Community Detection in Gene Network

Bowen Deng

Dept. of Prob. and Stat.

Optimization Problem

Relaxation of MWSP

Introduction

For directed graph, source community and terminal community could be detected with

$$(\hat{u}, \hat{v}) = \arg\max u^T Q v - \eta(\|u\|_0 + \omega\|v\|_0), s.t. \|u\|_2 = 1, \|v\|_2 = 1$$

Source community $SC = \{i|\hat{u}[i] \neq 0\}$,
Terminal community $TC = \{j|\hat{v}[j] \neq 0\}$.

Optimization Strategy

For a given vector z and a fixed constant $\rho > 0$, the solution of

$$\max u^T z - \rho \|u\|_0, s.t. \|u\|_2 = 1$$

is

$$u = z_l^h / \|z_l^h\|_2$$

Repeat

$$
\begin{aligned}
z \leftarrow Qv, \rho \leftarrow \eta \\
u \leftarrow z_I^h / \|z_I^h\|_2, \\
z \leftarrow Q^T u, \rho \leftarrow \eta\omega \\
v \leftarrow z_I^h / \|z_I^h\|_2
\end{aligned}
$$

Undirected Counterpart

For undirected graph, e.g. gene network, a community could be detected with the symmetric counterpart:

$$\min f(u) = -u^T Qu + \rho \|u\|_0, s.t. u^T u = 1$$

where Q be a fixed symmetric matrix, ρ be a positive number.

Mimic Algorithm

Borrowing the idea from the previous optimization solving,

$$z \leftarrow Qu$$
$$u \leftarrow z_I^h / \|z_I^h\|_2$$

Brute Force Algorithm

Fixing $\|u\|_0 = k$, the objective is to find a submatrix $|\hat{Q}| = k$ with the largest eigenvalue.
For the sake of computational cost, we could also sample the subset with Genetic Algorithm. Once sampled, we find the eigenvalue of the matrix.

Variation of the Problem

$$\min u^T M u + \rho \|u\|_1, s.t. u^T u = 1$$

Lagrangian Method

Consider

$$\min f(u, l) = u^T M u + \rho \|u\|_1 + l(u^T u - 1)$$

Repeat

$$u \leftarrow u - \lambda \nabla_u f(u, l)$$
$$l \leftarrow l - \lambda \nabla_l f(u, l)$$

Relaxation

The original simultaneously detection is

$$\max O(M_1, \cdots, M_t) = \sum_{\rho=1}^{t}\sum_{i=1}^{m}(2C_i(M_\rho) - \sum_{j=1}^{n} I_{M_\rho}(j)A_{ij})$$

$$s.t. \sum_{j=1}^{n} I_{M_\rho}(j)A_{ij} \geqslant C_i(M_\rho)$$

$$\sum_{\rho=1}^{t} I_{M_\rho}(j) \leqslant 1$$

where all variables are binary.
The relaxation counterpart is to relax $x \in \{0, 1\}$ to $0 \leqslant x \leqslant 1$.

Set 50 patients, 50 genes, genes 1-4; genes 5-8 are driver pathway genes. Background mutation rate 0.02.

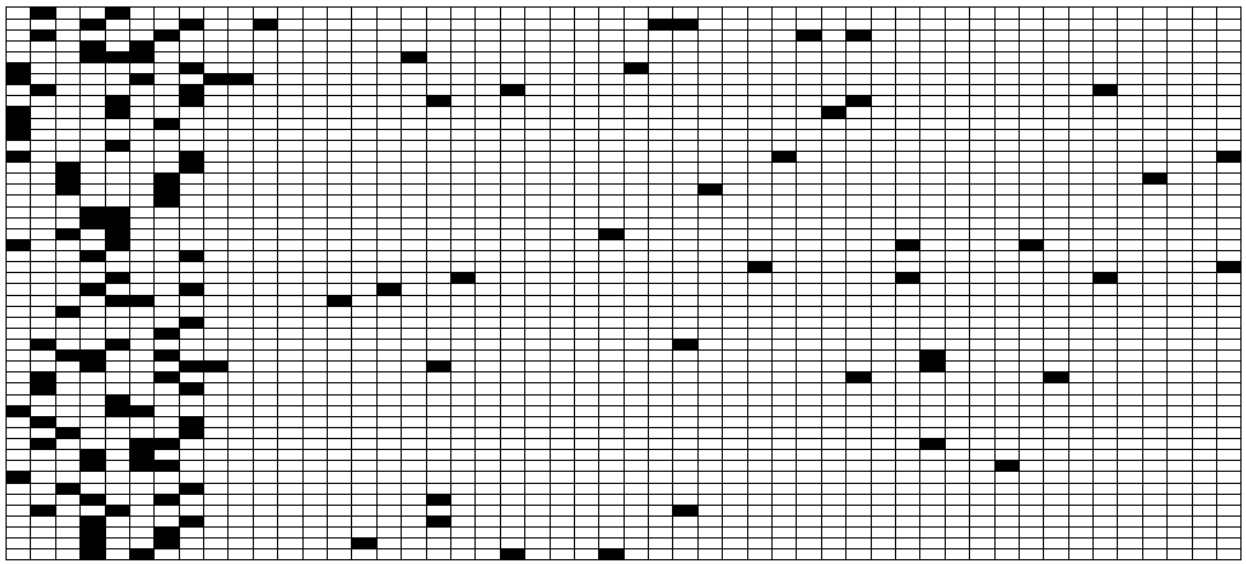

Figure: Simulated Mutation Data

Performance:
Set $t = 2$, $k_{\min} = 1$, $k_{\max} = 6$, under the same device (My PC), with relaxation of the problem, the linear programming was solved in about 1 minute, the result was not all integers.
First group:
0.5: 1,2,3,4,5,7,8,10,14,19.
1: 12.
0: others.
Second group:
0.5: 1,2,3,4,5,7,8,10,14,19.
1: 31.
0: others.
Without relaxation, time took: 0.05 second.
First group: 1,2,3,4,14,19.
Second group: 5,7,8,10,21,31.

Performance:
Set $t = 2$, $k_{\min} = 3$, $k_{\max} = 5$, under the same device (My PC), with relaxation of the problem, the linear programming was solved in about 20 seconds, the result was not all integers.
First group:
0.5: 1,2,3,4,5,7,8,10,14,19.
0: others.
Second group:
0.5: 1,2,3,4,5,7,8,10,14,19.
0: others.
Without relaxation, time took: 0.02 second.
First group:
1,2,3,4,14.
Second group:
5,7,8,10,31.

Set $k_{\min} = k_{\max} = 4$.
Relaxation Problem: time 42 seconds.
First group:
0.5: 1,2,3,4,5,6,7,8.
0: others.
Second group:
0.5: 1,2,3,4,5,6,7,8.
0: others.
Without relaxation: 0.03 second.
1,2,3,4; 5,7,8,10.