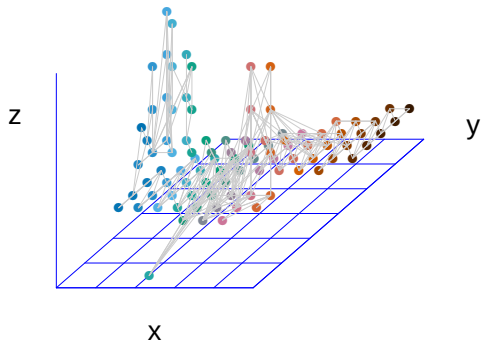
Testing independence between networks and nodal attributes via multiscale metrics

Youjin Lee

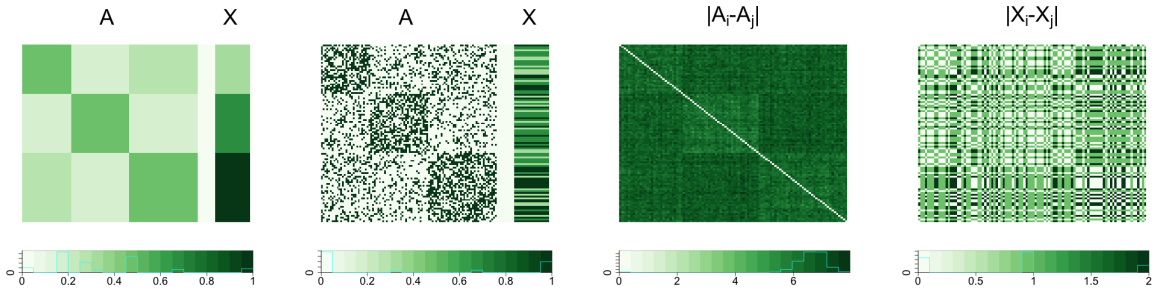
November 7, 2016

1. Introducing network topology



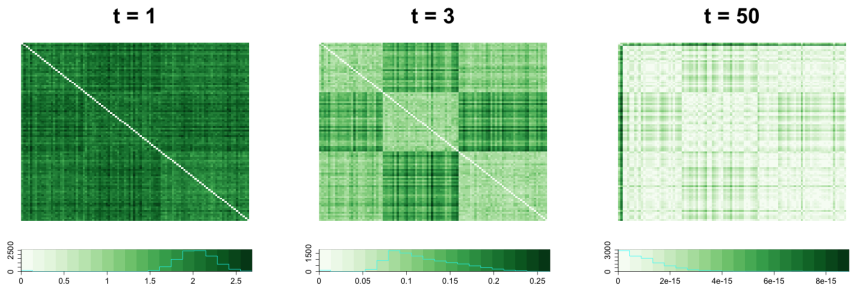
You can embed each subject (node) connected via edges upon Euclidean space, e.g. xyz 3D-space as above, according to their possessing attributes, e.g. physical location; while how to embed subjects from network into Euclidean space is not intuitive, which should consider their distance with respect to network relationship.

2. Problems in adopting valid distance metric defined over network



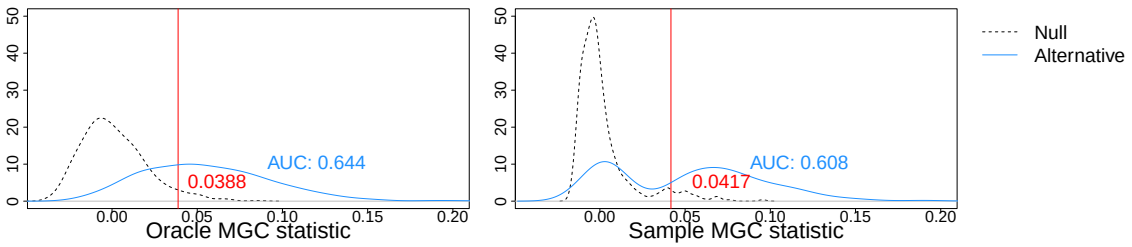
Assuming that a set of edges $\{A\}$ are distributed following certain stochastic block model, also depending on the distribution function of nodal attributes X(a), then with some amount of noise we have a realized adjacency matrix and a set of attribute outcomes(b) of which Euclidean distances (c)(d) are constructed to be possibly used in standard distance-based independence test.

3. introduce a family of network distance matrices



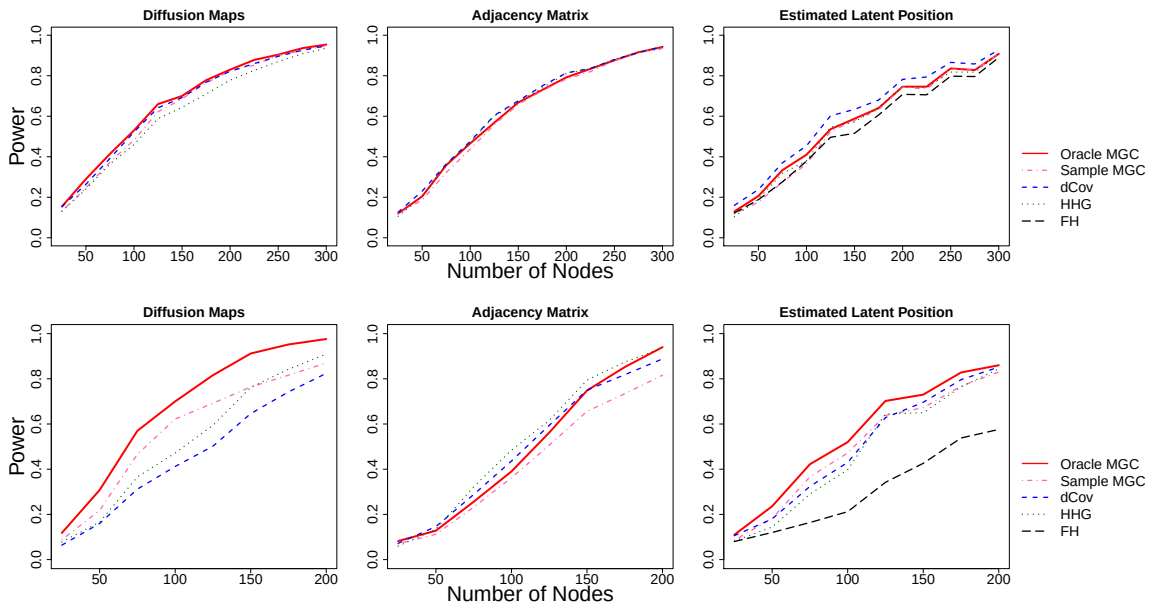
Diffusion matrix, as a proper alternative for Euclidean distance of A, provides one-parameter family of network-based distance where at early stage, e.g. at $t = 1$, distance matrix is very similar to Euclidean distance of A by taking into account very close relationship but as time goes by the pattern shown in the distance matrix changes, depending on time (t) spent on diffusion process.

4. Empirical power of Oracle MGC/Sample MGC



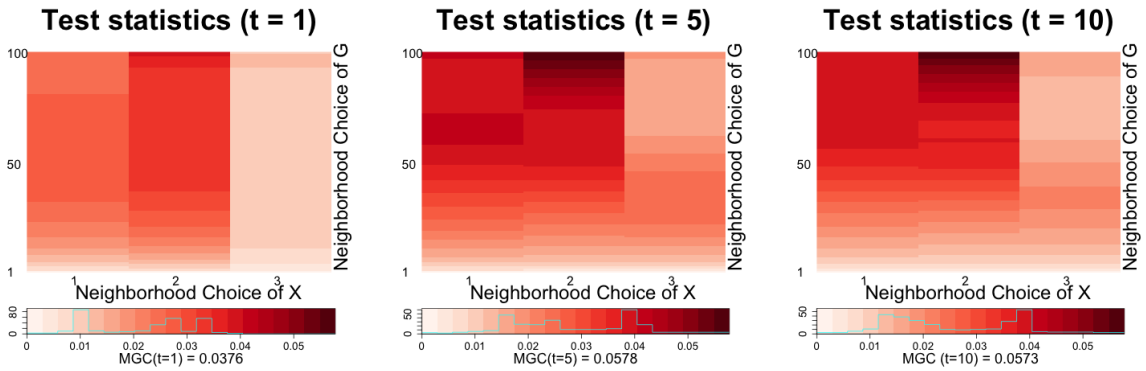
In the left panel we have empirical Null distribution of `Oracle MGC` illustrated by black line of which 95% sample quantile determines testing power of `Oracle MGC` by calculating area under the curve (AUC) of the empirical distribution under alternative beyond that quantile, and AUC of `Oracle MGC` (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.

5. Simplest Stochastic Block Model



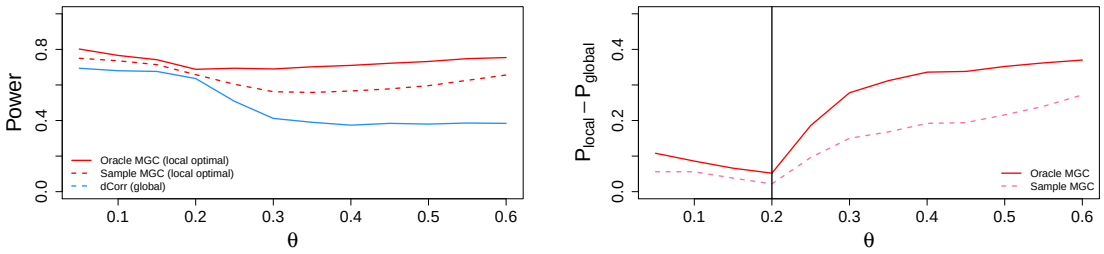
Top three figures are showing power of network independence test under two block SBM using diffusion maps, adjacency matrix, and estimated latent position as a network distance measure where results based on three metrics are similar as well as `MGC`, global distance-based tests (`dCov, HHG`) and `FH` tests are; whereas under three block SBM diffusion maps generally perform better than the other metrics across all type of distance-based tests and moreover `FH` test results worse than the others even in the metrics of latent position.

7. MGC statistic



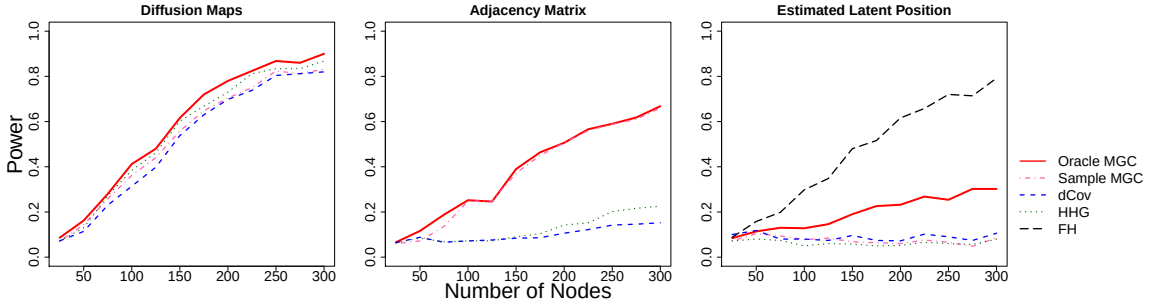
Under non-linear three SBM model, we obtain a family of multiscale statistic as a function of diffusion time t as above where for each time $t \in \{1, 5, 10\}$ we chose the optimal scale statistics $(k^*, l^*) = (100, 2)$ which is not global and of which superiority that the other local scales becomes distinct from $t = 1$ to $t = 5$.

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



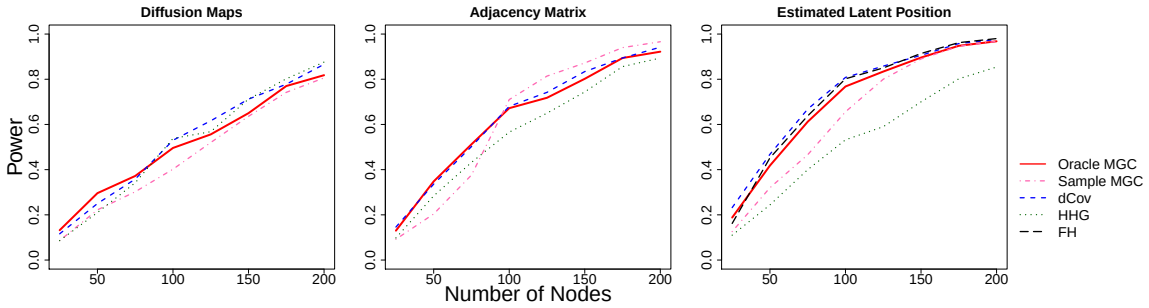
A x-axis of θ controls the existence/amount of non-linear dependency and in this particular case non-linearity exists when $\theta > 0.2$ and gets larger as it increases, and you can see the discrepancy in power between global and local scale tests also gets larger accordingly, mostly due to decreasing power of global test under non-linear dependency as presented in the left panel.

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



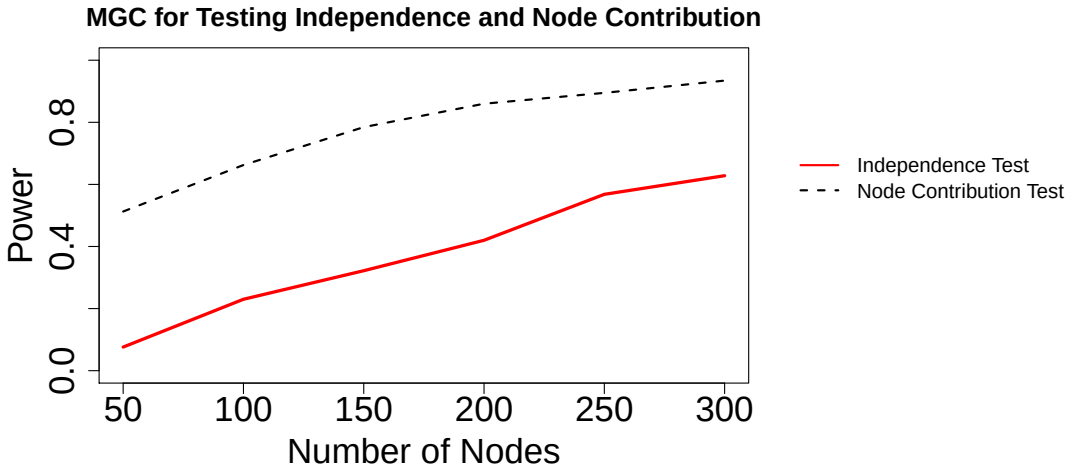
In DCSBM, the performance of all distance-based tests under distance maps used are similar but under adjacency matrix metric, global distance-based tests fail to detect the dependency due to an increased variability in A, which also possibly explains why `FH` outperforms the others under also more variable latent position metric.

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



Under additive and multiplicative model, where no specific test methods beat the others significantly, estimated latent positions provide most sensitive power for `MGC` and `FH`, and they result similar power even though this model is deliberately designed to favor `FH` tests.

11. Node Contribution



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.