Data Integrated Stochastic Block Models

Carter Allen; Dongjun Chung, PhD
Medical University of South Carolina Department of Public Health Sciences

ABSTRACT

A fundamental objective in analysis of genomic data is characterization of **gene networks** (hub genes and sub-networks) related to a given disease.
In many complex disease areas, signal from genomic experiments is **weak and widespread**. This poses a challenge for standard statistical network models, such as the stochastic block model (SBM).

We show that the issue of weak and widespread signal can be addressed through **data integration**.

MODEL

Definition: A random graph $\mathbf{G}$ is said to follow an $\mathbf{SBM}(n, \mathbf{P}, \mathbf{b})$ if

a. $\mathbf{G}$ has n nodes (vertices) denoted by $\mathcal{N} = \{\eta_1, \eta_2, \ldots \eta_n\}$.

b. Each node has exactly one label, denoted b_i for $i = 1,...,n$.

c. An edge exists between nodes η_i and η_j with probability $\mathbf{P}_{b_i,b_j}$.

Proposition: We propose **edge union**: a data integration scheme for the SBM.

1. Let $\mathbf{G}_1$ and $\mathbf{G}_2$ be two observed networks on the same set of nodes (genes).
2. Define $\mathscr{E}_1$ and $\mathscr{E}_2$ as the sets of edges in $\mathbf{G}_1$ and $\mathbf{G}_2$, respectively.
3. Form $\mathbf{G}$, the data-integrated network, by setting $\mathscr{E} = \mathscr{E}_1 \cup \mathscr{E}_2$, where $\mathscr{E}$ is the set of edges in $\mathbf{G}$.

SIMULATION STUDIES

We assess the performance of edge union data integration through simulation studies.

1. For each simulation, sample $\mathbf{G}_1$ and $\mathbf{G}_2$ from an $\mathrm{SBM}(n, \mathbf{P}, \mathbf{b})$. Each network has $B = 3$ true clusters.
2. Perform edge union data integration to obtain $\mathbf{G}$.
3. Fit Bayesian SBMs to each graph. Use MCMC sampling to obtain estimates of model parameters.
4. Plot posterior distributions of parameters of interest.

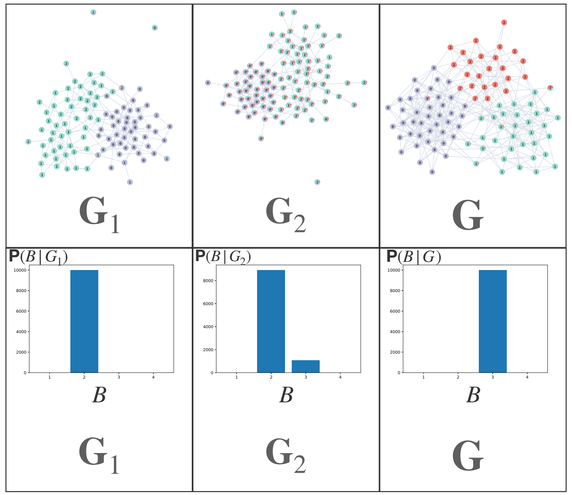
The Bayesian SBM models fit to $\mathbf{G}_1, \mathbf{G}_2$, and $\mathbf{G}$ are used to estimate the true number of communities B, and the community membership of each node $b_1, \ldots, b_n$.

We plot the posterior probabilities $\mathrm{P}(b_i \,|\, \mathbf{G})$ as pie charts on each node to assess node (gene) assignment performance.

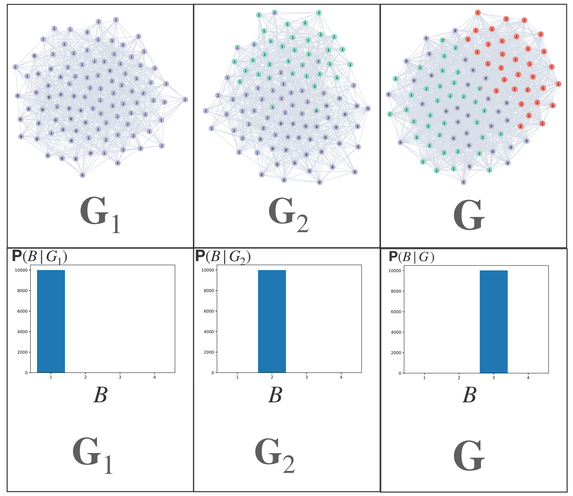
We plot the posterior probability $\mathrm{P}(B \,|\, \mathbf{G})$ to assess ability to estimate model dimension (number of sub-networks).

RESULTS

Sim. 1 Sparse networks with strong signal.



Sim. 2 Dense networks with weak signal.



FURTHER RESULTS

We implement two alternative data integration methods:

1. **Multigraph SBM:** We form $\mathbf{G}$ by combining all edges in $\mathbf{G}_1$ and $\mathbf{G}_2$, allowing for multiple edges between any two nodes.
2. **Weighted SBM:** We construct $\mathbf{G}$ with Binomial edge weights corresponding to the number of times the edge appeared in $\mathbf{G}_1$ and $\mathbf{G}_2$.

These approaches tend to overestimate B.

CONCLUSIONS

1. Data integration allows for reliable inference in the case of weak and widespread signal.
2. Compared to alternative approaches to data integration, edge union offers better performance.
3. Bayesian SBMs allow for quantification of uncertainty in model parameter estimates.

We plan to implement our proposed method to study hub genes and sub-networks related to systemic sclerosis (SSc).

REFERENCES

Peixoto, Tiago P. "Nonparametric weighted stochastic block models." *Physical Review E* 97.1 (2018): 012306.

Peixoto, Tiago P. "Nonparametric Bayesian inference of the microcanonical stochastic block model." *Physical Review E* 95.1 (2017): 012317.

Aicher, Christopher, Abigail Z. Jacobs, and Aaron Clauset. "Adapting the stochastic block model to edge-weighted networks." *arXiv preprint arXiv:1305.5782* (2013).

van der Pas, S. L., and A. W. van der Vaart. "Bayesian community detection." *Bayesian Analysis* 13.3 (2018): 767-796.

TL1 TR001451 & UL1 TR001450