Testing independence in networks via family of network metrics

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 As the network topology is a structured and often high-dimensional object, many traditional nonparametric tests are no longer applicable and parametric approaches are dominant in graph inferences.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: network dependence, distance correlation, exchangeable graph, diffusion maps

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 question 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.

Statisticians have long considered the problem of revealing the relationship between two data sets. The most classical approach is the Pearson's correlation ([12]), which determine 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, many new correlation measures and nonparametric statistics have been suggested to test independence between two random vectors [3–5, 10, 13–17]. In particular, the distance correlation by Szekely et al. [17] is the first correlation measure that is consistent against all possible dependencies (with finite moments), and the multiscale generalized correlation (MGC) statistic by Shen et al. [15] inherits the same consistency of distance correlation with significantly 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 for testing.

Despite of many recent development, the network data, different from a random vector, still suffers from a dearth of proper analysis, due to its structured and high-dimensional nature. Mathematically, a graph $\mathbf{G} = (V, E)$ consists of a set V of nodes (or vertices) together with a set E of edges, which is often represented via an adjacency matrix $A = \{A_{ij} : i, j = 1, .., n\}$, 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, 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 huge obstacle for directly applying conventional statistical methods.

When it comes to investigating relationships in 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 having a number of different attributes, e.g., occupations, sex, personal behaviors, etc., are interacting each other via the social network; in neuro-science, each brain region has its own functionality, but is connected with other regions in the brain map. Identifying dependency between network and nodal attributes has primarily focused on their relationship explained only by network model under the boundary of model assumption [2, 7, 18]. Even though Fosdick and Hoff [2] suggested estimating node-specific network factors without restricting the dimensions, a fundamental difficulty of model-based independence tests still lingers from the fact that not all networks exhibit the structures described by known network models. To our best knowledge so far, there is no principled method to compute a model-free correlation measure for testing network dependency.

To overcome the obstacles due to the distinct structure of network data and also due to the limitations of model-based method, we first define a family of distance metrics on network data via the diffusion maps, then apply nonparametric testing method of MGC utilizing 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. We further illustrate the advantages of diffusion maps and MGC over the existing benchmarks via comprehensive simulations under popular network models.

2 Results

2.1 Diffusion Maps and Diffusion Distances

In this section, we show that the diffusion maps of an exchangeable graph [1] can furnish conditional i.i.d. samples.

A graph $\mathbf{G}$ is called exchangeable if and only if its adjacency matrix $\mathbf{A}$ is jointly exchangeable [11], i.e. for every permutation σ of n, $(A_{ij}) \stackrel{d}{=} (A_{\sigma(i)\sigma(j)})$. Most statistical network models satisfy this property, including the Erdos-Renyi model, latent position model, stochastic block model, random dot product model, etc. [6, 8, 19, 20]. cs: check if the above claim for models is right? And distribute /add citations behind each model accordingly?

The diffusion maps are proposed in [1], defined by eigenvectors of Markov matrix constructed over a connected network. cs: rephrase the remaining paragraph on diffusion maps! This paragraph should clearly define $\mathbf{U}_t(i)$, i.e. how the eigenvalues and vectors are derived from the adjacency matrix to transition matrix. In the process of constructing such Markov matrix, we basically run random walks by iterating transition matrix and diffusion maps accordingly locates each node's position at each iteration time. Each of diffusion distance, i.e. C_t, can also be represented using a discrete set of real nonzero eigenvalues $\{\lambda_r\}$ and eigenvectors $\{\phi_r\}$ of a transition matrix [1, 9].

Under jointly exchangeable graph, Lemma 2.1 proves that the diffusion maps at each t can provide conditional *i.i.d* samples $\{\mathbf{U}_t\}_{t\in\mathbb{N}}$.

Lemma 2.1 (Exchangeability and *i.i.d* of diffusion maps $\mathbf{U}_t$)**.** Assume that $\mathbf{G}$ is an exchangeable random graph that is connected, undirected and unweighted. Then the diffusion maps $\mathbf{U}_t(i)$ are conditionally *i.i.d* given its underlying distribution.

cs: are connected, undirected and unweighted graph really necessary? Also removed 'Then its transition probability so thus diffusion maps at fixed time t is also exchangeable, conditioned on underlying distribution of graph. Furthermore, by *de Finetti's Theorem*'. They should be in proof section than theorem statement. Also $= \begin{pmatrix} \lambda_1^t\phi_1(i) & \lambda_2^t\phi_2(i) & \dots & \lambda_q^t\phi_q(i) \end{pmatrix}^T$ should be in the previous paragraph when defining the diffusion maps.

Then we can define the *diffusion distance* between each pair of nodes, by computing 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 diffusion time t increases, the corresponding diffusion distance C_t is 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 distance $\{C_t : t \in \mathbb{N}\}$. In practice, $t \in [5, 20]$ usually yields similar diffusion distance and suffices for later inferences. Compared to adjacent relation or geodesic distance which are two extremes, diffusion distance well reflects the connectivity since it takes into account every possible path between the two nodes. cs: is the above sentence correct? Adjacent relation is equivalent to $t = 0$ and geodesic corresponds to $t = \infty$?

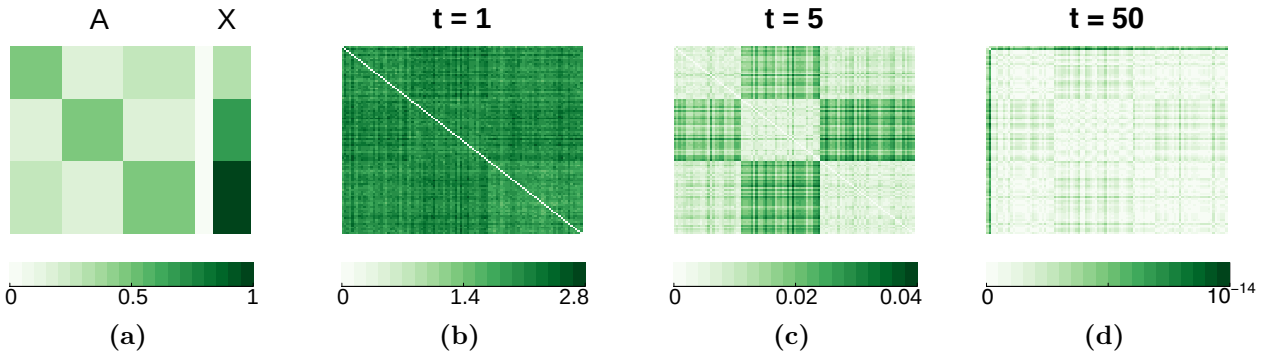

Figure 1: Figure (a) shows data generating probability of an adjacency matrix $\mathbf{A}$ and nodal attributes $\mathbf{X}$. Diffusion matrix, as a proposed network metric, provides one-parameter family of network-based distances where as time goes by the pattern shown in the distance matrix changes,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

2.2 Dependence Testing via MGC

cs: why use W? why not just X or U(corresponding to diffusion map)? The results in Section 2.1 allow us to cast the network dependency test into the following framework: given sample data $(\mathbf{W}, \mathbf{Y}) = \{(\mathbf{w}_i, \mathbf{y}_i); i = 1, 2, \ldots, n\}$ that are identically distributed as $(\mathbf{w}, \mathbf{y}) \in \mathbb{R}^{D \times D_y}$ (D and D_y are the respective feature dimension), we are looking to test whether their joint distribution equals the product of the marginals, i.e.,

$$H_0 : f_{\mathbf{wy}} = f_{\mathbf{w}} f_{\mathbf{y}},$$

$$H_A : f_{\mathbf{wy}} \neq f_{\mathbf{w}} f_{\mathbf{y}}.$$

If $(\mathbf{w}_i, \mathbf{y}_i)$ can be further assumed independently distributed for each i, a wide range of statistics are consistent for the above test, such as the distance correlation [17], HHG test [4], MGC statistic [15], etc. For example, the landmark distance correlation is computed as follows: denote $C_{ij} = \| \mathbf{w}_i - \mathbf{w}_j \|$ and $D_{ij} = \| \mathbf{y}_i - \mathbf{y}_j \|$ for $i, j = 1, 2, \ldots, n$, where $\| \cdot \|$ is the Euclidean distance. The distance covariance is defined as

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

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., $\texttt{dCorr}(\mathbf{W}, \mathbf{Y})$ 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. For example, neither the Euclidean distance of the adjacency vector nor the shortest-path distance can work together with distance correlation without breaking its consistency proof. Based on Lemma 2.1, we are able to prove that the both `dCorr` and MGC defined on the diffusion distance have the same consistency when extended to network dependency test.

Theorem 2.2. Assume that $\mathbf{G}$ is an exchangeable random graph that is connected, undirected and unweighted with n vertices with the diffusion maps $\mathbf{U}$ at certain t; and the nodal attributes $\mathbf{Y} = \{\mathbf{y}_i, i = 1, 2, \ldots, n\}$ is i.i.d. as a random vector $\mathbf{y}$ of finite moment.

Then $\texttt{dCorr}(\mathbf{U}, \mathbf{Y}) \longrightarrow 0$ as $n \to \infty$ if and only if $\mathbf{U}$ is independent of $\mathbf{Y}$. Then both `dCorr` and `MGC` are consistent for testing dependence between $\mathbf{U}$ and $\mathbf{Y}$.

cs: not sure if thm 2 is necessary or not, excluded for now.

Therefore, we successfully extend distance-based correlation measures to the graph domain, offering a principal approach to define correlations and testing dependency on network data. We will next investigate our approach via simulated models and empirical performances.

2.3 Measure for Node Contribution

On the other hand, some nodes often exert more reliance on their attributes than the others. Here we suggest the measure of node's contribution to detecting dependence as a byproduct of `MGC` statistic. Let (k^*, l^*) be the optimal neighborhood choice in distance matrix (C, D) respectively. Denote the contribution of node $v \in V(G)$ to the testing statistic by $c(\cdot) : v \to \mathbb{R}$

$$c(v) \propto \sum_{j=1}^{n} \tilde{C}_{jv}\tilde{D}_{jv} I\big(r(C_{jv}) \leq k^*\big) I\big(r(D_{jv}) \leq l^*\big), \tag{3}$$

which is proportional to v^{th} column-sum of the pre-summed test statistic **??**. Note that the deviation of non-negative `MGC` statistic from zero implies departure from the independence and also note that we truncate the correlation in `dCov` by column entry's rank. Thus $\tilde{C}_{jv}\tilde{D}_{jv}$ would not be truncated if node j ($\in \{1, 2, \ldots, n\} \setminus \{v\}$) is important to node v and its larger, positive value would contribute to $\mathcal{V}_n^{*2}$ more. The statistic $c(v)$ comes out from these observations. cs: change attributes notation from Y to X, in accordance with simulation notations.

3 Simulation Study

In our simulation studies, we compare the empirical testing powers of `MGC`, `mCorr`, Heller-Heller-Gorfine (`HHG`), and likelihood ratio test of Fosdick and Hoff (`FH`) [2]. We use type I error of $\alpha = 0.05$ and obtain p-values from each simulated network via the permutation test. The simulation models are shown by joint distribution of adjacent matrix $\mathbf{A}$ and nodal attributes $\mathbf{X}$.

3.1 Stochastic Block Model

We first consider a SBM with 3 blocks (model 4) where block affiliation for each node is correlated with its attributes X. Assume that for $i, j = 1, \ldots, n = 100$, we have

$$E(A_{ij}|X_i, X_j) = 0.5I(|X_i - X_j| = 0) + 0.2I(|X_i - X_j| = 1) + 0.3I(|X_i - X_j| = 2), \tag{4}$$

where $X_i \overset{i.i.d}{\sim} Multiple(1/3, 1/3, 1/3); i = 1, \ldots, 100$. When $X_i = X_j$, these two nodes are most likely to have an edge but when X_i and X_j differ by one, they are even less likely to have an edge, with probability of 0.2, than the most different pairs of nodes. This actually describes nonlinear dependence where `MGC` is

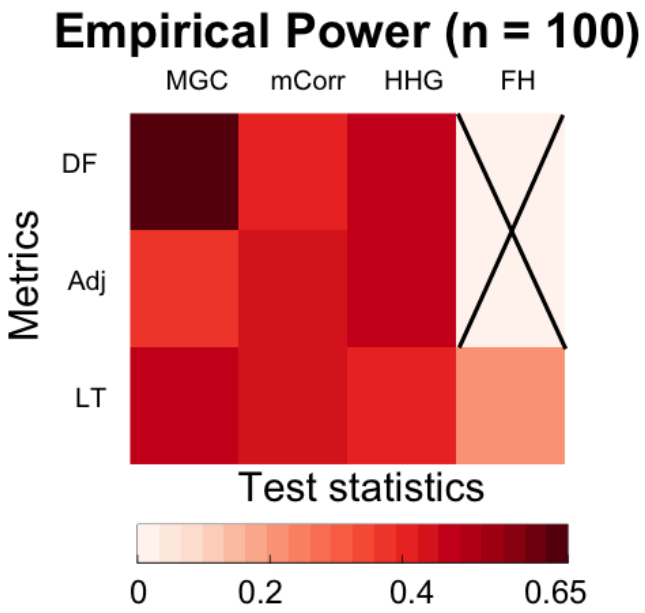

Figure 2: This power heatmap illustrates the superior power of multiscale generalized correlation (`MGC`) under diffusion distance matrix (`DF`) in three-block SBM (model 4), compared to under adjacency matrix distance (`Adj`) or latent factor distance (`LT`). This demonstrates one exemplary network where `MGC` statistic along with a family of diffusion distances catches non monotonic correlations efficiently than the other statistics and metrics.

believed to work better than the distance correlation. Figure 2 illustrates superior performance of `MGC` as a test statistic combined with diffusion maps (`DF`) as a network metric.

To scrutinize our conjecture on better performance of local optimal scaled `MGC` over global scale of `mCorr`, we control the amount of *nonlinear dependency* through changing the value of $\theta \in (0, 1)$ in the three block model 5.

$$E(A_{ij}|X_i, X_j) = 0.5I(|X_i - X_j| = 0) + 0.2I(|X_i - X_j| = 1) + \theta I(|X_i - X_j| = 2), \quad i, j = 1, \ldots, n = 100, \quad (5)$$

where $X_i \overset{i.i.d}{\sim} Bern(0.5); i = 1, \ldots, 100$. When $\theta > 0.2$, linear dependency of edge distribution in $\mathbf{A}$ upon nodal attribute of X is lost. If you see Figure 3, power of `mCorr` starts to drop from $\theta = 0.2$ while that of `MGC` almost stays clam, which implies `MGC` is significantly more sensitive to nonlinear dependency compared to `mCorr`.

On the other hand, the SBM connotes that all nodes within the same block have the same expected degree. Thus, this block model is limited by homogeneous distribution within block and provides a poor fit to networks with highly varying node degrees within block or community. Instead the Degree-Corrected Stochastic Block model (DCSBM) proposed by [8] add another random variable associated with each node to vary the node degrees. In the model 6, we controlled the amount of such variability by τ; the larger the value τ is, the more variability degree or edge distribution has. In Figure 4, power based on Euclidean distance of $\mathbf{A}$ or that of estimated network factors (locations) becomes less sensitive

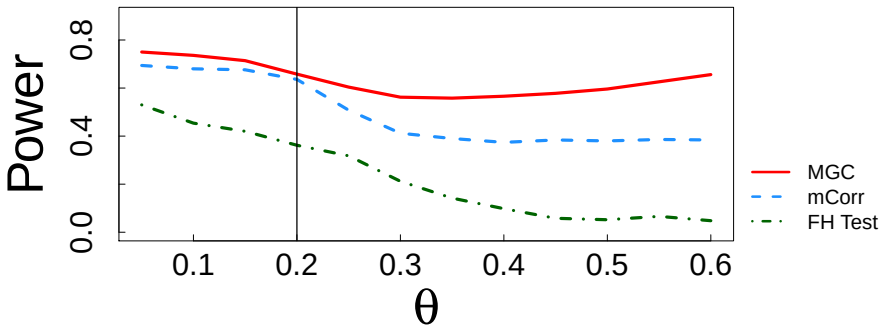

Figure 3: X-axis of θ controls the existence/amount of nonlinear dependency and in this particular case nonlinearity exists when $\theta > 0.2$ and gets larger as it increases. You can see the discrepancy in power between global and local scale tests also gets larger accordingly, mostly due to decreasing power of `mCorr` or `FH` test but relatively stable power of `MGC` under nonlinear dependency.

as τ increases. Compared to these two, diffusion maps are more robust to such variability.

$$E(A_{ij}|\mathbf{X}, \mathbf{V}) = 0.2V_iV_j \cdot I(|X_i - X_j| = 0) + 0.05V_iV_j \cdot I(|X_i - X_j| = 1), \tag{6}$$

where $X_i \overset{i.i.d.}{\sim} Bern(0.5); V_i \overset{i.i.d}{\sim} Uniform(1-\tau, 1+\tau), i = 1, \ldots, 250.$

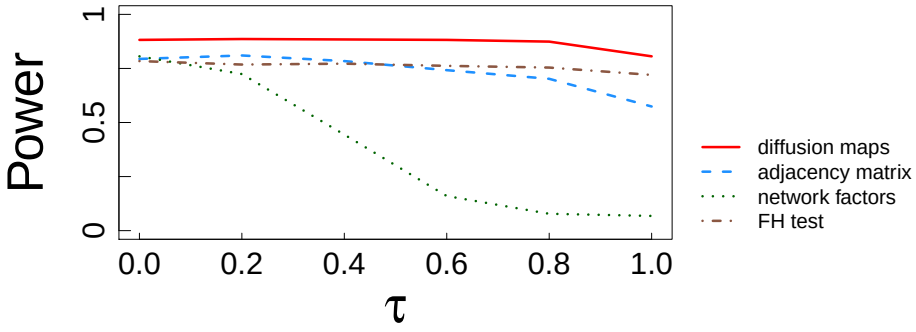

Figure 4: In degree-corrected SBM where the variability in degree distribution increases as τ increases, testing power of diffusion maps are more likely to be robust against increasing variability compared to other network metrics, e.g. adjacency matrix or latent positions. `FH` test statistics allowing different dimensions of network factors perform consistently well but still have less power than `MGC`.

3.2 Node Contribution Test

To examine the effectiveness of node contribution measure in testing dependency as presented in the statistic 3, we deliberately simulate the network and its nodal attributes as half of the nodes are

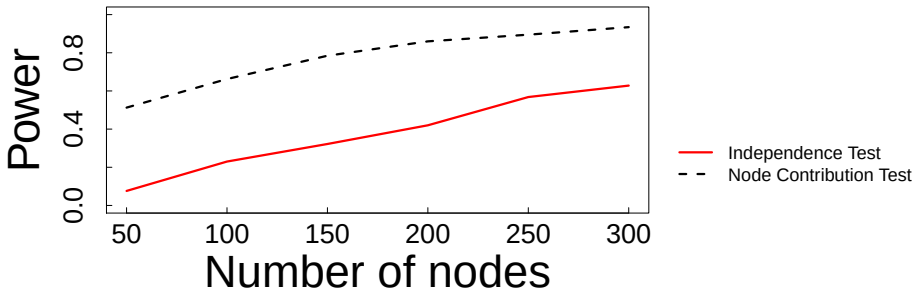

Figure 5: This plot describes that both power of `MGC` and the rate of correctly-ranked node contribution increase as the number of nodes increases when only half of the nodes for each simulation actually are set to be dependent on network, which validates the use of node contribution measure in independence test.

independent while the other half are dependent on network (model 7).

$$
\begin{aligned}
&X_i \overset{i.i.d}{\sim} Bern(0.5) \quad i = 1, \ldots, n/2, \ldots, n \\
&E(A_{ij}|X_i, X_j) \overset{d}{=} \begin{cases} 0.4I(|X_i - X_j| = 0) + Bern(0.1)I(|X_i - X_j| > 0) & i = 1, \ldots, n/2 \\ 0.25 & i = 1 + n/2, \ldots, n \end{cases}
\end{aligned} \tag{7}
$$

As an ad hoc test of node contribution, we rank the nodes in terms of decreasing order of $c(v)$ and count the ratio of dependent samples's ranks within the number of dependent nodes. If it works perfectly, all dependent nodes would take higher rank than every independent node so thus the rate equals to one. We call this rate as *inclusion rate*:

$$
\text{inclustion rate}\big(c(v)\big) = \sum_{v \in V(\mathbf{G})} \big\{ rank_{c(v)}\big(v\big) \leq m \big\}/m, \tag{8}
$$

where $m(\leq |V(\mathbf{G})|)$ is the number of nodes under network dependence. We set $m = n/2$ out of $n = |V(\mathbf{G})|$.

4 Conclusion

In this paper, we convince that `MGC`, merged with a family of diffusion distance, provides us powerful independence test statistics in network. Having multiscale statistics, i.e. one parameter family of statistics, is not avoidable because we regard distance between the nodes over network as a dynamic process. Through simulation studies, we demonstrate that our methods perform better than the others especially under nonlinear dependency, and we are able to measure each node's contribution to detecting dependency. Deriving the contributions is particularly important when there have possibly different amounts

of the dependencies among the nodes.

However obtaining a full family of statistics are computationally infeasible. Also we did not suggest any theoretically supported tools to select one metrics among them so thus we have one single statistic. As an ad hoc, we selected an *optimal* diffusion time t with highest power from $t = 1$ to $t = 10$ for our simulation since we could observe a stabilized empirical power within this period. Developing the adaptive method to find this optimal t where dependence is maximized would be a natural next step. Despite these shortcomings, we expect that we could also enjoy the properties of `MGC` and a family of diffusion distances in solving diverse problems which require to utilize local relationship of the data sets. For instance, we might be able to implement independence testing between two networks of same size by using diffusion distance of each network to investigate whether a pair of networks are topologically or structurally independent. This kind of work would shed light on revealing any relationship between the data sets which are not necessarily a random vector.

References

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

[2] 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.

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

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

[5] 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.

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

[7] 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.

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

[9] 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.

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

[11] 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.

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

[13] D. Reshef, Y. Reshef, H. Finucane, S. Grossman, G. McVean, P. Turnbaugh, E. Lander, M. Mitzenmacher, and P. Sabeti. Detecting novel associations in large data sets. *Science*, 334(6062):1518–1524, 2011.

[14] 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.

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

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

[17] 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.

[18] 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.

[19] 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.

[20] Y. Zhao, E. Levina, and J. Zhu. Consistency of community detection in networks under degree-corrected stochastic block modelstection 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 :

$$\mathbf{U}_t(i) = \begin{pmatrix} \lambda_1^t\phi_1(i) & \lambda_2^t\phi_2(i) & \cdots & \lambda_q^t\phi_q(i) \end{pmatrix} \in \mathbb{R}^q. \tag{9}$$

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 $i \neq j; i, j = 1, 2, \ldots, n$:

$$\left(P_{ij}\right) = \left(\frac{A_{ij}}{1 - A_{ij} + \sum_{j=1}^{n} A_{ij}}\right) \overset{d}{=} \left(\frac{A_{\sigma(i)\sigma(j)}}{1 - A_{\sigma(i)\sigma(j)} + \sum_{\sigma(j)=1}^{n} A_{\sigma(i)\sigma(j)}}\right) = \left(P_{\sigma(i)\sigma(j)}\right) \tag{10}$$

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 se-

quence of $(\mathbf{W}, \mathbf{Y}) = \{(\mathbf{w}_i, \mathbf{y}_i); i = 1, 2, \ldots, n\}$ which is identically distributed as $(\mathbf{w}, \mathbf{y})$ with finite second moment, we have

$$\mathcal{V}_n^2(\mathbf{W}, \mathbf{Y}) \longrightarrow \mathcal{V}^2(\mathbf{w}, \mathbf{y}) \qquad \text{as } n \to \infty \tag{11}$$

where $\mathcal{V}^2(\mathbf{w}, \mathbf{y}) := \|g_{\mathbf{w},\mathbf{y}}(t, s) - g_{\mathbf{w}}(t)g_{\mathbf{y}}(s)\|^2$, and $g_{\cdot}$ is a characteristic function, e.g., $g_{\mathbf{w},\mathbf{y}}(t, s) = E\{\exp\{i\langle t, \mathbf{w}\rangle + i\langle s, \mathbf{y}\rangle\}\}$. This follows exactly the same as *Theorem 1* in [17]. Note that this Lemma always holds without any assumption on $\{(\mathbf{w}_i, \mathbf{y}_i), i = 1, 2, ..., n\}$.

Followed by *de Finetti's Theorem*, if and only if $\{\mathbf{w}_i\}$ are (infinitely) exchangeble, there exists an underlying distribution $f_{\mathbf{w}}$ of $\mathbf{w}$ such that $\mathbf{w}_i \overset{i.i.d}{\sim} f_{\mathbf{w}}$. By the same logic there exists a random, we have an underlying distribution $f_{\mathbf{y}}$ where $\mathbf{y}_i \overset{i.i.d}{\sim} f_{\mathbf{y}}$. Let $(\mathbf{w}_i, \mathbf{y}_i) \overset{i.i.d}{\sim} f_{\mathbf{w},\mathbf{y}}$. 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 [17], i.e.,

$$\int_{D(\delta)} \|g^n_{\mathbf{w},\mathbf{y}}(t, s) - g^n_{\mathbf{w}}(t)g^n_{\mathbf{y}}(s)\|^2 dh \overset{n\to\infty}{\longrightarrow} \int_{D(\delta)} \|g_{\mathbf{w},\mathbf{y}}(t, s) - g_{\mathbf{w}}(t)g_{\mathbf{y}}(s)\|^2 dh, \tag{12}$$

where $D(\delta) = \{(t, s) : \delta \le |t|_p \le 1/\delta, \delta \le |s|_q \le 1/\delta\}$, and $h(t, s)$ is the weight function chosen in [17]. It follows that

$$\mathcal{V}_n^2(\mathbf{W}, \mathbf{Y}) \to 0 \qquad \text{as } n \to \infty \tag{13}$$

if and only if $g_{\mathbf{w},\mathbf{y}}(t, s) = g_{\mathbf{w}}(t)g_{\mathbf{y}}(s)$, i.e., $\mathbf{w}$ is independent of $\mathbf{y}$. Therefore, the `dCorr` or `mCorr` converges to 0 if and only if underlying distributions are independent; and its testing power converges to 1 under any joint distribution of finite moments. Since the multiscale generalized correlation based on any consistent global correlation is also consistent [15], `MGC` statistic constructed by `dCorr` or `mCorr` is also consistent in testing dependence. □