Introducing network topology

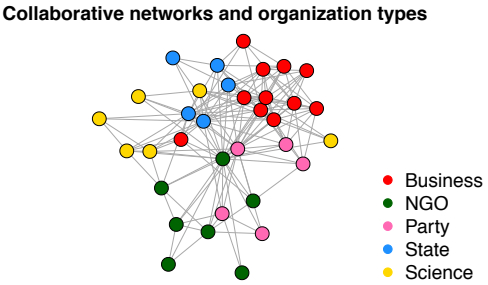


Figure 1: You may conjecture that organizations with the same type are more likely to collaborate each other at first glance; but there has been a lack of statistical method to **test if there exists any significant relationship between network topology and node-specific attributes and if any, which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

Introducing two simple Euclidean distance matrix in the context of network

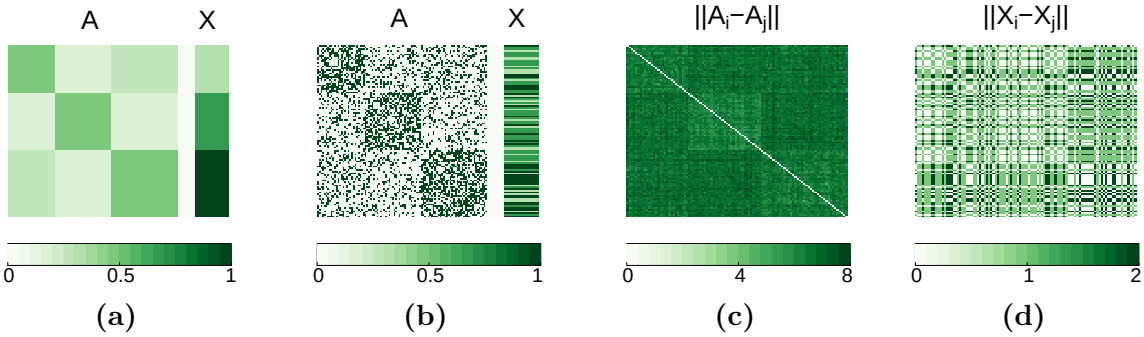


Figure 2: Assume that a set of edges follow certain stochastic block model, also depending on the distribution function of nodal attributes X (a), then with some amount of noise we have a realized adjacency matrix and a set of attribute outcomes (b) of which Euclidean distances ((c) and (d)) are suggested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

Introduce a family of network distance matrices

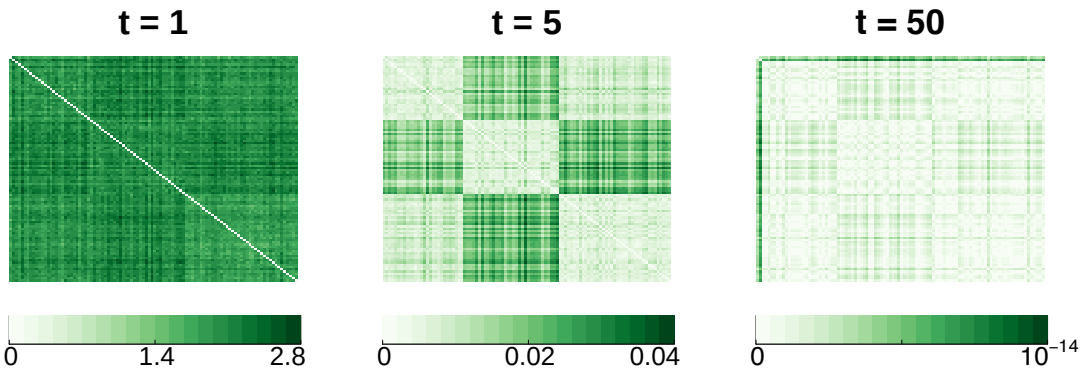


Figure 3: *Diffusion matrix*, as a proposed alternative for Euclidean distance of A, provides **one-parameter family of network-based distances** where at early stage, e.g. at $t = 1$, distance matrix is very similar to Euclidean distance of A but as time goes by the pattern shown in the distance matrix changes, and **at optimal time point $t^* = 5$ distance matrix shows most clear block structures and at the same time it exhibits most dependence to distance matrix of X.**

Highest power of MGC under diffusion maps

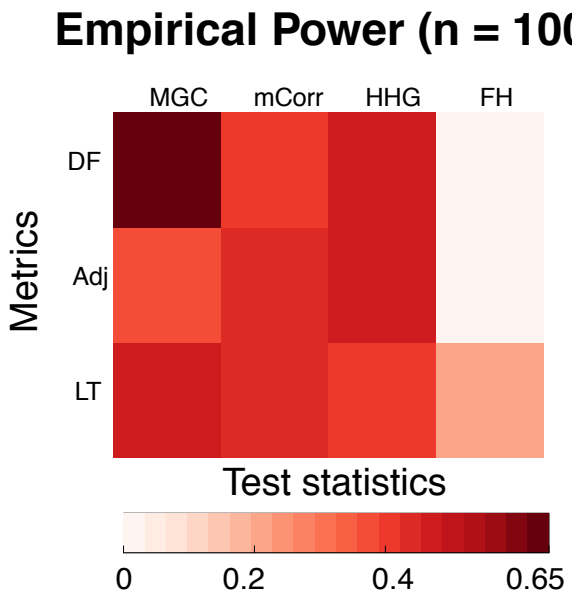


Figure 4: This power heatmap illustrates the superior power of multiscale generalized correlation (`MGC`) under diffusion distance matrix (`DF`) in three SBM (model **??**), compared to under adjacency matrix distance (`Adj`) or latent factor distance (`LT`). **This demonstrates that especially in the 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.**

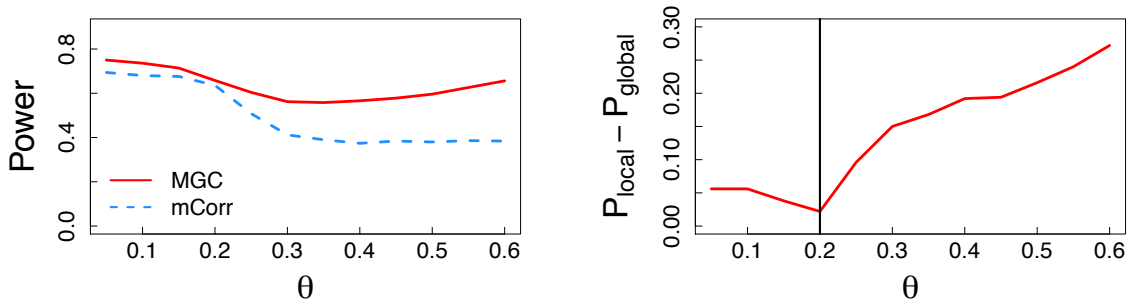
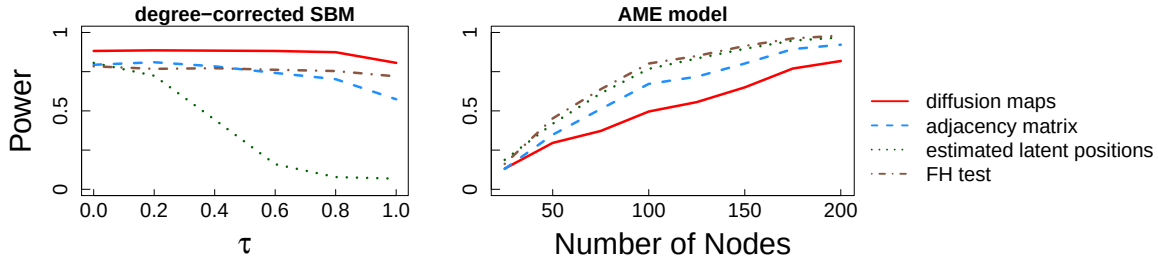


Figure 5: X-axis of θ controls the existence/amount of nonlinear dependency and in this particular case nonlinearity exists when $\theta > 0.2$ and gets larger as it increases. You can see the discrepancy in power between global and local scale tests also gets larger accordingly, **mostly due to decreasing power of global test but relatively stable power of `MGC` under nonlinear dependency** as presented in the left panel.

Superiority of the proposed method under non-linear dependency

Degree-corrected SBM with increased variability in node distribution

Validity of the method even under competitor's model



(a) **(b)**

Figure 6: (a) 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`. (b) `MGC` utilizing diffusion distances loses some power under additive and multiplicative model which favors estimated latent position metrics, but `MGC` does as good as `FH` tests under latent factor metrics which closes to the truth. This reveals the flexibility in distance-based matrix in `MGC` statistics, which can be chosen depending on model fit or preliminary knowledge.

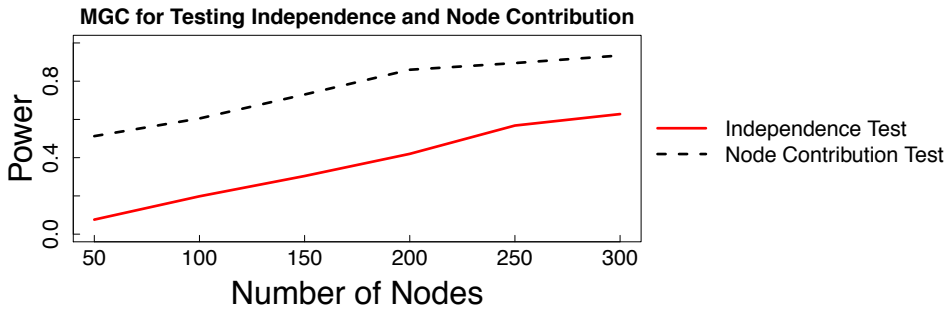


Figure 7: This plot describes both power of `MGC` and the rate of correctly-ranked node contribution increase as the number of nodes increases when only half of the nodes for each simulation actually are set to contribute to the independence test, **which validates the use of node contribution measure in independence test.**

Node Contribution

Political Network

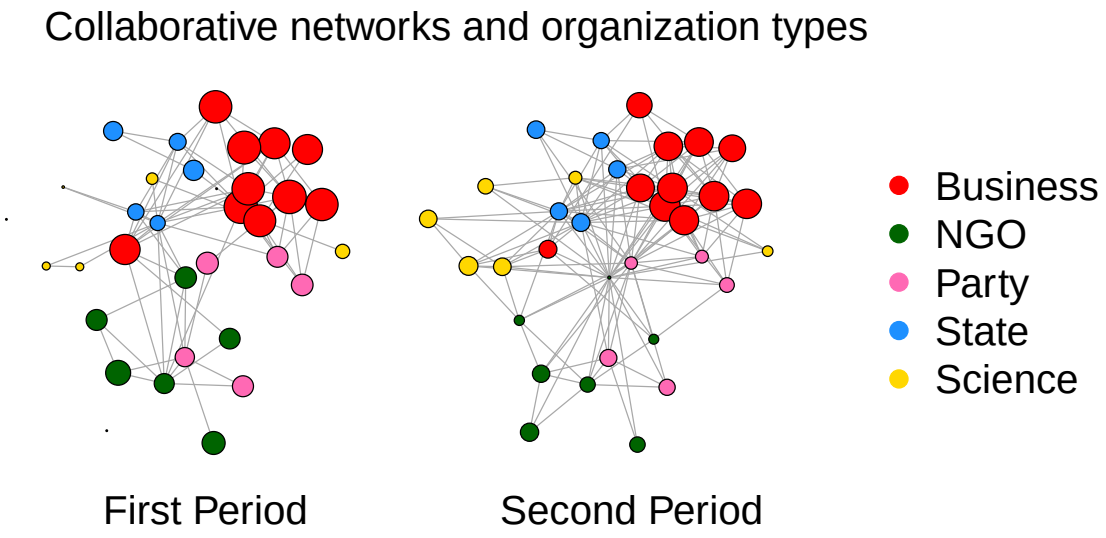


Figure 8: Both networks depict the collaboration network during the two time periods where it turns out significant network dependency in types of organizations. **Considering the size of nodes is proportional to the contribution to the statistic, you can notice that throughout all periods business sector is most actively collaborating within their group.**

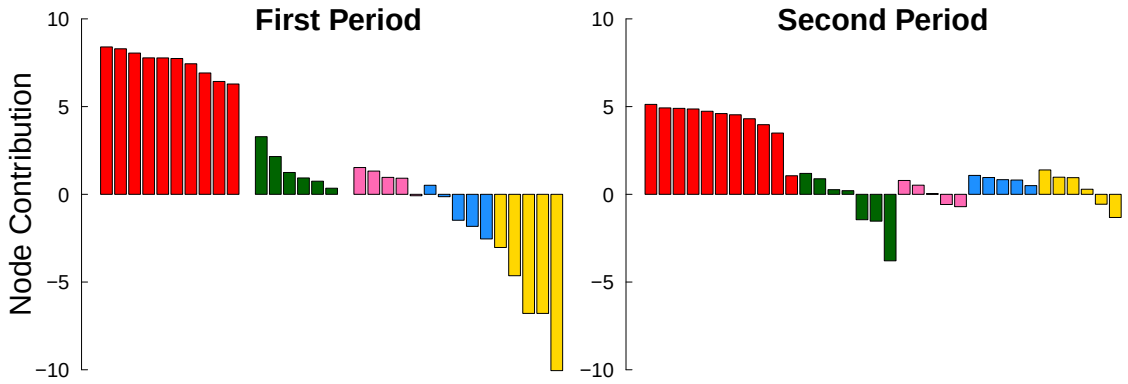
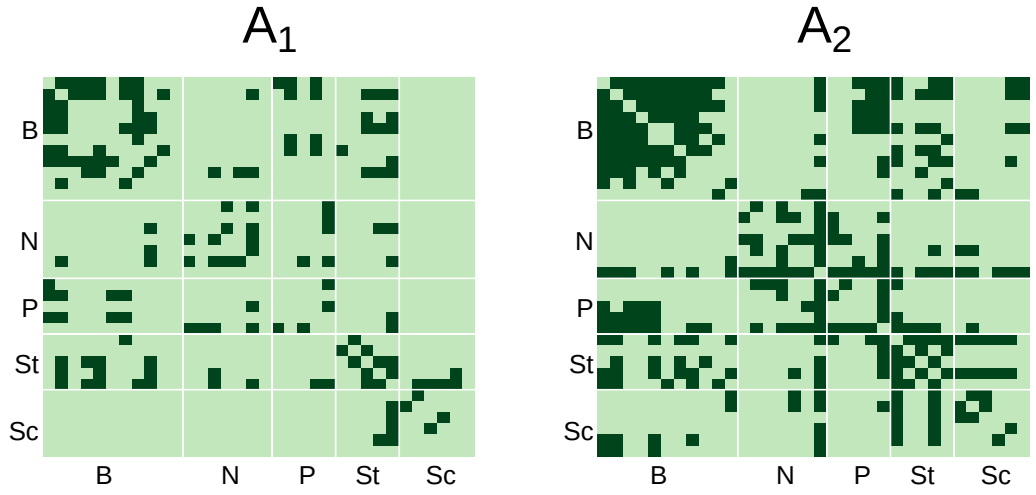


Figure 9: You can tell that collaboration within the same type is strongest among the Business group while scientist relatively collaborates less with any others, and generally organizations co-work more actively between different types but still their collaboration network is highly dependent on their organization types.

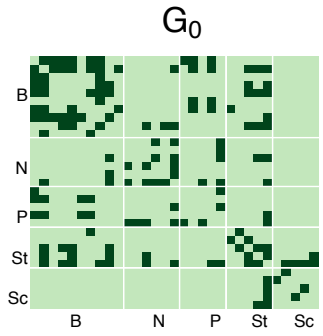


(a) Adjacency matrix of the first period **(b)** Adjacency matrix of the second period

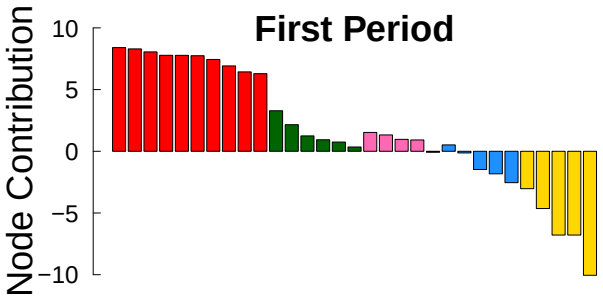
Figure 10: In general we have a denser adjacency matrix in the second period. Business organizations (B) are, except the last node, more likely to interact within their group, which becomes clearer in the second period. On the other hand, NGO (N), Party (P), and State (St) are relatively randomly collaborating within and between groups while Science (Sc) group exhibits very few edges so that their edge distribution just concentrates on two groups in the first period.

Political Network

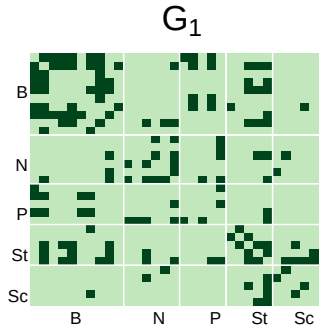
Adding edges one by one to Science group



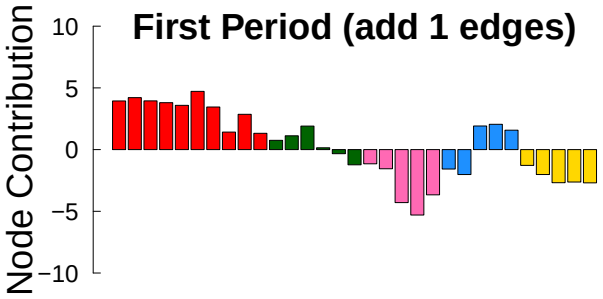
(a) Adjacency matrix of the first period (original)



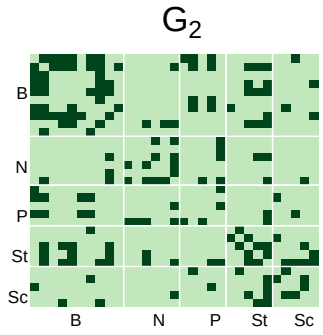
(b) Node contribution



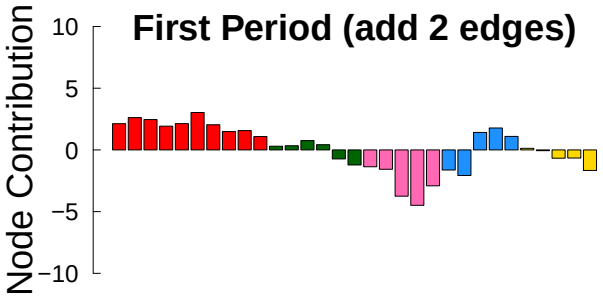
(a) Adjacency matrix of the first period (add one edge)



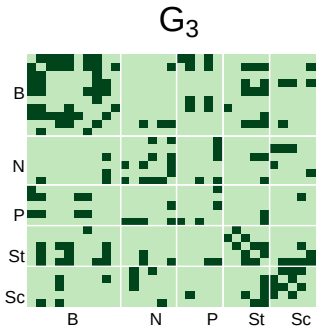
(b) Node contribution



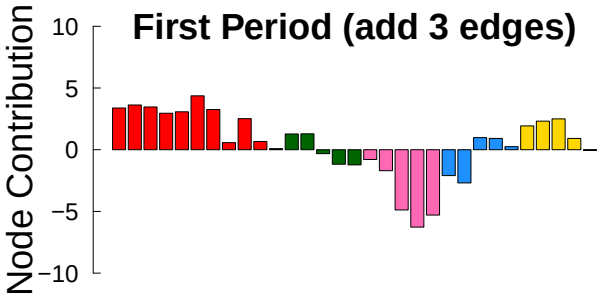
(a) Adjacency matrix of the first period (add 2 edges)



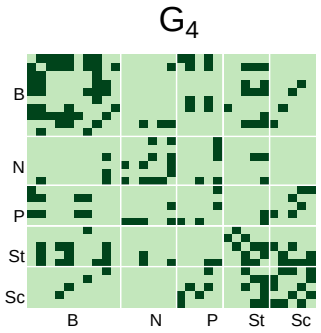
(b) Node contribution



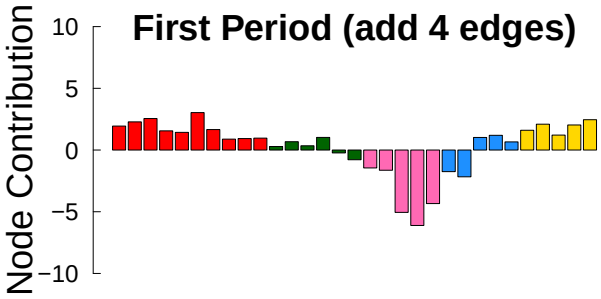
(a) Adjacency matrix of the first period (add 3 edges)



(b) Node contribution



(a) Adjacency matrix of the first period (add 4 edges)



(b) Node contribution

Simulations

Group 3 having sparse distribution

$$X_i \overset{i.i.d}{\sim} f_X(x) \overset{d}{=} Multi(1/3, 1/3, 1/3), i = 1, \ldots, n$$

$$Z_i|X_i \overset{i.i.d}{\sim} f_{Z|X}(z|x) \overset{d}{=} Multi(0.5, 0.25, 0.25)I(x = 1) + Multi(0.25, 0.5, 0.25)I(x = 2) + Multi(0.25, 0.25, 0.5)I(x = 3), \quad i = 1, \ldots, n$$

$$A_{ij}|Z_i, Z_j = \left[\begin{array}{c|c|c} 0.5 & 0.2 & 0.05 \\ \hline 0.2 & 0.5 & 0.05 \\ \hline 0.05 & 0.05 & 0.1 \end{array}\right] \tag{1}$$

In this case, similar to politic network example, I treated X as a categorical random variable having a value of 1, 2, or 3 and use its dissimilarity matrix.

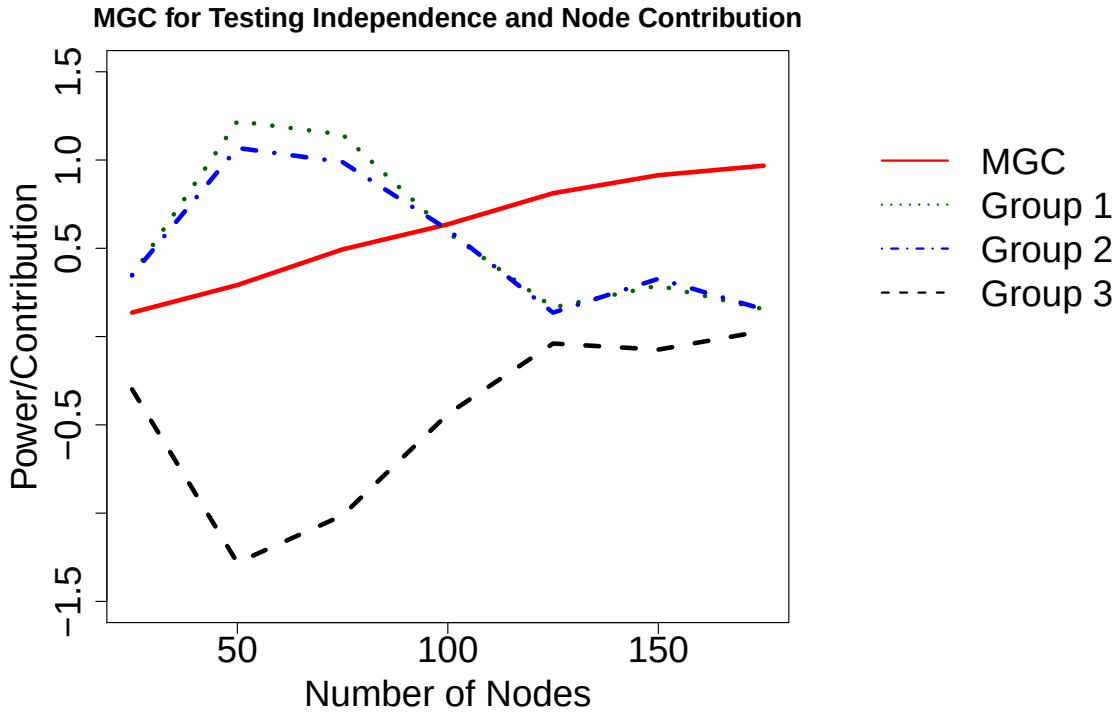


Figure 16: Group 3 has sparse edge distribution compared to Group 1 or Group2.

Group 3 having more relationship to other groups

$$X_i \overset{i.i.d}{\sim} f_X(x) \overset{d}{=} Multi(1/3, 1/3, 1/3), i = 1, \ldots, n$$

$$Z_i|X_i \overset{i.i.d}{\sim} f_{Z|X}(z|x) \overset{d}{=} Multi(0.5, 0.25, 0.25)I(x = 1) + Multi(0.25, 0.5, 0.25)I(x = 2) + Multi(0.25, 0.25, 0.5)I(x = 3), \quad i = 1, \ldots, n$$

$$A_{ij}|Z_i, Z_j = \left[\begin{array}{c|c|c} 0.5 & 0.1 & 0.4 \\ \hline 0.1 & 0.5 & 0.4 \\ \hline 0.4 & 0.4 & 0.2 \end{array}\right] \tag{2}$$

In this case, similar to politic network example, I treated X as a categorical random variable having a value of 1, 2, or 3 and use its dissimilarity matrix.

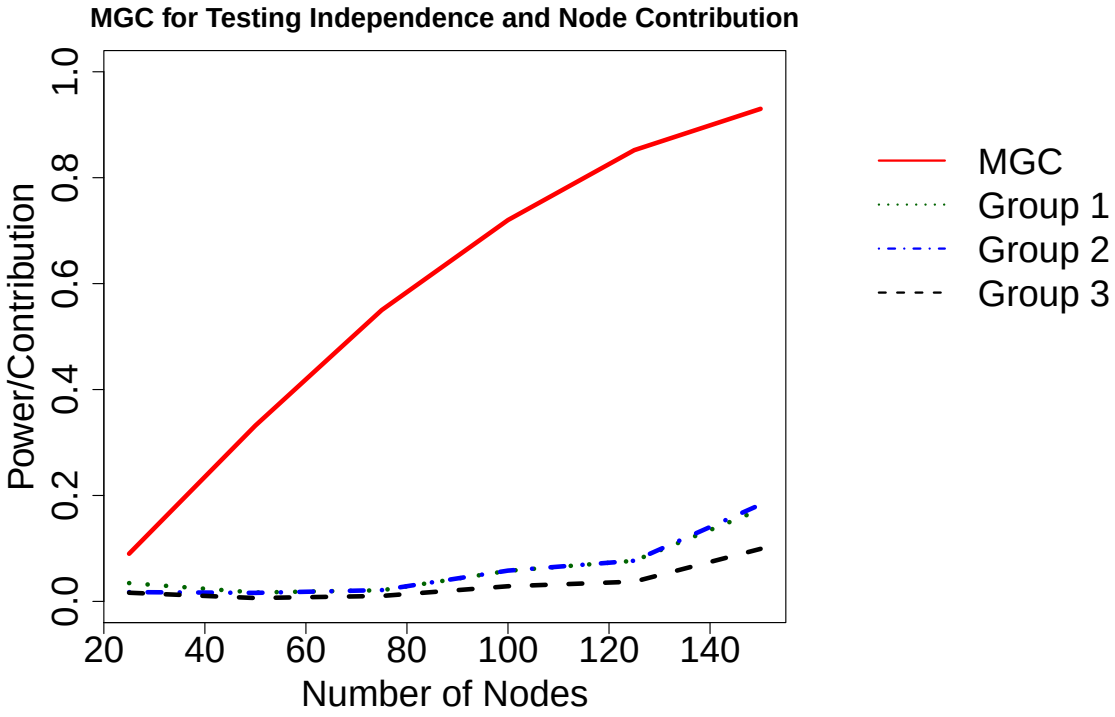


Figure 17: Group 3 has different network patterns compared to Group 1 or Group2.

1 Checklist

- **1. 1 sentence summary**

  We propose the nonparametric method to test independence between network and nodal attributes using multiscaled generalizing correlation with family of network-based distances applied.

  → **title** : Testing independence in networks via family of network metrics

- **2. 1 paragraph summary**

  (i) big opportunity sentence: Network data are ubiqutous.

  (ii) specific opportunity : We are interested in investigating relationship between network topology and a random variable associated with each node.

  (iii) challenge sentence : Network topology is high-dimensional data and there is often no standardized structure on network.

  (iv) gap sentence : Network model-based approach fails to deal with increasing information having no standard pattern of network.

  (v) action sentence : Multiscale generalized correlation (MGC) statistic successfully detects the nature dependence between two data sets and we are going to utilize a family of network-based distances when implementing MGC in testing independence.

(vi) resolution sentence : As a result we have multiscale test statistics which provide the evidence of dependence between network and nodal attributes, more robust to nonlinear dependence and increased variability compared to the other statistics and other metrics.

- **5 paragraph intro**

(i) bulleted list of 3-5 main factors that created an opportunity for your work.

- Statisticians have long considered the problem of revealing the relationship between two different data sets.

- Among the data having diverse types and dimensions, network is now an ubiquitous dataset which is very likely to possess the properties of both high dimensionality and nonlinearity.

- Beyond the interest in network topology, we consider the relationship between network topology and a random variable associated with each node.

(ii) one sentence summary of the gap, that is, the key ingredient that is missing.

Network model-based statistics presumed that all nodes would exhibit the same pattern of dependence on their network relationship as specified in the model, which cannot be assured in network data.

(iii) bulleted list of 3-5 main challenges that must be overcome.

- Whatever the true nature of network is and however network is correlated with its nodal attributes, our testing statistics should aim at detecting significant association between network and nodal attributes.

- We have to consider increasing amount of information inherent in network data as the number of nodes increases.

- There exists an dependency among columns of an adjacency matrix so theoretically we cannot directly use an Euclidean distance of an adjacency matrix.

2-3 sentence summary of what you did

First of all, considering the properties of network data, we propose applying multiscale generalized correlation statistics into testing network independence which deal with correlation between high-dimensional data sets which might have nonlinear dependence. However we confront with the difficulty in finding the valid iid node-specific coordinates of which Euclidean distance will be an ingredient for MGC statistics. We claim that diffusion maps derived from random walk on graph furnish robust network metrics against to increased variability in edge distributions as well as they satisfy theoretical constraints for a jointly exchangeable graph.

2-3 sentence summary on how your work changes the world.

Beyond the traditional setting of the data in independence test, i.e. two random vectors, we suggest the method to investigate unique but also ubiqutous data of network and its nodal attributes by utilizing the method testing two random vectors.

By suggesting using a family of network metrics which are equipped with favored theoretical and practical properties, we are able to better understand relative locations of each node over network and also how the amount of dependency changes along with diffusion time. Moveover now away from the assumption on globally existing monotonic dependence, MGC with diffusion distances performs better than the other statistics or metrics under nonlinear dependence and we measure each node's contribution to detecting the dependence.

- **Outline of the results : list the evidence that supports that you filled the gap**

  Through simulation studies, we demonstrate that superiority of `MGC` statistics becomes more evident when the relationship between network and nodal attributes is not linear. Moreover diffusion distance matrix maintains its stability under increased variability in network better than an adjacent matrix or latent network factors from model-based methods.

- **outline of discussion, to include:**

  (i) bulleted list of previous related work

  - We proposed the method to test independence in network data without model assumption nor estimations.

  - Our methods work better than the others especially under nonlinear dependency, and we are able to measure each node's contribution to detecting dependence.

  (ii) bulleted list of potential extensions

  - Since we did not suggest the method to derive the optimal diffusion time point but rather choose the one which shows highest power from $t = 1$ to $t = 10$, we can further study the algorithm that chooses the optimal diffusion time or diffusion matrix.

  - Both `MGC` and a family of diffusion distances are specialized in preserving local relationship between two data set. Based on this strength, we may preserve the true relationship between network and nodal attributes through investigating multiscale test statistics.

  - We may utilize the diffusion maps from two networks and use them to test independence of network topology of two different networks with same nodes.