MULTIDIMENSIONAL SCALING, CLUSTERING, AND NETWORK METHOD

LATENT SPACE MODEL FOR SOCIAL INFERENCE

Yi Chen

TEACHERS COLLEGE
COLUMBIA UNIVERSITY

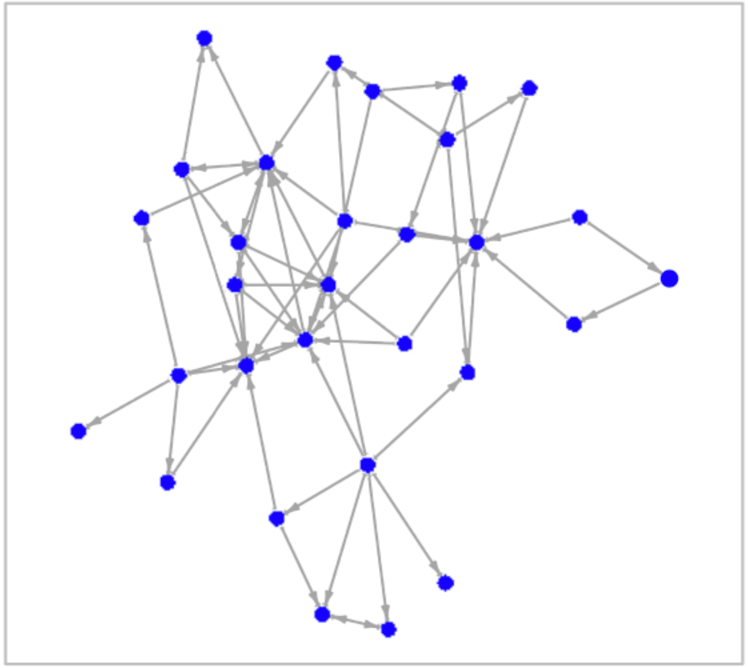
Objectives

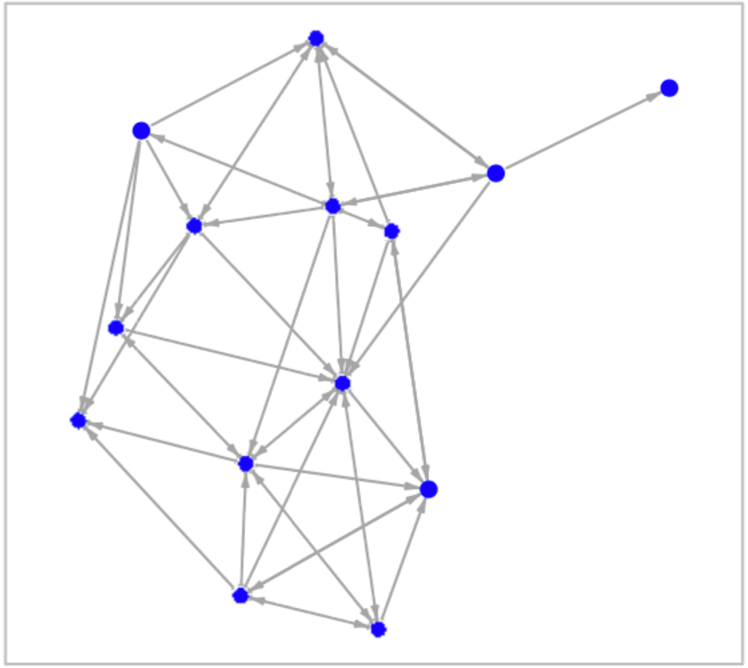
1. Introduce the statistical framework of Latent Space Network (LSN) model;

2. Show an example of using LSN in education using R

Latent Space Model for Social Inference

- Social Inference (also called spillover, contagion or diffusion): individuals become more similar to those with whom they interact or are most closely connected in their social network.





Latent Space Model for Social Inference

- Example of delinquency behavior delinquency:
  - Adolescent's delinquency behavior delinquency may be inferenced by the delinquency behavior delinquency of **friends** and also the delinquency behavior delinquency of the **popular student** (even they are *not* friend *directly*)

    ➔ There is a need to identify the latent influence of the unobserved connections.
  - At the same time, when there is homophonous selection based on this unobserved risk-taking tendency in the networks, such that *adolescents with similar level of risk-taking tendency are more likely to be friends*.

    ➔ The effect of connection in the social network is hard to separate from the effect of other salient individual behavior and unobserved psychological states.

    ➔ "possibility that there may be non-observed variables co-determining the probabilities of change in network and/or behavior" (Steglich, 2010).

- Benefit of using Latent Space Model (LSM) for social inference
  - Identify the underlying mechanism that generate the observed pattern in social network;
  - Specify the latent structure, which is able to identify the dependence/inference when there is no direct tie in the observed social network.
  - The latent structure is also able to represent the unobserved latent co-determining variables.

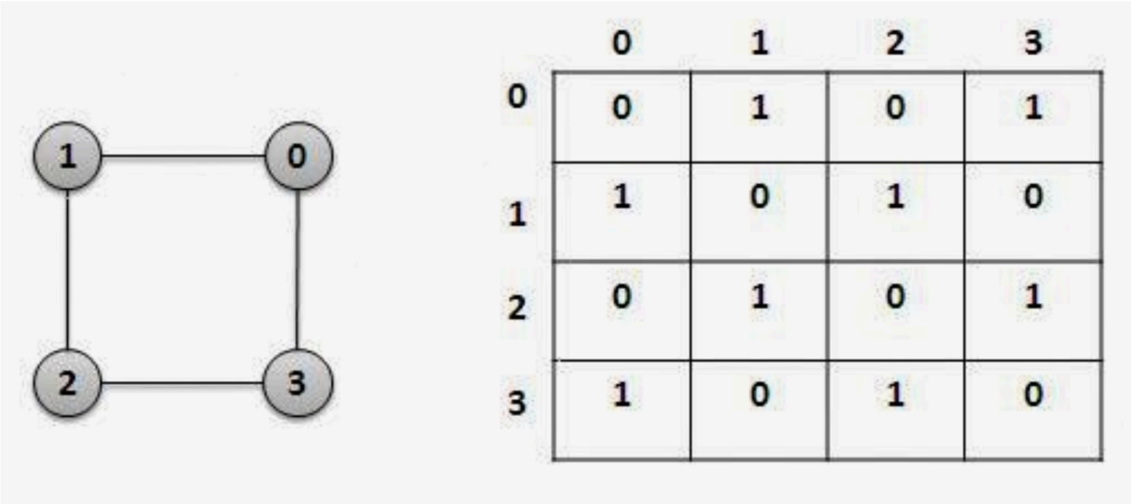
Latent Space Model for Social Inference

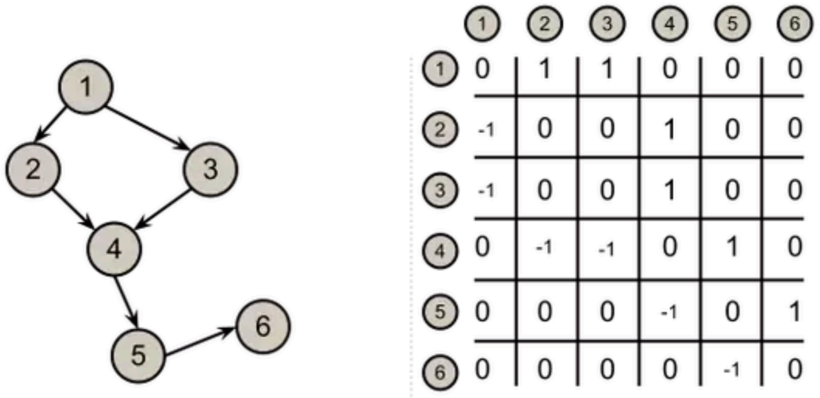
Model framework

1. Coding the social network using adjacent matrix

$$A = \begin{bmatrix} A_{11} & \dots & A_{1n} \\ \dots & \dots & \dots \\ A_{N1} & \dots & A_{NN} \end{bmatrix}$$

, where A_{ij} is the value of the edge from actor i to actor j, which is usually binary value to represent the ties absent or present (*we only focus on the unidirectional network in this presentation*).





Undirected Graph Adjacency Matrix

Latent Space Model for Social Inference

Model framework

2. *Conditional independence assumption*: the ties in the network are independent given the latent space variable.
    - Latent space variables represent the effect of unobserved co-determining variables.

$$P(\boldsymbol{A}|\alpha_0, \boldsymbol{Z}) = \prod_{i \neq j} p(A_{ij}|\boldsymbol{Z_i}, \boldsymbol{Z_i}, \alpha_0)$$

3. *Monotonous assumption*: closer two nodes in the latent space, more likely the tie between the two nodes will be 1 (present).

$$p\left(A_{ij}\middle|\boldsymbol{Z_i}, \boldsymbol{Z_i}, \alpha_0\right) = logit\ \left(P\left(A_{ij} = 1\right)\right) = \alpha_0 - \left|\left|\boldsymbol{Z_i} - \boldsymbol{Z_j}\right|\right|$$

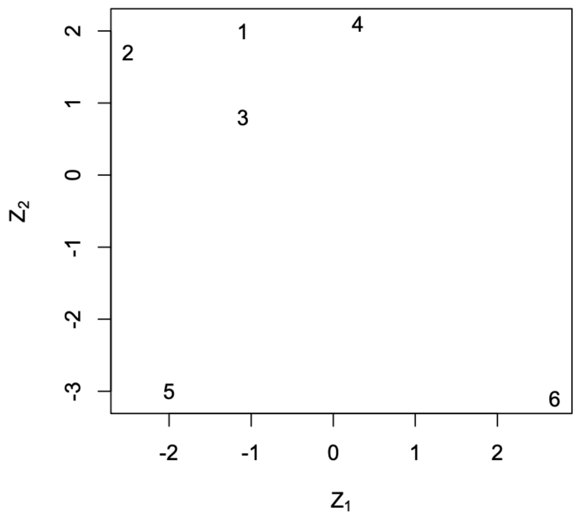
, where α_0 is the intercept term and represent the baseline probability of a tie for node in the network (overall density of the network), $\boldsymbol{Z_i}$ is a low d-dimension vector of latent space location, $\left|\left|\boldsymbol{Z_i} - \boldsymbol{Z_j}\right|\right|$ is the distance between node i and j in the latent space, which can be calculate based on Euclidean distance.

Latent Space Model for Social Inference

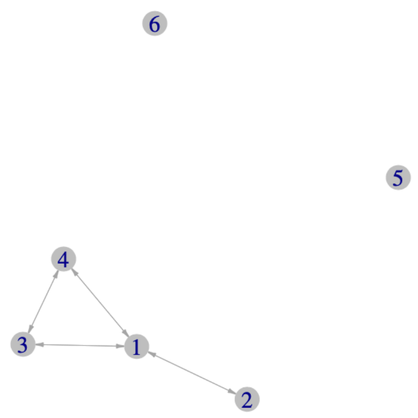
Model framework

$$P(\boldsymbol{A}|\alpha_0, \boldsymbol{Z}) = \prod_{i \neq j} p(A_{ij}|\boldsymbol{Z_i}, \boldsymbol{Z_i}, \alpha_0)$$

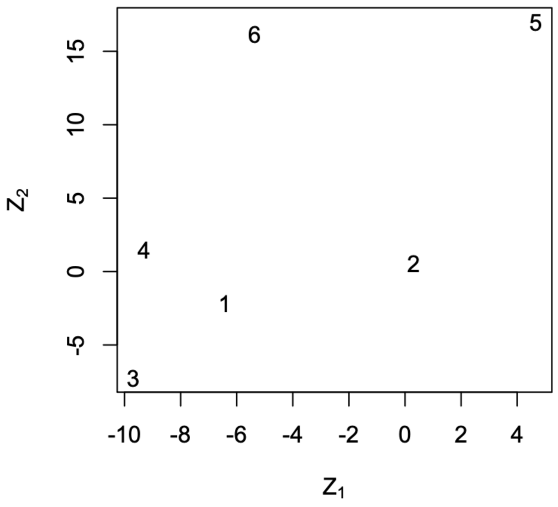
$$p(A_{ij}|\boldsymbol{Z_i}, \boldsymbol{Z_i}, \alpha_0) = logit\left(P(A_{ij} = 1)\right) = \alpha_0 - \left|\left|\boldsymbol{Z_i} - \boldsymbol{Z_j}\right|\right|$$



True latent space



Observed Network



Estimated Latent Space

Latent Space Model for Social Inference

Model framework

4. *LSM model for social inference*:

$$Y_i^t = \beta_0 + \beta_1 Y_i^{t-1} + \beta_2 \sum_{g \in G_i} (w_g Y_g^{t-1}) + \epsilon_i$$

$$w_g = \frac{||Z_g - Z_i||^{-1}}{\sum_{g \in G_i} \left|\left|Z_g - Z_i\right|\right|^{-1}}$$

, Y is the collection of nodal outcomes measured at two different times t and $t-1$, W_g is a weight matrix of neighbor g of the node i (closer the neighbor is, bigger the weight it will have), β_0 is the intercept coefficient, β_1 is the effect of node i outcome at the previous time on the outcome at the current time, and β_2 is the influence of the network on the outcome Y.

Latent Space Model for Social Inference

Bayesian Framework of model (Rjags or Stan in R)

$$
\begin{aligned}
A_{ij} &\sim \text{Bernoulli}\ (p_{ij}) \\
\text{logit}\,(p_{ij}) &= \alpha_0 - ||Z_i - Z_j|| \\
Z_i\ , Z_j &\overset{iid}{\sim} MVN_d(\vec{0}, \tau I) \\
Y^t &\sim N(\beta_0 + \beta_1 Y^{t-1} + \beta_2 N(Z) Y^{t-1}, \sigma^2) \\
\beta_i &\sim N(\mu_0, \sigma_0^2)\ i = 0, 1, 2 \\
\sigma^2 &\sim Inv - Gamma(a, b) \\
\tau_Z &\sim Inv - Gamma(c, d)\ ,
\end{aligned}
$$

The joint posterior distribution can be written as

$$
\begin{aligned}
& p(A, Z, \beta_0, \beta_1, \beta_2, Y^t, Y^{t-1}) \\
& \quad = p(A|Z, \tau) p(Y^t|Y^{t-1}, Z, \beta_0, \beta_1, \beta_2, \sigma^2) p(Z|\tau) p(p(\beta_0) p(\beta_1) p(\beta_2) p(\sigma^2)
\end{aligned}
$$

LSM example

Data Set

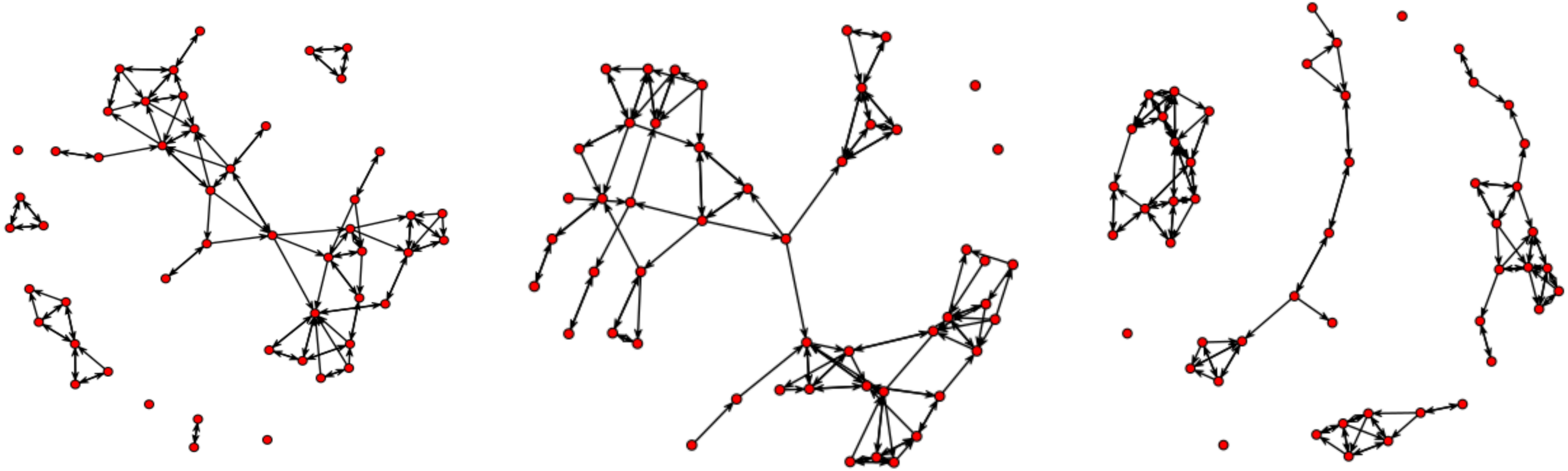
Source: Teenage Friends and Lifestyle Study data set (Michell 2000, Pearson and West 2003).

- https://www.stats.ox.ac.uk/~snijders/siena/s50_data.htm

Component:

- Friendship network: 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 a three year period 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 their attributes on *smoking (s), drug use (d), sport (sp),* and *alcohol use (a)*.

LSM example



Girls’ friendship network from 1995 to 1997

LSM example

Estimation of the latent space position, we 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 of sport (i.e., taking sport as fixed effects)

$$p\left(A_{ij} \middle| \boldsymbol{Z_i}, \boldsymbol{Z_i}, \alpha_0\right) = logit\ \left(P\left(A_{ij} = 1\right)\right) = \alpha_0 + sp - \left|\left|\boldsymbol{Z_i} - \boldsymbol{Z_j}\right|\right|$$

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

note:

1. g1 and g2 are network for year 1995 and 1996 associated with the attribute a, s, sp, and d for each note.
2. ergmm for for fit latent space model
3. euclidean(d=1) means the latent space is one dimension
4. absdiff for calculate the absolute difference

LSM example

Estimation of the latent space model for inference with other observed concurrent variables

$$Y_{it} = \beta_0 + \beta_1 Y_{it-1} + \beta_2 \sum w_g Y_{gt-1} + \beta_3 d_{it} + \beta_4 s_{it} + e_{it}$$

$$w_g = \frac{||Z_g - Z_i||^{-1}}{\sum_{g \in G} \left|\left|Z_g - Z_i\right|\right|^{-1}}$$

In this example, outcome of interest is attribute towards alcohol (a), other observed concurrent variables are their attribute towards drug (d) and smoke (s). Variables sport (sp) has been used in the last step to capture the friendship network and will be ignored in this step.

```
summary(lm(alcohol~lag_alc+w+smoke+drug,data=infl))
```

note:

1. lag_alc is the alcohol is the previous year
2. w is the weighted sum of alcohol for the neighbors, which is based on the latent position estimated in the last step.
3. smoke and drug are two controlling variables

LSM example

Estimation of the latent space model for inference with other observed concurrent variables

$$Y_{it} = \beta_0 + \beta_1 Y_{it-1} + \beta_2 \sum w_g Y_{gt-1} + \beta_3 d_{it} + \beta_4 s_{it} + e_{it}$$

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.60627    0.32076   1.890   0.0618 .
lag_alc      0.54505    0.08017   6.799 9.27e-10 ***
w            0.26380    0.12096   2.181   0.0317 *
smoke       -0.05454    0.11533  -0.473   0.6374
drug         0.16766    0.11544   1.452   0.1497
```

the effect of social inference (w) in this example is 0.264 and significant

R code

```
library(RSiena)
library(latentnet)

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)

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]

plot(g1)
plot(g2)
plot(g3)
```

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


latent_pos_1 <- m1$mkl$Z
latent_pos_2 <- m2$mkl$Z
w <- c()

for (i in 1:50){
  current_position <- latent_pos_2[i]
  neighbor_alcho <- s50a[,2][-i]
  neighbor <- latent_pos_2[-i]
  distances <- c()
  for (j in neighbor){ distances <- c(distances,1/abs(current_position-j))  }
  cur_w <- 0
  for (k in 1:length(distances)){ cur_w <- cur_w + (distances[k] / sum(distances)) * neighbor_alcho[k] }
  w <- c(w,cur_w)
}

for (i in 1:50){
  current_position <- latent_pos_1[i]
  neighbor_alcho <- s50a[,1][-i]
  neighbor <- latent_pos_1[-i]
  distances <- c()
  for (j in neighbor){distances <- c(distances,1/abs(current_position-j)) }
  cur_w <- 0
  for (k in 1:length(distances)){cur_w <- cur_w + (distances[k] / sum(distances)) * neighbor_alcho[k]  }
  w <- c(w,cur_w)
}

alcohol<-c(s50a[,3],s50a[,2])
lag_alc<-c(s50a[,2],s50a[,1])
drug<-c(s50d[,3],s50d[,2])
smoke<-c(s50s[,3],s50s[,2])
infl<-data.frame(cbind(alcohol,lag_alc,w,drug,smoke,rep(c(1:50),2),rep(c(1:2),each=50)))
summary(lm(alcohol~lag_alc+w+smoke+drug,data=infl))
```

Reference

- Sweet, T. and Adhikari S. (2020). A Latent Space Network Model for Social Influence. *Psychometrika*, https://doi.org/10.1007/s11336-020-09700-x.
- Xu, R. (2019) Estimating Social Influence Using Latent Space Adjusted Approach in R. *arXiv preprint,* https://arxiv.org/abs/1903.05999.

Thank you