

# Data Integrated Stochastic Block Models for Network Analysis of Genomic Data

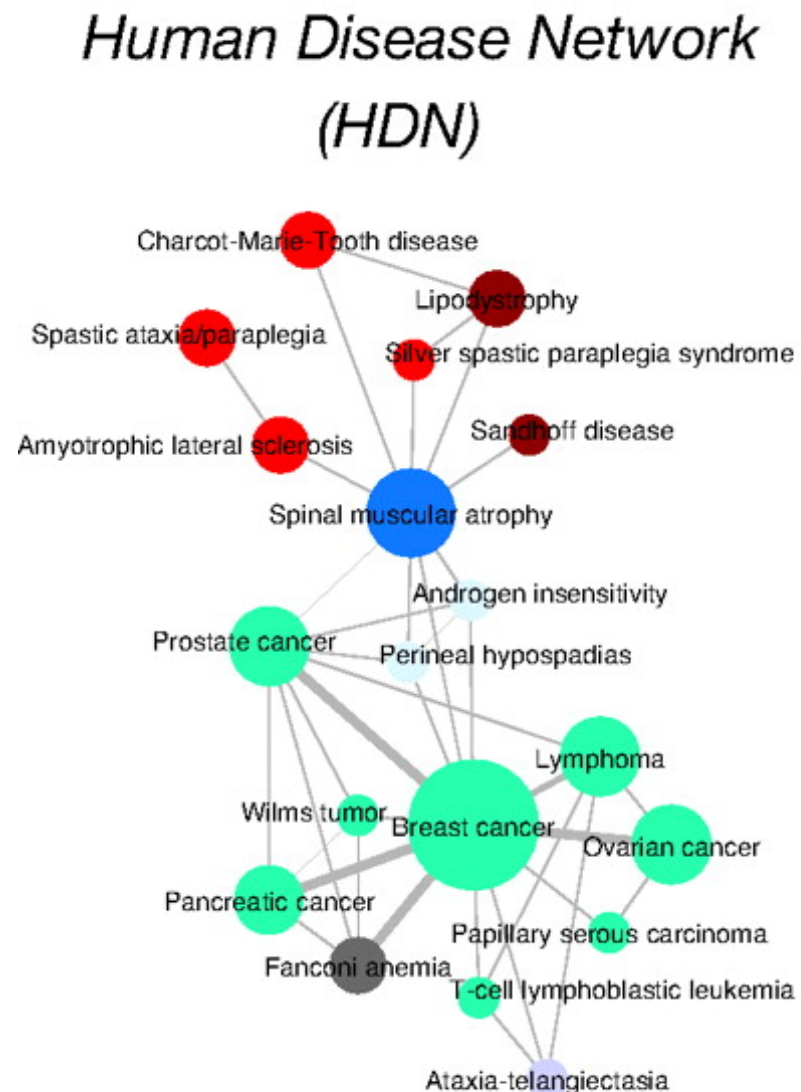
Carter Allen

Mentor: Dr. Dongjun Chung

# Introduction

- Characterization of **gene networks** is a fundamental objective in genomic studies.
  1. **Community detection** allows us to identify gene sub-networks.
  2. Detection of **hub genes** is important to facilitate understanding of biological mechanisms and develop targeted therapies.
- For complex diseases, **weak and widespread** signal presents challenges to these research goals.
- We propose to address this issue through **data integration**.

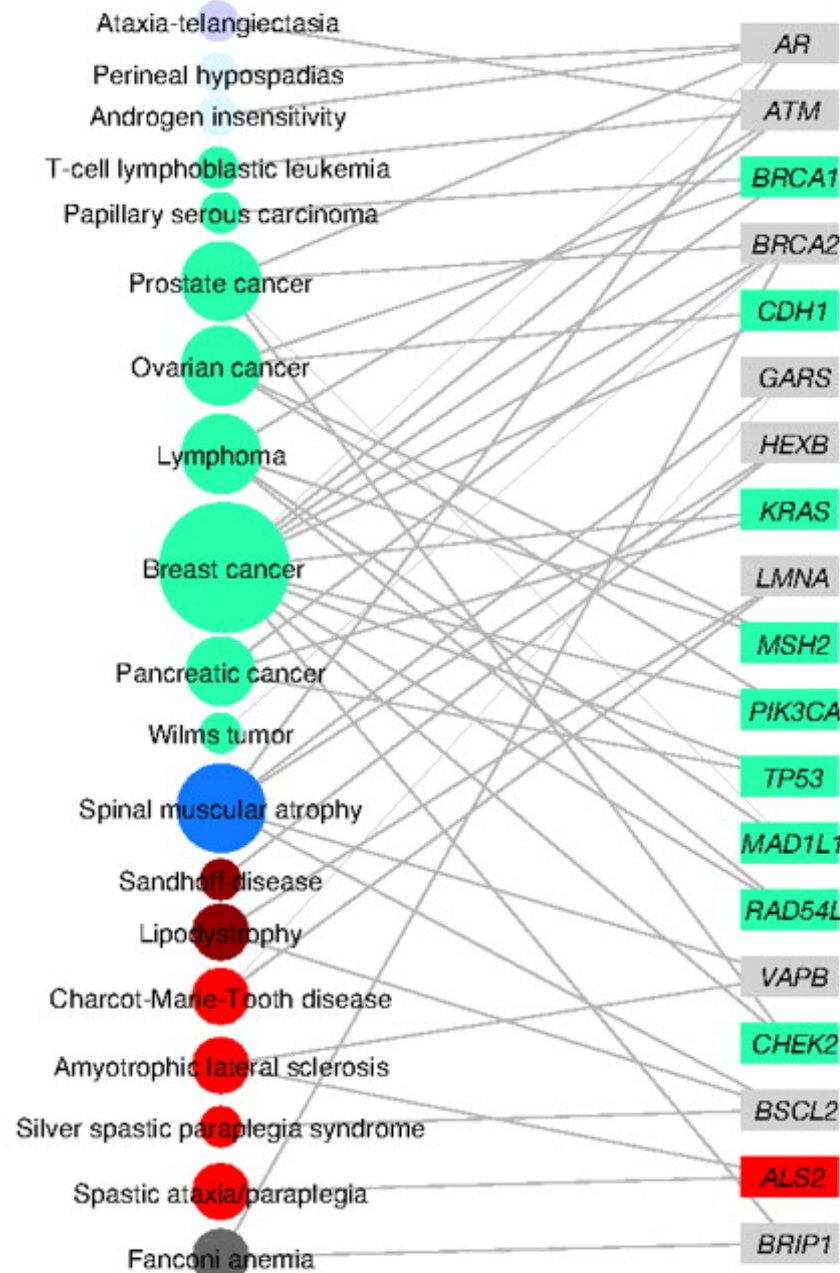
# Biological Networks



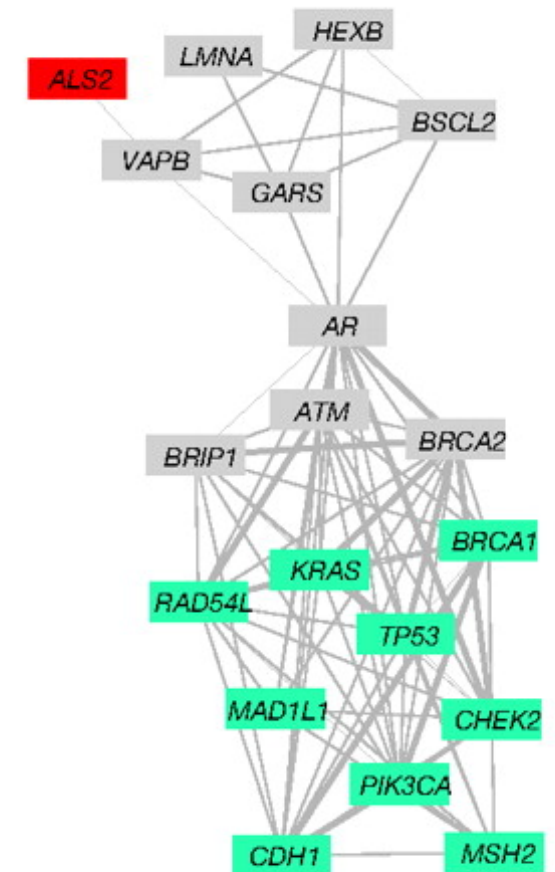
## DISEASOME

### disease phenome

### disease genome

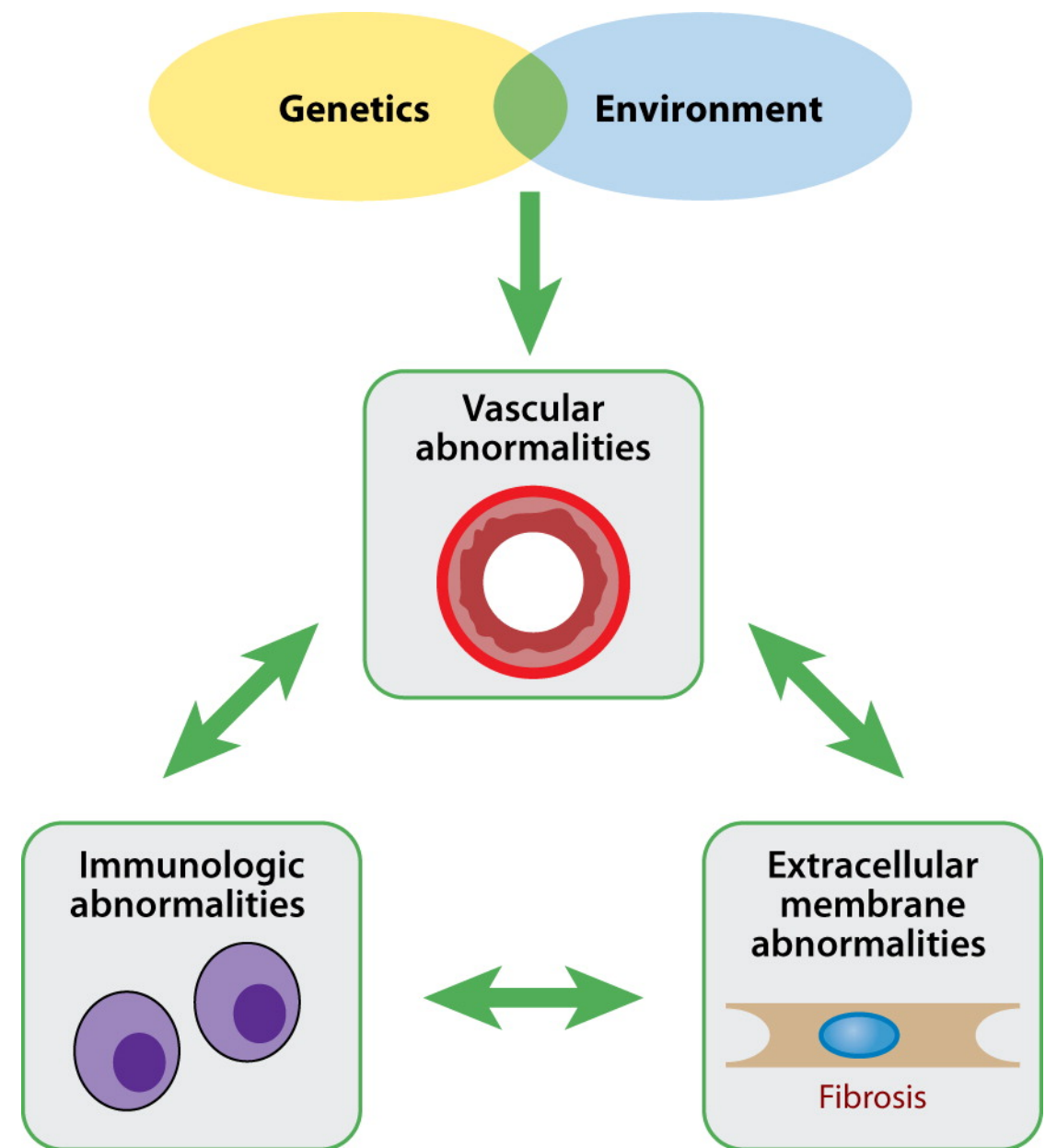



## Disease Gene Network (DGN)



# Motivation

- **Systemic sclerosis (SSc)** is an autoimmune disease involving multiple body systems.
- The genetic origins of SSc are weak and widespread.
- Our work is motivated by data from the **only cohort of twins with SSc** to date (Feghali-Bostwick *et al.*, 2003, *Arthr. & Rheum.: Official Journal of the ACR*).
- Multiple experiments on the SSc twin cohort have generated several sources of data.



 Katsumoto TR, et al. 2011.  
Annu. Rev. Pathol. Mech. Dis. 6:509–37

# Background

- The **stochastic block model** (SBM) is a generative model for network data (Holland *et al.*, 1983, Social Networks).
- Naturally allows for inference about community structure of a graph.
- Certain SBM variants allow for detection of hub nodes (genes) and subnetworks simultaneously within a unified framework.
- Bayesian SBMs allow for incorporation of prior information (Nowicki & Snijders, 2001, JASA).

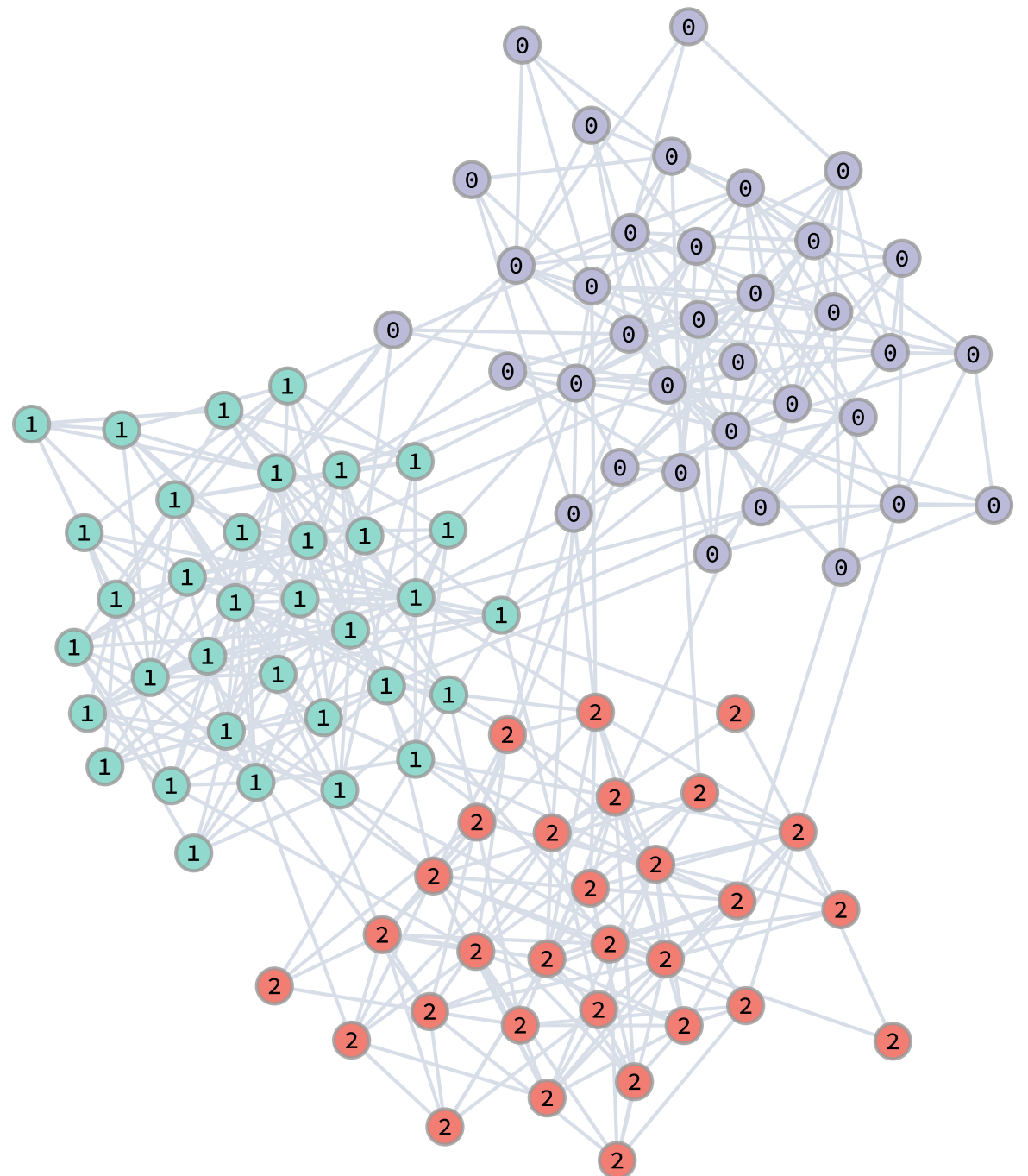


# Network Data

Suppose the observed data takes the form of an  $n \times n$  adjacency matrix  $\mathbf{A}$ , where  $n$  is the number of nodes (genes).

For **simple graphs**,  $A_{ij} = 1$  if an edge exists between nodes  $i$  and  $j$  and  $A_{ij} = 0$  otherwise.

For **symmetric graphs**,  $A_{ij} = A_{ji}$ . Thus edges are undirected.

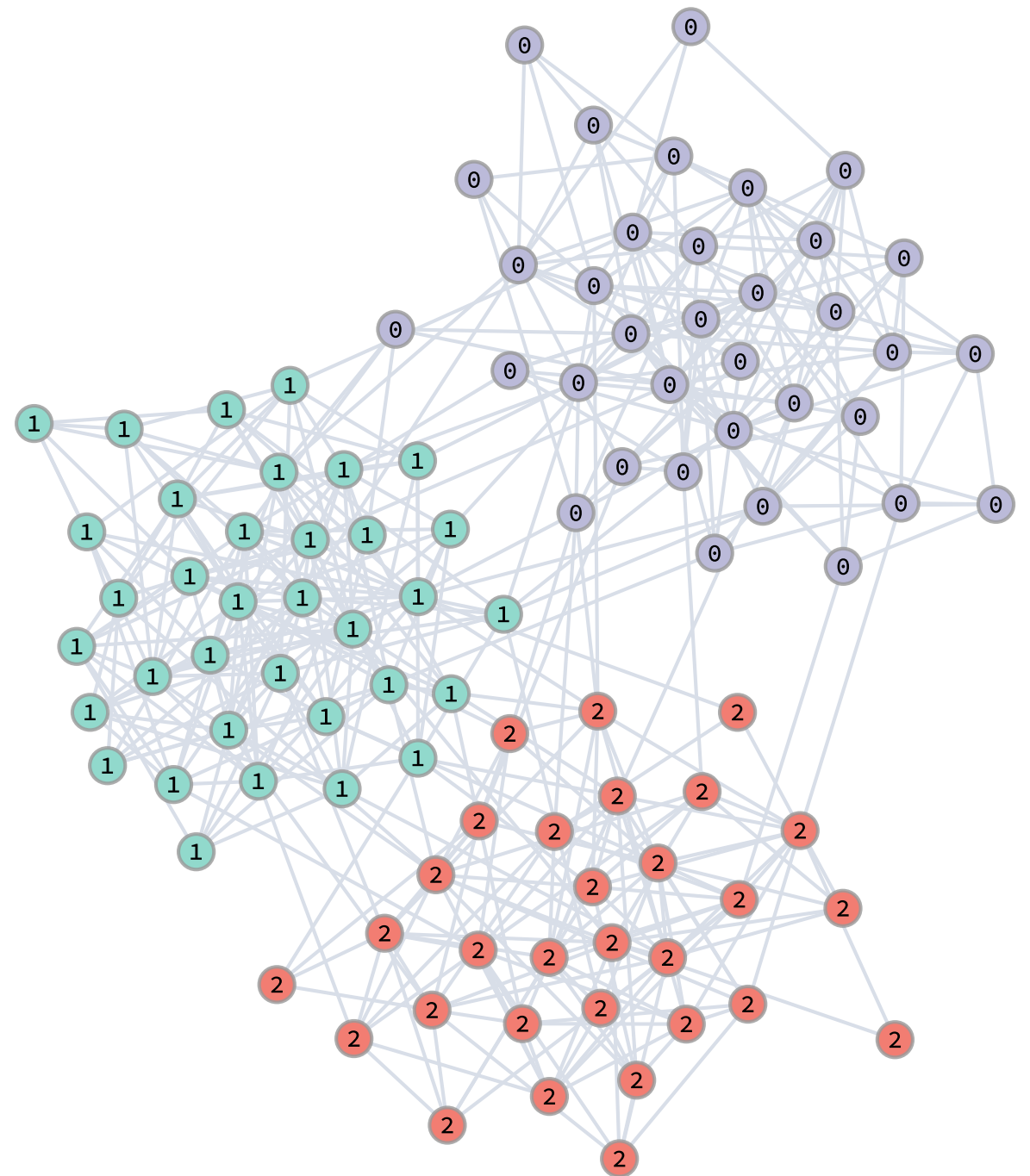


# Network Data

Community detection involves assigning nodes (genes) to labels (colors).

Connectivity (placement of edges) is expected to be dense within communities and sparse between communities.

We would like a statistical model to infer the community labels of a given graph.



# Bayesian SBM

The SBM assumes the probability of an edge between nodes  $i$  and  $j$  **depends only** on  $\sigma_i$  and  $\sigma_j$ . A simple model for  $A_{ij}$  is

$$A_{ij} \mid \boldsymbol{\sigma}, \boldsymbol{\theta} \sim \text{Bern}(\theta_{\sigma_i, \sigma_j}) \text{ for } i, j = 1, \dots, n; i < j$$

$$\sigma_i \sim \text{Multinom}(1, \boldsymbol{\pi}) \text{ for } i = 1, \dots, n,$$

where  $\boldsymbol{\theta}$  is a ragged array encoding the edge probability.

$$\boldsymbol{\theta} = \begin{bmatrix} \theta_{\sigma_1, \sigma_1} & & & \\ \theta_{\sigma_2, \sigma_1} & \theta_{\sigma_2, \sigma_2} & & \\ \vdots & \vdots & \ddots & \\ \theta_{\sigma_n, \sigma_1} & \theta_{\sigma_n, \sigma_2} & \dots & \theta_{\sigma_n, \sigma_n} \end{bmatrix}$$

(Peng & Carvalho, 2016, Electronic Journal of Statistics)



# A GLM Approach

The Bayesian SBM can be framed in terms of a **generalized linear model (GLM)**.

For **simple graphs** Peng & Carvalho (2016) adopt a logistic regression for  $A_{ij}$ .

$$A_{ij} | \boldsymbol{\sigma}, \boldsymbol{\gamma}, \boldsymbol{\eta} \sim \text{Bern}(\theta_{\sigma_i, \sigma_j}) \text{ for } i, j = 1, \dots, n; i < j$$
$$\text{logit}(\theta_{\sigma_i, \sigma_j}) = \gamma_{\sigma_i, \sigma_j} + \eta_i + \eta_j$$

Here  $\eta_i$  and  $\eta_j$  are node-specific intercepts that measure the **expected degree** of nodes  $i$  and  $j$ , and

$\gamma_{\sigma_i, \sigma_j}$  captures between and within **community association**.

(Peng & Carvalho, 2016, EJS)

# Other Considerations

The Bayesian SBM as stated has a few important drawbacks:

1. The number of clusters  $K$  must be determined *a priori*.
  - In practice, we can fit models over a range of  $K$  and assess fit of each to choose  $K$ .
  - Certain variants allow  $K$  to be inferred (Peixoto, 2014, Phys. Review)
2. In MCMC estimation, the community labeling vector  $\sigma$  suffers by label switching.
  - This is addressed through a **canonical remapping** of posterior samples (Peng & Carvalho, 2016, EJS).

# Weak and Widespread Signal

We have found through simulation studies that the SBM (and other clustering algorithms) suffers from poor performance when signal is **weak and widespread**.

Weak and widespread signal can manifest in networks as

1. Sparse networks with strong community structure
2. Dense networks with weak community structure

Little has been written regarding the performance of the SBM using data integration, though statistical power has been shown to be improved in certain graphical models (Kim *et al.*, 2018, Bioinformatics)

# Data Integration

Several experiments have been run on the SSc twin cohort, thus we have multiple data sources available such as

- Gene expression
- DNA methylation

In addition, we have recently developed GAIL, a novel research mining database containing associations between over 300 million pairs of genes (Couch *et al.*, 2019, PLOS One).

Integrating these data sources into one network may allow us to study a complex disease like SSc using SBMs.

# Edge Union Integration

Preliminary simulation studies have shown promising performance of data integration for the Bayesian SBM.

One preliminary simple approach to integrating multiple networks is **edge union**.

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

# Simulation Studies

In each simulation, we generate two random graphs  $\mathbf{G}_1$  and  $\mathbf{G}_2$  from an SBM with  $n = 100$  nodes and  $K = 3$  communities.

**Simulation 1:** The network is sparse but community structure is (relatively) strong.

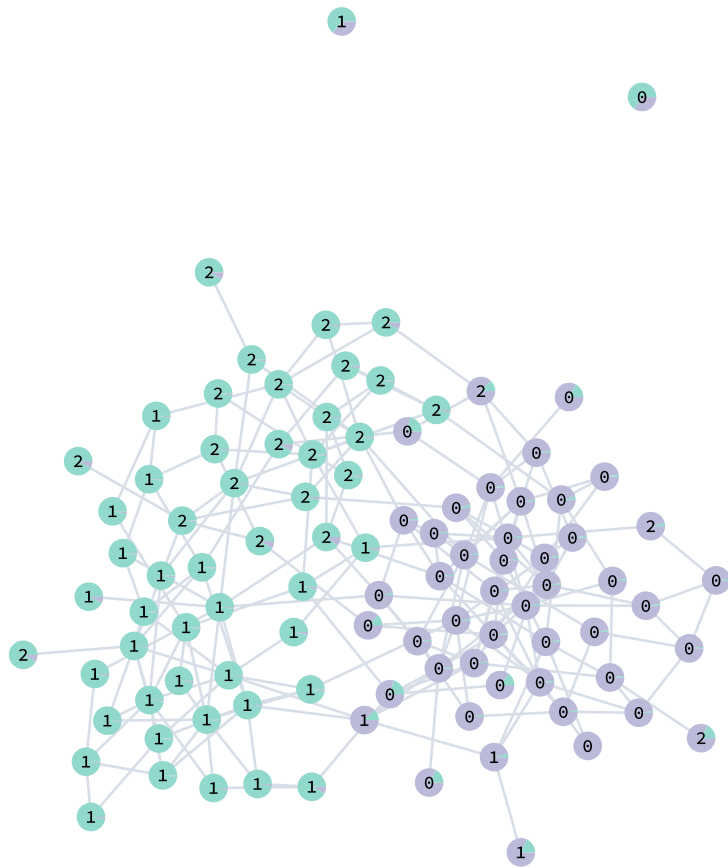
**Simulation 2:** The network is dense but community structure is (relatively) weak.

In each case, we use edge union to integrate  $\mathbf{G}_1$  and  $\mathbf{G}_2$  into  $\mathbf{G}$ , the data integrated network data.

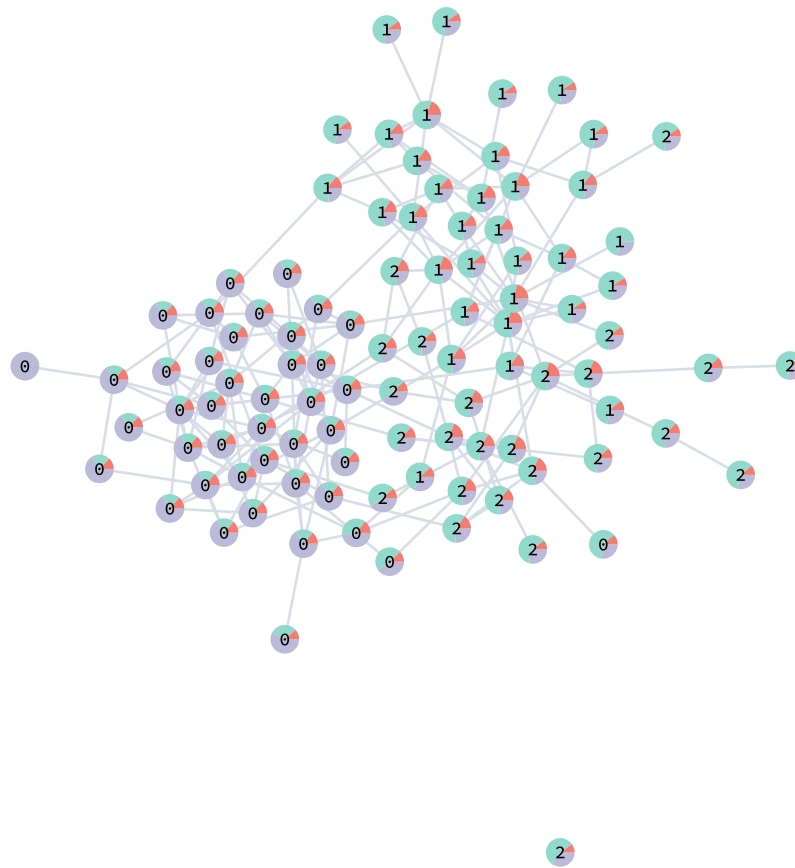
We compare SBMs fit to  $\mathbf{G}_1$ ,  $\mathbf{G}_2$ , and  $\mathbf{G}$ .



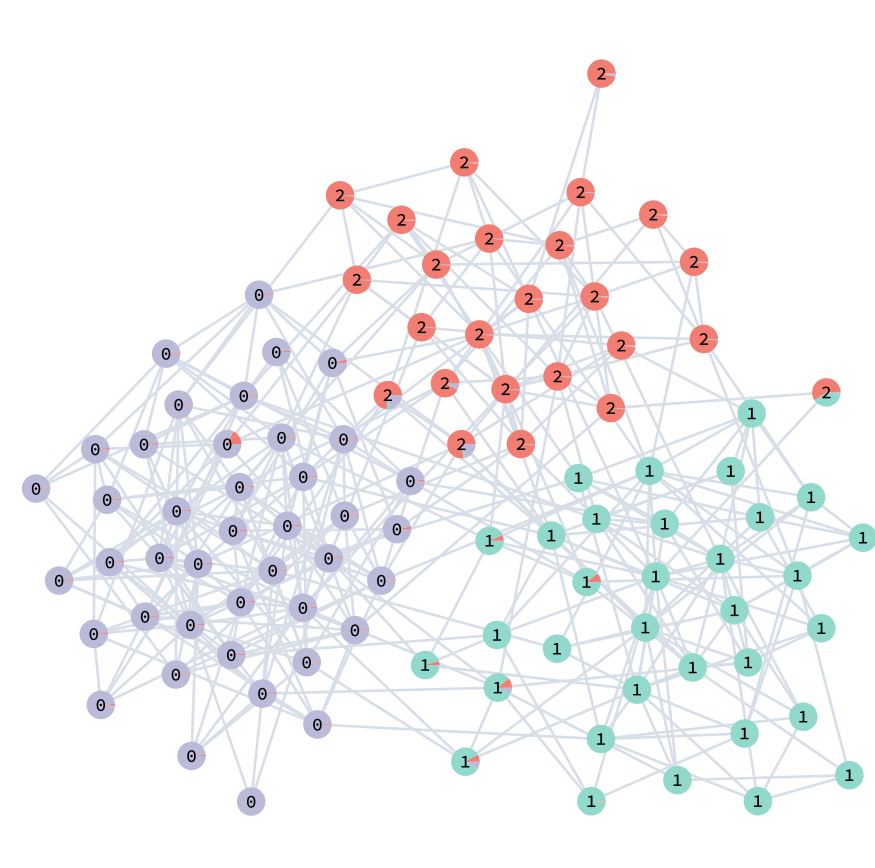
# Simulation 1



$G_1$



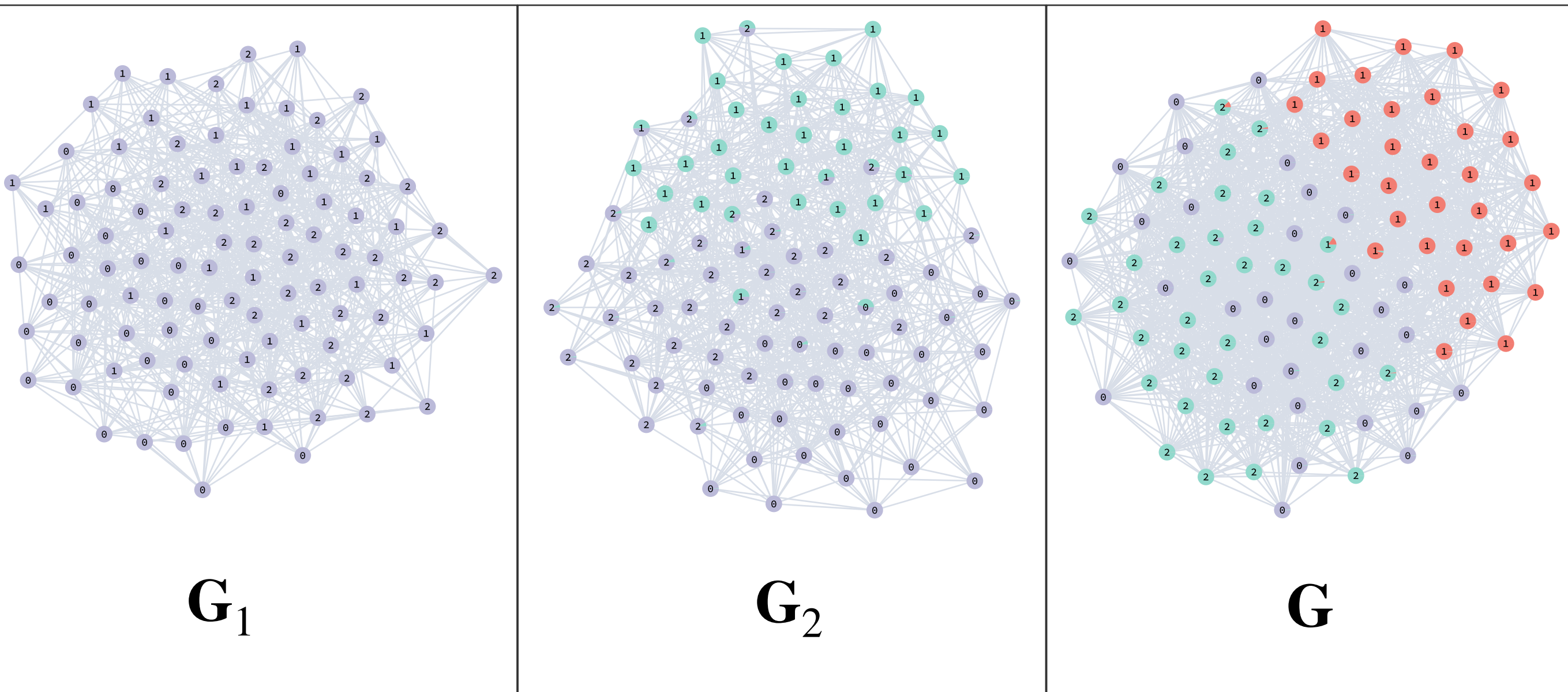
$G_2$



$G$

**Sim. 1** Sparse networks with strong signal. Posterior estimates of each community label is given (color) along with true community (number)

# Simulation 2



**Sim. 2** Dense networks with weak signal. Posterior estimates of each community label is given (color) along with true community (number)

# Future Work

- Continue to develop data integration methods for data that is not in network form:
  - How to best transform rectangular data (e.g., count data) to network data? How to define association matrices? (Mrzelj & Poličar, 2017, arXiv).
- How to better integrate data that is already in network form?
- Relax assumptions made by SBM:
  - Allow for weighted/directed edges (Aicher *et al.*, 2013, arXiv).
  - Allow for multi-edge graphs, i.e., multiple edges between nodes (Peixoto, 2014, Phys. Review) .

# Future Work

- Apply to the twin SSc data to implement community detection and detect hub genes.
- Community detection may uncover subsets of genes responsible for certain phenotypes of SSc.
- Hub gene detection identifies most crucial genes for function of the network.

# Future Work

- Apply to single cell data to improve identification of cell clusters by integrating multiple datasets:
- SBMs offer more informative statistical inference than traditional heuristic clustering algorithms.
- However, SBMs often do not scale as well as heuristic algorithms.
- Can we possibly use a two stage approach to get best of both worlds?

Prabhakaran *et al.*, 2016, ICML; Barkas *et al.*, 2019, Nat. Meth.

DePasquale *et al.*, 2019, NAR; Haghverdi *et al.*, 2018, Nat. Biotech.

# References

Peng L, Carvalho L. *Bayesian degree-corrected stochastic blockmodels for community detection*. Electronic Journal of Statistics. 2016;10(2):2746-79.

Feghali-Bostwick C, Medsger Jr TA, Wright TM. *Analysis of systemic sclerosis in twins reveals low concordance for disease and high concordance for the presence of antinuclear antibodies*. Arthritis & Rheumatism: Official Journal of the American College of Rheumatology. 2003 Jul;48(7):1956-63.

Peixoto TP. *Efficient Monte Carlo and greedy heuristic for the inference of stochastic block models*. Physical Review E. 2014 Jan 13;89(1):012804.

Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. *Integrating single-cell transcriptomic data across different conditions, technologies, and species*. Nature biotechnology. 2018 May;36(5):411-20.

Prabhakaran S, Azizi E, Carr A, Pe'er D. *Dirichlet process mixture model for correcting technical variation in single-cell gene expression data*. International Conference on Machine Learning. 2016 Jun 11 (pp. 1070-1079).

Barkas N, Petukhov V, Nikolaeva D, Lozinsky Y, Demharter S, Khodosevich K, Kharchenko PV. *Joint analysis of heterogeneous single-cell RNA-seq dataset collections*. Nature methods. 2019 Aug;16(8):695-8.

DePasquale EA, Schnell D, Dexheimer P, Ferchen K, Hay S, Chetal K, Valiente-Alandí Í, Blaxall BC, Grimes HL, Salomonis N. *cellHarmony: cell-level matching and holistic comparison of single-cell transcriptomes*. Nucleic acids research. 2019 Dec 2;47(21):e138-.

Haghverdi L, Lun AT, Morgan MD, Marioni JC. *Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors*. Nature biotechnology. 2018 May;36(5):421-7.



# Thank You!

Email: [allen.2554@osu.edu](mailto:allen.2554@osu.edu)