# 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 As the network topology is structured and often high-dimensional, many traditional nonparametric tests are no longer applicable and instead parametric approaches are dominant in network inferences. Here 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.

*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. 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.

Mathematically, a graph (or equivalently a network) $\mathbf{G} = (V, E)$ consists of a set $V$ of nodes (or vertices) together with a set $E$ of edges, which is often represented via an adjacency matrix $\mathbf{A} = \{A_{ij} : i, j = 1, .., n = |V|\}$, 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 for directly applying conventional statistical methods. Therefore, graph inferences have long relied on specifying a particular statistical model, 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 $q$ when assuming a latent position model. 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].

When it comes to investigating the relationships among network data, a core problem is to detect dependency between network topology and nodal attributes, i.e., certain properties 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 functionality, and is connected with other regions in the brain map. Identifying dependency between network and nodal attributes has also 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 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 and overcomes all existing restraints on network analysis.

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 $\pm 1$ 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), and the multiscale generalized correlation statistic 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, via defining a family of distance-based local correlations and efficiently searching the optimal correlation in testing. Since all 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.

To overcome the theoretical barricades by the distinct structure of network data, and to relax the limitations of model-based method for network testing, we come up with a solution that is theoretically sound and numerically superior: we first define a family of distance metrics on network data via the diffusion maps, then employ `MGC` to compute the optimal local correlation between the diffusion distance of the network topology and the Euclidean distance of the nodal attributes. Theoretical results show that the diffusion maps, acting as a node-wise representation, can allow distance-based correlation measures to be consistent in testing network dependencies under very mild condition, which includes almost all existing generative graph models, and regardless of the connected-ness, weighted-ness, and directed-ness of the graph. Moreover, the `MGC` statistic offers major power improvement under various scenarios in finite-sample testing. The combined advantages of diffusion maps and `MGC` over the existing benchmarks are illustrated via comprehensive simulations under popular network models.

# 2 Results

## 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], and show that they can yield node-wise conditional *i.i.d.* samples for an exchangeable graph as sample size increases to infinity.

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. We are going to define such transition matrix $\mathbf{P}$ as $P_{ij} = A_{ij} / \sum_{j=1}^{n} A_{ij}, i, j = 1, \ldots, n$, but that it can be based on any reasonable kernels that represent the similarity between the node while satisfying the assumptions mentioned in [26]. The diffusion maps at time $t$ are computed as follows :

$$\begin{aligned} \mathbf{U}_t &= [\mathbf{u}_t(1), \ldots, \mathbf{u}_t(n)] \\ &= \left( \lambda_1^t \phi_1(i), \lambda_2^t \phi_2(i), \cdots, \lambda_q^t \phi_q(i) \right)^T \in \mathbb{R}^{n \times q}. \end{aligned}$$

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, $\lambda_j^t$ is the $t^{\text{th}}$ power of the eigenvalue, and $\cdot^T$ is the matrix transpose. Then diffusion maps locate each node's position at every diffusion time and provide node-wise multivariate coordinates through $\{\mathbf{u}_t(i)\}$.

A graph $\mathbf{G}$ is called exchangeable if and only if its adjacency matrix $\mathbf{A}$ is jointly exchangeable [28], i.e., for every permutation $\sigma$ 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, 29]. Lemma 2.1 proves that the node-wise multivariate coordinates $\{\mathbf{U}_t\}_{t \in \mathbb{N}}$ can furnish conditional *i.i.d.* samples for nodes by an exchangeable graph, with the proof supplied in the Appendix.

**Lemma 2.1** (Conditional *i.i.d.* of diffusion maps $\{\mathbf{u}_t(i)\}$)**.** Assume that $\mathbf{G}$ is an exchangeable random graph. Then as $n \to \infty$, the diffusion maps $\{\mathbf{u}_t(i) : i = 1, \ldots, n\}$ are conditionally *i.i.d.* given its underlying distribution.

The *diffusion distance* between each pair of nodes is then computed as the Euclidean distance of the diffusion maps.

$$C_t^2[i,j] := \| \mathbf{u}_t(i) - \mathbf{u}_t(j) \| \quad i,j = 1,2,\ldots,n. \tag{1}$$

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. Figure 1 shows how well diffusion distance exhibits the community structure in a graph (generated by the stochastic block model by Equation 3), 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, diffusion distance better reflects the connectivity since it takes into account every possible path between the two nodes.

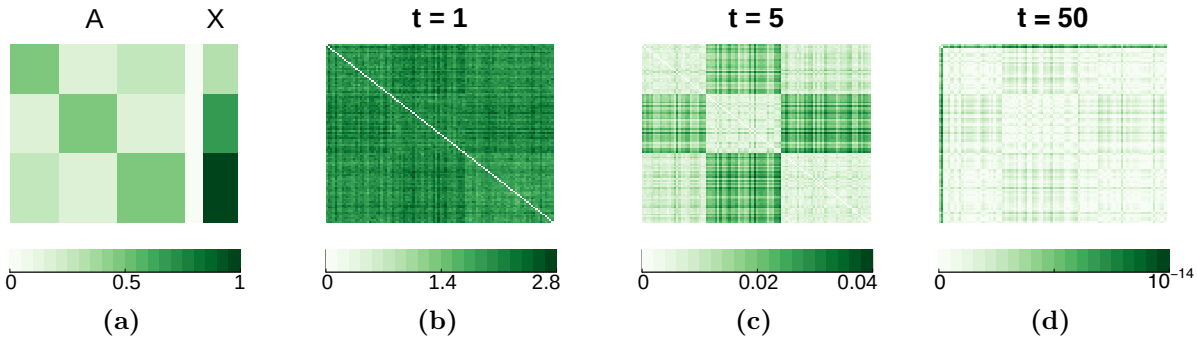


**Figure 1:** Panel (a) shows the adjacency matrix $\mathbf{A}$ and nodal attributes $\mathbf{X}$ Equation 3. Panel (b), (c), (d) shows the diffusion distances of the graph, as a proposed network metric to provide a one-parameter family of network-based distances. As $t$ increases, there is a slight change in pattern, and the diffusion distance at $t = 5$ illustrates a very distinct block structures and thus a very clear dependency to the attributes $\mathbf{X}$.

Note that the diffusion distance can always be defined regardless whether the graph is connected or not, weighted or not, and directed or not. Although the parameter $t$ may seem like another model parameter to tune, in practice $t \in [3, 10]$ usually yields similar inference results. Moreover, when combined with the `MGC` statistic later, the selection of $t$ only has a very small inference effect in testing power (due to the capability of `MGC` to locate the optimal local correlation). Therefore, throughout the paper we always take $t = 5$ in 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 $(\mathbf{U}, \mathbf{X}) = \{(\mathbf{u}_i, \mathbf{x}_i); i = 1, 2, \ldots, n\}$ that are 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 $(\mathbf{u}_i, \mathbf{x}_i)$ 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

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

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 $n$ large, for any joint distributions of finite moment. 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 correlations $c_{k,l}$ 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. Then we locate the optimal local correlation, i.e., optimal choice of included neighborhood $(k^*, l^*)$, which yields MGC statistic.

However, as the *i.i.d.* assumption is not satisfied under network topology, the consistency of distance correlation is no longer guaranteed when applied to the arbitrary distance metric of the graph. In particular, neither the Euclidean distance of the adjacency vector nor the shortest-path

distance can work together with distance correlation without breaking its consistency proof.

Using Lemma 2.1, 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 2.2** (`MGC` Consistency via Diffusion Distance)**.** Assume that $\mathbf{G}$ is an exchangeable random graph and its diffusion maps are $\mathbf{U}$ at certain $t \in \mathbf{N}$ with finite moment; and the nodal attributes $\mathbf{X} = \{\mathbf{x}_i, i = 1, 2, \ldots, n\}$ is *i.i.d.* as a random vector $\mathbf{x}$ of finite moment.

Then $\texttt{dCorr}(\mathbf{U}, \mathbf{X}) \longrightarrow 0$ 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. The proof of this theorem is postponed to the Appendix 5.1.

## 3 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`, `dCorr`, `HHG`, and the likelihood ratio test 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 all latent factors based methods require 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 0.05. The testing power of each method equals the percentage of correct rejection. We will mainly

consider the stochastic block model (SBM) and its degree-corrected version for the simulation models, which are two major models used in community detection.

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 class label $\mathbf{x}_i$ takes values in $0, 1, 2$ equally likely. The edge probability is designed as

$$E(A_{ij}|X_i, X_j) = 0.5I(|X_i - X_j| = 0) + 0.2I(|X_i - X_j| = 1) + 0.3I(|X_i - X_j| = 2), \quad i, j = 1, \ldots, n = 100. \tag{3}$$

Namely, within-block edge probability is 0.5, between-block edge probability is 0.2 or 0.3 depending on the communities. This 3-block model describes a nonlinear dependency, where `MGC` has been shown to work better than the `dCorr` 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 one sample graph is offered in Figure 1. After repeatedly data generation and hypothesis testing by all methods, the powers are computed and shown in Figure 2, for which `MGC` combined with diffusion maps indeed yields the most superior power comparing to all other benchmarks.

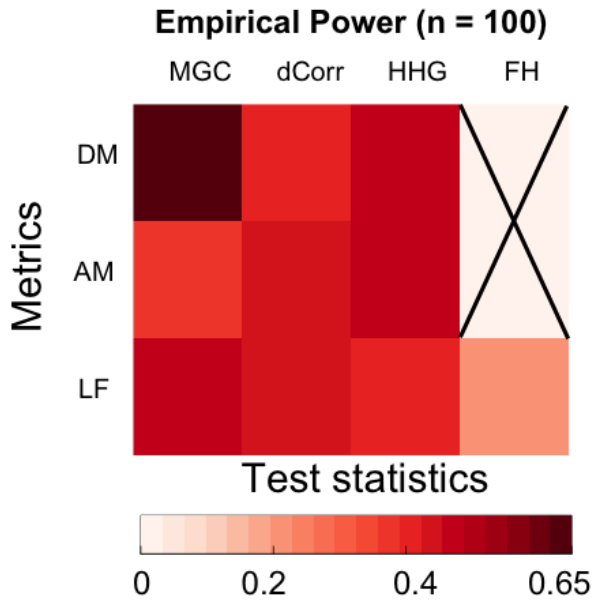


**Figure 2:** The power heatmap under the SBM with three blocks, for all possible combinations of test statistics with distance metrics. `MGC` with the diffusion maps yields the best power comparing to all other methods.

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

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

$$E(A_{ij}|X_i, X_j) = 0.5I(|X_i - X_j| = 0) + 0.2I(|X_i - X_j| = 1) + \theta I(|X_i - X_j| = 2), \quad i, j = 1, \ldots, n = 100. \tag{4}$$

When $\theta > 0.2$, the network dependency changes from a close to linear relationship to strongly nonlinear. Figure 3 shows the testing power with respect to increasing $\theta$, and there is a clear trend that both the `dCorr` and `FH` tests have deteriorating power while `MGC` has a very stable performance against varying $\theta$. The same phenomenon holds by varying other edge probabilities. Therefore, `MGC` can better capture the nonlinear dependencies for network dependence testing, and is the best method to couple with the diffusion distance.

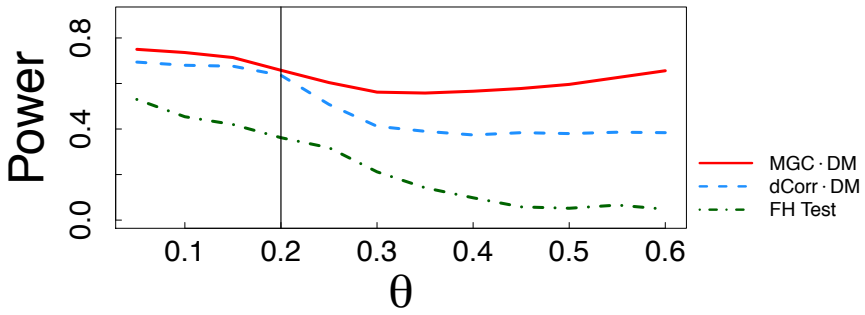


**Figure 3:** The power curve with respect to increasing $\theta$ in the SBM with three blocks, for `MGC`, `dCorr`, and `FH`. Larger $\theta$ implies stronger nonlinear dependency, while $\theta < 0.2$ has close-to-linear dependency. `MGC` is the best performing method throughout all possible $\theta$.

Our next simulation shifts to the degree-corrected stochastic block model (DCSBM) with two blocks. DCSBM 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 / class label $X_i$ takes values in 0 and 1 equally likely, and the edge probabilities are specified by

$$E(A_{ij}|\mathbf{X}, \mathbf{V}) = 0.2V_iV_j \cdot I(|X_i - X_j| = 0) + 0.05V_iV_j \cdot I(|X_i - X_j| = 1), \tag{5}$$

where $V_i \overset{i.i.d}{\sim} Uniform(1 - \tau, 1 + \tau)$ for $i = 1, \ldots, n$, and $\tau$ is a parameter to control the amount of variability of the edge distribution. Again, `MGC` $\circ$ `DM` is the best method in power throughout $\tau$; and in Figure 4 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.

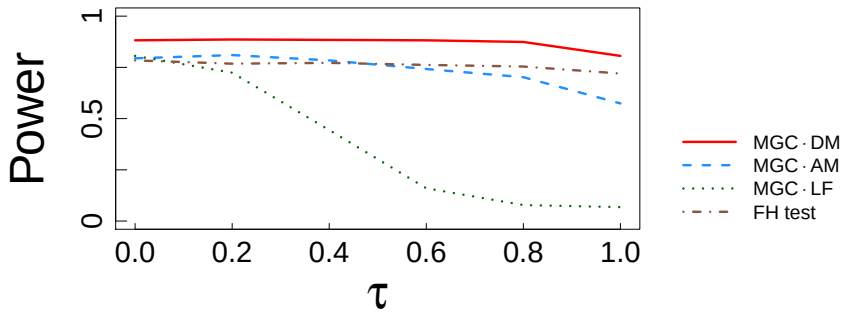


**Figure 4:** The power curve with respect to increasing $\tau$ in DCSBM with two blocks, for `MGC` with different distance metrics and `FH`. The diffusion distance exhibits the best testing power comparing to other metrics, and is better than the `FH` test.

## 4 Conclusion And Future Work

In this paper, we combined recent progress in dependency testing and metric learning into the graph domain, and 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 it has superior power over all benchmarks, with `MGC` and the diffusion distance being the core elements behind the success.

There are a number of additional potential extensions of this work. First, how to choose a better diffusion time $t$, or find a $t$ with provable finite-sample performance, may provide further insight into and establish a more solid foundation of this approach. Second, 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. Third, 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] 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.

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

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

[31] V. S. Koroljuk and Yu. V Borovskich. *Theory of U-statistics*. Springer Science+Business Media Dordrecht, 1994.

# 5 Appendix

## 5.1 Proofs

***Proof of Lemma 2.1.*** To prove conditional *i.i.d.* of $\{\mathbf{u}(i)\}$ for $i = 1, \ldots, n$ as $n \to \infty$, by the celebrated *de Finetti's Theorem* [30], it suffices to prove that $\mathbf{u}(i)$ for $i = 1, \ldots, n$ are exchangeable, i.e., for any permutation $\sigma$, the permuted sequence $\mathbf{u}(\sigma(1)), \mathbf{u}(\sigma(2)), \ldots, \mathbf{u}(\sigma(n))$ distributes the

same as the original sequence $\mathbf{u}(1), \mathbf{u}(2), \ldots, \mathbf{u}(n)$. Denote the permutation matrix as $\pi$, it is to show $\mathbf{U}$ always distributes the same as $\mathbf{U}\pi^T$ in matrix notation.

Recall that the diffusion map at time $t$ is represented as follows :

$$\begin{aligned}\mathbf{U} &= [\mathbf{u}(1), \ldots, \mathbf{u}(n)] \\ &= \Lambda\Phi^T \\ &= \begin{pmatrix} \lambda_1^t\phi_1(i) & \lambda_2^t\phi_2(i) & \cdots & \lambda_q^t\phi_q(i) \end{pmatrix}^T \in \mathbb{R}^{n\times q},\end{aligned}$$

where $\Lambda = diag\{\lambda_1, \lambda_2, \ldots, \lambda_q\}$ and $\Phi = [\phi_1, \phi_2, \cdots, \phi_q]$ are the diagonal matrix of non-zero eigenvalues and the corresponding matrix of eigenvectors of the transition matrix $\mathbf{P}$, i.e, $\mathbf{P} = \Phi\Lambda\Phi^T$.

Given the graph $G$ 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

$$\begin{aligned}\pi\mathbf{P}\pi^T &\overset{d}{=} \mathbf{P} \\ \Rightarrow (\pi\Phi)\Lambda(\pi\Phi)^T &\overset{d}{=} \Phi\Lambda\Phi^T \\ \Rightarrow \mathbf{U}\pi^T = \Lambda(\pi\Phi)^T &\overset{d}{=} \Lambda\Phi^T = \mathbf{U}\end{aligned}$$

Therefore, the diffusion maps are exchangeable, and also conditional *i.i.d.* asymptotically. □

***Proof of Theorem 2.2*** *MGC Consistency via Diffusion Distance.* Assume that the pairs of observations $(\mathbf{U}, \mathbf{X}) = \{(\mathbf{u}_i, \mathbf{x}_i); i = 1, 2, \ldots, n\}$ is identically distributed as $(\mathbf{u}, \mathbf{x})$. By *Theorem 1* in [19], we immediately have

$$\texttt{dCov}(\mathbf{U}, \mathbf{X}) \longrightarrow \int_{D(\delta)} \|g^n_{\mathbf{u},\mathbf{x}}(t,s) - g^n_{\mathbf{u}}(t)g^n_{\mathbf{x}}(s)\|^2 dw \qquad \text{as } n \to \infty \tag{6}$$

where $g^n_{\mathbf{u},\mathbf{x}}(t,s), g^n_{\mathbf{u}}(t), g^n_{\mathbf{x}}(s)$ are the sample characteristic functions, e.g., $g^n_{\mathbf{u},\mathbf{x}}(t,s) = \frac{1}{n}\sum_{j=1}^n \exp\{i\langle t,\mathbf{u}_j\rangle + i\langle s,\mathbf{x}_j\rangle\}$, and $w(t,s)$ is the weight function chosen in [19].

Lemma 2.1 shows that $\{\mathbf{u}_i\}$ 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 $n \to \infty$. By the strong law of large number for V-statistics [31], under finite moment assumption of $\mathbf{u}$ and $\mathbf{x}$, we have

$$\int_{D(\delta)} \|g^n_{\mathbf{u},\mathbf{x}}(t,s) - g^n_{\mathbf{u}}(t)g^n_{\mathbf{x}}(s)\|^2 dw \overset{n\to\infty}{\longrightarrow} \int_{D(\delta)} \|g_{\mathbf{u},\mathbf{x}}(t,s) - g_{\mathbf{u}}(t)g_{\mathbf{x}}(s)\|^2 dw, \tag{7}$$

within a bounded circle $D(\delta) = \{(t,s) : \delta \le |t|_p \le 1/\delta, \delta \le |s|_q \le 1/\delta\}$ for any $\delta > 0$, where $g_\cdot(\cdot)$ is the population characteristic function, i.e., $g_{\mathbf{u},\mathbf{x}}(t,s) = E\{\exp\{i\langle t,\mathbf{u}\rangle + i\langle s,\mathbf{x}\rangle\}\}$. The same convergence still holds out of the circle $D(\delta)$, and a technical proof is in *Theorem 2* of [19].

Combining equation 6 and equation 7 yields

$$\texttt{dCov}(\mathbf{U},\mathbf{X}) \longrightarrow \int \|g_{\mathbf{u},\mathbf{x}}(t,s) - g_{\mathbf{u}}(t)g_{\mathbf{x}}(s)\|^2 dw, \tag{8}$$

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

$$\texttt{dCorr}(\mathbf{U},\mathbf{X}) \longrightarrow 0, \tag{9}$$

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

By [25], Equation 9 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 $\mathbf{X}$. □