Outline of "Testing independence between networks and nodal attributes via multiscale metrics"

Youjin Lee

October 30, 2016

1. Introducing network topology

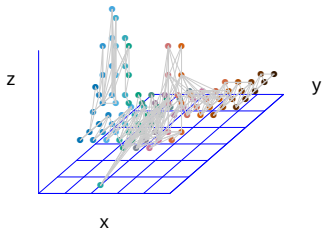

Figure 1: Physical location of one component of human brain network and its tracts that connect one vertex to another.

We introduce a concept of each node's location over their underlying network to bring up the problem of testing independence between distance in terms of network in distance-based test.

2. Multiscale Generalized Correlation

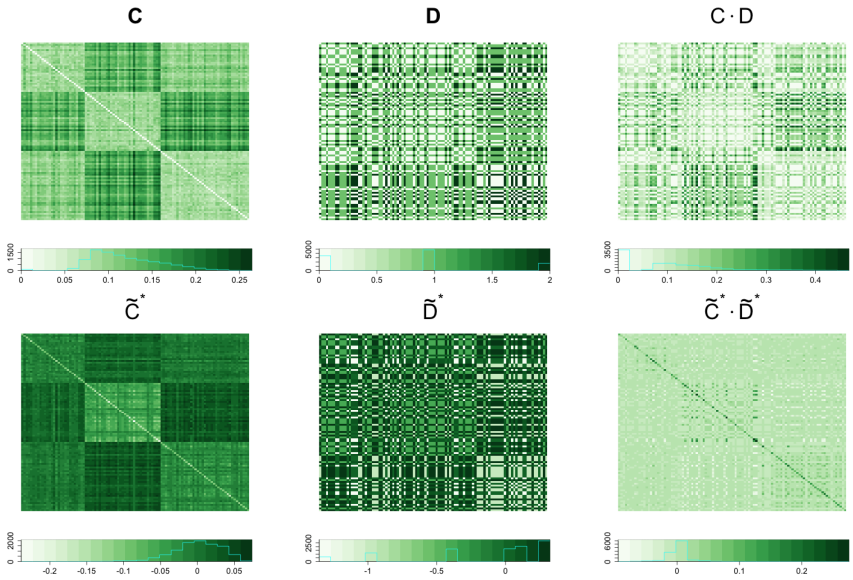

Figure 2: (a) Top : Euclidean distance of diffusion maps at time $t = 3$, C; Euclidean distance of X, D; element-wise product of C and D. (b) Bottom : truncated double-centered $\tilde{C}$ by k^*th nearest neighbor in C; truncated double-centered $\tilde{D}$ by l^*th nearest neighbor in D; element-wise product of two truncated matrices $\tilde{C}^*$ and $\tilde{D}^*$.

As an efficient distance-based test in existence of nonlinear dependency and high-dimensionality, we adopt using local scale distance correlation which truncates each component of distance matrices up to a certain rank.

3. Problems in adopting valid distance metric defined over network

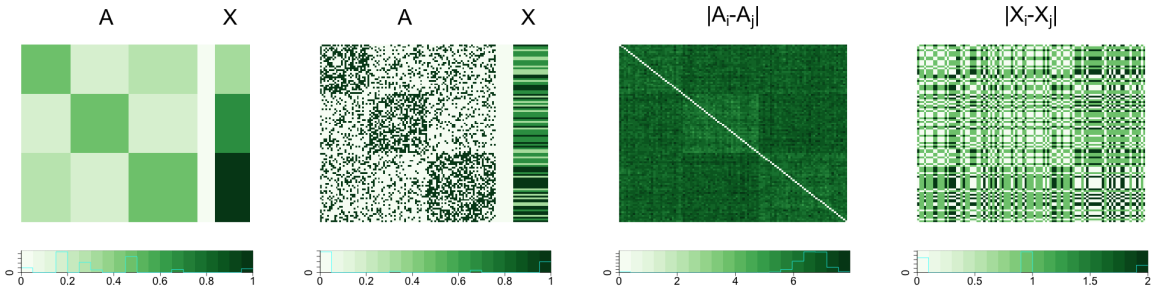

Figure 3: Population probability distribution and realized values of A and X (scaled by 1/3) in the left two panels. Euclidean distance applied to realized A and X are presented in the right two panels.

Every information on edge distribution is denoted in an adjacency matrix so we can consider its Euclidean distance matrix as an ingredient of the test statistics as well as that of nodal attributes.

4. demonstrate the validity of diffusion matrix

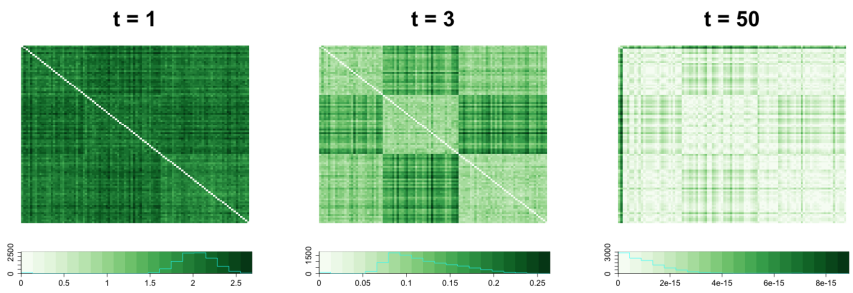

Figure 4: Diffusion distance, i.e. Euclidean distance of diffusion maps at $t = 1$, $t = 3$ and $t = 50$ of sample graph G from Stochastic Block Model.

Out of concerns on theoretical and also practical shortcomings of using an adjacency matrix, we introduce a one-parameter family of network metrics called *diffusion matrix* which keeps every information of adjacent relation and also effectively captures the clustering of networks.

5. Introducing the simulation study

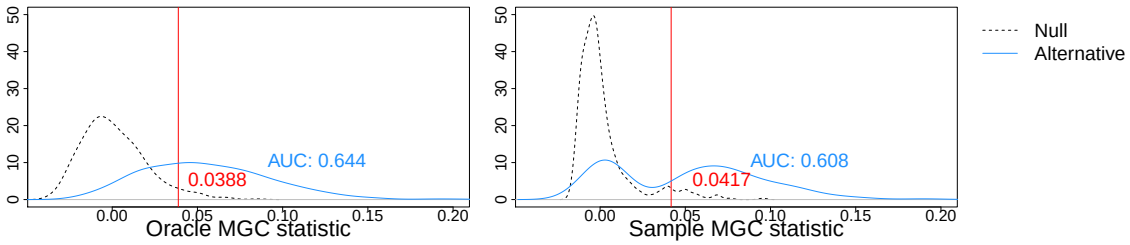

Figure 5: Statistics under null distribution and dependent distribution based on $M = 500$ independently generated SBM with three Blocks

Throughout the simulation study, we are going to make a comparison between the proposed statistics from null simulated networks and also attribute-dependent simulated networks to calculate the empirical power, based on `Oracle MGC` and `Sample MGC`.

6. Simplest Stochastic Block Model

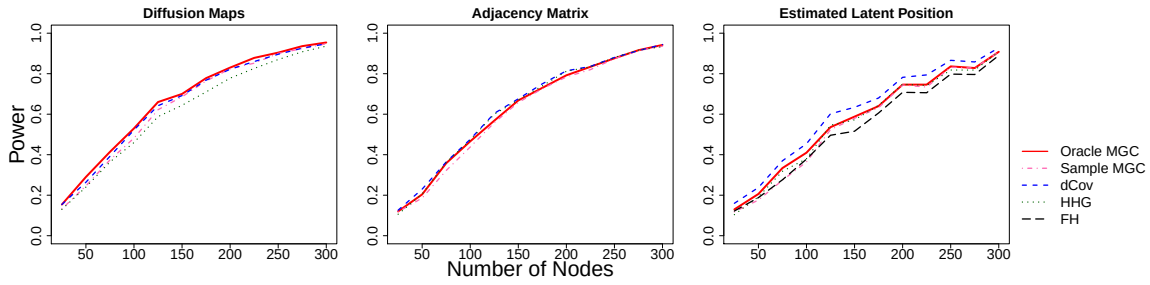

Figure 6: Empirical power based on $M = 500$ independently generated two-block SBM using diffusion maps (left) and Euclidean of adjacency matrix (middle) and estimated latent position(right). The most right figure contains the results of `FH` test as well.

First simplest SBM with two blocks illustrates the typical case of linear-dependence so that we have similar results for each distance-based tests as well as `FH`-test.

7. Stochastic Block Model with non-linearly Dependent Attributes

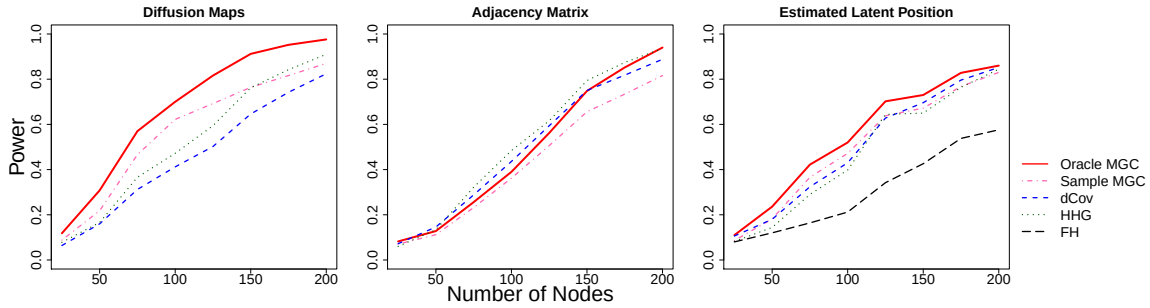

Figure 7: Empirical power based on $M = 500$ independently generated three-block SBM using diffusion maps (left) and Euclidean of adjacency matrix (middle) and estimated latent position(right). The most right figure contains the results of `FH` test as well.

Next SBM is our punch line that `MGC` shows its superiority over other statistics especially in diffusion maps metrics and also results higher power in estimated latent position metric than `FH`.

8. Superiority of the proposed method under non-linear dependency

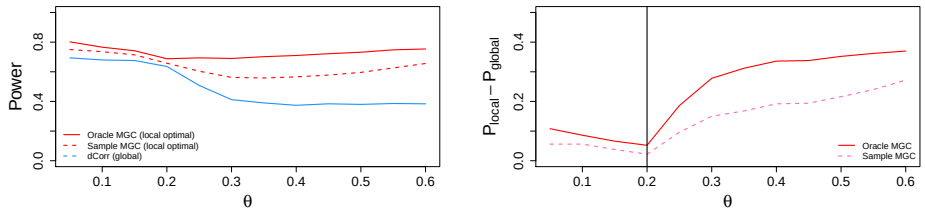

Figure 8: Change of empirical power across θ in both local and global scale of distance correlation (left). Change of difference between these two powers in Oracle and Sample `MGC`. Superiority of optimal local scale become evident from $\theta > 0.2$, when distribution of edges have non-linear dependence on X.

This plot deeps into when exactly our proposed tests exerts better performance; in the existence of non-linearity which can be formalized into conditional distribution of A_{ij} given Euclidean distance between $\mathbf{X}_i$ and $\mathbf{X}_j$.

9. Degree-corrected SBM with increased variability in node distribution

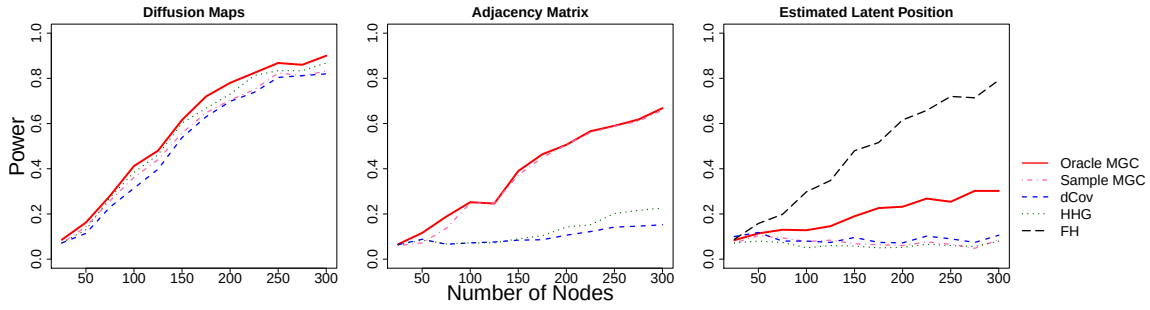

Figure 9: Empirical power based on $M = 500$ independently generated degree-corrected SBM using diffusion maps (left) and Euclidean of adjacency matrix (middle) and estimated latent position(right). The most right figure contains the results of `FH` test as well.

Since previous two SBMs might not demonstrate the real example, we introduce a SBM but with increase variability in edge distribution and especially claim the improved power of diffusion maps when variance in an adjacency matrix is relatively higher.

10. Validity of the method even under competitor's model

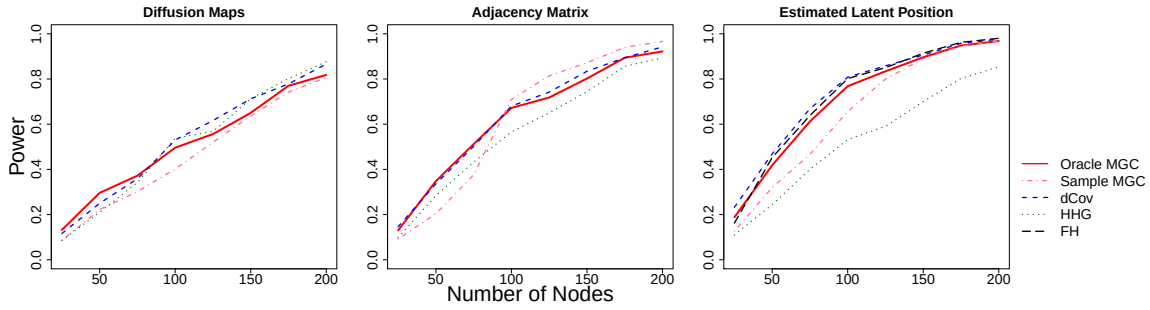

Figure 10: Empirical power based on $M = 500$ independently generated additive and multiplicative graph model using diffusion maps (left) and Euclidean of adjacency matrix (middle) and estimated latent position(right). The most right figure contains the results of `FH` test as well.

However it would be fair to include others' mode-based tests and we can still suggest using estimated latent factors when the model is correct; but within that metrics our method does as good as their method.

11. Node Contribution

Include one plot which shows validity of our measure of node-specific contribution to the tests.