Nonparametric Network Dependence Testing By Diffusion Maps

Abstract

Investigating how network structures are associated with nodal attributes of interest is a core problem in network science. As the network topology is a structured and often high-dimensional object, many traditional nonparametric tests are no longer applicable and parametric approaches are dominant in graph inferences. Here we propose a new procedure to testing dependence between graphs and attributes, via diffusion distances and distance-based correlation testing. We demonstrate that our nonparametric method not only yields a consistent test statistic 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 $\mathbf{A} = \{A_{ij} : i, j = 1, .., n\}$, 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. Therefore, $\mathbf{A}$ is a symmetric square matrix that 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., which is a notable obstacle for directly applying conventional statistical methods. Therefore, graph inferences have long relied on specifying a particular statistical model, such as the Erdos-Renyi model [2, 4], stochastic block model [8, 12, 18, 20] and its degree-corrected version [10, 28], the latent position model [3, 24], the random dot product model [21, 27], etc.

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 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 [3, 9, 26]. For example, the test by Fosdick and Hoff [3] estimates the latent factor of each node, then proceeds to test network dependence on the covariance via assuming a multivariate normal distribution of the latent factors. To our best knowledge so far, there is no principled method to compute a model-free correlation measure for testing network dependency.

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 [15] is the most classical approach, which determines the existence of linear relationship via a correlation coefficient in the range of $[-1, 1]$, with 0 indicating no linear association while ± 1 indicating perfect linear association. To capture all types of dependencies not limited to linear relationship, new correlation measures and nonparametric statistics have been suggested recently, such as the Mantel coefficient [13], RV coefficient [17], distance correlation and energy statistic [16, 22, 23], kernel-based independence test [5], Heller-Heller-Gorfine (`HHG`) test [6, 7], and multiscale generalized correlation (`MGC`) [19]. In particular, the distance correlation by Szekely et al. [23] 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. [19]

inherits the same consistency of distance correlation with remarkable 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 are non-parametric and do not depend on particular models, 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 under very mild condition, the diffusion maps, acting as a node-specific random vector, can allow distance-based correlation measures to be consistent in testing network dependencies. 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 of a graph [1], and show that they can yield node-wise conditional *i.i.d.* samples for an exchangeable graph.

Coifman and Lafon [1, 11] proposed multiscale geometries of data called diffusion maps, which is constructed by iterating the transition matrix that determines random walk on graph. Given the $n \times n$ adjacency matrix $\mathbf{A}$, the $n \times n$ transition matrix of $\mathbf{P}$ is defined by $P_{ij} = A_{ij} / \sum_{j=1}^{n} A_{ij}$, indicating the probability of moving forward from node i to node j for

$i, j = 1, \ldots, n$. The diffusion maps at time t are computed as follows :

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

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, and λ_j^t is the tth power of the eigenvalue. Then diffusion maps locate each node's position at every diffusion time and provide node-specific 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 [14], i.e. for every permutation σ of n, $(A_{ij}) \stackrel{d}{=} (A_{\sigma(i)\sigma(j)})$. 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 [18, 21, 25]. Lemma 2.1 proves that the node-specific multivariate coordinates $\{\mathbf{U}_t\}_{t\in\mathbb{N}}$ can furnish conditional *i.i.d* samples for vertices by an exchangeable graph, with a short 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 that is connected and unweighted. Then 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 diffusion distance can better reflects the connectivity and exhibits the community

structure in a graph, 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. In practice, $t \in [3, 10]$ usually yields similar inference results, and we take $t = 5$ in the simulations.

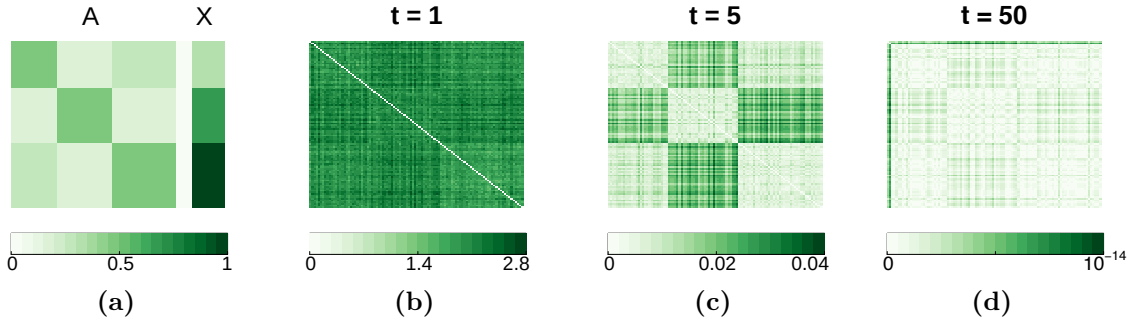

Figure 1: Panel (a) shows data generating probability of an adjacency matrix $\mathbf{A}$ and nodal attributes $\mathbf{X}$. Diffusion distances, as a proposed network metric, provides one-parameter family of network-based distances where as time goes by the pattern shown in the distance matrix changes, and at time point $t = 5$, distance distance in panel (c) illustrates most clear block structures and the most distinct dependency to the attributes $\mathbf{X}$.

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 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 locating the optimal local correlation, i.e., excluding far away distances in the computation of distance correlation.

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 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 the both `dCorr` and `MGC` defined on the diffusion distance can have the same consistency when extended to network dependency test of exchangeable graphs. This offers a principal approach to define correlations and testing dependency on network data.

Theorem 2.2 (`MGC` Consistency via Diffusion Distance)**.** Assume that $\mathbf{G}$ is an exchangeable random graph that is connected and unweighted; 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 for any t, $\texttt{dCorr}(\mathbf{U}_t, \mathbf{X}) \longrightarrow 0 \quad$ as $n \to \infty$ if and only if $\mathbf{U}$ is independent of $\mathbf{X}$. Therefore, both `dCorr` and `MGC` are consistent for testing dependence between $\mathbf{U}$ and $\mathbf{X}$.

The proof is postponed to the Appendix.

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`) [3]. For the first three statistics, we further consider three different metrics of the network topology: the Euclidean distances of the diffusion maps at $t = 5$ (`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.

For each simulation model and each test, we repeatedly generate sample data 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 and its degree-corrected version for the simulation models, which are often used in community detection.

Let us first consider SBM with 3 blocks, i.e., partition the vertices into 3 communities, and generate the edges by a Bernoulli random variable whose probability that is determined by the communities of the connecting vertices. Assume $n = 100$ vertices 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 between the network topology and the class labels, where `MGC` should work the best when coupled with a proper metric. Indeed, Figure 2 illustrates that `MGC` combined with diffusion maps yields the most superior power comparing to all other benchmarks.

To further understand the advantage of `MGC`, next we fix the distance as the diffusion distance, and control the amount of *nonlinear dependency* through changing the value of

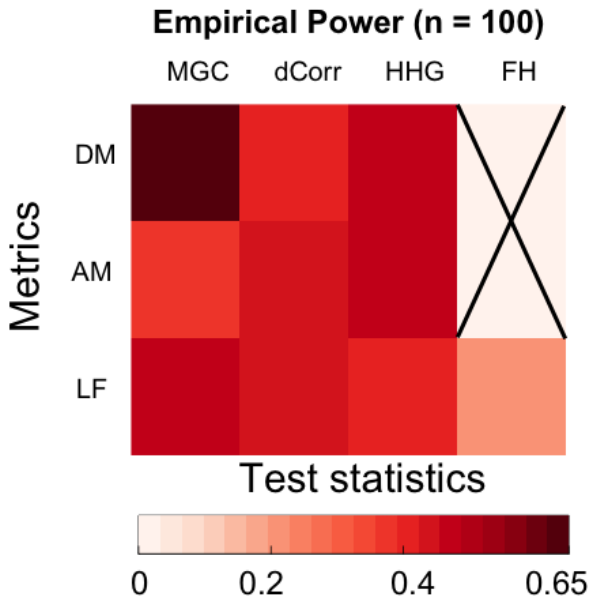

Figure 2: The power heatmap in SBM with three blocks, for all possible combinations of test statistics with distance metrics. `MGC` with the diffusion maps yields the best power comparing to all other methods.

$\theta \in (0, 1)$ in the edge probability:

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

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 θ, and there is a clear trend that both the `dcorr` and `FH` tests have deteriorating power while `MGC` is very stable against varying θ. The same phenomenon holds by varying other edge probabilities. Therefore, the `MGC` capability to better capture the nonlinear dependencies shown in [19] carries over to network dependence testing.

Our next simulation is based on the degree-corrected stochastic block model (DCSBM) with two blocks. This time we fix the test statistic to `MGC`, but varying the distance metrics. DC-SBM adds another random variable V_i associated with each node to vary the node degrees, which is a generalization of the stochastic block model and provides a better fit to real networks. Let $n = 250$, suppose the nodal attributes / class label X_i takes values in 0 and 1 equally likely, the edge probabilities are specified by

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

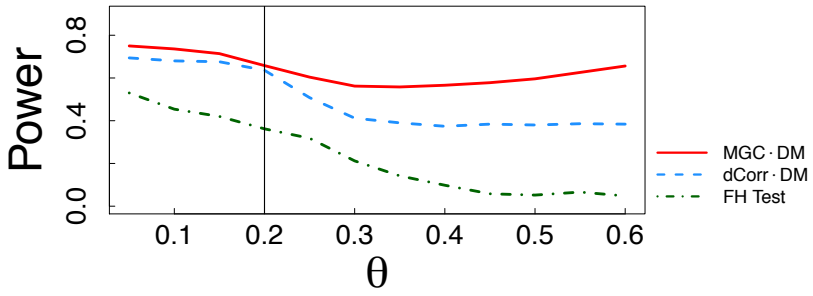

Figure 3: The power curve with respect to increasing θ in SBM with three blocks, for `MGC`, `dcorr`, and `FH`. Larger θ implies stronger nonlinear dependency, while $\theta < 0.2$ has close-to-linear dependency. `MGC` is the best performing method throughout all possible θ.

where $V_i \overset{i.i.d}{\sim} Uniform(1-\tau, 1+\tau)$ for $i = 1, \ldots, n$, and τ is a parameter to control the amount of variability of the edge distribution. Figure 4 shows the testing power based on the above model with respect to different metrics, and the diffusion distance is clearly the most superior one regardless of τ.

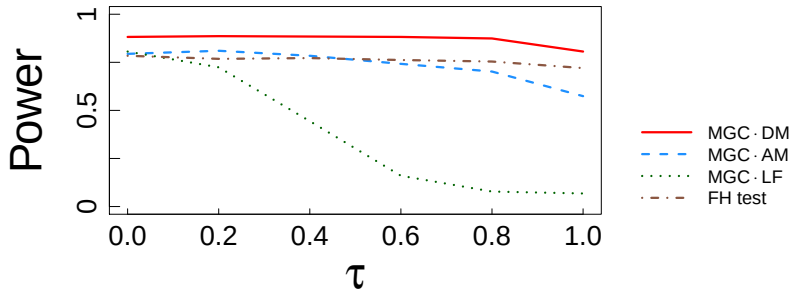

Figure 4: The power curve with respect to increasing τ in DCSBM with two blocks, for `MGC` with different distance metrics and `FH`. The diffusion distance exhibits the best testing power comparing to other metrics, and is better than the `FH` test.

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 offer an elegant and powerful solution to the network dependency problem. We proved that this method is consistent under most popular graph models; and empirically demonstrated its superior power comparing to using other distance metrics, or other correlation measures and tests.

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 good 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.

References

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

[2] P Erds and A Rnyi. On random graphs. i. *Publicationes Mathematicae*, 6:290–297, 1959.

[3] 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.

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

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

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

[7] 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.

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

[9] 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.

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

[11] 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.

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

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

[14] 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.

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

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

[17] 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.

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

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

[20] 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.

[21] 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.

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

[23] 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.

[24] 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.

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

[26] 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.

[27] 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.

[28] 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.

5 Appendix

5.1 Lemmas and Theorems

Proof of Lemma 2.1. Diffusion map at time t is represented as follows :

$$\mathbf{U}_t(i) = \begin{pmatrix} \lambda_1^t\phi_1(i) & \lambda_2^t\phi_2(i) & \cdots & \lambda_q^t\phi_q(i) \end{pmatrix} \in \mathbb{R}^q. \tag{6}$$

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

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

$$\left(P_{ij}\right) = \left(\frac{A_{ij}}{1 - A_{ij} + \sum_{j=1}^{n} A_{ij}}\right) \overset{d}{=} \left(\frac{A_{\sigma(i)\sigma(j)}}{1 - A_{\sigma(i)\sigma(j)} + \sum_{\sigma(j)=1}^{n} A_{\sigma(i)\sigma(j)}}\right) = \left(P_{\sigma(i)\sigma(j)}\right) \tag{7}$$

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

Proof of Theorem 2.2 *Consistency of* `dCorr` *applied to exchangeable variables.* For exchangeable sequence of $(\mathbf{U}, \mathbf{X}) = \{(\mathbf{u}_i, \mathbf{x}_i); i = 1, 2, \ldots, n\}$ which is identically distributed as

$(\mathbf{u}, \mathbf{x})$ with finite second moment, we have

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

where $\mathcal{V}^2(\mathbf{u}, \mathbf{x}) := \|g_{\mathbf{u},\mathbf{x}}(t,s) - g_{\mathbf{u}}(t)g_{\mathbf{x}}(s)\|^2$, and $g.$ is a characteristic function, e.g., $g_{\mathbf{u},\mathbf{x}}(t,s) = E\{\exp\{i\langle t, \mathbf{u}\rangle + i\langle s, \mathbf{x}\rangle\}\}$. This follows exactly the same as *Theorem 1* in [23]. Note that this Lemma always holds without any assumption on $\{(\mathbf{u}_i, \mathbf{x}_i), i = 1, 2, ..., n\}$.

Followed by *de Finetti's Theorem*, if and only if $\{\mathbf{u}_i\}$ are (infinitely) exchangeable, there exists an underlying distribution $f_{\mathbf{u}}$ of $\mathbf{u}$ such that $\mathbf{u}_i \overset{i.i.d}{\sim} f_{\mathbf{u}}$. By the same logic there exists a random, we have an underlying distribution $f_{\mathbf{x}}$ where $\mathbf{x}_i \overset{i.i.d}{\sim} f_{\mathbf{x}}$. Let $(\mathbf{u}_i, \mathbf{x}_i) \overset{i.i.d}{\sim} f_{\mathbf{u},\mathbf{x}}$. Then under the assumption of finite second moment of the underlying distributions and measurable, conditioned random functions, we have a strong large number for V-statistics followed by [23], i.e.,

$$\int_{D(\delta)} \|g^n_{\mathbf{u},\mathbf{x}}(t,s) - g^n_{\mathbf{u}}(t)g^n_{\mathbf{x}}(s)\|^2 dh \quad \overset{n\to\infty}{\longrightarrow} \int_{D(\delta)} \|g_{\mathbf{u},\mathbf{x}}(t,s) - g_{\mathbf{u}}(t)g_{\mathbf{x}}(s)\|^2 dh, \tag{9}$$

where $D(\delta) = \{(t,s) : \delta \leq |t|_p \leq 1/\delta, \delta \leq |s|_q \leq 1/\delta\}$, and $h(t,s)$ is the weight function chosen in [23]. It follows that

$$\mathcal{V}_n^2(\mathbf{U}, \mathbf{X}) \quad \to 0 \qquad \text{as } n \to \infty \tag{10}$$

if and only if $g_{\mathbf{u},\mathbf{x}}(t,s) = g_{\mathbf{u}}(t)g_{\mathbf{x}}(s)$, i.e., $\mathbf{u}$ is independent of $\mathbf{x}$. Therefore, the `dCorr` or `mCorr` converges to 0 if and only if underlying distributions are independent; and its testing power converges to 1 under any joint distribution of finite moments. Since the multiscale generalized correlation based on any consistent global correlation is also consistent [19], MGC statistic constructed by `dCorr` or `mCorr` is also consistent in testing dependence. □