Introducing network topology

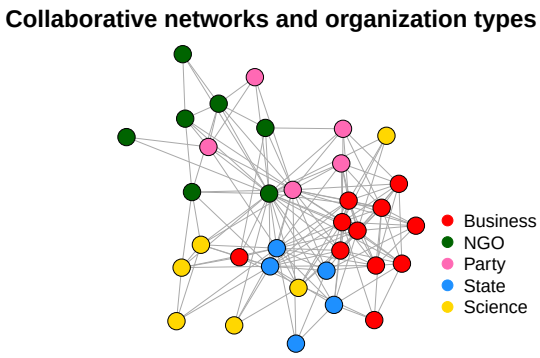


Figure 1: You may conjecture that organizations with the same type are more likely to collaborate each other at first glance; but there has been a lack of statistical method to **test if there exists any significant relationship between network topology and node-specific attributes and if any, which node exerts the most dependency on network.**

Introducing two simple Euclidean distance matrix in the context of network

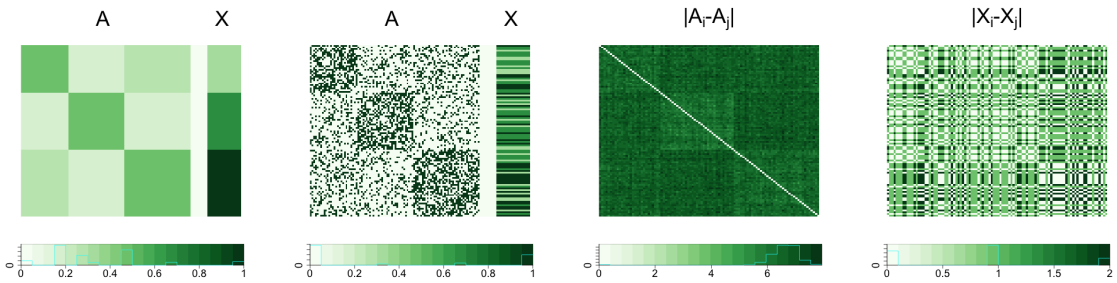


Figure 2: Assume that a set of edges follow certain stochastic block model, also depending on the distribution function of nodal attributes X (a), then with some amount of noise we have a realized adjacency matrix and a set of attribute outcomes (b) of which Euclidean distances (c & d) are suggested to be used in standard distance-based independence test **but neither of them manifests block structures evident in the data generating model.**

Introduce a family of network distance matrices

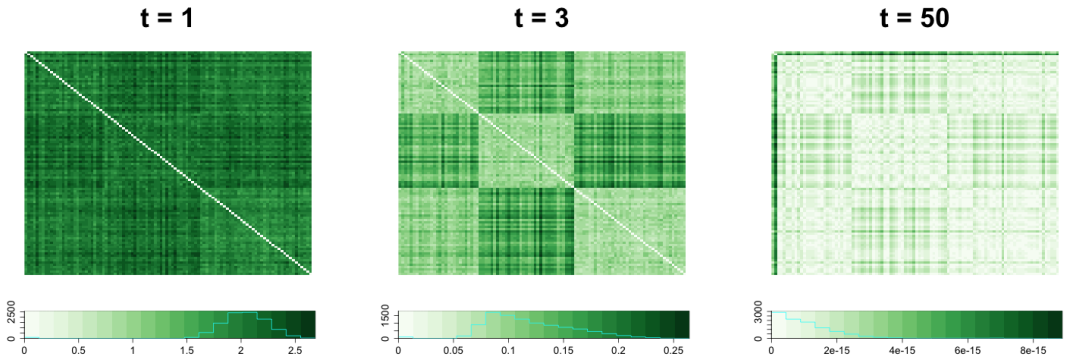


Figure 3: *Diffusion matrix*, as a proposed alternative for Euclidean distance of A, provides **one-parameter family of network-based distances** where at early stage, e.g. at $t = 1$, distance matrix is very similar to Euclidean distance of A but as time goes by the pattern shown in the distance matrix changes, and **at optimal time point $t^* = 3$ distance matrix shows most clear block structures and at the same time it exhibits most dependence to distance matrix of X.**

Empirical power of Oracle MGC/Sample MGC

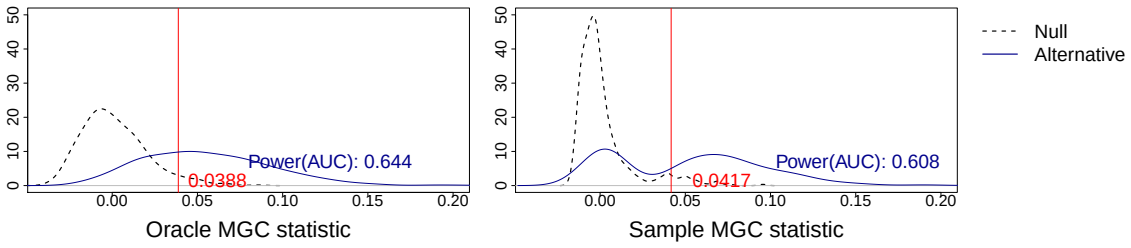


Figure 4: In the left panel we have empirical Null distribution of `Oracle MGC` illustrated by dotted line of which 95% sample quantile determines testing power of `Oracle MGC` by calculating area under the curve (AUC) of the empirical distribution under alternative beyond that quantile, and AUC of `Oracle MGC` (0.644) looks similar to that of `Sample MGC` (0.608), as presented in the right panel, even though the shape of its distributions under null and alternative look different, which supports the use of `Sample MGC` as a substitute for `Oracle MGC` in real data.

Simplest Stochastic Block Model

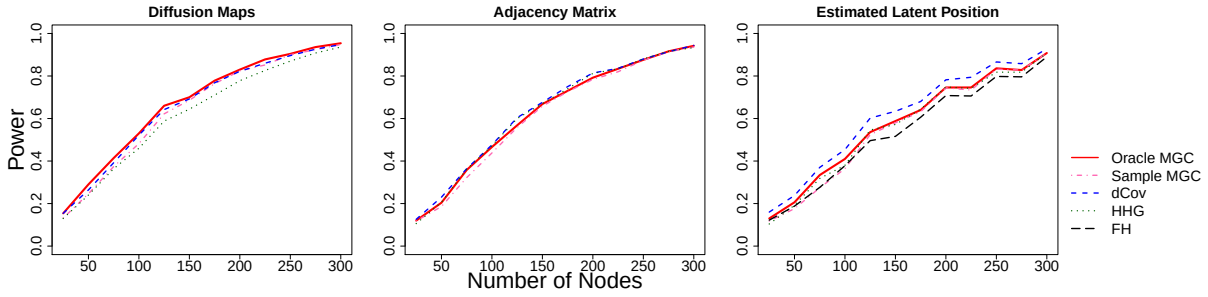
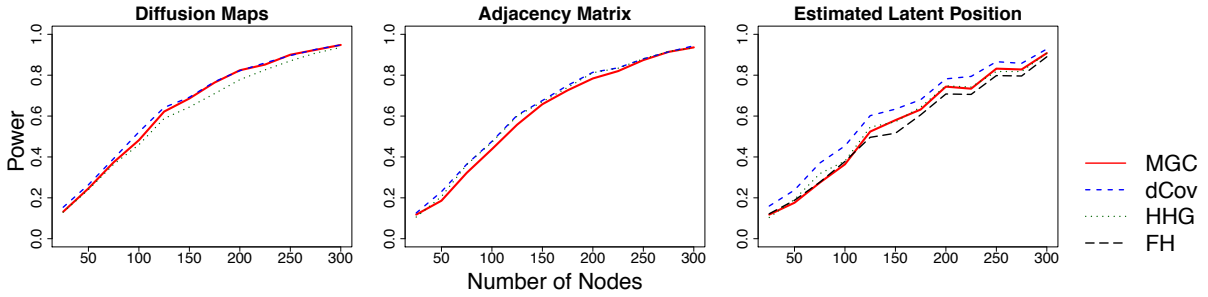


Figure 5: Above three figures are showing power of three different test statistics under two block SBM using diffusion maps, adjacency matrix, and estimated latent position as a network distance measure, and also `FH` test results are shown in the left. **This demonstrates that in the simplest stochastic block model, diffusion distance is generally better than the other two metric; while the performance of `MGC` statistic is very similar to the others in this case.**

Alternate : Let `MGC` = `Sample MGC`.



Highest power of MGC under diffusion maps

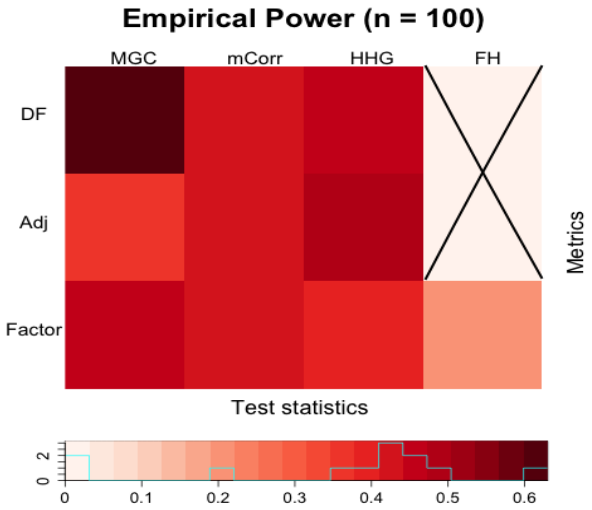


Figure 6: This power heatmap illustrates the superior power of multiscale generalized correlation (`MGC`) under diffusion distance matrix (`DF`) in three SBM (model **??**), compared to under adjacency matrix distance (`Adj`) or latent factor distance (`Factor`). **This demonstrates that especially in the presence of nonlinear network dependency, `MGC` statistic along with a family of diffusion distances catches non monotonic correlations efficiently than the other statistics and metrics.**

Superiority of the proposed method under non-linear dependency

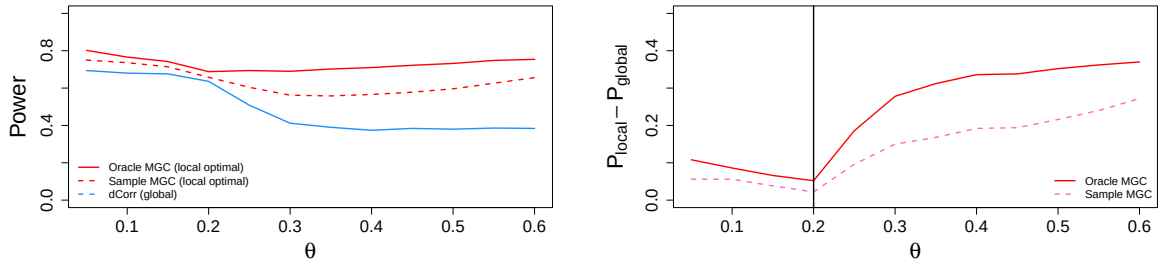
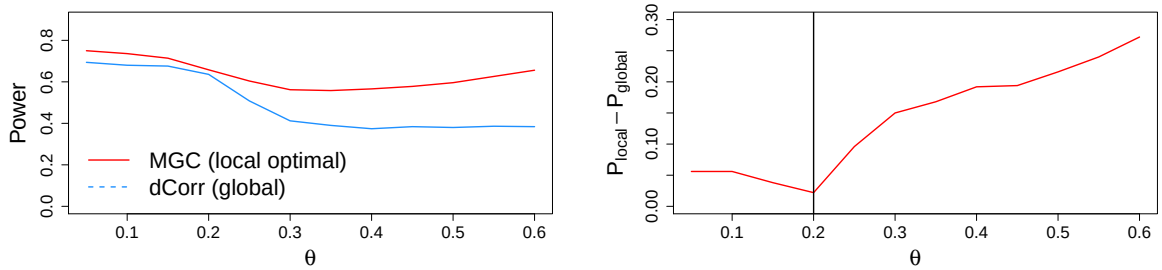


Figure 7: X-axis of θ controls the existence/amount of nonlinear dependency and in this particular case nonlinearity exists when $\theta > 0.2$ and gets larger as it increases. You can see the discrepancy in power between global and local scale tests also gets larger accordingly, **mostly due to decreasing power of global test but relatively stable power of `MGC` under nonlinear dependency** as presented in the left panel.

Alternate : Let `MGC` = `Sample MGC`



Degree-corrected SBM with increased variability in node distribution

Increasing variance in DCSBM

$$\begin{aligned} &\theta_i \overset{i.i.d}{\sim} Uniform(1-\tau, 1+\tau), i = 1, \ldots, n; \quad \tau = 0, 0.2, \ldots, 1 \\ &A_{ij}|\mathbf{Z}, \theta \overset{i.i.d}{\sim} f_{A|Z,\theta}(a_{ij}|z_i, z_j, \theta_i, \theta_j) \overset{d}{=} Bern(0.2 \cdot \theta_i\theta_j) I(|z_i - z_j| = 0) \\ &\quad + Bern(0.05 \cdot \theta_i\theta_j) I(|z_i - z_j| = 1), \quad i, j = 1, \ldots, n; i < j. \end{aligned} \tag{1}$$

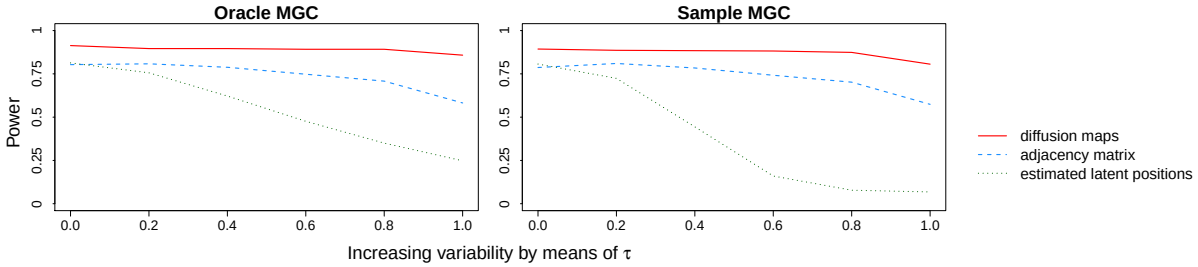
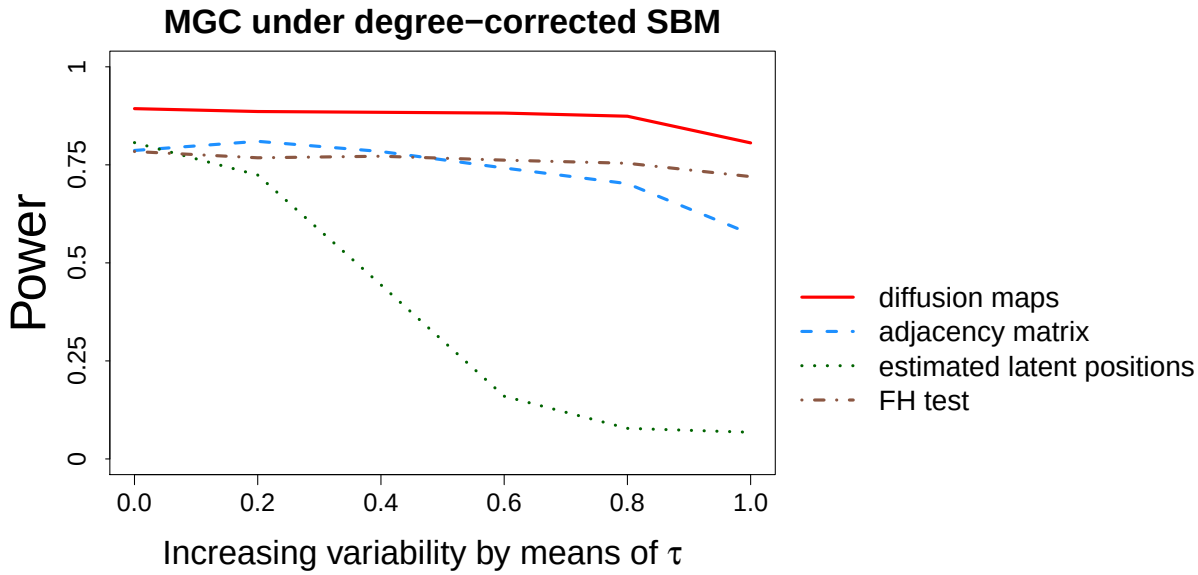


Figure 8: In degree-corrected SBM where the variability in degree distribution increases as τ increases, **power of diffusion maps are more likely to be robust against increasing variability compared to adjacency matrix and latent positions.**

Alternate : Let `MGC` = `Sample MGC`



Validity of the method even under competitor's model

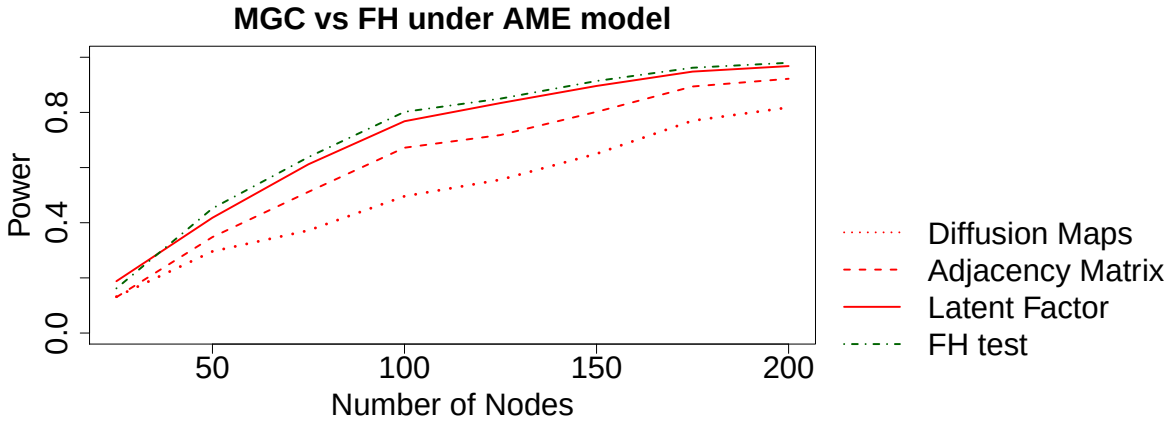


Figure 9: Even under additive and multiplicative model which favors estimated latent position metrics, `MGC` lost some power when using diffusion maps and adjacency matrix but `MGC` **does as good as** `FH` **tests under latent factor metrics which is the truth, which supports excellent ability of** `MGC` **in diverse, nearly true network metrics.**

Node Contribution

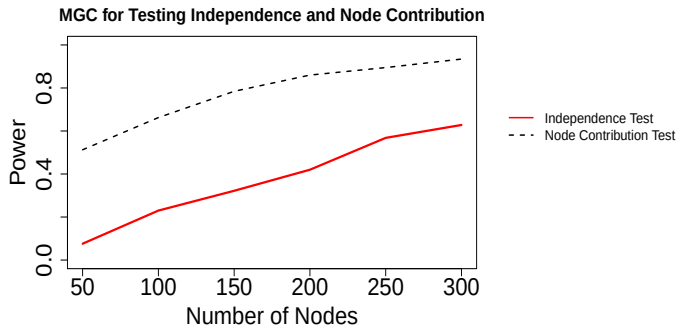


Figure 10: This plot describes both power of `MGC` and the rate of correctly-ranked node contribution increase as the number of nodes increases when only half of the nodes for each simulation actually are set to contribute to the independence test, **which validates the use of node contribution measure in independence test.**

Political Network

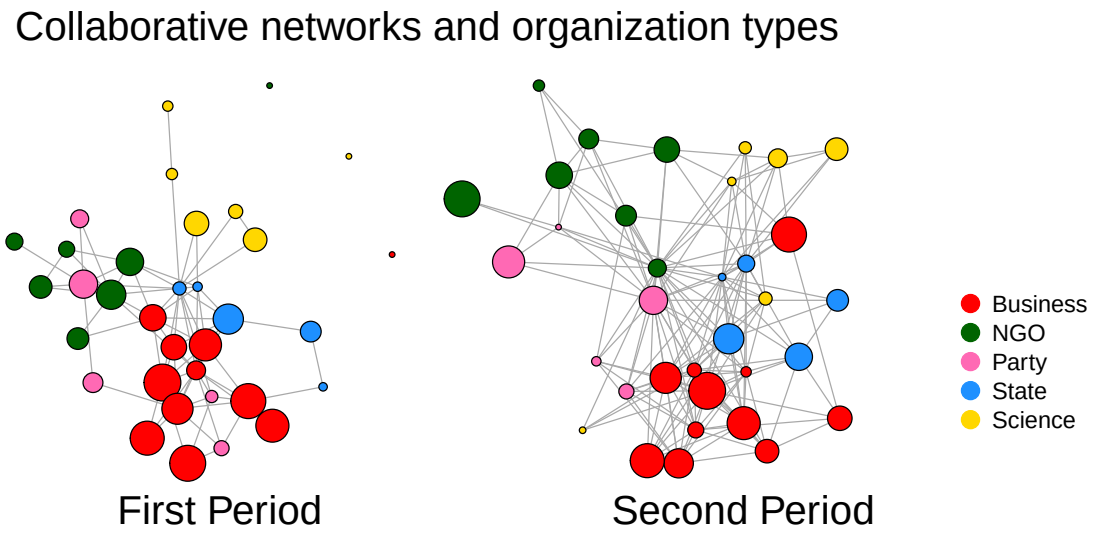


Figure 11: Both networks depict the collaboration network during the two time periods where it turns out significant network dependency in types of organizations. **Using `MGC` statistics, we are not only able to test network independence but also rank each node in terms of the amount of contribution to detecting dependence, which is proportional to node size here.**