Testing Independence between Observations from a Single Network

Youjin Lee & Elizabeth (Betsy) Ogburn ● Department of Biostatistics, Johns Hopkins School of Public Health

ylee160@jhu.edu ● eogburn@jhsph.edu

Question

Testing the existence of correlation in our observations depending on their social relationship.

Motivation

Observations in a study are often collected from a single network within the target population. In this case, it is likely that the observed value of subjects' outcome is dependent on whom they are closely related to. Ignoring this dependence and implementing standard statistical analysis could lead to invalid statistical inference. We propose a test statistic for testing independence between observations from a single network.

Methods

We extend the existing test used for spatial autocorrelation :

- Moran's I (continuous observations Y)

Weighted correlation coefficient: close to zero without dependence.

$$I \propto \sum_{i,j=1}^{n} W_{ij}(y_i - \bar{y})(y_j - \bar{y})$$

- ϕ (categorical observations Y)

Weighted concordant(+1) and discordant(-1) pairs.

$$\phi \propto \sum_{i,j=1}^{n} W_{ij}\{2I(y_i = y_j) - 1\} \,/\, \{P(Y = y_i)P(Y = y_j)\}$$

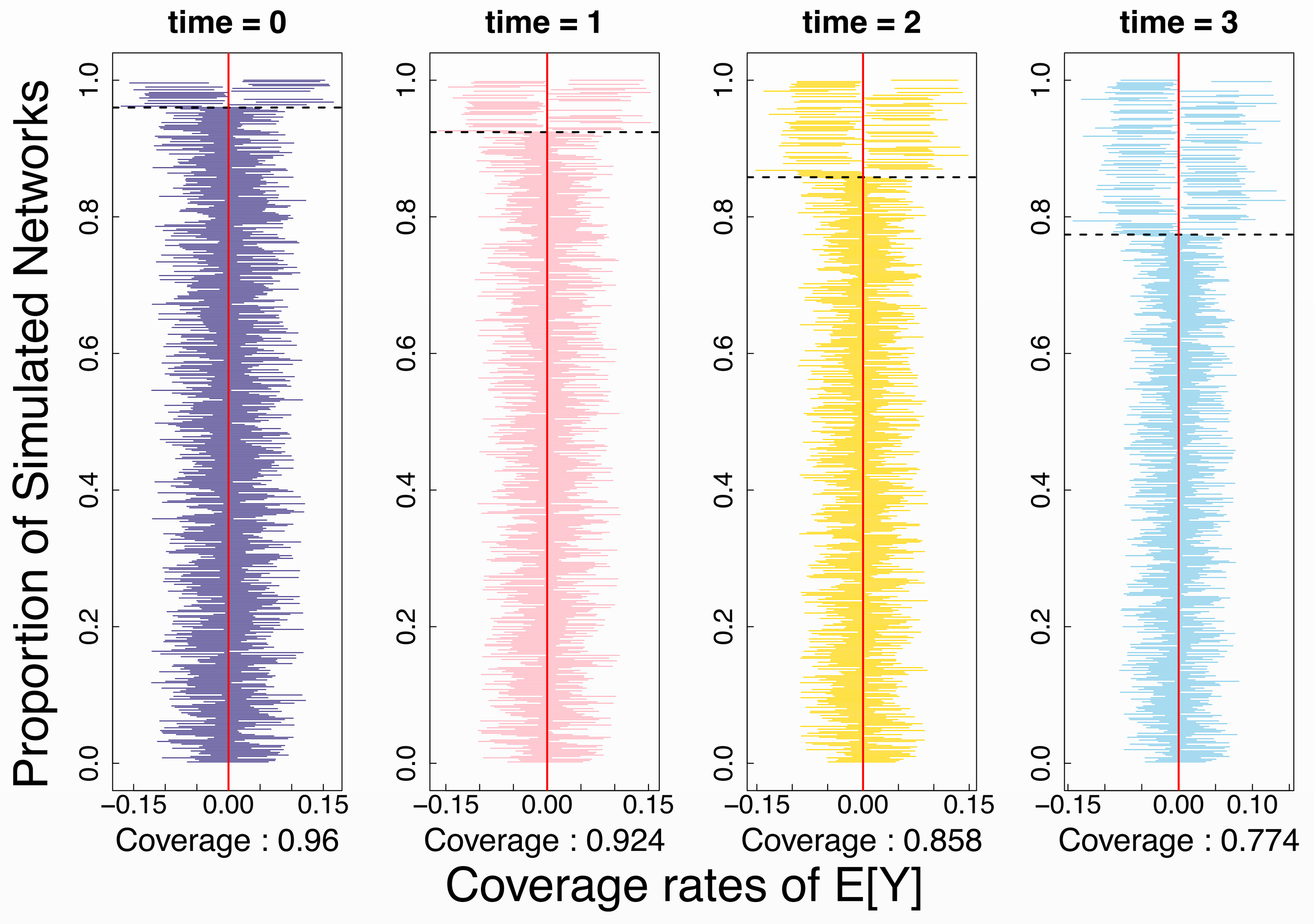
- W_{ij} : closeness measure between i and j.
- Statistics are weighted more when the observations are rare.
- When Y is binary, $I = \phi$.
- Asymptotic normality of both I and ϕ under some conditions.

Numerical Study

1. Peer-dependent continuous outcome of Y

$$Y_i^t = (1-\theta)Y_i^{t-1} + \theta\, \overline{Y_i^{t-1}},$$

where $\overline{Y_i^{t-1}}$ is the average of i 's friends at time $t-1$. As time t increases, the amount of dependence increases.



Standard 95% confidence interval assuming independence on Y leads to decrease in coverage rates. We are more likely to reject the null when confidence interval coverage rates drops; assumption i.i.d. outcomes should not be used in inference.

Testing for Network Dependence

time	95% CI coverage rate	Moran' I	Power(%)
t = 0	0.96	0.06	6.40
t = 1	0.92	0.99	27.80
t = 2	0.86	2.09	60.60
t = 3	**0.77**	3.42	**81.80**

Numerical Study

2. Peer-dependent categorical outcome of Y

$$Y_i^t = (1-\theta)Y_i^{t-1} + \theta Z_i^t$$

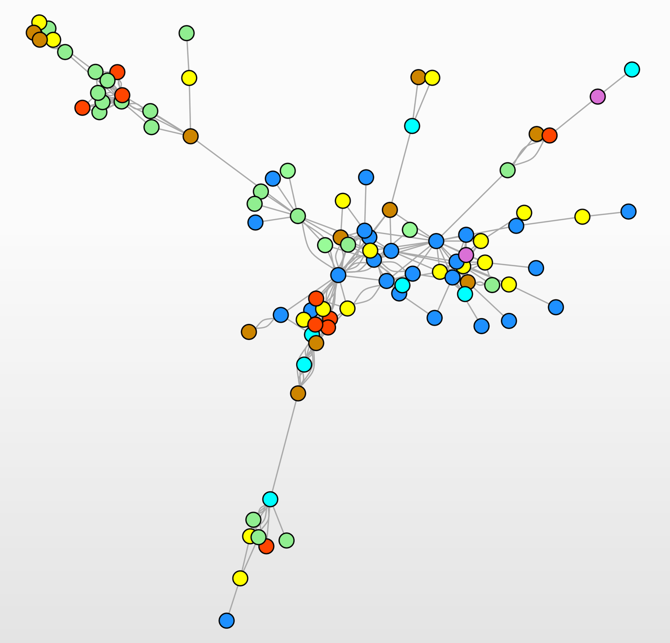
where $Y_i^0 \sim Multi(0.1, 0.2, 0.3, 0.25, 0.15)$ and $P(Z_i^t = k) = \sum I\,(y_i^{t-1} = k)/n$. Assume that we are interested in estimating the simultaneous confidence interval for population proportion for five categories. As time t increases, the amount of correlation in Y also increases, which leads to decreasing coverage rate of confidence interval.

Testing for Network Dependence

time	95% CI coverage rate	ϕ	Power(%)
t = 0	0.94	0.06	7.00
t = 1	0.85	1.66	48.20
t = 2	**0.75**	2.93	**82.20**

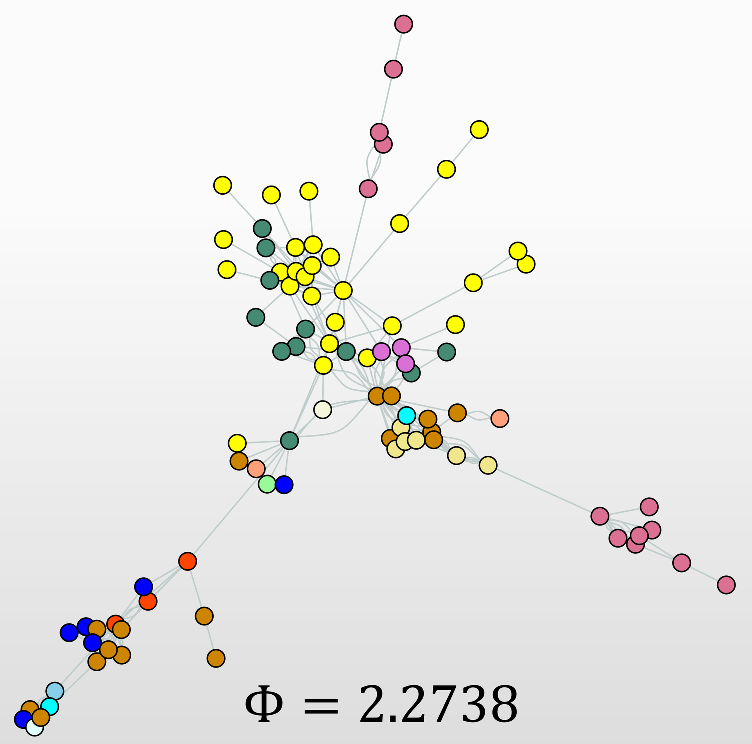
Example of Collaborative Network

Collaboration Network and Position



$\Phi = 0.6448$
(p-value : 0.26)

Collaboration Network and Department



$\Phi = 2.2738$
(p-value : 0.01)

Discussion

- If we reject the null hypothesis, it is highly recommended not to make i.i.d. assumption in the observations.
- We used an adjacency matrix as a weight matrix W, and this is a robust choice for weight which does not depend on the exact form of dependence.
- If we have substantive knowledge of the dependence mechanism, other weight matrices that incorporate this information might be more efficient.
- Even though we use a coverage rate of confidence intervals as a measure of dependence in our numerical studies, the existence of correlation between the observations does not necessarily lead to lower coverage rate than expected.
- The test statistic will be finally applied to the observations from Framingham Heart Study where individuals are socially related to each other.

Reference

- Moran, Patrick AP. "The interpretation of statistical maps." *Journal of the Royal Statistical Society. Series B (Methodological)* 10, no. 2 (1948): 243-251.
- Sen, Ashish. "Large Sample-Size Distribution of Statistics Used In Testing for Spatial Correlation." *Geographical analysis* 8, no. 2 (1976): 175-184.
- Christakis, Nicholas A., and James H. Fowler. "The spread of obesity in a large social network over 32 years." *New England journal of medicine* 357, no. 4 (2007): 370-379.