Estimating Social Influence Effects in Networks Using A Latent Space Adjusted Approach in R

by Ran Xu

Abstract Social influence effects have been extensively studied in various empirical network research. However, many challenges remain in estimating social influence effects in networks, as influence effects are often entangled with other factors, such as homophily in the selection process and the common social-environmental factors that individuals are embedded in. Methods currently available either do not solve these problems or require stringent assumptions. Recent works by Xu (2018) and others have shown that a latent space adjusted approach based on the latent space model has the potential to disentangle the influence effects from other processes, and the simulation evidence has shown that this approach outperforms other state-of-the-art approaches in terms of recovering the true social influence effect when there is an unobserved trait co-determining influence and selection. In this paper, I will further illustrate how the latent space adjusted approach can account for bias in the estimation of social influence effects and how this approach can be easily implemented in R.

Introduction

Social influence effects, sometimes referred to as spillover or contagion effects, have long been central to the field of social science (Asch, 1972; Erbring and Young, 1979; Bandura, 1986). It is defined as the propensity for the behavior of an individual to vary along with the prevalence of that behavior in some reference group (Manski, 1993), such as one's social contacts. With the availability of social network data, social influence effects have received much attention and have been widely used to study various phenomena such as the spread of health behavior (e.g., obesity and smoking) (Christakis and Fowler, 2007, 2008), psychological states (Cacioppo et al., 2008; German et al., 2012), professional practices (Frank et al., 2004) and information diffusion (Valente, 1995, 1996).

However, many challenges remain in estimating social influence effects, especially from observational network data, because it is difficult to separate the effect of social influences from other processes that operate simultaneously. That is, when we observe that people in close relationships or interactions tend to be similar in their behaviors or states, it is difficult to identify the underlying mechanisms that generate these patterns. One mechanism could be influence or contagion (Friedkin and Johnsen, 1999; Friedkin, 2001; Oetting and Donnermeyer, 1998), whereby individuals assimilate the behavior of their network partners. Another mechanism could be selection – in particular, homophily (Mcpherson and Smith-Lovin, 1987; Mcpherson et al., 2001), in which individuals seek to interact with similar others. Furthermore, there could be some common social-environmental factors – individuals with previous similarities select themselves into the same social settings (e.g., hospital or alcoholics anonymous (AA) support group), and actual network formation just reflects the opportunities of meeting in this social setting (Feld, 1981, 1982; Kalmijn and Flap, 2001)[1].

Entanglement among these different mechanisms unavoidably induces bias when we estimate social influence effects (Shalizi and Thomas, 2011). Various statistical methods and recent advancements in the field of social network analysis have attempted to reduce the bias in estimating social influence effects, such as instrumental variable (IV) methods (Bramoulle et al., 2009), propensity score methods (Aral et al., 2009), and stochastic actor-oriented models (SAOM) (Snijders et al., 2010). Although each potentially leverages extra information in the data to reduce bias, none can claim to eliminate all sources of bias.

Recent works by Xu (2018) and others (Shalizi and McFowland III, 2018) have shown that a latent space adjusted approach based on the latent space model (Hoff et al., 2002) has the potential to disentangle the social influence effects from other processes operating at the same time, and simulation evidence has shown that this approach outperforms some other state-of-the-art methods (e.g., instrumental variable method, structural equation model) in terms of recovering the true social influence effects. In this paper, I will illustrate how the latent space adjusted approach can account for

[1] There are also structural constraints such as transitivity and preferential attachment which could cause people to become friends. However, these mechanisms in themselves do not entangle with influence (e.g., one can befriend another having high popularity but different behavior). In these cases, another mechanism must be present to induce similarities between these friends (e.g., selection of common friends based on similarity in attributes), and thus the entanglement goes back to the original three mechanisms, namely influence, selection based on homophily, and social-environmental factors.

bias in the estimation of social influence effects and demonstrate how it can be easily incorporated with various models in R to estimate social influence effects. In the following sections, I will first explain the challenges in estimating social influence effects and how they can be framed as an omitted variable bias problem. Then I will formally introduce the latent space adjusted approach and explain how it can account for bias in the estimation of social influence effects. Finally, I will demonstrate how this approach can be easily implemented in R and how it can be incorporated with various models to estimate social influence effects, including a dynamic linear-in-mean influence model and a stochastic actor-oriented model (SAOM).

Identification of Social Influence as An Omitted Variable Bias Problem

Similarities of behavior, state, and characteristics of two individuals in a network relationship can be caused by three primary mechanisms: influence, homophilous selection, or common social-environmental factors (Vanderweele and An, 2013). While it is possible to rule out some mechanisms through random treatment assignment or networks in experiments, entanglement among these different mechanisms makes it difficult to correctly estimate social influence effects from observational data (Xu, 2020). The challenges in estimation caused by entanglement among social influence effects and common social-environmental factors can be easily framed as an omitted variable bias problem (e.g., ignoring the group or environment individuals belong to when estimating the social influence model). What is less obvious is that entanglement between the influence and the homophilous selection can also be framed as an omitted variable bias problem. As pointed out by Steglich et al. (2010), one of the important concerns of SAOM is the "possibility that there may be non-observed variables co-determining the probabilities of change in network and/or behavior". Shalizi and Thomas (2011) have shown that when there is an unobserved trait that co-determines both behavior and network choice, social influence effects are generally unidentifiable as social influence and homophily (selection) are generically confounded through this unobserved trait.

To give an example, assuming that adolescent i's alcohol use at time t, $alcohol_{it}$, is the outcome of interest, and it is a function of his/her previous alcohol use, $alcohol_{it-1}$, his/her friend j's previous alcohol use, $alcohol_{jt-1}$ (i.e., social influence), and an unobserved tendency for substance-abuse (arrow D in Figure 1). At the same time, there is a homophilous selection based on this unobserved substance-abuse tendency in the network – individuals with similar levels of substance-abuse tendency are more likely to be friends (arrow A in Figure 1). As a result, person j's alcohol use, which is a function of person j's substance-abuse tendency (arrow B_j), will be correlated with person i's substance-abuse tendency through homophilous selection (arrow C in Figure 1). However, as the substance-abuse tendency is unobserved, this violates the key assumption of most estimation methods (i.e., the omitted variable should not correlate with the independent variables) such that the estimates of the social influence effects will be biased and inconsistent.

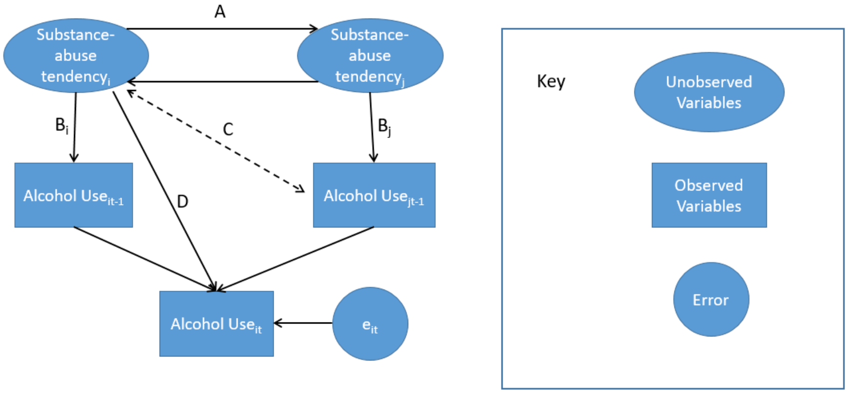

Figure 1: Demonstration of the omitted variable bias.

Latent Space Adjusted Approach

Xu (2018) recently proposed a latent space adjusted approach, and simulation evidence has shown that it has the potential to correctly identify social influence effects when there is an unobserved variable that co-determines the influence and the selection process. Specifically, the behavioral (influence) model can be represented as

$$Y_i = f(Z_{ij}, Y_j, X_i, c_i) \tag{1}$$

where the behavior of person i, Y_i, is a function of the behaviors of his/her network partners j, Y_j, the network relations between i and j, Z_{ij}, person i's observed characteristics X_i as well as a time-invariant unobserved trait c_i.[2] For example, adolescent i's alcohol use at time t, Y_{it}, can be a function of his/her previous alcohol use, Y_{it-1}, his/her close friends' previous alcohol use Y_{jt-1}, his/her own cigarette use, X_{it}, and a time-invariant unobserved tendency for substance abuse c_i.

The selection model can be represented as

$$P(Z_{ij} = 1) = g(X_{ij}, D(c_i, c_j)) \tag{2}$$

where the probability that person i and person j has a network relation is a function of the individual and dyadic level observed variables X_{ij}, and a distance function of the unobserved trait c between i and j, such that i and j are more likely to have a network relation when they are close to each other in terms of c. For example, the probability that adolescent i and j are friends at time t Z_{ijt} can be a function of the absolute value of the differences between their cigarette use at time t $|X_{it} - X_{jt}|$ (observed homophily) and the absolute value of the differences between their unobserved tendencies for substance abuse $|c_i - c_j|$ (latent homophily), where i and j are more likely to become friends when they are similar to each other in terms of the cigarette use (X) or the tendency for substance abuse (c).

Ideally, if there is any information about this unobserved trait c from the selection process in (2), it can be leveraged and used in the estimation of the behavioral model in (1), and in principle this will reduce the bias when estimating the social influence effects. However, the estimations of most selection models are based on the observed variables and thus do not attend to those factors that are unobserved. Xu (2018) extended this idea and built on the theoretical logic of latent space models (Hoff et al., 2002). Latent space models assume that each individual has a "latent position" that lies in an unobserved n-dimensional social space, and the probability of interaction between any two actors depends on the latent positions of these two actors. Specifically, the model takes a logistic form and can be specified as

$$logodds(Z_{ij} = 1 \| c_i, c_j, x_{ij}, \alpha, \beta) = \alpha + \beta' x_{ij} - |c_i - c_j| \tag{3}$$

Here, Z_{ij} indicates whether there is a network relation from i to j, x_{ij} is a vector of the observed covariates (at the dyadic level or node level), c indicates the latent social position of i and j, and $|c_i - c_j|$ represents the Euclidean distance between i and j's latent positions (it could also be replaced by other distance functions). When i and j are closer to each other in terms of the latent position c, they will have a higher probability of having a network relation. And these latent positions can represent determinants of the network relations that have not been accounted for by the observed variables in the selection process. The parameters α and β are estimated using either Maximum-Likelihood Estimation (MLE) or Markov Chain Monte Carlo (MCMC) methods, and the latent positions c can be estimated by Minimum Kullback-Leibler (MKL) estimates (Shortreed et al., 2006).

It is not difficult to see that the latent space model in (3) is very similar to the selection process in (2), except that c represents the latent position in the latent space model, while c represents the individual's unobserved trait in (2).[3] For any pair of i and j, a smaller distance between the latent social positions or the unobserved traits will result in a higher likelihood of having network relations. Therefore, when two individuals are close to each other in terms of the unobserved traits, they are more likely to have a network relation, and they should also be close to each other in terms of the latent positions (and vice versa).

Furthermore, if these latent positions from the latent space model are estimated accurately enough, the estimates of these latent positions can be used as the proxies for the unobserved traits that determine the homophily in the selection process. In fact, for two one dimensional variables X and Y, if the distance correlation (e.g., correlation between $|X_i - X_j|$ and $|Y_i - Y_j|$) is 1, then Y can be written as a linear function of X: $Y = a + bX$ (Szekely et al., 2007), which means the correlation between the two variables are either 1 or -1. **Thus, the estimated latent positions from the latent space model can be used as the proxies (Wooldridge, 2011) for the unobserved traits that co-determine influence and selection, and including the latent positions as additional covariates in the behavioral model will reduce the bias in the estimation of social influence effects**. For example, to model adolescents'

[2] Here Y, X, Z are assumed to be time-variant and c is assumed to be time-invariant, but the assumption can be relaxed.

[3] Here I only choose one-dimensional latent social positions to mimic the unobserved trait that drives the homophily in the selection process. The arguments can easily be extended to multi-dimensional latent positions.

social influences on their alcohol use, we can first use a latent space model to model the friendship network of adolescents and acquire the estimated "latent positions" for each individual, and then use these estimates as the proxies for the unobserved substance-abuse tendencies in the behavioral model, and thus achieve a better estimation of the true social influence effects. If the social network data is longitudinal, estimated latent positions from each time point can be included as separate covariates in the behavioral model to better approximate the unobserved trait.

Shalizi and McFowland III (2018) have shown that if the network grows according to a continuous latent space model, the latent positions can be consistently estimated. Controlling for these latent positions allows for unbiased and consistent estimation of the social-influence effects in additive influence models. Simulation evidence from Xu (2018) has shown that when there is a time-invariant unobserved variable that co-determines selection and influence, the estimated latent positions can be good proxies for the unobserved variable, and the latent space adjusted approach outperforms other methods that are commonly used to deal with the unobserved variables, including a structural equation based estimator (implemented using **lavaan** package in R (Rosseel and Jorgensen, 2019)) and an instrumental variable estimator (implemented using **plm** package in R (Croissant et al., 2021)), in producing the smallest bias and standard error of the social influence effect using a dynamic linear-in-mean influence model. The results are robust to the inclusion of additional covariates, structural properties (e.g., transitivity) in networks, different scaling of the latent space model, or even misspecifications (Xu, 2018).

Finally, there are a couple of things to note: (1) for the estimated latent positions to better approximate the unobserved traits, we need to control for other mechanisms that are likely to drive the selection process in the latent space model, such as homophily based on the observed variables, transitivity, alter, and ego effects. (2) In principle this method can apply to any functional form of the behavioral/influence model (e.g., stochastic actor-oriented models), as essentially this approach just adds additional covariates as the proxies for the unobserved traits. (3) As the scales and the actual positions of the estimated latent positions are essentially arbitrary (Hoff et al., 2002), the actual values of the latent positions might be very different from the actual values of the unobserved traits that co-determine influence and selection. However, as long as the estimated latent positions are highly correlated with the unobserved traits (i.e., actors who are close to each other on the latent positions are also close to each other in terms of the unobserved traits), the social influence effects can still be consistently estimated. (4) This approach works in scenarios where there are unobserved traits that co-determine influence and selection (homophily).[4] It does not improve the estimation of social influence effects when the unobserved traits are only present in one process but not the other.

An Empirical Example in R

In this section, I present an empirical example illustrating how to implement the latent space adjusted approach to estimate the social influence effect using R 3.5.2. The data comes from the social network data collected in the Teenage Friends and Lifestyle Study data set (Michell, 2000; Pearson and West, 2003). Friendship network data and substance use were recorded for a cohort of 50 female pupils in a school in the West of Scotland. The panel data were recorded over three years, starting in 1995, when the pupils were aged 13, and ending in 1997. The friendship networks were formed by allowing the pupils to name up to twelve best friends. Pupils were also asked about substance use and adolescent behavior associated with, for instance, lifestyle, sporting behavior, tobacco, alcohol, and cannabis consumption. The question on sporting activity asked if the pupil regularly took part in any sport, or went training for sport, out of school (e.g., football, gymnastics, skating, mountain biking). The school was representative of others in the region in terms of social class composition (Pearson and West, 2003). The key variables used in this example were measured three times from 1995-1997 and included pupils' friendship networks (binary variable representing each possible directed pair, 1 if nominated and 0 otherwise), smoking (measured on a 1-3 scale), drug use (measured on a 1-4 scale), alcohol use (measured on a 1-5 scale) and sport activity (measured on a 1-2 scale). The dataset is available here.

First, I install and load all the packages needed in R. **latentnet** is the package that is used to estimate the latent space model (Krivitsky and Handcock, 2020). And **statnet** is the package to manipulate and create the network object (Handcock et al., 2019).

```
> library(latentnet)
> library(RSiena)
> library(sna)
> library(statnet)
```

[4] In principle this approach could also account for unobserved social-environmental factors that drive influence and selection.

The network data comes with the **RSiena** package (Ripley et al., 2018). I load the attribute data into the current session and create network objects over 3 time points:

```
##Load girls' attributes on smoking, drug use, sport and alcohol use
> s50s<-read.table(”s50-smoke.dat”,header=FALSE)
> s50d<-read.table(”s50-drugs.dat”,header=FALSE)
> s50sp<-read.table(”s50-sport.dat”,header=FALSE)
> s50a<-read.table(”s50-alcohol.dat”, header=FALSE)

## Create network object with attributes for each time point
> g1<-network(s501,directed=TRUE)
> g1%v%”a” <- s50a[,1]
> g1%v%”s” <- s50s[,1]
> g1%v%”sp” <- s50sp[,1]
> g1%v%”d” <- s50d[,1]

> g2<-network(s502,directed=TRUE)
> g2%v%”a” <- s50a[,2]
> g2%v%”s” <- s50s[,2]
> g2%v%”sp” <- s50sp[,2]
> g2%v%”d” <- s50d[,2]

> g3<-network(s503,directed=TRUE)
> g3%v%”a” <- s50a[,3]
> g3%v%”s” <- s50s[,3]
> g3%v%”sp” <- s50sp[,3]
> g3%v%”d” <- s50d[,3]
```

We can plot each network and observe how they have changed over time. Figure 2 shows how these girls′ friendship networks have changed from 1995 to 1997. The network graphs show that there have been considerable network changes over time, and distinct components/clusters have emerged over time.

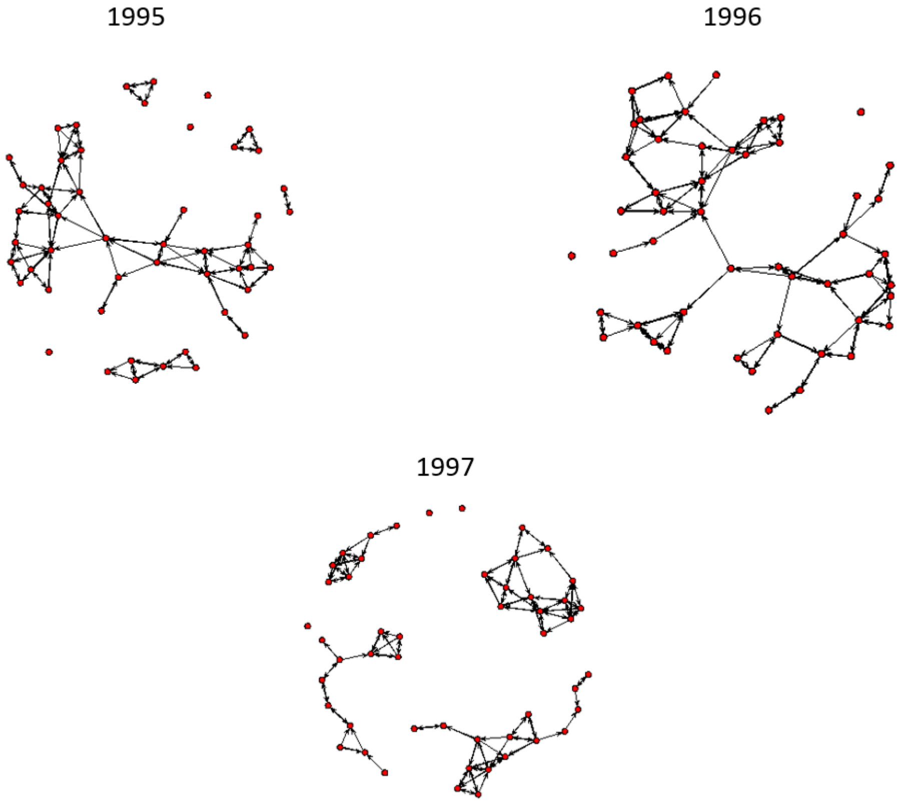

Figure 2: Girls′ friendship network from 1995 to 1997.

My primary research question is whether these girls influence each other′s alcohol use. Here I demonstrate how to estimate the social influence effect by incorporating the latent space adjusted approach with a dynamic linear-in-mean model (Friedkin and Johnsen, 1990) using the "lm" function and a stochastic actor-oriented model using **RSiena** package (Snijders et al., 2010; Ripley et al., 2018) in R. I start by estimating the latent space models using the "ergmm" function to extract the estimated latent positions. Specifically, I estimate two latent space models based on networks in 1995 and 1996 with one dimensional latent space, while controlling for homophily based on observed variables such as alcohol, smoking, drug use and sport:

```
> m1<-ergmm(g1 ~ euclidean(d = 1)+absdiff("a")+absdiff("s")+absdiff("sp")+absdiff("d"),
+ control=ergmm.control(sample.size=5000,burnin=20000,interval=10,Z.delta=5))
> m2<-ergmm(g2 ~ euclidean(d = 1)+absdiff("a")+absdiff("s")+absdiff("sp")+absdiff("d"),
+ control=ergmm.control(sample.size=5000,burnin=20000,interval=10,Z.delta=5))
```

Once the latent space models are estimated, I can extract the latent positions and add them as additional covariates when estimating the behavioral/influence model. First I estimate a dynamic linear-in-mean influence model, which can be represented as (Friedkin and Johnsen, 1990):

$$Y_{it} = \beta_0 + \beta_1 Y_{it-1} + \beta_2 \frac{\sum Z_{ijt-1} Y_{jt-1}}{\sum Z_{ijt-1}} + \beta_3 X_{it} + e_{it}, \tag{4}$$

where Y_{it} is the behavior of i at time t, Y_{it-1} is the previous behavior of i, Z_{ijt-1} is a dummy variable indicating if there is a link from i to j at time $t-1$, i.e., 1 if yes and 0 otherwise, and $\frac{\sum Z_{ijt-1} Y_{jt-1}}{\sum Z_{ijt-1}}$ is the average behaviors at time $t-1$ among the network neighbors of i, and β_2 represents the social influence effect. X_{it} represents other concurrent variables of i that might affect the behavioral outcome Y. To estimate the dynamic linear-in-mean influence model, I first need to construct the dataset used by this model:

```
## create the average alcohol use of each person's friends
> E<-matrix(0,50,3)
for (i in 1:50)
{
if (sum(s501[i,])!=0)
E[i,1]<-(s501[i,]%*%s50a[,1])/sum(s501[i,])
if (sum(s502[i,])!=0)
E[i,2]<-(s502[i,]%*%s50a[,2])/sum(s502[i,])
if (sum(s503[i,])!=0)
E[i,3]<-(s503[i,]%*%s50a[,3])/sum(s503[i,])
}

## create the dataset to estimate the dynamic linear-in-mean influence model
> alcohol<-c(s50a[,3],s50a[,2])
> lag_alc<-c(s50a[,2],s50a[,1])
> expo<-c(E[,2],E[,1])
> drug<-c(s50d[,3],s50d[,2])
> smoke<-c(s50s[,3],s50s[,2])
> sport<-c(s50sp[,3],s50sp[,2])
> latent_pos2<-rep(m2$mkl$Z,2)
> latent_pos1<-rep(m1$mkl$Z,2)
> infl<-data.frame(cbind(alcohol,lag_alc,expo,drug,smoke,sport,
+ latent_pos1,latent_pos2,rep(c(1:50),2),rep(c(1:2),each=50)))
> head(infl)

   alcohol lag_alc     expo drug smoke sport latent_pos1 latent_pos2 V9 V10
1       3       1 4.333333    1     1     1   -5.997364   -8.472008  1   1
2       2       2 4.000000    3     3     1   -7.324663   -2.941830  2   1
3       3       3 2.500000    1     1     1    6.734962    9.064313  3   1
4       2       3 3.000000    1     1     1    6.734962    9.197778  4   1
5       4       3 3.500000    3     1     2    1.945568    7.413702  5   1
6       4       4 5.000000    1     3     2   18.585402    1.648355  6   1
```

We can also look at the correlations between the estimated latent positions and the observed variables:

```
> cor(infl[,1:8])
```

```
             alcohol  lag_alc   expo   drug  smoke   sport  latent_pos1  latent_pos2
alcohol       1.0000    0.699  0.458  0.455  0.386  -0.092    -0.387       -0.317
lag_alc       0.6992    1.000  0.461  0.455  0.465  -0.165    -0.403       -0.364
expo          0.4585    0.461  1.000  0.348  0.416  -0.221    -0.550       -0.241
drug          0.4553    0.455  0.348  1.000  0.592  -0.382    -0.283       -0.453
smoke         0.3863    0.465  0.416  0.592  1.000  -0.224    -0.340       -0.463
sport        -0.0922   -0.165 -0.221 -0.382 -0.224   1.000     0.145        0.162
latent_pos1 -0.3872   -0.403 -0.550 -0.283 -0.340   0.145     1.000        0.150
latent_pos2 -0.3173   -0.364 -0.241 -0.453 -0.463   0.162     0.150        1.000
```

From the correlation table, strong network autocorrelations are observed – one's alcohol use alcohol, previous alcohol use lag_alc, and friends' alcohol use expo are all highly correlated with each other. Furthermore, the estimated latent positions in 1995 and 1996 latent_pos1 and latent_pos2 have sizable correlations with both girls' alcohol use and their friends' alcohol use. As the calculations of the latent positions are already conditioned on homophily based on the observed variables such as alcohol, drug, smoking, and sport, the results suggest that there might be some unobserved variables (e.g., an unobserved tendency for substance abuse) that drive both girls' alcohol use and choice of friends.

To estimate the dynamic linear-in-mean influence model, I first estimate an influence model with the latent positions as the additional covariates and then estimate another model without the latent positions:

```
> summary(lm(alcohol~lag_alc+expo+smoke+sport+drug+latent_pos1+latent_pos2,data=infl))

Call:
lm(formula = alcohol ~ lag_alc + expo + smoke + sport + drug +
    latent_pos1 + latent_pos2, data = infl)

Residuals:
    Min      1Q  Median      3Q     Max
-2.2031 -0.5060  0.1155  0.5177  1.6341

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.625653   0.453098   1.381   0.1707
lag_alc      0.525024   0.084136   6.240 1.32e-08 ***
expo         0.128865   0.083382   1.545   0.1257
smoke       -0.071843   0.120567  -0.596   0.5527
sport        0.235112   0.168289   1.397   0.1658
drug         0.251790   0.122337   2.058   0.0424 *
latent_pos1 -0.007709   0.011007  -0.700   0.4854
latent_pos2 -0.004525   0.014706  -0.308   0.7590
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7715 on 92 degrees of freedom
Multiple R-squared:  0.5426,       Adjusted R-squared:  0.5078
F-statistic: 15.59 on 7 and 92 DF,  p-value: 2.541e-13

> summary(lm(alcohol~lag_alc+expo+smoke+sport+drug,data=infl))

Call:
lm(formula = alcohol ~ lag_alc + expo + smoke + sport + drug,
    data = infl)

Residuals:
    Min      1Q  Median      3Q     Max
-2.2382 -0.4876  0.0384  0.4935  1.6371

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.47833    0.40080   1.193   0.2357
lag_alc      0.53760    0.08160   6.588 2.54e-09 ***
expo         0.15298    0.07516   2.035   0.0446 *
```

```
smoke       -0.05760    0.11602  -0.496   0.6207
sport        0.23565    0.16698   1.411   0.1615
drug         0.26057    0.11873   2.195   0.0307 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7655 on 94 degrees of freedom
Multiple R-squared:  0.5398,        Adjusted R-squared:  0.5154
F-statistic: 22.06 on 5 and 94 DF,  p-value: 1.473e-14
```

Results show that if I only include previous alcohol use and other observed covariates, the social influence effect on alcohol use is significant (coef=.152, se=.075, p=.045) – that is, if these girls' friends use more alcohol, they will also use more alcohol. However, when I include the latent positions as the additional covariates in the model, the social influence effect (coef=.129, se=.083, p=.126) is no longer significant.[5] These results suggest that there are likely to be unobserved variables that drive both girls' alcohol use and choice of friends (e.g. unobserved substance-abuse tendency), and ignoring them will lead to 18% overestimation of the social influence effect in this case, which can lead to erroneous statistical inferences.

Next, I estimate a Stochastic Actor-Oriented Model (SAOM) using **RSiena** to test if there is any social influence effect on girls' alcohol use. SAOM is a class of simulation-based statistical models that can model behavioral and network change simultaneously. In the simulation process, SAOM assumes the underlying time is continuous and that actors control their behavior and outgoing ties. At a given moment, one probabilistically selected actor has the opportunity to change one outgoing tie or small step in his or her behavior. The change follows a Markov process in which small changes in networks and behavior are accumulated in each micro-step, and large differences can then be observed between initial and final networks (Snijders et al., 2010). For statistical inference, the parameter values of the simulation algorithms are selected such that the simulated and observed data resemble each other most closely, and the parameters can be estimated by matching key statistics of the simulated and observed networks via the method of moments, generalized method of moments, or likelihood-based methods (Steglich et al., 2010). SAOM is appealing as it intuitively incorporates both the influence and network-selection process from an individual-level perspective, such that the network-selection effects are adjusted for in the estimation of influence effects. However, estimates from SAOMs are still likely to be biased when unobserved variables are co-determining the probabilities of change in network and/or behavior (Steglich et al., 2010) and thus may benefit from the latent space adjusted approach in these scenarios.

I start by constructing a dataset that can be used by SAOM models for estimation:

```
## create data structure that can be used to estimate SAOM
> friend.data.w1 <- s501
> friend.data.w2 <- s502
> friend.data.w3 <- s503
> drink <- s50a
> smoke <- s50s
> drug  <- s50d
> sport <- s50sp
> friendship <- sienaDependent( array( c( friend.data.w1, friend.data.w2,
+                                         friend.data.w3 ),
+                                 dim = c( 50, 50, 3 ) )
> drinkingbeh <- sienaDependent( drink, type = "behavior" )
> smokingbeh <- varCovar( as.matrix(smoke))
> drugbeh <- varCovar( as.matrix(drug))
> sportbeh <- varCovar( as.matrix(sport))
> lat1<-coCovar(as.vector(m1$mkl$Z)) ## latent position from 1995
> lat2<-coCovar(as.vector(m2$mkl$Z)) ## latent position from 1996
> myCoEvolutionData <- sienaDataCreate( friendship, drinkingbeh,
+                       smokingbeh,drugbeh,sportbeh,lat1,lat2 )
```

To specify the SAOM model, the following codes can be used. Specifically, in the selection part of the model, I include structural effects such as reciprocity, transitivity, popularity, geometrically weighted degree, and homophily based on alcohol, drug use, smoking, sport, and the latent positions. In the behavioral part of the model, I model girls' alcohol use as a function of the linear and quadratic

[5] Latent space model uses a MCMC estimation and thus the results will be slightly different each time. It is suggested to estimate latent space model with longer burn-in, larger sample size, and over multiple times to acquire the final estimates (e.g., using mean or mode of the estimates).

shapes, average similarity effect (i.e., social influence effect), the observed covariates such as drug use, smoking, sport, as well as the latent positions as the additional covariates:

```
> myCoEvolutionEff2 <- getEffects( myCoEvolutionData )
>
> effectsDocumentation(myCoEvolutionEff2)
>
## specify predictors to model selection/network in SAOM
> myCoEvolutionEff2 <- includeEffects( myCoEvolutionEff2, transTrip,
+                                     cycle3,gwespFF,inPop,outPop)
> myCoEvolutionEff2 <- includeEffects( myCoEvolutionEff2, simX,
+                                     interaction1 = "smokingbeh" )
> myCoEvolutionEff2 <- includeEffects( myCoEvolutionEff2, simX,
+                                     interaction1 = "drugbeh" )
> myCoEvolutionEff2 <- includeEffects( myCoEvolutionEff2, simX,
+                                     interaction1 = "sportbeh" )
> myCoEvolutionEff2 <- includeEffects(myCoEvolutionEff2,  simX,
+                                     interaction1 = "drinkingbeh" )
> myCoEvolutionEff2 <- includeEffects(myCoEvolutionEff2,  simX,
+                                     interaction1 = "lat1" )
> myCoEvolutionEff2 <- includeEffects(myCoEvolutionEff2,  simX,
+                                     interaction1 = "lat2" )

## specify predictors to model behavior (alcohol use) in SAOM
> myCoEvolutionEff2 <- includeEffects( myCoEvolutionEff2,
+                                      name = "drinkingbeh",
+                                      avSim,
+                                      interaction1 = "friendship" )
> myCoEvolutionEff2 <- includeEffects( myCoEvolutionEff2,
+                                      name = "drinkingbeh", effFrom,
+                                      interaction1 = "smokingbeh")
> myCoEvolutionEff2 <- includeEffects( myCoEvolutionEff2,
+                                      name = "drinkingbeh", effFrom,
+                                      interaction1 = "drugbeh")
> myCoEvolutionEff2 <- includeEffects( myCoEvolutionEff2,
+                                      name = "drinkingbeh", effFrom,
+                                      interaction1 = "sportbeh")
> myCoEvolutionEff2 <- includeEffects( myCoEvolutionEff2,
+                                      name = "drinkingbeh", effFrom,
+                                      interaction1 = "lat1")
> myCoEvolutionEff2 <- includeEffects( myCoEvolutionEff2,
+                                      name = "drinkingbeh", effFrom,
+                                      interaction1 = "lat2")
```

To estimate the SAOM model, I type:

```
> betterCoEvAlgorithm <- sienaAlgorithmCreate( projname = 's50CoEv_3',
+                              diagonalize = 0.2, doubleAveraging = 0)
>
>
> (ans2 <- siena07( betterCoEvAlgorithm, data = myCoEvolutionData,
+                   effects = myCoEvolutionEff2))

Estimates, standard errors and convergence t-ratios

                                               Estimate   Standard   Convergence
                                                            Error      t-ratio
Network Dynamics
   1. rate constant friendship rate (period 1)  6.7804  ( 1.9520   )   -0.0247
   2. rate constant friendship rate (period 2)  5.5804  ( 1.5074   )   -0.0446
   3. eval outdegree (density)                 -3.8226  ( 0.4710   )   -0.0268
   4. eval reciprocity                          2.2901  ( 0.4352   )   -0.0036
   5. eval transitive triplets                 -1.1221  ( 0.9042   )   -0.0228
   6. eval 3-cycles                             1.1517  ( 0.5529   )   -0.0201
   7. eval GWESP I -> K -> J (69)               2.7667  ( 1.8913   )   -0.0204
   8. eval indegree - popularity               0.1178  ( 0.1121   )   -0.0313
   9. eval outdegree - popularity             -0.5368  ( 0.1570   )   -0.0286
  10. eval lat1 similarity                     0.6507  ( 0.5186   )   -0.0439
  11. eval lat2 similarity                     7.7015  ( 1.3888   )   -0.0198
  12. eval drinkingbeh similarity              0.6110  ( 0.6676   )    0.0086
  13. eval smokingbeh similarity               0.0755  ( 0.2577   )    0.0176
  14. eval drugbeh similarity                  0.8831  ( 0.5177   )   -0.0197
  15. eval sportbeh similarity                 0.2044  ( 0.1843   )    0.0627

Behavior Dynamics
  16. rate rate drinkingbeh (period 1)         1.2506  ( 0.3943   )    0.0629
  17. rate rate drinkingbeh (period 2)         1.7510  ( 0.5416   )    0.0174
  18. eval drinkingbeh linear shape            0.3880  ( 0.1903   )   -0.0095
  19. eval drinkingbeh quadratic shape        -0.1304  ( 0.1459   )   -0.0447
  20. eval drinkingbeh average similarity      3.0265  ( 2.2662   )    0.0059
  21. eval drinkingbeh: effect from lat1      -0.0240  ( 0.0235   )    0.0612
  22. eval drinkingbeh: effect from lat2      -0.0169  ( 0.0324   )    0.0166
  23. eval drinkingbeh: effect from smokingbeh -0.3243  ( 0.3157   )   -0.0706
  24. eval drinkingbeh: effect from drugbeh     0.0538  ( 0.2728   )   -0.0154
  25. eval drinkingbeh: effect from sportbeh    0.3266  ( 0.3720   )    0.0066

Overall maximum convergence ratio:    0.1721


Total of 3944 iteration steps.
```

Results show that there is strong homophily based on the latent positions in the selection process. Furthermore, the estimate for average similarity (i.e., social influence effect) effect is 3.03, and the standard error is 2.27. Next, I compare it with a SAOM model that excludes the latent positions in both selection and behavioral models. Results are shown below:

```
Estimates, standard errors and convergence t-ratios

                                              Estimate   Standard   Convergence
                                                           Error      t-ratio
Network Dynamics
   1. rate constant friendship rate (period 1)  5.6744  ( 1.4262   )    0.0123
   2. rate constant friendship rate (period 2)  4.4861  ( 0.9524   )   -0.0206
   3. eval outdegree (density)                -2.3732  ( 0.2822   )    0.0193
   4. eval reciprocity                         3.0429  ( 0.4632   )    0.0205
   5. eval transitive triplets                -1.4128  ( 0.8951   )    0.0193
   6. eval 3-cycles                            1.7027  ( 0.5465   )    0.0205
   7. eval GWESP I -> K -> J (69)              3.6722  ( 1.7601   )    0.0104
   8. eval indegree - popularity               0.0872  ( 0.1019   )   -0.0118
   9. eval outdegree - popularity             -0.6361  ( 0.1700   )    0.0215
  10. eval drinkingbeh similarity              1.2178  ( 0.7357   )    0.0277
  11. eval smokingbeh similarity              -0.0006  ( 0.2812   )   -0.0166
  12. eval drugbeh similarity                  0.9889  ( 0.4224   )   -0.0231
  13. eval sportbeh similarity                 0.1628  ( 0.1859   )   -0.0149

Behavior Dynamics
  14. rate rate drinkingbeh (period 1)         1.2869  ( 0.3117   )    0.0219
  15. rate rate drinkingbeh (period 2)         1.7214  ( 0.4520   )    0.0173
  16. eval drinkingbeh linear shape            0.3975  ( 0.1840   )   -0.0159
  17. eval drinkingbeh quadratic shape        -0.0542  ( 0.1209   )    0.0014
  18. eval drinkingbeh average similarity      4.0685  ( 2.0968   )    0.0147
  19. eval drinkingbeh: effect from smokingbeh -0.2452  ( 0.3031   )    0.0476
  20. eval drinkingbeh: effect from drugbeh     0.0836  ( 0.2829   )    0.0277
  21. eval drinkingbeh: effect from sportbeh    0.3029  ( 0.3710   )   -0.0065

Overall maximum convergence ratio:    0.1545


Total of 3743 iteration steps.
```

The estimate for the social influence effect is now 4.07, with a standard error of 2.10. As a result, ignoring the latent position will likely lead to 34% overestimation of the social influence effect in this case using the SAOM models.

Discussion and Conclusion

Social influence effects are generally difficult to identify, as influence processes are often entangled with other processes such as selection and social-environmental factors. Here I have shown that this entanglement/difficulty can essentially be framed as an omitted variable bias problem, and a latent space adjusted approach holds promise to correctly identify social influence effects in this case. And I have demonstrated how to use the latent space adjusted approach to estimate various social influence models with existing packages in R. Results show that models that ignore the unobserved variables that drive both influence and selection are likely to overestimate the true social influence effect, while the latent space adjusted approach holds promise to correct that bias and serves as a more conservative test of the true social influence effect.

Although the latent space adjusted approach proposed in this paper is flexible enough to be incorporated with any functional form of the behavioral/influence model, and holds much promise as an alternative approach to identify the social influence effect, several limitations also come with this approach: (1) As previously mentioned, the latent space adjusted approach requires that the same unobserved traits occur in both the influence and the selection process. It can not account for the unobserved traits that are only present in one of the processes but not the other. (2) The choice of the dimensions for the latent positions in the latent space model is not clear. Although I have chosen one-dimensional latent positions in all of the simulations and empirical examples, this does not need to be the case and there is no clear rule deciding how many dimensions users should use. (3) The computation of latent positions is very time-consuming, and the computation time increases significantly with the increase of data or the number of dimensions of the latent positions.

Nevertheless, the latent space adjusted approach proposed here provides a useful and more plausible estimate of the true social influence effect, especially when the entanglement between influence and selection is of concern. This paper contributes to the literature by further illustrating how the latent space adjusted approach may account for bias in the estimation of the social influence effect, as well as how this approach can be easily implemented in R.

Bibliography

S. Aral, L. Muchnik, and A. Sundararajan. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences*, 106(51): 21544–21549, 2009. URL https://doi.org/10.1073/pnas.0908800106. [p1]

S. Asch. Group forces in the modification and distortion of judgments. *Social psychology.*, pages 450–501, 1972. URL https://doi.org/10.1037/10025-016. [p1]

A. Bandura. *Social foundations of thought and action*. Englewood Cliffs, NJ, 1986. [p1]

Y. Bramoulle, H. Djebbari, and B. Fortin. Identification of peer effects through social networks. *Journal of Econometrics*, 150(1):41–55, 2009. URL https://doi.org/10.1016/j.jeconom.2008.12.021. [p1]

J. T. Cacioppo, J. H. Fowler, and N. A. Christakis. Alone in the crowd: The structure and spread of loneliness in a large social network. *SSRN Electronic Journal*, 2008. URL https://doi.org/10.2139/ssrn.1319108. [p1]

N. A. Christakis and J. H. Fowler. The spread of obesity in a large social network over 32 years. *New England Journal of Medicine*, 357(4):370–379, 2007. URL https://doi.org/10.1056/nejmsa066082. [p1]

N. A. Christakis and J. H. Fowler. The collective dynamics of smoking in a large social network. *New England Journal of Medicine*, 358(21):2249–2258, 2008. URL https://doi.org/10.1056/nejmsa0706154. [p1]

Y. Croissant, G. Millo, and K. Tappe. plm: Linear models for panel data, 2021. URL https://CRAN.R-project.org/package=plm. R package version 2.4-1. [p4]

L. Erbring and A. Young. Individuals and social structure. *Sociological Methods and Research*, 7(4): 396–430, 1979. URL https://doi.org/10.1177/004912417900700404. [p1]

S. L. Feld. The focused organization of social ties. *American Journal of Sociology*, 86(5):1015–1035, 1981. URL https://doi.org/10.1086/227352. [p1]

S. L. Feld. Social structural determinants of similarity among associates. *American Sociological Review*, 47(6):797, 1982. URL https://doi.org/10.2307/2095216. [p1]

K. A. Frank, Y. Zhao, and K. Borman. Social capital and the diffusion of innovations within organizations: The case of computer technology in schools. *Sociology of Education*, 77(2):148–171, 2004. URL https://doi.org/10.1177/003804070407700203. [p1]

N. E. Friedkin. Norm formation in social influence networks. *Social Networks*, 23(3):167–189, 2001. URL https://doi.org/10.1016/s0378-8733(01)00036-3. [p1]

N. E. Friedkin and E. C. Johnsen. Social influence and opinions. *The Journal of Mathematical Sociology*, 15(3-4):193–206, 1990. URL https://doi.org/10.1080/0022250x.1990.9990069. [p6]

N. E. Friedkin and E. C. Johnsen. Social influence networks and opinion change. *Advances in Group Processes*, 16(1):1–29, 1999. [p1]

D. German, C. G. Sutcliffe, B. Sirirojn, S. G. Sherman, C. A. Latkin, A. Aramrattana, and D. D. Celentano. Unanticipated effect of a randomized peer network intervention on depressive symptoms among young methamphetamine users in thailand. *Journal of Community Psychology*, 40(7):799–813, Jul 2012. URL https://doi.org/10.1002/jcop.21488. [p1]

M. S. Handcock, D. R. Hunter, C. T. Butts, S. M. Goodreau, P. N. Krivitsky, S. Bender-deMoll, and M. Morris. statnet: Software tools for the statistical analysis of network data, 2019. URL https://CRAN.R-project.org/package=statnet. R package version 2019.6. [p4]

P. D. Hoff, A. E. Raftery, and M. S. Handcock. Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97(460):1090–1098, 2002. URL https://doi.org/10.1198/016214502388618906. [p1, 3, 4]

M. Kalmijn and H. Flap. Assortative meeting and mating: Unintended consequences of organized settings for partner choices. *Social Forces*, 79(4):1289–1312, Jan 2001. URL https://doi.org/10.1353/sof.2001.0044. [p1]

P. N. Krivitsky and M. S. Handcock. latentnet: Latent position and cluster models for statistical networks, 2020. URL https://CRAN.R-project.org/package=latentnet. R package version 2.10.1. [p4]

C. F. Manski. Identification of endogenous social effects: The reflection problem. *The Review of Economic Studies*, 60(3):531, 1993. URL https://doi.org/10.2307/2298123. [p1]

J. M. Mcpherson and L. Smith-Lovin. Homophily in voluntary organizations: Status distance and the composition of face-to-face groups. *American Sociological Review*, 52(3):370, 1987. URL https://doi.org/10.2307/2095356. [p1]

M. Mcpherson, L. Smith-Lovin, and J. M. Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27(1):415–444, 2001. URL https://doi.org/10.1146/annurev.soc.27.1.415. [p1]

M. P. L. Michell. Smoke rings: social network analysis of friendship groups, smoking and drug-taking. *Drugs: Education, Prevention and Policy*, 7(1):21–37, 2000. URL https://doi.org/10.1080/dep.7.1.21.37. [p4]

E. R. Oetting and J. F. Donnermeyer. Primary socialization theory: The etiology of drug use and deviance. i. *Substance Use & Misuse*, 33(4):995–1026, 1998. URL https://doi.org/10.3109/10826089809056252. [p1]

M. Pearson and P. West. Drifting smoke rings. *Connections*, 25(2):59–76, 2003. [p4]

R. Ripley, K. Boitmanis, T. A. Snijders, and F. Schoenenberger. Rsiena: Siena - simulation investigation for empirical network analysis, 2018. URL https://CRAN.R-project.org/package=RSiena. R package version 1.2-12. [p5, 6]

Y. Rosseel and T. D. Jorgensen. lavaan: Latent variable analysis, 2019. URL https://CRAN.R-project.org/package=lavaan. R package version 0.6-5. [p4]

C. R. Shalizi and E. McFowland III. Estimating causal peer influence in homophilous social networks by inferring latent locations. *arXiv*, 33, 2018. URL arXiv:1607.06565. [p1, 4]

C. R. Shalizi and A. C. Thomas. Homophily and contagion are generically confounded in observational social network studies. *Sociological Methods & Research*, 40(2):211–239, 2011. URL https://doi.org/10.1177/0049124111404820. [p1, 2]

S. Shortreed, M. S. Handcock, and P. Hoff. Positional estimation within a latent space model for networks. *Methodology*, 2(1):24–33, 2006. URL https://doi.org/10.1027/1614-2241.2.1.24. [p3]

T. A. Snijders, G. G. V. D. Bunt, and C. E. Steglich. Introduction to stochastic actor-based models for network dynamics. *Social Networks*, 32(1):44–60, 2010. URL https://doi.org/10.1016/j.socnet.2009.02.004. [p1, 6, 8]

C. Steglich, T. A. B. Snijders, and M. Pearson. 8. dynamic networks and behavior: Separating selection from influence. *Sociological Methodology*, 40(1):329–393, 2010. URL https://doi.org/10.1111/j.1467-9531.2010.01225.x. [p2, 8]

G. Szekely, M. L. Rizzo, and N. K. Bakirov. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6):2769–2794, 2007. URL https://doi.org/10.1214/009053607000000505. [p3]

T. W. Valente. Network models and methods for studying the diffusion of innovations. *Models and Methods in Social Network Analysis*, pages 98–116, 1995. URL https://doi.org/10.1017/cbo9780511811395.006. [p1]

T. W. Valente. Social network thresholds in the diffusion of innovations. *Social Networks*, 18(1):69–89, 1996. URL https://doi.org/10.1016/0378-8733(95)00256-1. [p1]

T. J. Vanderweele and W. An. Social networks and causal inference. *Handbooks of Sociology and Social Research Handbook of Causal Analysis for Social Research*, pages 353–374, 2013. URL https://doi.org/10.1007/978-94-007-6094-3_17. [p2]

J. M. Wooldridge. *Econometric analysis of cross section and panel data*. MIT Press, 2011. [p3]

R. Xu. Alternative estimation methods for identifying contagion effects in dynamic social networks: A latent-space adjusted approach. *Social Networks*, 54:101–117, 2018. URL https://doi.org/10.1016/j.socnet.2018.01.002. [p1, 2, 3, 4]

R. Xu. Statistical methods for the estimation of contagion effects in human disease and health networks. *Computational and Structural Biotechnology Journal*, 18:1754–1760, 2020. URL https://doi.org/10.1016/j.csbj.2020.06.027. [p2]

Ran Xu
Department of Allied Health Sciences, University of Connecticut
Storrs, CT. 06269
USA
(ORCiD:0000-0002-5832-9226)
ran.2.xu@uconn.edu