Data Integrated Stochastic Block Models for Gene Networks

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 for targeted therapies.

- For complex diseases, **weak and widespread** presents challenges to these research goals.

- We address this issue through **data integration**.

Motivation

- **Systemic sclerosis (SSc)** is an autoimmune disease involving fibrosis in multiple body systems.

- A broad understanding of the genetic underpinnings of SSc has not been achieved.

- Our study is motivated by data from the **only cohort of twins with SSc** to date (Feghali-Bostwick *et al.* 2003).

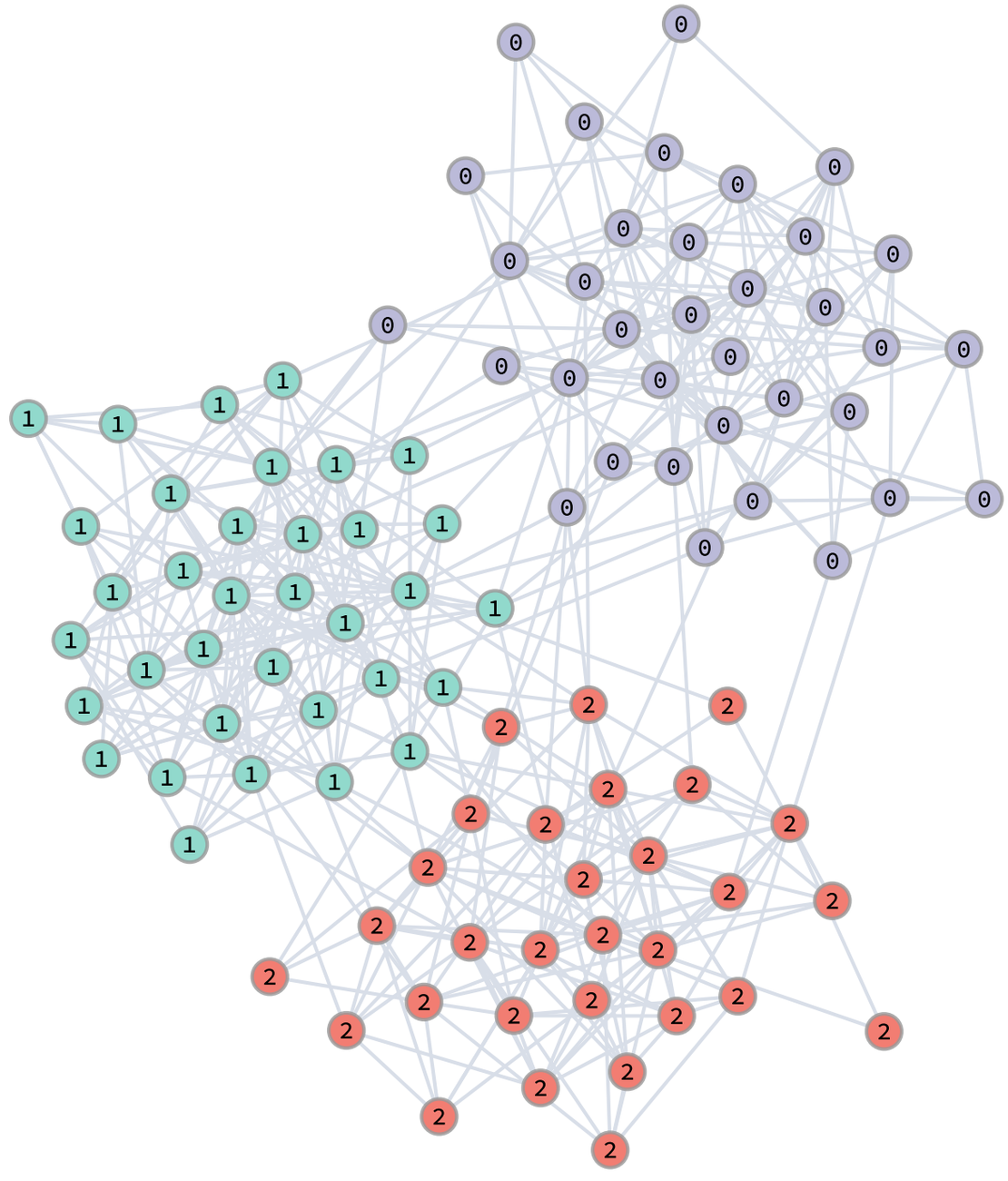
- Multiple experiments on the SSc twin cohort have generated several sources of data.

Background

- The **stochastic block model** (SBM) is a generative model for network data.
    - Naturally allows for inference about community structure of a graph.
    - Certain SBM variants allow for detection of hub nodes (genes).
- Bayesian SBMs allow for incorporation of prior information.

Network Data

Suppose the observed data takes the form of an $n \times n$ adjacency matrix **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.

Bayesian SBM

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

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

$$\boldsymbol{\sigma}_i \sim \text{Multinom}(1,\boldsymbol{\pi}) \text{ for } i = 1,\ldots,n,$$

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

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

A GLM Approach

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

For **simple graphs** we adopt a logistic regression for A_{ij}.

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

$$\text{logit}(\theta_{\sigma_i,\sigma_j}) = \gamma_{\sigma_i,\sigma_j} + \eta_i + \eta_j$$

Here η_i and η_j are node-specific intercepts that measure the **expected degree** of nodes i and j, and

$\gamma_{\sigma_1,\sigma_2}$ captures between and within **community association**.

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 model fit.

2. In MCMC estimation, the community labeling vector $\boldsymbol{\sigma}$ suffers by label switching. This is addressed through a **canonical mapping** which ensures labels are comparable across MCMC draws.

Weak and Widespread Signal

We have found through simulation studies that the Bayesian SBM suffers from poor performance when signal is weak and widespread.

Weak and widespread signal manifests in networks as

1. Sparse networks with strong community structure

2. Dense networks with weak community structure

These cases can be addressed through **data integration**.

Data Integration

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

- RNA sequencing
- DNA methylation

In addition, we have recently developed GAIL, a novel research mining database containing associations between over 300 million pairs of genes.

Integrating these data sources into one network will allow us to study a complex disease like SSc more effectively.

Edge Union Integration

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

One 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 $\mathscr{E}_1$ 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

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.

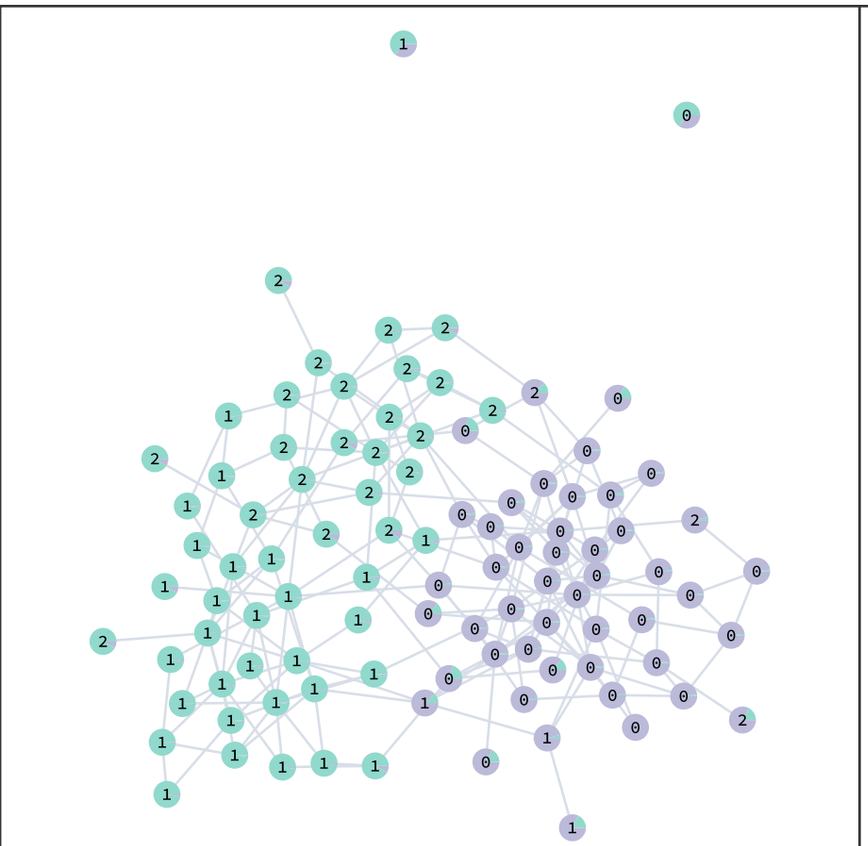
Simulation 1: The network is sparse but community structure is (relatively) strong.

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

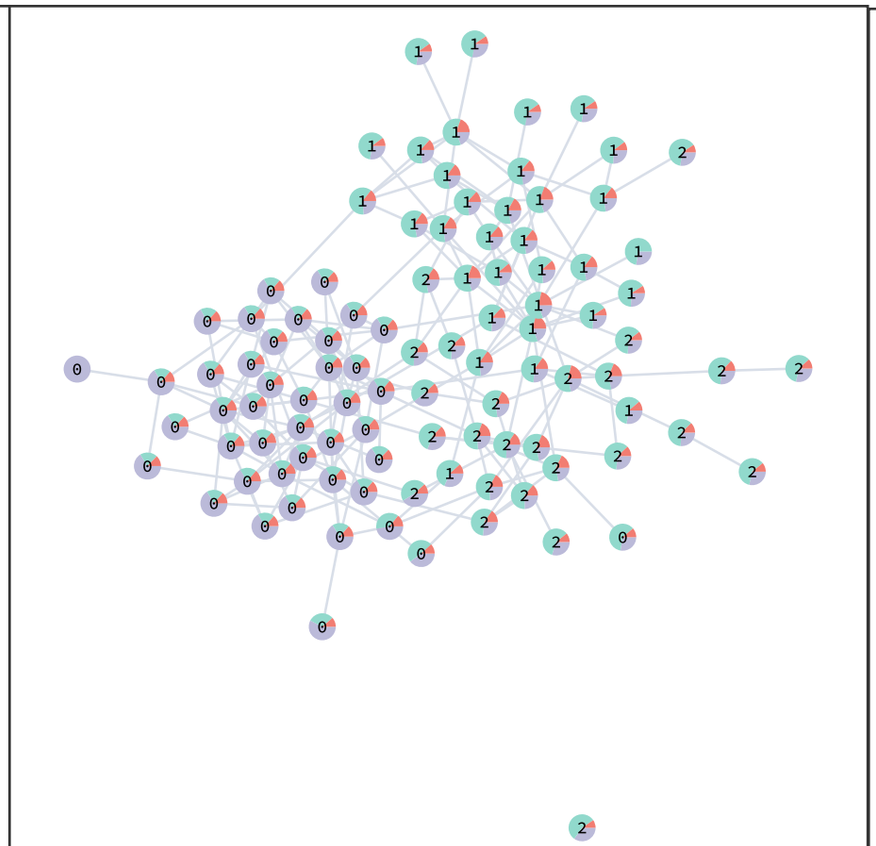
In each case, we use edge union to integrate $\mathbf{G}_1$ and $\mathbf{G}_2$ into $\mathbf{G}$, the data integrated network data.

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

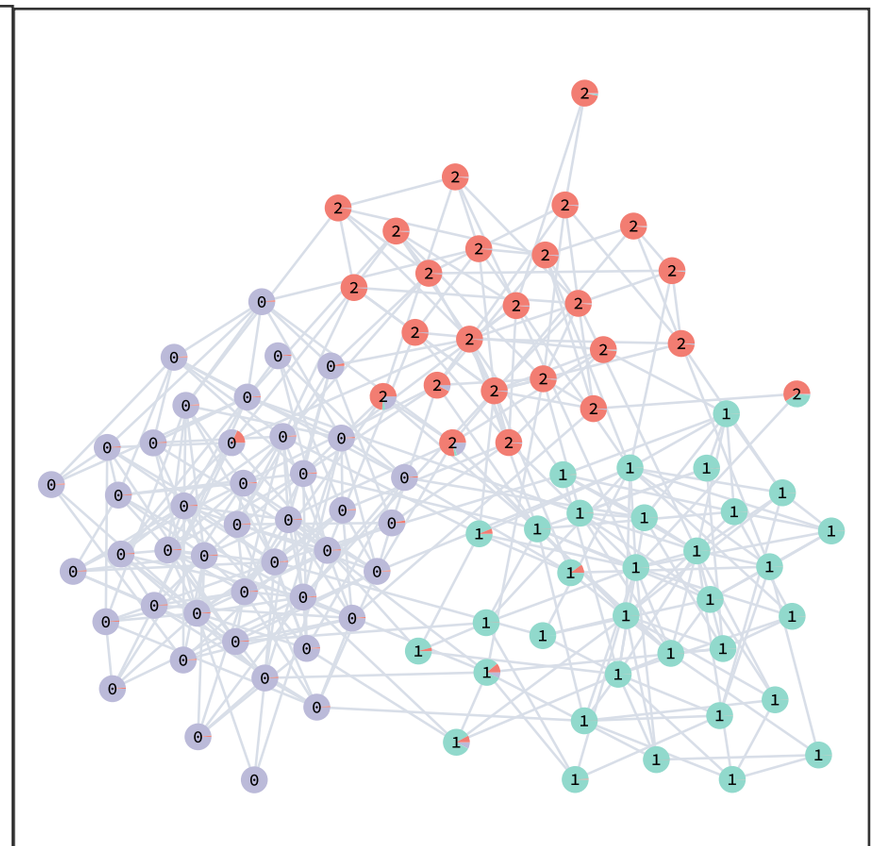
Simulation 1



$\mathbf{G}_1$



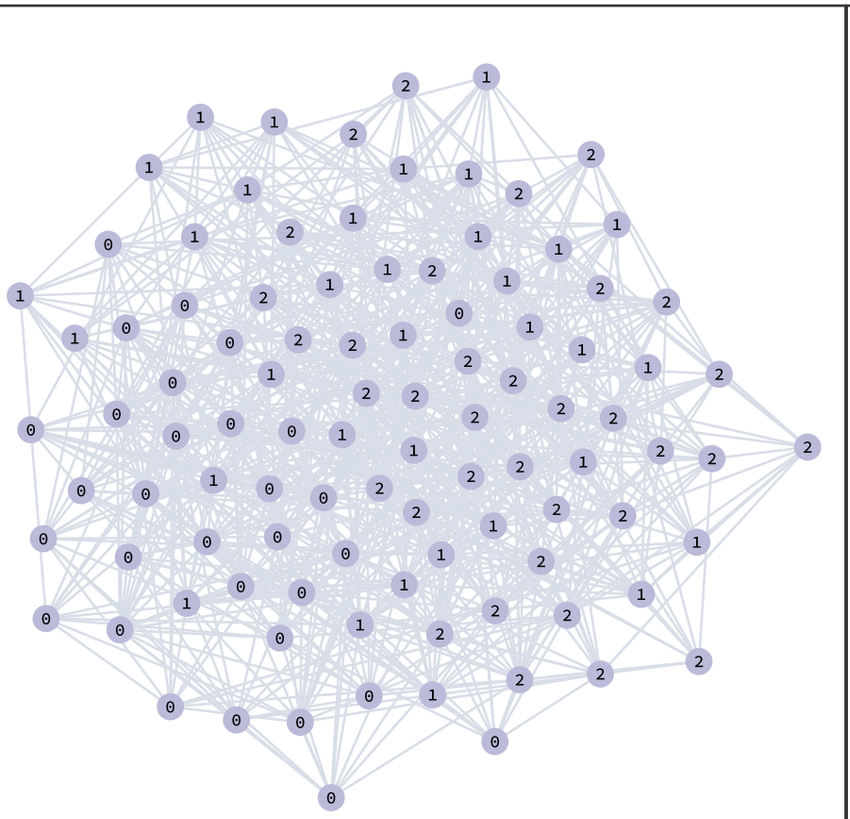
$\mathbf{G}_2$



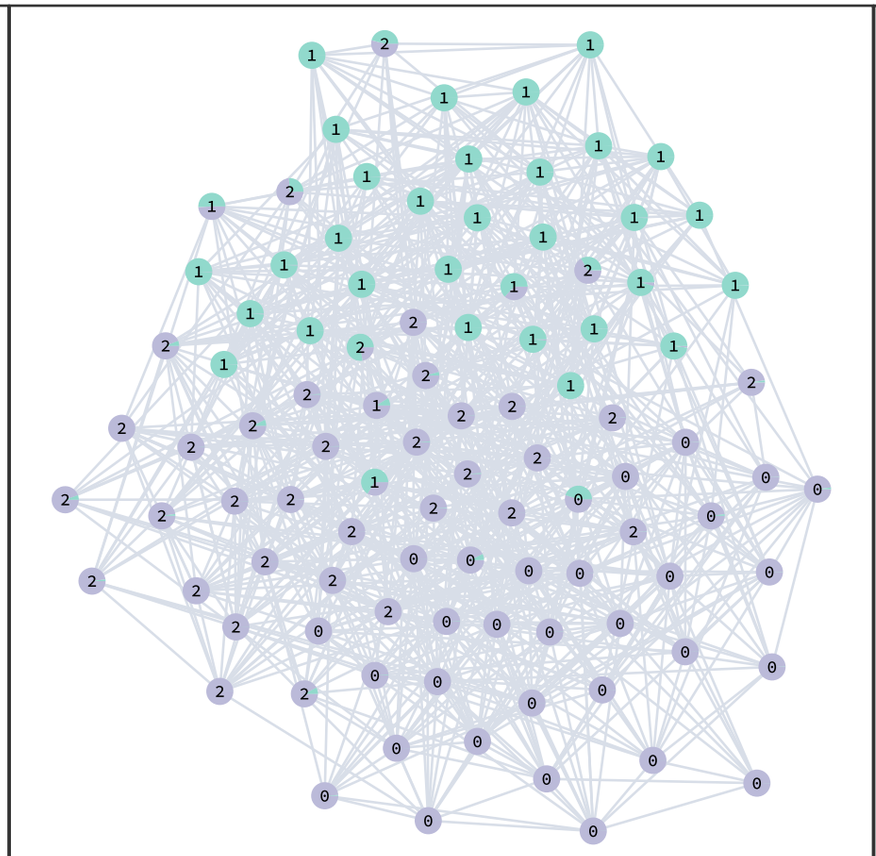
$\mathbf{G}$

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

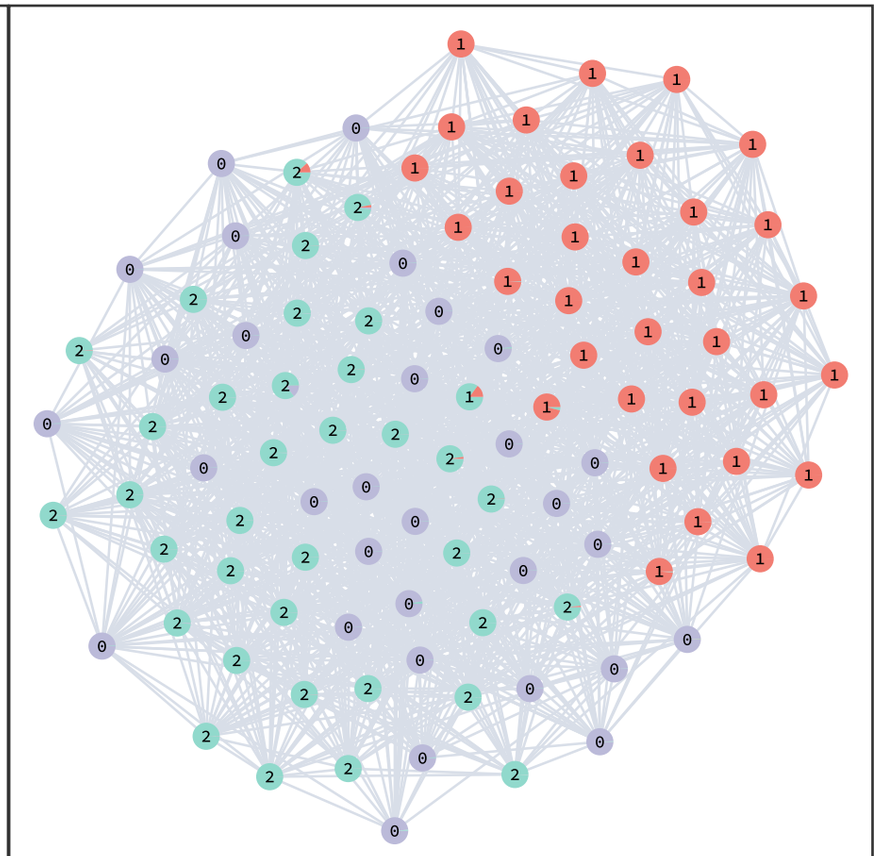
Simulation 2



$\mathbf{G}_1$



$\mathbf{G}_2$



$\mathbf{G}$

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 transform rectangular data to network data?
- Relax assumptions made by SBM (i.e., simple, undirected graphs.
- Apply to the twin SSc data to implement community detection and detect hub genes.

Thank You!