embedding q. Theoretical properties always hold for finite t and q, but their choices may affect the testing power. Empirically, it turns out that the diffusion map is relatively insensitive to choice of t, and able to well capture the geometric structure in a wide range of $t > 1$, from which we choose to fix $t = 3$ throughout; upon fixing t, the diffusion distance is also relatively indifferent to the choice of q, for which we opt to select q as the third elbow of the eigenvalue scree plot by the profile likelihood method from Zhu and Ghodsi (2006). This is a widely-used automatic algorithm for selecting the number of important features, whenever eigenvalues or singular values are involved.

To support our parameter choices, we provide empirical evidence in Figure 2: an adjacency matrix of $n = 100$ is generated by the three-block stochastic block model in Equation 8 (details in Section 4), followed by the diffusion distances at different t and q. The diffusion distance matrix is always able to preserve the block structures for $t \in [2, 10]$ and

$q \in [2, n]$. Recognizing such block structures is important in detecting dependence between the graph structure and block-membership.

In comparison, standard graph embedding methods are often sensitive to the choice of q. Figure 3 shows the Euclidean distance of the adjacency spectral embedding (ASE) (Sussman et al., 2012) applied to the same adjacency matrix. For ASE, the correct dimensional choice equals the number of blocks, i.e., the distance matrix at $q = 3$ shows a clear block structure. However, a slight mis-specification of q can cause the embedding to have a more obscure block structure, and the elbow method often fails to find the correct q for ASE.

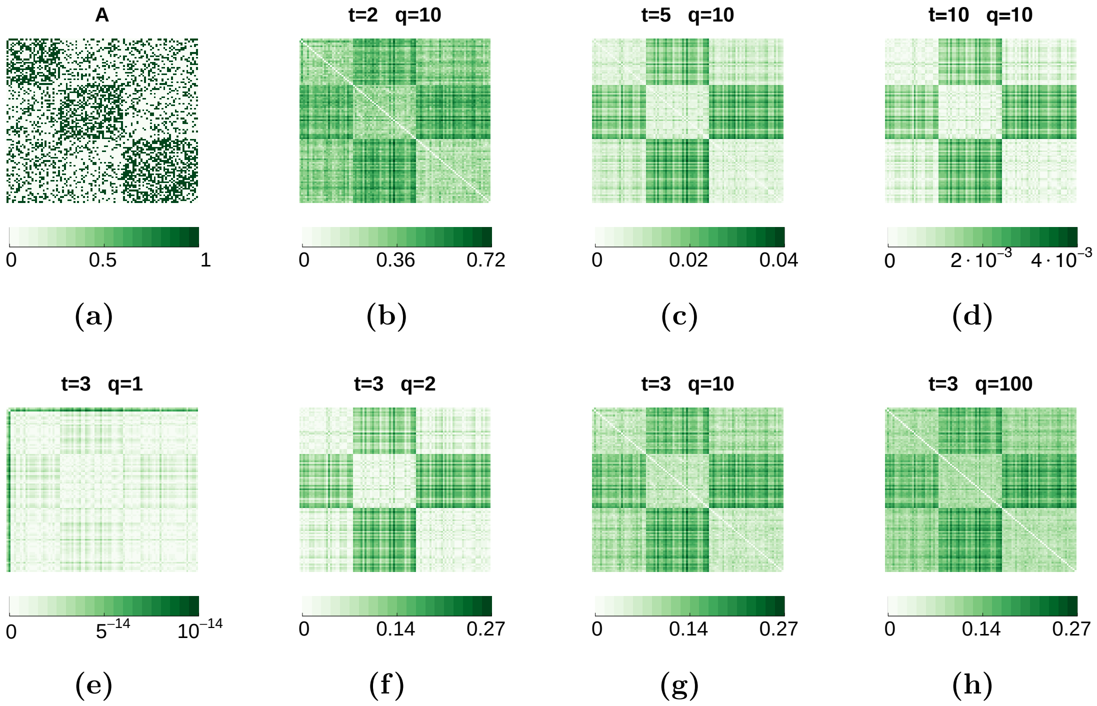
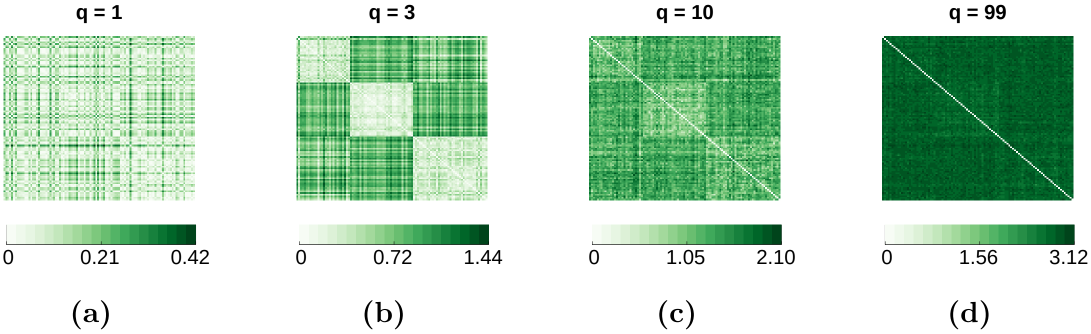


Figure 2: Diffusion maps are robust against parameter mis-specifications. Panel (a) illustrates the three-block adjacency matrix $\mathbf{A}$ generated by Equation 8. Panel (b)-(d) show the diffusion distance matrix at $q = 10$ and increasing t, while panel (e)-(h) present the diffusion distance at $t = 3$ and increasing q. The distance matrix exhibits a clear block structure, for $t \in [2, 10]$ and $q \in [2, n]$. Note that at $t = 3$, the first three elbows of eigenvalues are $(1, 3, 84)$ respectively.



(a) (b) (c) (d)

Figure 3: Adjacency spectral embedding is not robust against dimension misspecification particularly compared to the diffusion maps. Panel (a)-(d) show the Euclidean distance matrix of ASE at increasing q, using the same adjacency matrix in Figure 2(a). Only ASE at $q = 3$, namely the correct dimension, is able to display a clear block structure. Note that the first three elbows are $(1, 45, 70)$, so that it is difficult for ASE to recognize the block structure when the dimension is chosen via the scree plot.

Therefore, our motivation to use diffusion maps is three-fold: it yields an embedding of the graph structure into the Euclidean space; it is relatively robust against parameter mis-specification; and it allows a seamless integration with distance-based correlations for consistent dependency testing. On the other hand, it is tempting to use other embedding methods, like adjacency spectral embedding, Laplacian spectral embedding (Rohe et al., 2011), the latent factor embedding (Fosdick and Hoff, 2015), etc. Those embeddings can also be directly combined with `dCorr` or `MGC` to test dependence, but they may lose the theoretical properties without more strict model assumption and are often more sensitive to the parameter choice, thereby leading to sub-optimal performance. For comparison purposes, we include other embedding choices for benchmark in the simulations.

4 Numerical Studies

In this section, we demonstrate the advantages of our approach via simulations and real data experiment. Throughout the numerical studies, we mainly compare our approach to the likelihood ratio test proposed by Fosdick and Hoff (FH) (Fosdick and Hoff, 2015), as well as a number of variants mentioned in Section 3: in place of MGC, we consider applying mCorr and HHG to the distances; and instead of diffusion distance as network metric, we consider using the Euclidean distances of the adjacency matrix (AM) or the latent factors (LF), prior to computing the distance-based correlation. Thus the main approach is denoted as $\texttt{MGC} \circ \texttt{DM}$ (MGC statistic applied to diffusion maps), whereas the other variants are denoted as $\texttt{MGC} \circ \texttt{AM}$, $\texttt{mCorr} \circ \texttt{DM}$, $\texttt{HHG} \circ \texttt{LF}$, etc.

For each simulation, we generate sample graph and the corresponding attributes (or two sample graphs in Section 4.4), carry out the permutation test for each method, 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 out of 500 Monte-Carlo replicates, and a higher power implies a better method. For the real data experiment, we directly report the p-value of the permutation test, since the testing power is not available, and a lower p-value indicates a better method assuming the existence of relationships.

Whenever dimension selection is required, i.e., selection of q for DM and LF, we always pick the third elbow of the scree plot bounded above by $\min(n, 100)$.

4.1 Stochastic Block Model

The first simulation generates graphs by the stochastic block model (SBM), which is one of the most popular network models which many network methodologies have been built on. The SBM assumes that each of n nodes in $\mathbf{G}$ must belong to one of $K \in \mathbb{N}$ blocks, and determines the edge probability based on the block-membership of the connecting nodes: For $i = 1, \dots, n$, assume that a latent variable of $\mathbf{z}_i \overset{i.i.d.}{\sim} Multinomial\big(\pi_1, \pi_2, ..., \pi_K\big)$ denotes the block-membership of each node, and $p_{kl} \in [0, 1]$ implies the edge probability between any two nodes of class k and l respectively; then the upper triangular entries of $\mathbf{A}$ are independently and identically distributed conditioned on $\{\mathbf{z}_i\}$:

$$\mathbf{A}_{ij}\big|\mathbf{z}_i, \mathbf{z}_j \overset{i.i.d.}{\sim} Bernoulli\big(\sum_{k,l=1}^{K} p_{kl}\mathbf{I}\big(\mathbf{z}_i = k, \mathbf{z}_j = l\big)\big); \quad \forall i < j,\ i, j = 1, 2, \dots, n, \tag{7}$$

where $\mathbf{I}(\cdot)$ is the indicator function.

It is of common interest to detect whether the block structure has anything to do with the graph connectivity; and noisy information on the block structure often better reflects reality and adds difficulty to the testing. Thus we consider testing dependency between the graph having the adjacency matrix $\mathbf{A}$ and a noisy block-membership $\mathbf{X}$, which are correlated through true block-membership $\mathbf{Z}$. We set $n = 100$ and $K = 3$, select $\mathbf{z}_i$ uniformly from $\Omega = \{1, 2, 3\}$ for each i, generate the edge probability by

$$E(\mathbf{A}_{ij}|\mathbf{z}_i, \mathbf{z}_j) = 0.5\mathbf{I}(|\mathbf{z}_i - \mathbf{z}_j| = 0) + 0.2\mathbf{I}(|\mathbf{z}_i - \mathbf{z}_j| = 1) + 0.3\mathbf{I}(|\mathbf{z}_i - \mathbf{z}_j| = 2), \quad i, j = 1, \dots, n, \tag{8}$$

and contaminate the true block-membership by: for each i, $\mathbf{x}_i = \mathbf{z}_i$ with probability 0.5, and equally likely to take other values in Ω, i.e., the true block-membership are observed half of the time.

To elaborate Equation 8, within-block edge probability is always 0.5, while between-block edge probability is 0.2 when the block labels differ by 1, and 0.3 when the block labels differ by 2. A visualization of the sample data is shown in Figure 2. Notably, although within-block edge probability is the largest, the between-block edge probability is not linearly related to the distance of the block-membership, i.e., the edge probability between a node of block 1 and a node of block 3 is higher than the edge probability between block 1 and block 2.

Therefore, this three-block SBM generates a noisy and nonlinear dependency structure between **A** and **X**, for which `MGC` is expected to work better than `mCorr`, `HHG`, and the standard likelihood ratio test; moreover, the diffusion map should be more robust than the other two metrics of adjacency matrix (`AM`) and latent factors (`LF`). Indeed, Figure 2 shows that the main approach (i.e., `MGC` $\circ$ `DM`) prevails in the testing powers among all the methods.

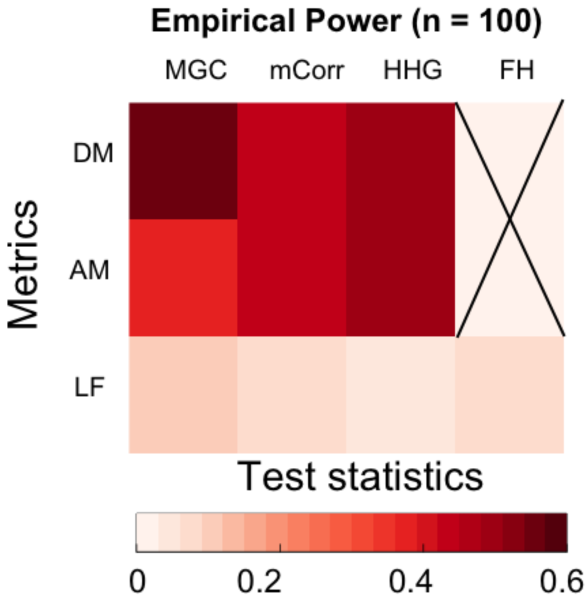


Figure 4: The power heatmap under the three-block SBM (Equation 8) demonstrates that for all possible combinations of test statistics with distance metrics, `MGC` with the diffusion maps (`DM`) provides the best power compared to all other methods.

4.2 SBM with Linear and Nonlinear Dependencies

To better understand the advantage of our main approach under different scenarios, here we use the same three-block SBM and its block-membership $\{\mathbf{z}_i : i = 1, 2, \ldots, n = 100\}$ as in the previous section, except that the edge probability is now controlled by $\theta \in (0, 1)$ as follows:

$$E(\mathbf{A}_{ij}|\mathbf{z}_i, \mathbf{z}_j) = 0.5\mathbf{I}(|\mathbf{z}_i - \mathbf{z}_j| = 0) + 0.2\mathbf{I}(|\mathbf{z}_i - \mathbf{z}_j| = 1) + \theta\mathbf{I}(|\mathbf{z}_i - \mathbf{z}_j| = 2); \quad i, j = 1, \ldots, n. \tag{9}$$

The noisy block-membership $\mathbf{X}$ is generated in the same way as before.

When $\theta = 0.2$, the three-block SBM is the same as a two-block SBM, where within-block edge probability equals 0.5 while the between-block edge probability is always 0.2, i.e., it represents a linear association between the adjacency matrix and the block-membership; when $\theta < 0.2$, the association is still close to linear; when $\theta > 0.2$ and gets further away, the relationship becomes strongly nonlinear.

Figure 5 plots the power against θ of `DM` based on all methods. All of `MGC`, `mCorr`, and `HHG` perform almost the same at linear or close to linear dependency (i.e, $\theta \leq 0.2$), with `MGC` being significantly more powerful as the dependency shifts to strongly nonlinear. This observation confirms that the `MGC` is able to capture nonlinear dependencies in network testing, suggesting it is a better test statistic for general usage.

4.3 Degree-corrected Stochastic Block Model

Next we investigate the degree-corrected stochastic block model (DC-SBM), which better reflects many sparse networks in reality. The DC-SBM is an extension of SBM by introducing an additional random variable $\mathbf{c}_i$ to control the degree of each node.

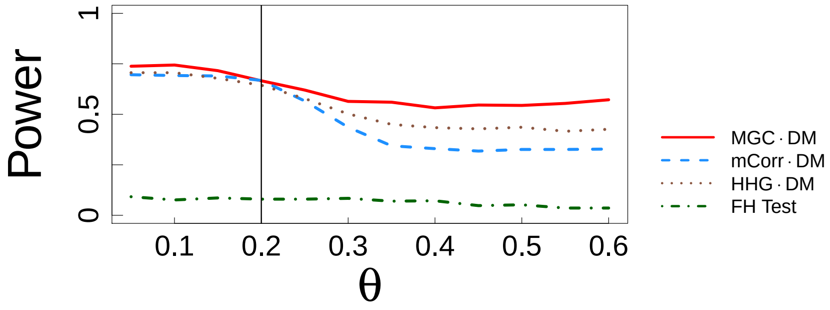


Figure 5: The power curve with respect to increasing θ under three-block SBM (Equation 9) where $\{\theta : \theta \leq 0.2\}$ represents a close to linear dependency between the adjacency matrix and the block structure; while $\{\theta : \theta > 0.2\}$ generates nonlinear dependency and it becomes strongly nonlinear as θ gets further away from 0.2. Among other benchmarks utilizing diffusion maps, `MGC` is the best performing method throughout all possible θ, implying that it is able to better capture nonlinear dependencies.

We set $n = 200$ and $K = 2$, select the block-membership $\mathbf{z}_i$ uniformly in $\Omega = \{0, 1\}$, and generate the edge probability by

$$E(\mathbf{A}_{ij}|\mathbf{z}_i, \mathbf{z}_j, \mathbf{c}_i, \mathbf{c}_j) = 0.2\mathbf{c}_i\mathbf{c}_j \cdot \mathbf{I}(|\mathbf{z}_i - \mathbf{z}_j| = 0) + 0.05\mathbf{c}_i\mathbf{c}_j \cdot \mathbf{I}(|\mathbf{z}_i - \mathbf{z}_j| = 1), \tag{10}$$

where $\mathbf{c}_i \overset{i.i.d}{\sim} Uniform(1-\tau, 1+\tau)$ for $i = 1, \ldots, n$, and $\tau \in [0, 1]$ is a parameter to control the amount of variability in the edge degree, e.g., as τ increases, the model becomes more complex as the variability of the edge probability becomes larger; when $\tau = 0$, the above model reduces to a two-block SBM without any variability induced by $\{\mathbf{c}_i\}$.

Similar as before, we generate the nodal attributes $\mathbf{X}$ as a noisy version of the true block-membership via Bernoulli distribution, i.e., for each i, $\mathbf{x}_i = \mathbf{z}_i$ with probability 0.6, and equals the wrong label with probability 0.4.

Under the DC-SBM model, the main approach also has the most superior power. In particular, Figure 6 compares different metrics using `MGC`, which illustrates that the diffusion

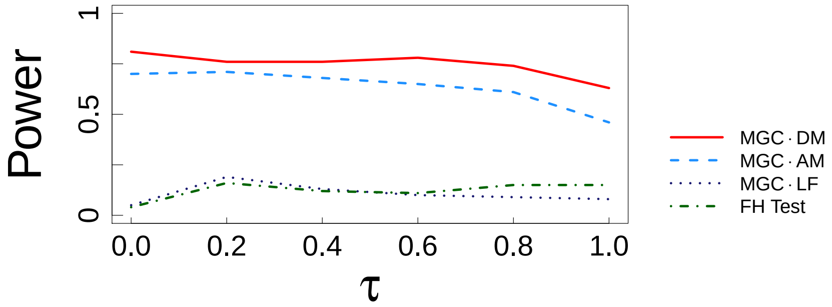


Figure 6: The power curve with respect to increasing τ under DC-SBM (Equation 10). As the edge variability increases as τ does, the testing power of diffusion maps are relatively stable against increasing variability compared to the adjacency matrix; while the latent positions fail to detect the dependency across all levels of τ.

map consistently reflects the network topology of DC-SBM better throughout different choices of τ.

4.4 Two-graph Testing under the Random Dot Graph Model

The next simulation applies the network dependency test approach to two graphs, i.e., instead of testing dependencies between one graph and the corresponding attributes, we test two graphs of the same node set with different edges, which is a scenario included in Corollary 1.

Assume two graphs $\mathbf{G}_1$ and $\mathbf{G}_2$ are generated via a latent variable $\mathbf{u}_i$ as follows:

$$\begin{aligned}
\mathbf{u}_i &\overset{i.i.d.}{\sim} Uniform(0,1), \quad i = 1, 2, \ldots, n \\
\mathbf{w}_i &:= \mathbf{u}_i^2, \quad i = 1, 2, \ldots, n \\
\mathbf{A}_{ij}^{(1)} \big| \mathbf{u}_i, \mathbf{u}_j &\sim Bernoulli\big(< \mathbf{u}_i, \mathbf{u}_j > \big), \quad \forall i < j,\ i, j = 1, 2, \ldots, n \\
\mathbf{A}_{ij}^{(2)} \big| \mathbf{w}_i, \mathbf{w}_j &\sim Bernoulli\big(< \mathbf{w}_i, \mathbf{w}_j > \big), \quad \forall i < j\ i, j = 1, 2, \ldots, n.
\end{aligned} \tag{11}$$

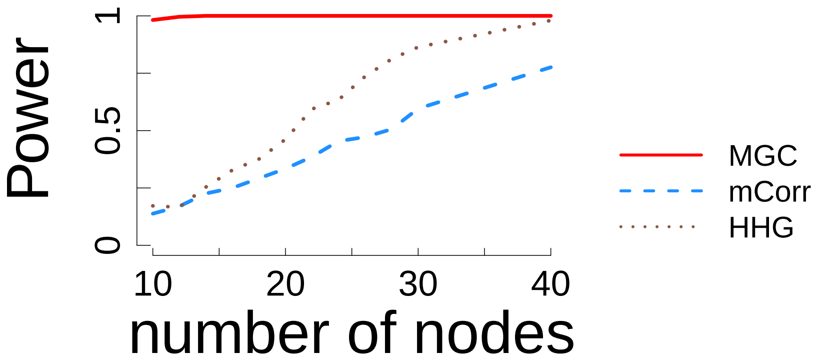


Figure 7: The power curve with respect to increasing number of nodes for the two-graph dependency testing simulation. The proposed approach quickly attains perfect power at a very small node size, while other benchmarks often require a much larger graph for perfect testing.

Namely, each graph is generated by a random dot product graph, and the underlying dependency is reflected via the quadratic relationship of the latent variables.

Figure 7 shows the testing power of `MGC`, `mCorr`, and `HHG` against the number of nodes n, all based on the diffusion maps. Although all methods achieve power of 1 as the number of nodes increases (which is expected for consistent test statistics), the proposed approach is able to achieve perfect testing power at a very small size while all other alternatives require a larger graph size. Note that if noise is included in the set-up, or the nonlinear relationship is more complex than quadratic, the proposed approach still enjoys the same advantage, i.e., the testing power converges to 1 faster than all other methods (similar to the simulations in Shen et al. (2017)), though the actual number of nodes to achieve perfect power will likely increase under noisy and complex dependency.

4.5 Real Data Experiment with Contamination

To demonstrate the application of the proposed method in real data, we test independence on the brain network, where each node indicates voxel in the brain and edges represent brain fibers connecting each region (Kiar et al., 2016b). Along with the information of the functional connectivity between the voxels, the data also provides the physical locations of each voxel in the brain represented by 3-dimensional voxel-wise coordinates. The brain network has been constructed by combining dMRI and sMRI data using the tool called `ndmg` (Kiar et al., 2016a), and the data is publicly available online [2].

The whole brain network data itself contains millions of nodes, so it is highly time-consuming to use all data for testing. As it is a common practice to sub-sample a small portion of big data for testing (under the assumption that significant relationship within the sub-sample is representative of the full data), we randomly selected one connected component having 95 nodes and 337 edges to test dependency between the functional connectivity versus each voxel's physical location. Moreover, we consider data contamination by edge flipping of the adjacency matrix: at contamination rate $c\%$, we randomly select $c\%$ of ones in the adjacency matrix and make them zeros, then randomly select $c\%$ of non-diagonal zeros and make them ones. Therefore, $c = 0$ equals the original data, while $c = 100$ flips all edges and it is merely a transformation of the original graph.

Figure 8 shows the p-value against the level of contamination $c\%$. The diffusion maps-based tests perform the best, with `MGC` and `mCorr` being slightly better than `HHG`. All other tests are less robust as the contamination level increases, e.g., `FH` test is able to yield a significant p-value at $c = 0$, but fails at $c = 100$ while the proposed approach is still able to test significant relationship. Note that there is only one p-value to report at $c = 0$ and

[2] http://openconnecto.me/data/public/MR/m2g_v1_1_1/KKI2009/derivatives/bg/

$c = 100$, while at other choices of c we randomly contaminate the data for 100 replicates and report the mean p-value of them.

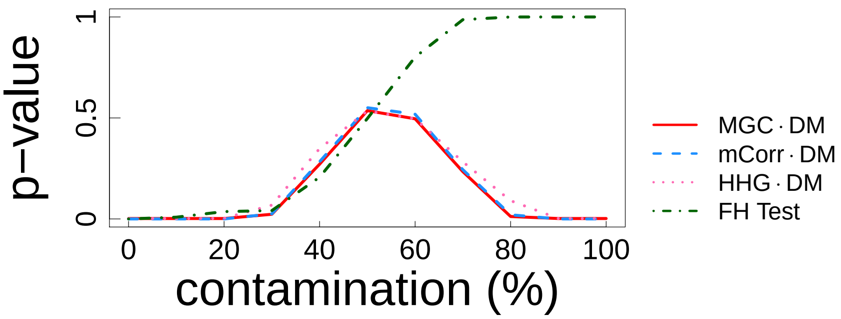


Figure 8: P-values are obtained through 100 random contamination for each of contamination level $c\% \in \{10\%, 20\%, \ldots, 90\%\}$. As the level of contamination increases, it is less likely to reject the independence hypothesis with the `FH Test` than with our proposed approach.

5 Conclusion

In this study, we propose a new method for testing dependency on network data via diffusion maps and distance-based correlations. The utilization of the diffusion maps not only warrants the integration with various types of distance-based correlations, but also makes the testing method robust against parameter mis-specifications. Moreover, applying multiscale generalized correlation allows the test to detect a wide range of nonlinear, high-dimensional, and noisy dependencies. Therefore, the proposed approach is able to achieve significant power improvement over other alternatives in network dependency testing.

There are several follow-ups that would further advance a number of frontiers in this area. First, it would be desirable to theoretically prove the robustness of the diffusion maps, and compare to other embedding methods under certain network models. Secondly, as network data in practice are often of huge size, the time and space requirement for dependence testing can be formidable without sub-sampling (Zhang et al., 2017), thus it would be necessary to further investigate the testing performance via sub-sampling. Third, it will be worthwhile to further investigate dependence testing or two-sample testing on networks, in particular graphs with only partial correspondence. Last but not least, the framework proposed in this paper also defines a good correlation measure on networks, thereby enabling certain popular statistical techniques to be directly applied to network data such as feature screening via distance correlation (Li et al., 2012).

References

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

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

Coifman, R. R., Lafon, S., Lee, A. B., Maggioni, M., Nadler, B., Warner, F., and Zucker, S. W. (2005). 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.

Diaconis, P. and Freedman, D. (1980). Finite exchangeable sequences. *The Annals of Probability*, pages 745–764.

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

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

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

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

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

Heller, R., Heller, Y., Kaufman, S., Brill, B., and Gorfine, M. (2016). Consistent distribution-free k-sample and independence tests for univariate random variables. *Journal of Machine Learning Research*, 17(29):1–54.

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

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

Inoue, H. and Taylor, R. (2006). Laws of large numbers for exchangeable random sets in kuratowski-mosco sense. *Stochastic Analysis and Applications*, 24(2):263–275.

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

Kiar, G., Gorgolewski, K. J., Kleissas, D., Roncal, W. G., Litt, B., Wandell, B., Poldrack, R. A., Wiener, M., Vogelstein, R. J., Burns, R., et al. (2016a). Science in the cloud (sic): A use case in mri connectomics. *arXiv preprint arXiv:1610.08484*.

Kiar, G., Roncal, W. G., Mhembere, D., Bridgeford, E., Burns, R., and Vogelstein, J. (2016b). ndmg: Neurodatas mri graphs pipeline.

Koroljuk, V. S. and Borovskich, Y. V. (1994). *Theory of U-statistics*. Springer Science+Business Media Dordrecht.

Lafon, S. and Lee, A. B. (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.

Lei, J. and Rinaldo, A. (2015). Consistency of spectral clustering in stochastic block models. *The Annals of Statistics*, 43(1):215–237.

Li, R., Zhong, W., and Zhu, L. (2012). Feature screening via distance correlation learning. *Journal of American Statistical Association*, 107:1129–1139.

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

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

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

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

Rohe, K., Chatterjee, S., and Yu, B. (2011). Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics*, pages 1878–1915.

Shen, C., Priebe, C. E., Maggioni, M., and Vogelstein, J. T. (2017). Discovering relationships across disparate data modalities. *arXiv preprint arXiv:1609.05148*.

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

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

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

Székely, G. and Rizzo, M. (2014). Partial distance correlation with methods for dissimilarities. *Annals of Statistics*, 42(6):2382–2412.

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

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

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

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

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

Zhang, Q., Filippi, S., Gretton, A., and Sejdinovic, D. (2017). Large-scale kernel methods for independence testing. *Statistics and Computing*, pages 1–18.

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

Zhu, M. and Ghodsi, A. (2006). Automatic dimensionality selection from the scree plot via the use of profile likelihood. *Computational Statistics and Data Analysis*, 51:918–930.

Acknowledgment

This work was partially supported by the National Security Science and Engineering Faculty Fellowship (NSSEFF), the Johns Hopkins University Human Language Technology Center of Excellence (JHU HLT COE), the Defense Advanced Research Projects Agency's (DARPA) SIMPLEX program through SPAWAR contract N66001-15-C-4041, the XDATA program of DARPA administered through Air Force Research Laboratory contract FA8750-12-2-0303. The authors thank Dr. Daniel Sussman, Dr. Minh Tang, and Dr. Carey Priebe for insightful suggestions.

A Proofs

Proof of Theorem 1. To prove conditional i.i.d. of the diffusion map $\mathbf{U} = \{\mathbf{u}_i : i = 1, \ldots, n\}$ as $n \to \infty$, by the celebrated *de Finetti's Theorem* (Diaconis and Freedman, 1980; Orbanz and Roy, 2015) it is equivalent to prove that $\{\mathbf{u}_i : i = 1, \ldots, n\}$ are infinitely exchangeable, i.e., for any n and all possible permutation σ, the permuted sequence $\{\mathbf{u}_{\sigma(1)}, \mathbf{u}_{\sigma(2)}, \ldots, \mathbf{u}_{\sigma(n)}\}$ always distributes the same as the original sequence $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n\}$.

From now on, denote $\mathbf{U}$ as the $q \times n$ matrix having $\mathbf{u}_i$ as its i^{th} column, and the permuted matrix as $\mathbf{U}\pi$ with π being the permutation matrix. Transforming Equation 1 into matrix notation yields

$$\mathbf{U} = \Lambda^t \Phi',$$

where $\Lambda = \text{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_q\}$ is the diagonal matrix having selected eigenvalues of the transition matrix $\widetilde{\mathbf{K}}$; $\Phi = [\phi_1, \phi_2, \cdots, \phi_q]$ consists of the corresponding eigenvectors; $\cdot^t$ denotes t^{th} power; and $\cdot'$ is the matrix transpose. Therefore, it suffices to show that $\mathbf{U}$ always distributes as same as $\mathbf{U}\pi$ for any n and π.

Given that the graph $\mathbf{G}$ is infinitely exchangeable, i.e., $\mathbf{A}_{\sigma(i)\sigma(j)} \overset{d}{=} \mathbf{A}_{ij}$, we have

$$\begin{aligned}
\widetilde{\mathbf{K}}_{\sigma(i)\sigma(j)} &= \max\{\mathbf{A}_{\sigma(i)\sigma(j)}, \mathbf{A}_{\sigma(j)\sigma(i)}\} / \sum_j \max\{\mathbf{A}_{\sigma(i)\sigma(j)}, \mathbf{A}_{\sigma(j)\sigma(i)}\} \\
&\overset{d}{=} \max\{\mathbf{A}_{ij}, \mathbf{A}_{ji}\} / \sum_j \max\{\mathbf{A}_{ij}, \mathbf{A}_{ji}\} \\
&= \widetilde{\mathbf{K}}_{ij}.
\end{aligned}$$

Thus the transition matrix $\widetilde{\mathbf{K}}$ is also infinitely exchangeable, i.e., $\pi'\widetilde{\mathbf{K}}\pi \overset{d}{=} \widetilde{\mathbf{K}}$ for any permutation π and any size n.

By eigen-decomposition, the first q eigenvectors and the corresponding eigenvalues of $\pi'\widetilde{\mathbf{K}}\pi$ are Λ and $\pi'\Phi$, so it follows that

$$\begin{aligned} & \Phi \overset{d}{=} \pi'\Phi \\ \Leftrightarrow \quad & \mathbf{U} = \Lambda^t\Phi' \overset{d}{=} \Lambda^t\Phi'\pi = \mathbf{U}\pi \end{aligned}$$

Therefore, the diffusion maps are also infinitely exchangeable so that there exists a latent variable u such that the diffusion map is conditional i.i.d. asymptotically. □

Proof of Lemma 1. To prove that u is of finite second moment, it suffices to show that $\|\mathbf{u}_i\|_2$ is always bounded for all $i \in [1, n]$.

By Equation 1, we have

$$\begin{aligned} \|\mathbf{u}_i\|_2^2 &= \sum_{j=1}^{q} \lambda_j^{2t}\phi_j^2(i) \\ &\leq \sum_{j=1}^{q} \lambda_j^{2t} \\ &\leq q, \end{aligned}$$

where the second line follows by noting $\phi_j(i) \in [-1, 1]$ (the eigenvector ϕ_j is always of unit norm), and the third line follows by observing that $|\lambda_j| \leq \|\widetilde{\mathbf{K}}\|_\infty = 1$.

Therefore, all of $\mathbf{u}_i$ are bounded in ℓ_2 norm as $n \to \infty$, so the latent variable u must be of finite second moment. □

Proof of Theorem 2. We first state two useful lemmas:

Lemma 2. Let $\mathcal{V}_n^2(\mathbf{U},\mathbf{X})$ be the distance covariance ($\texttt{dCov}$) of $(\mathbf{U},\mathbf{X}) = \{(\mathbf{u}_i,\mathbf{x}_i) : i = 1,\ldots,n\}$ defined as Equation 3. It follows that

$$\mathcal{V}_n^2(\mathbf{U},\mathbf{X}) \ = \|g^n_{\mathbf{U},\mathbf{X}}(t,s) - g^n_{\mathbf{U}}(t)g^n_{\mathbf{X}}(s)\|^2, \tag{12}$$

where $g^n_{\cdot}$ is the empirical characteristic function of $\{(\mathbf{u}_i,\mathbf{x}_i) : i = 1,2,...,n\}$ or the marginals, e.g., $g^n_{\mathbf{U},\mathbf{X}}(t,s) = \frac{1}{n}\sum_{i=1}^n \exp\{i\,\langle t,\mathbf{u}_i\rangle + i\,\langle s,\mathbf{x}_i\rangle\}$, with the understanding that the i ahead of the inner product denotes the imaginary unit.

Lemma 3. Assume $\{\mathbf{u}_i : i = 1,2,\ldots,n\}$ are conditional i.i.d. based on u, and $\{\mathbf{x}_i \overset{i.i.d.}{\sim} x,\ i = 1,2,\ldots,n\}$, where u and x are both of finite second moment. It follows that

$$\mathcal{V}_n^2(\mathbf{U},\mathbf{X}) \ \longrightarrow \mathcal{V}^2(u,x) \qquad \text{as } n \to \infty \tag{13}$$

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

By Theorem 1, each observation in the diffusion map $\mathbf{U}$ is asymptotic i.i.d. conditioned on u under an infinitely exchangeable graph, while the nodal attributes $\mathbf{X}$ are assumed i.i.d. as x. The finite moment of u is guaranteed by Lemma 1, and the finite moment of x is assumed in (C2).

Therefore a direct application of Lemma 2 and Lemma 3 yields that

$$\texttt{dCov}(\mathbf{U},\mathbf{X}) \ \longrightarrow \|g_{u,x}(t,s) - g_u(t)g_x(s)\|^2, \tag{14}$$

which equals 0 if and only if u is independent of x.

As distance correlation is just a normalized version of distance covariance, it further leads to

$$\texttt{dCorr}(\mathbf{U},\mathbf{X}) \ \longrightarrow 0, \tag{15}$$

if and only if independence.

By Shen et al. (2017), whenever Equation 15 holds for `dCorr`, it automatically holds for `MGC`. Therefore, `MGC` is consistent for testing dependence between the diffusion maps **U** and the nodal attributes **X**. □

Proof of Lemma 2. This lemma follows exactly from Theorem 1 in Székely et al. (2007), which holds without any assumption on $(\mathbf{U},\mathbf{X}) = \{(\mathbf{u}_i,\mathbf{x}_i) : i = 1, 2, ..., n\}$, e.g., it holds without assuming exchangeability, nor identically distributed, nor finite second moment. □

Proof of Lemma 3. The conclusion of this lemma is equivalent to Theorem 2 in Székely et al. (2007). The only difference is that the original set-up assumes $\{(\mathbf{u}_i,\mathbf{x}_i) : i = 1,\ldots,n\}$ is independently identically distributed as (u, x) with finite second moment; while here $\{\mathbf{u}_i : i = 1,\ldots,n\}$ is asymptotic conditionally i.i.d. and of finite second moment.

All major steps in proving Theorem 2 of Székely et al. (2007) are essentially the same for either i.i.d. random variables or infinitely exchangeable random variables (Inoue and Taylor, 2006). In particular, under the finite second moment assumption, we still have the large number for V-statistics (Koroljuk and Borovskich, 1994), i.e.,

$$\int_{D(\delta)} \|g^n_{u,x}(t,s) - g^n_u(t)g^n_x(s)\|^2 dh \ \overset{n\to\infty}{\longrightarrow} \int_{D(\delta)} \|g_{u,x}(t,s) - g_u(t)g_x(s)\|^2 dh, \tag{16}$$

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

Proof of Corollary 1. (1) All the proofs remain the same regardless of the choice of t and q, as long as they are finite positive integers.

(2) When the nodal attribute is replaced by a second graph with the same node set as the first graph, we simply compute the diffusion map of the second graph by the same procedure as for the first graph. The resulting diffusion map of the second graph clearly satisfies both Theorem 1 and Lemma 1, and the proof of Theorem 2 works for two asymptotic conditional i.i.d. latent variables without any change.

(3) Changing the test statistic only affects Theorem 2, for which Equation 15 is the key step. As `mCorr` converges to `dCorr` by Székely and Rizzo (2013b), Equation 15 also holds for `mCorr`. It has been further shown in Shen et al. (2017) that `MGC` satisfies the same equation whenever it holds for the respective global correlation. □