Testing independence between networks and nodal attributes via multiscale metrics and multiscale distance correlation

John Smith,[1*] Jane Doe,[1] Joe Scientist[2]

[1]Department of Chemistry, University of Wherever,
An Unknown Address, Wherever, ST 00000, USA
[2]Another Unknown Address, Palookaville, ST 99999, USA

[*]To whom correspondence should be addressed; E-mail: ylee160@jhu.edu.

Network dependence over network space, which refers to the dependence between network topology and its nodal attributes, often exhibits nonlinear dependent properties. Unfortunately, without knowledge on specific neighborhood structures, no statistic has been suggested to test network dependence further than globally linear dependence. In this study we propose a multiscale dependence test statistic, which borrows the idea of diffusion maps and Multiscale Generalized Correlation (MGC). Through simplest network simulation and the supportive theory, we have found that the newly proposed test is a consistent test and achieves higher power than other available ones without any parametric assumptions neither on graphical model nor dependency structure, but only with the restriction of exchangeability.

Contents

1 Introduction

1.1 intro1

- *[intro1] Introduce interests in association between network and nodal attributes and the related literature.*

Network, a collection of nodes and edges between them, has been a celebrated area of study over a field of psychology, information theory, biology, statistics, economics, etc. The relationship between the way a pair of nodes are connected and the values of their attribute is a common interest in network analysis. There has been a lot of efforts to represent a network as a function of nodal attributes or model an outcome of nodal attribute variables through their underlying network structures. However, it is very obscure to determine which one should be put as a dependent variable. And most of all network often does not have a natural structure. This is why there exist a plethora of works on latent structure of network, which also depends on the characteristics of each node (Hoff et al. (2002) , Austin, Linkletter, and Wu (2013)). In a latent space model, local independence between network and nodal attributes, conditional on a latent variable is often assumed (Lazarsfeld and Henry (1968)), which makes easier to interpret the dependency mechanism. In real network data, however, it is almost impossible to estimate such latent variable without any knowledge on true network generative model and also we cannot guarantee that direction or amount of association between network and nodal attributes keeps consistent across latent variables, i.e. we cannot guarantee linear dependence.

1.2 intro2

- *[intro2] Introduce notations we are going to use and introduce a common network model of nodal attributes.*

Throughout this paper, assume that we are given an unweighted and undirected, connected network $\boldsymbol{G}$ comprised of n nodes, for a fixed $n \in \mathbb{N}$. Suppose that our observation of $\boldsymbol{G}$ is one random sample from true, population distribution of $\mathcal{G}$. Even though we assume that $\boldsymbol{G}$ is an undirected and unweighted network but we are able to extend all of the theory here to directed and even weighted network. An adjacency matrix of a given network, denoted by $\boldsymbol{A} = \{A_{ij} : i, j = 1, .., n\}$, is often introduced to formalize this relational data of network. Let us introduce a m-variate ($m \in \mathbb{N}$) variable for nodal attributes $\boldsymbol{X} \in \mathbb{R}^m$ which we are interested in. Investigating correlation between $\boldsymbol{G}$ and $\boldsymbol{X}$ and testing whether their distributions are independent or not is the key focus in our study.

Even though distribution of graph $\boldsymbol{G}$ is often formalized through modeling adjacent relations $\{a_{ij} : i, j = 1, ..., n\}$, we cannot directly use $\mathbf{A}$ in testing independence to $\mathbf{X}$. Rather than regressing a_{ij} on observed attribute values x_i or x_j, Fosdick and Hoff proposed a latent variable model to estimate node-specific network factor which provides a one-to-one correspondence between $\boldsymbol{G}$ and $\boldsymbol{X}$ as well as reduces a dimension of network data. In their paper, Fosdick and Hoff also used these factors to test independence between $\boldsymbol{G}$ and $\boldsymbol{X}$. However, the performance of this test would not be good enough when the relational data $\boldsymbol{A}$ does not have linear relationship to network factors or has a latent mixture model. Since we never know the structure of networks and the way they are related to other variables, there always exist a limitation on testing based on modeling network.

2 Multiscale Distance Metrics

2.1 [MDM1]

- *[MDM1] represent a network structure as a one-parameter family of multivariate variables*

There have been a lot of efforts to represent the network in terms of a summarizing network factor([Fosdick & Hoff (2015)]) or some meaningful coefficients, e.g. centrality or connectivity. However, there has been no vertex-wise variable which provides a configuration of vertex over network space without losing any information. [Coifman & Lafon] demonstrated that diffusion maps provide a meaningful multiscale geometries of data while keeping information on every local relation. Diffusion maps is constructed via Markov chain on graph. Here an adjacency matrix $\boldsymbol{A}$ acts as a kernel, representing a similarity between each node in $\boldsymbol{G}$. The adjacency matrix also takes into account every single relationships between nodes, rather than estimating or summarizing network structures.

Let $(\boldsymbol{G}, \mathcal{A}, \mu)$ be a measure space. Throughout all of the arguments, assume that we have a countable vertex set with size of $n \in \mathbb{N}$. The vertex set of network $\boldsymbol{G}$ is the data set of vertices and edges and $\mathcal{A}$ is a set of a pair of nodes $\{(i, j) : v_i, v_i \in V(\boldsymbol{G})\}$. A measure of μ which represents a distribution of the vertices on $\boldsymbol{G}$, is equivalent to an adjacency matrix $\boldsymbol{A}$. A transition matrix $P = \{P[i, j] : i, j = 1, ..., n\}$ in Markov chain on $\boldsymbol{G}$, which represents the probability that flow or signal goes from Node i to Node j, is defined as below:

$$P[i, j] = A_{ij} \Big/ \sum_{j=1} A_{ij} \tag{1}$$

A transition matrix P is a new kernel of a Markov chain of which element $P[i, j]$ represents the probability of travel from Node i to Node j in one time step. On the other hand, corresponding probability in t steps is given by the t th ($t \in \mathbb{N}$) power of P. How to derive diffusion distance over a directed network or weighted network is provided in Tang & Trosset (2010)]. Other than a transition matrix, we need a stationary probability $\boldsymbol{\pi} = \{\pi(1), \pi(2), ..., \pi(n)\}$ of which $\pi(i)$ represents the probability that the chain stays in Node i regardless of the starting state. In our setting, $\pi(i)$ is proportional to the degree of Node i, i.e. $\pi(i) = \sum_{j=1}^{n} A_{ij} \big/ \sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij}$ $(i = 1, 2, ..., n)$.

For each time point $t \in \mathbb{N}$, we can define a diffusion distance C_t given by :

$$
\begin{aligned}
C_t^2[i,j] &= \sum_{w=1}^{n} \left(P^t[i,w] - P^t[j,w]\right)^2 \frac{1}{\pi(w)} = \sum_{w=1}^{n} \left(\frac{P^t[i,w]}{\sqrt{\pi(w)}} - \frac{P^t[j,w]}{\sqrt{\pi(w)}}\right)^2 \\
&= \| P^t[i,\cdot] - P^t[i,\cdot] \|^2_{L^2(\boldsymbol{G}, d\mu/\pi)}
\end{aligned}
\tag{2}
$$

As diffusion time t increases, distance matrix C_t is more likely to take into account distance between two nodes far away in terms of the length of the path. Key idea behind such diffusion distance at fixed time t is that it measures the chance that we are likely to stay between Node i and Node j at t step on our journey of all other possible paths. The higher chance is, the smaller distance between two is. This distance well reflects the connectivity between two nodes. Connectivity between two nodes is higher if we need to eliminate more number of vertices to disconnect these two. Unlike an adjacency relation or geodesic distance, a connectivity between two nodes depends on their relationship to other vertices in a given network so it is more robust to the unexpected edges. Often a set of nodes with higher connectivity have a higher propensity of having edges within this set and they are likely to form a cluster. Thus diffusion distance is very robust measure and also very sensitive to the clustering structure of network.

2.2 [MDM2]

- *[MDM2] explain spectral properties of diffusion maps*

Diffusion distance of $\boldsymbol{G}$ defined as above can be represented via a spectral decomposition of its transition matrix P. That is, we can derive diffusion distance using its eigenvectors and eigenvalues. The spectral analysis on diffusion distance or diffusion maps have been studied for its usefulness for nonlinear dimensionality reduction ([Coifman & Lafon (2006)], [Lafon & Lee (2006)]).

Recall that diffusion distance at time t, C_t is a functional L^2 distance, weighted by $1/\pi$. If we transform the way to represent $C_t[i,j]$ slightly, we are able to obtain an orthonormal basis of $L^2(\mathbf{G}, d\mu/\pi)$ via eigenvalues and eigenvectors.

Keeping mind that a symmetry of an adjacency matrix A does not guarantee a symmetric of P, define a symmetric kernel $\boldsymbol{Q} = \boldsymbol{\Pi^{1/2} P \Pi^{-1/2}}$, where $\boldsymbol{\Pi}$ is a $n \times n$ diagonal matrix of which ith diagonal element is $\pi(i)$. Under compactness of P, $\boldsymbol{Q}$ has a discrete set of real nonzero eigenvalues $\{\lambda_r\}_{r=\{1,2,\dots,q\}}$ and a set of their corresponding orthonormal eigenvectors $\{\psi_r\}_{r=\{1,2,\dots,q\}}$, i.e. $Q[i,j] = \sum_{r=1}^{q} \lambda_r \psi_r(i)\psi_r(j)$ $(1 \le q \le n)$. Since $P[i,j] = \sqrt{\pi(j)/\pi(i)}Q[i,j]$, $P[i,j] = \sum_{r=1}^{q} \lambda_r \{\psi_r(i)/\sqrt{\pi(i)}\}\{\psi_r(j)\sqrt{\pi(j)}\} := \sum_{r=1}^{q} \lambda_r \phi_r(i)\{\psi_r(j)\sqrt{\pi(j)}\}$, where $\phi_r(i) := \psi_r(i)/\sqrt{\pi(i)}$. Then from $\sum_{r=1}^{q} \psi_r(j)\sqrt{\pi(j)} = 1$ for all $j \in \{1, 2, ..., n\}$, we can represent the diffusion distance as:

$$
\begin{aligned}
C_t^2[i,j] &= \sum_{r=1}^{n} \lambda_r^{2t} \big(\phi_r(i) - \phi_r(j)\big)^2 \\
&= \| \ P^t[i,\cdot] - P^t[i,\cdot] \ \|^2_{L^2(\boldsymbol{G}, d\mu/\pi)}
\end{aligned} \tag{3}
$$

That is,

$$
C_t[i,j] = \| \ \boldsymbol{U}_t(i) - \boldsymbol{U}_t(j) \ \| \tag{4}
$$

, where

$$
\boldsymbol{U}_t(i) = \begin{pmatrix} \lambda_1^t \phi_1(i) \\ \lambda_2^t \phi_2(i) \\ \vdots \\ \lambda_q^t \phi_q(i) \end{pmatrix} \in \mathbb{R}^q. \tag{5}
$$

2.3 [MDM3]

- *[MDM3] introduce exchangeable graph*

Due to the inter-correlated construction of U, e.g. ith subject's diffusion depends on others in the given network, it is hard to say that the observed diffusion coordinates of n subjects are independent samples without any restrictions on networks. Independence within sample of one variable is important in testing independence with other variable. (why? example)

To talk about independence of U, we need more general concept of exchangeable graph. Formally speaking, an exchangeable sequence of random variables is a finite or infinite sequence $U_1, U_2, ...$ of random variables such that for any finite permutation σ of the indices 1,2,3, ... , the joint probability distribution of the permutated sequence $U_{\sigma(1)}, U_{\sigma(2)}, U_{\sigma(3)}, ...$ is the same as joint probability distribution of the original sequence.

The property of exchangeability is closely related to the use of independent and identically-distributed(i.i.d) random variable. A sequence of random variables that are i.i.d. conditional on some underlying distributional form is exchangeable. Moreover, the converse can be established for "infinite sequence" by [Bruno de Finetti](7.1)'s representation theorem. The extended versions of the theorem show that in any infinite sequence of exchangeable random variables, random variables are conditionally i.i.d, given the underlying distributional form.

It is straightforward to check that $\mathbf{G}$ is an exchangeable graph if and only if its adjacency matrix $\mathbf{A}$ is jointly exchangeable.

Definition 2.1. A random 2-array (A_{ij}) is called jointly exchangeable if

$$(A_{ij}) \overset{d}{=} (A_{\sigma(i)\sigma(j)})$$

for every permutation σ of n, and separately exchangeable if

$$(A_{ij}) \overset{d}{=} (A_{\sigma(i)\sigma'(j)})$$

for every pair of permutation σ, σ' of n.

joint exchangeability of A is what we need. Widely used graphical model, e.g. Stochastic Block Model (SBM), Random Dot Product Graph (RDPG), satisfies this definition of jointly exchangeable graph under some constraints.

2.4 [MDM4]

- *[MDM4] Derive exchangeability and iid representation of diffusion maps*

The main advantage from such representation is that it becomes now possible to represent given network $\boldsymbol{G}$ as a one-parameter family of $q(< n)$-coordinates called diffusion coordinate, of which metric well reflects how corresponding vertices are connected each other at each diffusion time point. A set of diffusion maps for each vertex have some desirable properties which will be helpful in testing. The followings are concerning about a few conditions for earning them.

Lemma 2.1 (Exchangeability and i.i.d. of A in graphon)**.** Assume that a connected, undirected and unweighted graph $\mathbf{G}$ is a graphon. Then 2-array of $\{A_{ij} : i = 1, 2, ..., n, i < j\}$ are i.i.d. conditioning on random link function $g : [0, 1]^2 \to [0, 1]$. Thus for fixed row (column) of $\mathbf{A}$, $\{A_{i1}, A_{i2}, ..., A_{in}\}$, $i \in \{1, 2, ..., n\}$ are i.i.d. on random link function g.

On the other hand, a graphon, despite its advantage on simple representation, is either empty or dense. Thus, it fails to represent real network data where the sparsity or scale-free distribution is fairly common. Thus we introduce a concept of graphex introduced by Veitch and Roy, 2015, which is more generalized version of graphon. The following claim says that we can still have i.i.d. sequence from $\mathbf{A}$ conditioning on random link function $g : \mathbb{N}^2 \to [0, 1]$:

Lemma 2.2 (Exchangeability and i.i.d. of A in graphex)**.** Assume that a connected, undirected and unweighted graph $\mathbf{G}$ is a graphex. Then 2-array of $\{A_{ij} : i = 1, 2, ..., n, i < j\}$ are i.i.d.

conditioning on random link function $h : \mathbb{N}^2 \to [0,1]$. Thus for fixed row (column) of $\mathbf{A}$, $\{A_{i1}, A_{i2}, ..., A_{in}\}$, $i \in \{1, 2, ..., n\}$ are conditionally i.i.d. on random link function g.

From the above Lemma 2.1, 2.2, we can also prove exchangeability and conditional i.i.d. of diffusion maps at each time point.

Lemma 2.3 (Exchangeability and iid of U)**.** Assume that a connected, undirected and unweighted graph $\mathbf{G}$ is a graphon or graphex, i.e. any exchangeable random graph from an infinite graph. Then its transition probability P_{ij} so thus diffusion maps at fixed time t also exchangeable conditional on link function of graph. Furthermore, by [de Finetti]'s Theorem, we can say that such diffusion maps at t are conditionally i.i.d given random probability measure η on U_t.

Lemma 2.3 above provides us i.i.d. one-parameter family of $\{\mathbf{U}_t\}_{t\in\mathbb{N}}$ conditional on a random probability measure of $\mathbf{U}_t$, which is free of any further model assumptions on network structures.

3 Multiscale Generalized Correlation

3.1 [MGC1]

- *[MGC1] Introduce a Distance Correlation and Multiscale version*

Relationship between network and nodal attributes often exhibits local or nonlinear properties. Moreover, dimension of spectrum of network increases as a sample size increases. Unfortunately, widely used correlation measures often fail to characterize non-linear associations or multivariate associations so they fail to provide a consistent test statistic against all types of dependencies. [Szekely et al. (2007)] extended pairwise constructed generalized correlation coefficient and developed a novel statistics called distance correlation (`dCor`) as a measure for all types of dependence between two random vectors in any dimension. Let us start from a general setting that we are given $n \in \mathbb{N}$ pairs of random samples $\{(x_i, y_i) : x_i \in \mathbb{R}^p, y_i \in \mathbb{R}^q, i = 1, ..., n\}$. Define $C_{ij} = \parallel x_i - x_j \parallel$ and $D_{ij} = \parallel y_i - y_j \parallel$ for $i, j = 1, ..., n$.

Distance correlation (`dCorr`) is defined via distance covariance (`dCov`) $\mathcal{V}_n^2$ of $\boldsymbol{X}$ and $\boldsymbol{Y}$, which is the following:

$$\mathcal{V}_n^2(\boldsymbol{X}, \boldsymbol{Y}) = \frac{1}{n^2} \sum_{i,j=1}^{n} \tilde{C}_{ij} \tilde{D}_{ij} \tag{6}$$

, where $\tilde{C}$ and $\tilde{D}$ is a doubly-centered C and D respectively, by its column mean and row mean. Distance correlation $\mathcal{R}_n^2(\boldsymbol{X}, \boldsymbol{Y})$ is a standardized `dCov` by $\mathcal{V}_n^2(\boldsymbol{X}, \boldsymbol{X})$ and $\mathcal{V}_n^2(\boldsymbol{Y}, \boldsymbol{Y})$.

$$\mathcal{R}_n^2(\boldsymbol{X}, \boldsymbol{Y}) = \frac{\mathcal{V}_n^2(\boldsymbol{X}, \boldsymbol{Y})}{\sqrt{\mathcal{V}_n^2(\boldsymbol{X}, \boldsymbol{X}) \mathcal{V}_n^2(\boldsymbol{Y}, \boldsymbol{Y})}} \tag{7}$$

On the other hand, a modified distance covariance (`MCov`) $\mathcal{V}_n^*$ and a modified distance correlation (`MCorr`) $\mathcal{R}_n^*$ for testing high dimensional random vectors were also proposed in [Szekely et al. (2013)]. However, `dCorr` and even MCorr still perform not very well in various nonlinear settings and under existence of outliers ([Cencheng et al. (2016)]). Out of this concern, Cencheng at al. developed Multiscale Generalized Correlation (`MGC`) by adding local scale on correlation coefficients.

3.2 [MGC2]

- *[MGC2] choice of proper metrics for testing network independence*

Returning to the problem of network setting, the fundamental problem is in measuring all types of dependence between $\boldsymbol{G}$ and $\boldsymbol{X}$, we are required a vertex-wise coordinates of which Euclidean distance measures a distance between them. You might first propose directly using a column of an adjacency matrix so that we have a n-pair of observations $\big\{\big(\boldsymbol{A}_{i\cdot}, \boldsymbol{X}_i\big) : \boldsymbol{A}_{i\cdot} = (A_{i1}, ..., A_{in}), \boldsymbol{X}_i \in \mathbb{R}^m, i = 1, ..., n\big\}$. In the context of network, however, it is almost impossible to assume $\{\boldsymbol{A}_{1\cdot}, \boldsymbol{A}_{2\cdot}..., \boldsymbol{A}_{n\cdot}\}$ is an independent observation from a common distribution. Since an adjacency matrix $\boldsymbol{A}$ is formed by relational data, one row is dependent on the other. Even if it is not, Euclidean distance between $\{\boldsymbol{A}_{i\cdot} : i = 1, ..., n\}$ is not a proper metric over network space. For simplest example, assume that a given network $\boldsymbol{G}$ is an undirected network so that its adjacency matrix $\boldsymbol{A}$ must be a symmetric matrix. Then for any $i \neq j$, $\boldsymbol{A}_{i\cdot}$ and $\boldsymbol{A}_{j\cdot}$ cannot be independent, and under no self-loop, $A_{ii} = 0$ for all $i \in \{1, ..., n\}$. Moreover, as for the validity of its Euclidean distance, let us introduce a simple example. Let a given network $\boldsymbol{G}$ having 8 nodes be an unweighted, directed network and possibly having self-loop. Let $\boldsymbol{A}$ be its 8×8 binary adjacency matrix. Assume `Node 1`, `Node 4` and `Node 8` have the following row entries:

$$\boldsymbol{A}_{1\cdot} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix} \tag{8}$$

$$\boldsymbol{A}_{4\cdot} = \begin{pmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \end{pmatrix} \tag{9}$$

$$\boldsymbol{A}_{8\cdot} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \tag{10}$$

,which results $\| \boldsymbol{A}_{1\cdot} - \boldsymbol{A}_{4\cdot} \|^2 = 4$, $\| \boldsymbol{A}_{1\cdot} - \boldsymbol{A}_{8\cdot} \|^2 = 7$,and $\| \boldsymbol{A}_{4\cdot} - \boldsymbol{A}_{8\cdot} \|^2 = 3$. Accordingly, $\| \boldsymbol{A}_{4\cdot} - \boldsymbol{A}_{8\cdot} \| < \| \boldsymbol{A}_{1\cdot} - \boldsymbol{A}_{4\cdot} \|$. However, you can easily see that this does not make sense because `Node 4` and `Node 8` are connected each other only through `Node 1`.

Therefore instead of using an adjacency matrix directly, we are considering embedding a vertex $v \in V(\boldsymbol{G})$ into its diffusion map of $\boldsymbol{U}$ and apply Euclidean distance metric, which is exactly a diffusion distance. As explained before, its Euclidean distance takes into account all possible paths between every pair of node and measure the connectivity between them. Unlike in the other metrics in network, i.e. adjacency matrix or geodesic distance, triangle inequality holds in diffusion distance (proof in [Appendix]).

Corollary 3.0.1 (Triangle inequality)**.** For fixed time t,

Thanks to these properties of diffusion maps, we have better interpretation of its Euclidean distance so that we will use it to the distance matrix in `MGC`.

3.3 [MGC3]

- *[MGC3] present the main theorem*

We have discussed one-parameter family i.i.d. representation of network structures, called diffusion maps, and also when given i.i.d. pair of observations of two variables, how distance correlation and its multiscale version test independence between these two. Now it is time to combine these two to test independence between network structures and nodal attributes.

Theorem 3.1 (MGC of testing independence)**.** Assume that a connected, undirected and unweighted graph $\mathbf{G}$ is an exchangeable graph. Assume that we are given n-pair of observations

$\{(u_i^{(t)}, x_i) : i = 1, 2, ..., n, t \in \mathbb{N}\}$. Then $u_i^{(t)} \overset{i.i.d.}{\sim} f_u^{(t)}\big(\eta, g\big)$, $t \in \mathbb{N}$ and $x_i \overset{i.i.d.}{\sim} f_x$, where $f_u^{(t)}\big(\eta, g\big)$ are conditional distribution function given a link function g and a random probability measure η_t of $U_{(t)}$. Then `MGC` applied to these pair of data is theoretically consistent against all dependent alternatives in testing :

$$H_0 : f_{U(t)\cdot X} = f_U^{(t)} \cdot f_X$$

Since $\boldsymbol{U}$ provides a configuration of vertices in $\boldsymbol{G}$, the above hypothesis implies testing independence between the configuration of vertices in network space and in attribute space, but as a function of a link function g and a random function (variable) of η of $\mathbf{U}$.

Remark 1. Roughly speaking, we can say that diffusion maps are i.i.d. function of a link function g and a random function of $\mathbf{U}$, η. Thus testing independence between conditional U and X can be considered as testing independence between $f\big(g, \eta\big)$ and X. A link function g concerns the distribution of edges and a random function η concerns nature distribution of diffusion maps. Our testing basically examines whether how edges are constructed and how diffusions(propagation) process are correlated to nodal attributes.

4 Simulation Study

4.1 [sim1]

- *[sim1] present simplest examples of stochastic block model (graphon)*

We mentioned in the Introduction that latent network model is very common followed by the assumption of local independence. Stochastic Block Model (SBM) is one of the most popular and also useful network generative model, especially as a tool for community detection ([Karrerl and Newman]). The SBM, in the simplest setting, assumes that each of n vertices in graph $\boldsymbol{G}$ belongs to one of $K \in \mathbb{N}(\leq n)$ blocks or groups. Block affiliation is important because the probability of having edges between a pair of vertices depends on which blocks they are in.

Assume that the latent variable $Z_1, Z_2, ..., Z_n \overset{i.i.d.}{\sim} Multinomial\big(\pi_1, \pi_2, ..., \pi_K\big)$. Then the uppwer triangular entries of $\mathbf{A}$ are independent with $A_{ij} \overset{i.i.d.}{\sim} Bern\big(\sum\limits_{k,l=1}^{K} p_{kl} I\big(Z_i = k, Z_j = l\big)\big), \forall i < j$.

Let $W_1, W_2, ..., W_n \overset{i.i.d.}{\sim} Unif[0,1]$. Then such $\mathbf{A}$ can be represented with respect to $\{W_i\}$ as follows:

$$A_{ij} \overset{i.i.d.}{\sim} Bern\big(g(W_i, W_j)\big), \forall i < j \tag{11}$$

, where $g\big(W_i, W_j\big) = \sum\limits_{k,l=1}^{K} p_{kl} I\big(W_i \in [\sum\limits_{j=1}^{k-1} \pi_j, \sum\limits_{j=1}^{k} \pi_j], W_j \in [\sum\limits_{j=1}^{l-1} \pi_j, \sum\limits_{j=1}^{l} \pi_j]\big)$

4.1.1 [sim1.1]

$$X_i \overset{i.i.d}{\sim} Bern(0.5), i = 1, ..., 100 \tag{12}$$

$$Z_i \sim \begin{cases} Bern(0.6) & X_i = 0 \\ Bern(0.4) & X_i = 1 \end{cases} \tag{13}$$

$$A_{z_i,z_j} \sim Bern \begin{bmatrix} \mathbf{0.4} & \mathbf{0.1} \\ \mathbf{0.1} & \mathbf{0.4} \end{bmatrix} \tag{14}$$

4.2 [sim2]

- *[sim2] present simplest examples of sparse block model (graphex)*

5 Real Data Examples

5.1 c.elegans

6 Discussions

6.1 [dis1]

- *[dis1] discuss merits of using MNT*

Throughout this study, we demonstrate that multiscale network test statistic to test network independence performs well in diverse settings, being supported by thorough theory on distance correlation and diffusion maps. Testing independence is often the very first step in investigating relationship between network topology and nodal variables in our interest. It is more likely that we want to know more than binary decision of rejecting or not rejecting the hypothesis. Multiscale test statistics due to both neighborhood choice and time spent in diffusion processes provides us a hint on a latent dependence structure as well.

6.2 [dis2]

- *[dis2] discuss limitations of using MNT*

6.3 [dis3]

- *[dis3] discuss possible contributions of MNT and future works*

7 Appendix

7.1 [apd1]

- *[apd1] additional plots*

7.2 [apd2]

- *[apd2] (proof of) lemmas and theorems*

Theorem 7.1 (de Finetti's Theorem)**.** 1. Let $X_1, X_2, ...$ be an infinite sequence of random variables with values in a space $\mathbf{X}$. The sequence $X_1, X_2, ...$ is exchangeable if and only if there is a random probability measure η on $\mathbf{X}$ such that the X_i are conditionally i.i.d. given η.

2. If the sequence is exchangeable, the empirical distributions

$$\hat{S}_n(.) := \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}(.), n \in \mathbb{N}$$

converges to η as $n \to \infty$ with probability 1.

Theorem 7.2. Let $\mathbf{A} = \{A_{ij}\}, 1 \leq i, j \leq \infty$ be a jointly exchangeable binary array if and only if there exists a random measurable function $f : [0, 1]^3 \to \mathbf{A}$ such that

$$\big(A_{ij}\big) \overset{d}{=} \big(f\big(U_i, U_j, U_{ij}\big)\big) \tag{15}$$

where $(U_i)_{i\in\mathbb{N}}$ and $(U_{ij})_{i,j>i\in mathbbN}$ with $U_{ij} = U_{ji}$ are a sequence and matrix, respectively, of i.i.d. Uniform[0,1] random variables.

Proof of Lemma 2.1. By [Aldous-Hoover]Theorem, a random array (A_{ij}) is jointly exchangeable if and only if it can be represented as follows : there is a random function $g : [0, 1]^2 \to [0, 1]$ such that

$$(A_{ij}) \stackrel{d}{=} Bern(g(W_i, W_j))$$

, where $W_i \stackrel{i.i.d.}{\sim} Uniform(0,1)$.

Thus if $\mathbf{A}$ is an adjacency matrix of an undirected, exchangeable network, for any $i < j$, $i, j = 1, ..., n$:

$$\begin{aligned} P\big(A_{ij} = a_{ij}\big) &= \int P\big(A_{ij} \big| w_i, w_j\big) Pr(W_i = w_i) Pr(W_j = w_j) dw_i dw_j \\ &= \int_0^1 \int_0^1 g(w_i, w_j)^{a_{ij}} \big(1 - g(w_i, w_j)\big)^{1-a_{ij}} dw_i dw_j \end{aligned} \tag{16}$$

For fixed row index to $i \in \{1, 2, ..., n\}$,

$$P(A_{i1} = a_{i1}, A_{i2} = a_{i2}, ..., A_{in} = a_{in}) = \prod_{j=1}^{n} P(A_{ij} = a_{ij})$$

□

Proof of Lemma 2.2. Based on Kallenberg and Exchangeable Graph (KEG) frameworks, introduced in **??**Veitch and Roy, a random array (A_{ij}) is jointly exchangeable if and only if it can be represented as follows : there is a random function $g : [0,1]^2 \to [0,1]$ such that

$$(A_{ij}) \stackrel{d}{=} (A_{v_i, v_j}) \stackrel{d}{=} Bern(g(\vartheta_i, \vartheta_j)) \tag{17}$$

, where $v_i \stackrel{i.i.d.}{\sim} Poisson(1), \vartheta_i \stackrel{i.i.d.}{\sim} Poisson(1), v_i \leq \nu, i = 1, 2, ..., n$, for some pre-specified $\nu > 0$ so that finite size graphs can include vertices only if they participate in at least one edges.

Thus if $\mathbf{A}$ is an adjacency matrix of an undirected, exchangeable network, for any $i < j$,

$i, j = 1, ..., n$:

$$P\big(A_{ij} = a_{ij}\big) = \int P\big(A_{ij}\big|v_i, v_j\big)Pr(V_i = v_i)Pr(V_j = v_j)Pr(\vartheta_i = \vartheta_i)Pr(\vartheta_j = \vartheta_j)dv_i dv_j d\vartheta_i d\vartheta_j$$
$$= \int_0^\tau \int_0^\tau \int_0^\infty \int_0^\infty g(\vartheta_i, \vartheta j)^{a_{ij}}\big(1 - g(\vartheta_i, \vartheta_j)\big)^{1-a_{ij}} \times dPois_1(x_1) \times dPois_1(x_2) \times dPois_1(x_3) \times \tag{18}$$

□

Proof of Lemma 2.3. We have shown that for fixed time t, diffusion distance is defined as an Euclidean distance of diffusion maps. Diffusion maps is represented as follows :

$$\boldsymbol{U}_t(i) = \begin{pmatrix}\lambda_1^t\phi_1(i) & \lambda_2^t\phi_2(i) & \cdots & \lambda_q^t\phi_q(i)\end{pmatrix} \in \mathbb{R}^q. \tag{19}$$

Recall that $\Phi = \Pi^{-1/2}\Psi$ and $\mathbf{Q} = \boldsymbol{\Psi\Lambda\Psi}^T = \boldsymbol{\Pi}^{1/2}\mathbf{P}\boldsymbol{\Pi}^{-1/2}$. Thus, $\mathbf{P}\boldsymbol{\Pi}^{-\mathbf{1/2}}\boldsymbol{\Psi} = \boldsymbol{\Pi}^{-\mathbf{1/2}}\boldsymbol{\Psi\Lambda}$.

Then for any $r \in \{1, 2, ..., q\}$ th row $(q \leq n)$, we can see that $P\phi_r = \lambda_r\phi_r$, where $\phi_r = \begin{pmatrix}\frac{\psi_r(1)}{\sqrt{\pi(1)}} & \frac{\psi_r(2)}{\sqrt{\pi(2)}} & \cdots & \frac{\psi_r(n)}{\sqrt{\pi(n)}}\end{pmatrix}$.

Therefore for exchangeability (or i.i.d.) of $\mathbf{U}_t$, it suffices to show exchangeability (or i.i.d.) of $\mathbf{P}$.

Assume joint exchangeability of $\mathbf{G}$, i.e. $(A_{ij}) \overset{d}{=} \big(A_{\sigma(i)\sigma(j)}\big)$.

Then $\frac{A_{ij}}{\sum\limits_{ij} A_{ij}} = \frac{A_{ij}}{1+\sum\limits_{l\neq j} A_{il}}$ since A_{ij} is binary. Moreover, A_{ij} and $(1 + \sum\limits_{l\neq j} A_{il})$ are independent given its link function g, and $A_{\sigma(i)\sigma(j)}$ and $(1 + \sum\limits_{l\neq j} A_{\sigma(i)\sigma(l)})$ are independent also given g.

Then the following joint exchangeability of transition probability holds:

$$\big(P_{ij}\big) = \left(\frac{A_{ij}}{1 - A_{ij} + \sum\limits_{j=1}^{n} A_{ij}}\right) \overset{d}{=} \left(\frac{A_{\sigma(i)\sigma(j)}}{1 - A_{\sigma(i)\sigma(j)} + \sum\limits_{\sigma(j)=1}^{n} A_{\sigma(i)\sigma(j)}}\right) = \big(P_{\sigma(i)\sigma(j)}\big) \tag{20}$$

Thus, transition probability is exchangeable. This results exchangeable eigenfunctions $\{\Phi(1), \Phi(2), , ..., \Phi(n)$
where $\Phi(i) := \begin{pmatrix} \phi_1(i) & \phi_2(i) & \cdots & \phi_q(i) \end{pmatrix}^T$. Thus diffusion maps at fixed t, $\mathbf{U}_t = \begin{pmatrix} \Lambda^t\Phi(1) & \Lambda^t\Phi(2) & \cdots & \Lambda^t\Phi(\end{pmatrix}$
are exchangeable.

By de Finetti's Theorem(7.1), we can say that $\mathbf{U}(t) = \{U_1^{(t)}, U_2^{(t)}, ..., U_n^{(t)}\}$ are conditionally independent on a random probability measure η. □

Proof of Theorem 3.1. □

Proof of corollary 3.0.1. □

Let $x, y, z \in V(G)$.

$$
\begin{aligned}
D_t^2(x,z) &= \sum_{w \in V(G)} \big(P^t(x,w) - P^t(z,w)\big)^2 \frac{1}{\pi(w)} \\
&= \sum_{w \in V(G)} \big(P^t(x,w) - P^t(y,w) + P^t(y,w) - P^t(z,w)\big)^2 \frac{1}{\pi(w)} \\
&= \sum_{w \in V(G)} \big(P^t(x,w) - P^t(y,w)\big)^2 \frac{1}{\pi(w)} + \sum_{w \in V(G)} \big(P^t(y,w) - P^t(z,w)\big)^2 \frac{1}{\pi(w)} \\
&+ 2 \sum_{w \in V(G)} \big(P^t(x,w) - P^t(y,w)\big)\big(P^t(y,w) - P^t(z,w)\big) \frac{1}{\pi(w)} \\
&= D_t^2(x,y) + D_t^2(y,z) + 2 \sum_{w \in V(G)} \big(P^t(x,w) - P^t(y,w)\big)\big(P^t(y,w) - P^t(z,w)\big) \frac{1}{\pi(w)}
\end{aligned} \tag{21}
$$

Thus it suffices to show that

$$
\sum_{w \in V(G)} \big(P^t(x,w) - P^t(y,w)\big)\big(P^t(y,w) - P^t(z,w)\big) \frac{1}{\pi(w)} \leq D_t(x,y) \cdot D_t(y,z). \tag{22}
$$

Let $a_w = \big(P^t(x,w) - P^t(y,w)\big)\sqrt{1/\pi(w)}$ and $b_w = \big(P^t(y,w) - P^t(z,w)\big)\sqrt{1/\pi(w)}$. Then the above inequality is equivalent to :

$$\sum_{w \in V(G)} a_w \cdot b_w \leq \sqrt{\sum_{w \in V(G)} a_w^2 \cdot \sum_{w \in V(G)} b_w^2}. \tag{23}$$

,which is true by Cauchy-Schwarz inequality.

References and Notes

1. Hoff, P. D., Raftery, A. E., & Handcock, M. S. (2002). Latent space approaches to social network analysis. Journal of the american Statistical association, 97(460), 1090-1098.

2. Austin, A., Linkletter, C., & Wu, Z. (2013). Covariate-defined latent space random effects model. Social Networks, 35(3), 338-346.

3. Lazarsfeld, P. F, & Henry, N. W. (1968). Latent structure analysis. New York: Houghton, Mifflin.

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

5. Szkely, G. J., Rizzo, M. L., & Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. The Annals of Statistics, 35(6), 2769-2794.

6. Szkely, G. J., & Rizzo, M. L. (2013). The distance correlation t-test of independence in high dimension. Journal of Multivariate Analysis, 117, 193-213.

7. Cencheng at al, Revealing the structure of dependency between multimodal datasets via multiscale generalized correlation.

8. Lyons, R. (2013). Distance covariance in metric spaces. The Annals of Probability, 41(5), 3284-3305.

9. Lee, J. R. (2006). Distance scales, embeddings, and metrics of negative type. University of California, Berkeley.

10. Tang, Minh, and Michael Trosset. ”Graph metrics and dimension reduction.” Indiana University, Indianapolis, IN (2010).

11. Lafon, S., & Lee, A. B. (2006). Diffusion maps and coarse-graining: A unified framework for dimensionality reduction, graph partitioning, and data set parameterization. IEEE transactions on pattern analysis and machine intelligence, 28(9), 1393-1403.

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