Network Homework5

Chenrui Xu

I Background

This time we are going to build some EGRMs. The EGRM, which is the short form of exponential random graph model. An EGRM is a true generative statistical model of network structure and characteristics. It is used to handle some complex dependencies of networks. Also, it handles different types of predictors/covariates to predict overall the network characteristics. It is used through well-implemented software. ERGMs are fit using Monte Carlo Markov Chain (MCMC) maximum-likelihood estimation (MLE).

In this project, we are going to use EGRMs in terms of different kinds of data and variables. By comparing the p-values and AIC/BICs, we can get the most significant model.

The data is a network data with four different times. It describes the network of smokers and non-smokers and how those smokers effect non-smokers.

II Data I

```
[']
fr_w1 <- Coevolve$fr_w1
class(fr_w1)
fr_w1 <- as.undirected(fr_w1)
fr_w1 <- asNetwork(fr_w1)

fr_w2 <- Coevolve$fr_w2
fr_w2 <- as.undirected(fr_w2)
fr_w2 <- asNetwork(fr_w2)

fr_w3 <- Coevolve$fr_w3
fr_w3 <- as.undirected(fr_w3)
fr_w3 <- asNetwork(fr_w3)

fr_w4 <- Coevolve$fr_w4
fr_w4 <- as.undirected(fr_w4)
fr_w4 <- asNetwork(fr_w4)
```

After loading data from package, I turned the data from 'igraph' format to network format so that we can continue to the next step.

```{r}
fr_w1
```

```
Network attributes:
  vertices = 37
  directed = FALSE
  hyper = FALSE
  loops = FALSE
  multiple = FALSE
  bipartite = FALSE
  total edges= 125
    missing edges= 0
    non-missing edges= 125

Vertex attribute names:
    gender smoke vertex.names

No edge attributes
```

By picking one dataset, we can find that there are two attributes of vertex. They are gender and smoke. Both of them are factors.

Model 1

So, at first, we'll build null model.

```
Call:
ergm(formula = fr_w1 ~ edges, control = control.ergm(seed = 40))

Iterations:  5 out of 20

Monte Carlo MLE Results:
      Estimate Std. Error MCMC % z value Pr(>|z|)
edges -1.46511    0.09924      0  -14.76   <1e-04 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

      Null Deviance: 923.3  on 666  degrees of freedom
 Residual Deviance: 643.2  on 665  degrees of freedom

 AIC: 645.2    BIC: 649.7    (Smaller is better.)
```

From the model summary, we can see that the AIC is 645 and BIC is 649.

Model 2

Then, we will add two factors to the model.

```r
w1mod1 <- ergm(fr_w1 ~ edges + nodefactor('gender') + nodefactor('smoke'), control=control.ergm(seed = 40))
summary(w1mod1)
```

Starting maximum pseudolikelihood estimation (MPLE):
Evaluating the predictor and response matrix.
Maximizing the pseudolikelihood.
Finished MPLE.
Stopping at the initial estimate.
Evaluating log-likelihood at the estimate.
Call:
ergm(formula = fr_w1 ~ edges + nodefactor("gender") + nodefactor("smoke"),
    control = control.ergm(seed = 40))

Iterations: 4 out of 20

Monte Carlo MLE Results:
                   Estimate Std. Error MCMC % z value Pr(>|z|)
edges   -1.38314   0.17062   0 -8.106   <1e-04 ***
nodefactor.gender.2 -0.05212   0.14550   0 -0.358   0.72
nodefactor.smoke.1 -0.09479   0.17583   0 -0.539   0.59

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

     Null Deviance: 923.3 on 666 degrees of freedom
 Residual Deviance: 642.7 on 663 degrees of freedom

AIC: 648.7   BIC: 662.3   (Smaller is better.)

However, from the statistics, it is not hard to find that both of them are not significant, so we do not need to add them. Instead, I'll use nodematch() function.

Starting maximum pseudolikelihood estimation (MPLE):
Evaluating the predictor and response matrix.
Maximizing the pseudolikelihood.
Finished MPLE.
Stopping at the initial estimate.
Evaluating log-likelihood at the estimate.
Call:
ergm(formula = fr_w1 ~ edges + nodematch("gender", diff = TRUE) +
    nodematch("smoke", diff = TRUE), control = control.ergm(seed = 40))

Iterations: 5 out of 20

Monte Carlo MLE Results:
                   Estimate Std. Error MCMC % z value Pr(>|z|)
edges   -4.8318   0.6036   0 -8.005   <1e-04 ***
nodematch.gender.1 3.9484   0.5970   0   6.614   <1e-04 ***
nodematch.gender.2 4.5252   0.6124   0   7.390   <1e-04 ***
nodematch.smoke.0   0.2033   0.2393   0   0.850   0.395
nodematch.smoke.1   0.2256   0.5960   0   0.379   0.705

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

     Null Deviance: 923.3 on 666 degrees of freedom
 Residual Deviance: 467.9 on 661 degrees of freedom

AIC: 477.9   BIC: 500.4   (Smaller is better.)

From this model, we can get the AIC/BIC equal 477.9/500.4 and only gender node match is

significant. So next step we should delete smoke.

Model 3

In this model, we will add gwesp.

```
Call:
ergm(formula = fr_w1 ~ edges + nodematch("gender", diff = TRUE) +
    gwesp(0.7, fixed = TRUE), control = control.ergm(seed = 40))

Iterations:  2 out of 20

Monte Carlo MLE Results:
                    Estimate Std. Error MCMC % z value Pr(>|z|)
edges                -5.2174     0.4805      0 -10.858   <1e-04 ***
nodematch.gender.1    2.2068     0.5073      0   4.350   <1e-04 ***
nodematch.gender.2    2.6484     0.5312      0   4.986   <1e-04 ***
gwesp.fixed.0.7       0.9065     0.1823      0   4.972   <1e-04 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

     Null Deviance: 923.3  on 666  degrees of freedom
 Residual Deviance: 442.9  on 662  degrees of freedom

AIC: 450.9     BIC: 468.9     (Smaller is better.)
```

The AIC/BIC are much smaller this time and all variables are very significant. So, model 3 is the best model given data 1.

III Data II

Model 1

```
Starting maximum pseudolikelihood estimation (MPLE):
Evaluating the predictor and response matrix.
Maximizing the pseudolikelihood.
Finished MPLE.
Stopping at the initial estimate.
Evaluating log-likelihood at the estimate.
Call:
ergm(formula = fr_w2 ~ edges, control = control.ergm(seed = 40))

Iterations:  5 out of 20

Monte Carlo MLE Results:
      Estimate Std. Error MCMC % z value Pr(>|z|)
edges -1.44553    0.09864      0  -14.65   <1e-04 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

     Null Deviance: 923.3  on 666  degrees of freedom
 Residual Deviance: 649.0  on 665  degrees of freedom

AIC: 651     BIC: 655.5     (Smaller is better.)
```

Here is the null model for data 2.

Model 2
At first, we take the look at the factor addition.

```
Starting maximum pseudolikelihood estimation (MPLE):
Evaluating the predictor and response matrix.
Maximizing the pseudolikelihood.
Finished MPLE.
Stopping at the initial estimate.
Evaluating log-likelihood at the estimate.
Call:
ergm(formula = fr_w2 ~ edges + nodefactor("gender") + nodefactor("smoke"),
    control = control.ergm(seed = 40))

Iterations:  5 out of 20

Monte Carlo MLE Results:
                   Estimate Std. Error MCMC % z value Pr(>|z|)
edges               -1.30606    0.18075      0  -7.226   <1e-04 ***
nodefactor.gender.2 -0.02582    0.14544      0  -0.178   0.859
nodefactor.smoke.1  -0.23070    0.16642      0  -1.386   0.166
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

     Null Deviance: 923.3  on 666  degrees of freedom
 Residual Deviance: 647.0  on 663  degrees of freedom

AIC: 653     BIC: 666.5     (Smaller is better.)
```

They are not significant.

```
Starting maximum pseudolikelihood estimation (MPLE):
Evaluating the predictor and response matrix.
Maximizing the pseudolikelihood.
Finished MPLE.
Stopping at the initial estimate.
Evaluating log-likelihood at the estimate.
Call:
ergm(formula = fr_w2 ~ edges + nodematch("gender", diff = TRUE) +
    nodematch("smoke", diff = TRUE), control = control.ergm(seed = 40))

Iterations:  5 out of 20

Monte Carlo MLE Results:
                    Estimate Std. Error MCMC % z value Pr(>|z|)
edges                -4.5237     0.4810      0  -9.405   <1e-04 ***
nodematch.gender.1    3.4577     0.4741      0   7.293   <1e-04 ***
nodematch.gender.2    4.0387     0.4924      0   8.203   <1e-04 ***
nodematch.smoke.0     0.5355     0.2384      0   2.246   0.0247 *
nodematch.smoke.1     0.4176     0.4547      0   0.918   0.3584
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

      Null Deviance: 923.3  on 666  degrees of freedom
 Residual Deviance: 479.9  on 661  degrees of freedom

AIC: 489.9    BIC: 512.4    (Smaller is better.)
```

So, there is the model 2. AIC and BIC are smaller. Half of the smoke match is not significant but the rest is significant.

Model 3

```
Call:
ergm(formula = fr_w2 ~ edges + nodematch("gender", diff = TRUE) +
    nodematch("smoke", diff = TRUE) + gwesp(0.7, fixed = TRUE),
    control = control.ergm(seed = 40))

Iterations:  2 out of 20

Monte Carlo MLE Results:
                    Estimate Std. Error MCMC % z value Pr(>|z|)
edges                -5.3131     0.4171      0 -12.737   <1e-04 ***
nodematch.gender.1    1.8367     0.3640      0   5.046   <1e-04 ***
nodematch.gender.2    2.2693     0.3971      0   5.715   <1e-04 ***
nodematch.smoke.0     0.3899     0.1903      0   2.049   0.0404 *
nodematch.smoke.1     0.5677     0.4033      0   1.407   0.1593
gwesp.fixed.0.7       1.0118     0.1940      0   5.216   <1e-04 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

      Null Deviance: 923.3  on 666  degrees of freedom
 Residual Deviance: 447.3  on 660  degrees of freedom

AIC: 459.3    BIC: 486.3    (Smaller is better.)
```

Here is the result if we keep all match variables. The AIC/BIC is the smallest one. So, the model 3 is the best one.

IV Data III
Model 1

```
Starting maximum pseudolikelihood estimation (MPLE):
Evaluating the predictor and response matrix.
Maximizing the pseudolikelihood.
Finished MPLE.
Stopping at the initial estimate.
Evaluating log-likelihood at the estimate.
Call:
ergm(formula = fr_w3 ~ edges, control = control.ergm(seed = 40))

Iterations:  5 out of 20

Monte Carlo MLE Results:
       Estimate Std. Error MCMC % z value Pr(>|z|)
edges -1.41660    0.09776      0  -14.49   <1e-04 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

     Null Deviance: 923.3  on 666  degrees of freedom
 Residual Deviance: 657.6  on 665  degrees of freedom

AIC: 659.6     BIC: 664.1     (Smaller is better.)
```

Here is the null model. It is not hard to find that all null models are similar according to AIC/BIC.

Model 2

```
Starting maximum pseudolikelihood estimation (MPLE):
Evaluating the predictor and response matrix.
Maximizing the pseudolikelihood.
Finished MPLE.
Stopping at the initial estimate.
Evaluating log-likelihood at the estimate.
Call:
ergm(formula = fr_w3 ~ edges + nodefactor("gender") + nodefactor("smoke"),
    control = control.ergm(seed = 40))

Iterations:  5 out of 20

Monte Carlo MLE Results:
                    Estimate Std. Error MCMC % z value Pr(>|z|)
edges                -1.3046     0.1764      0  -7.395   <1e-04 ***
nodefactor.gender.2  -0.0266     0.1431      0  -0.186    0.853
nodefactor.smoke.1   -0.1437     0.1523      0  -0.943    0.345
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

     Null Deviance: 923.3  on 666  degrees of freedom
 Residual Deviance: 656.6  on 663  degrees of freedom

AIC: 662.6     BIC: 676.1     (Smaller is better.)
```

Still not significant in terms of two factors.

```
Call:
ergm(formula = fr_w3 ~ edges + nodematch("gender", diff = TRUE) +
    nodematch("smoke", diff = TRUE), control = control.ergm(seed = 40))

Iterations:  5 out of 20

Monte Carlo MLE Results:
                    Estimate Std. Error MCMC % z value Pr(>|z|)
edges                -4.5227     0.4464      0 -10.132  < 1e-04 ***
nodematch.gender.1    3.3222     0.4379      0   7.587  < 1e-04 ***
nodematch.gender.2    3.9601     0.4610      0   8.591  < 1e-04 ***
nodematch.smoke.0     0.7540     0.2395      0   3.148  0.00164 **
nodematch.smoke.1     1.0547     0.3859      0   2.733  0.00627 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

     Null Deviance: 923.3  on 666  degrees of freedom
 Residual Deviance: 482.9  on 661  degrees of freedom

 AIC: 492.9    BIC: 515.4    (Smaller is better.)
```

If using match function, the model can fit well as all variables are all significant. And there is a great reduce for AIC/BIC.

Model 3

```
Call:
ergm(formula = fr_w3 ~ edges + nodematch("gender", diff = TRUE) +
    nodematch("smoke", diff = TRUE) + gwesp(0.7, fixed = TRUE),
    control = control.ergm(seed = 40))

Iterations:  2 out of 20

Monte Carlo MLE Results:
                    Estimate Std. Error MCMC % z value Pr(>|z|)
edges                -5.2769     0.4057      0 -13.008  < 1e-04 ***
nodematch.gender.1    1.8850     0.3517      0   5.360  < 1e-04 ***
nodematch.gender.2    2.3718     0.3973      0   5.970  < 1e-04 ***
nodematch.smoke.0     0.6126     0.1901      0   3.222 0.001274 **
nodematch.smoke.1     1.0321     0.3101      0   3.329 0.000872 ***
gwesp.fixed.0.7       0.9052     0.1884      0   4.806  < 1e-04 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

     Null Deviance: 923.3  on 666  degrees of freedom
 Residual Deviance: 456.4  on 660  degrees of freedom

 AIC: 468.4    BIC: 495.4    (Smaller is better.)
```

By adding gwesp, AIC/BIC still going down. So the model 3 is always the best model.

V Data IV
Model 1

```
Call:
ergm(formula = fr_w4 ~ edges, control = control.ergm(seed = 40))

Iterations:  5 out of 20

Monte Carlo MLE Results:
      Estimate Std. Error MCMC % z value Pr(>|z|)
edges -1.38817    0.09693      0  -14.32   <1e-04 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

     Null Deviance: 923.3  on 666  degrees of freedom
 Residual Deviance: 666.0  on 665  degrees of freedom

AIC: 668     BIC: 672.5     (Smaller is better.)
```

Model 2

```
Call:
ergm(formula = fr_w4 ~ edges + nodefactor("gender") + nodefactor("smoke"),
    control = control.ergm(seed = 40))

Iterations:  5 out of 20

Monte Carlo MLE Results:
                     Estimate Std. Error MCMC % z value Pr(>|z|)
edges                -1.34838    0.18193      0  -7.412   <1e-04 ***
nodefactor.gender.2  -0.12060    0.14284      0  -0.844    0.398
nodefactor.smoke.1    0.07309    0.14291      0   0.511    0.609
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

     Null Deviance: 923.3  on 666  degrees of freedom
 Residual Deviance: 665.0  on 663  degrees of freedom

AIC: 671     BIC: 684.5     (Smaller is better.)
```

Not possible for factors to be significant in this data.....lol

```
call:
ergm(formula = fr_w4 ~ edges + nodematch("gender", diff = TRUE) +
    nodematch("smoke", diff = TRUE), control = control.ergm(seed = 40))

Iterations:  5 out of 20

Monte Carlo MLE Results:
                    Estimate Std. Error MCMC % z value Pr(>|z|)
edges               -4.5008     0.4445       0 -10.126  < 1e-04 ***
nodematch.gender.1   3.4330     0.4394       0   7.813  < 1e-04 ***
nodematch.gender.2   3.9412     0.4629       0   8.514  < 1e-04 ***
nodematch.smoke.0    0.6028     0.2427       0   2.484 0.012990 *
nodematch.smoke.1    1.2456     0.3366       0   3.701 0.000215 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

     Null Deviance: 923.3  on 666  degrees of freedom
 Residual Deviance: 485.8  on 661  degrees of freedom

AIC: 495.8    BIC: 518.3    (Smaller is better.)
```

All good, without any hesitation.

Model 3

```
mcmc.diagnostics() function.
call:
ergm(formula = fr_w4 ~ edges + nodematch("gender", diff = TRUE) +
    nodematch("smoke", diff = TRUE) + gwesp(0.7, fixed = TRUE),
    control = control.ergm(seed = 40))

Iterations:  2 out of 20

Monte Carlo MLE Results:
                    Estimate Std. Error MCMC % z value Pr(>|z|)
edges               -5.4540     0.4138       0 -13.179  < 1e-04 ***
nodematch.gender.1   1.8288     0.3357       0   5.448  < 1e-04 ***
nodematch.gender.2   2.2202     0.3598       0   6.171  < 1e-04 ***
nodematch.smoke.0    0.5710     0.1922       0   2.970 0.00297 **
nodematch.smoke.1    1.0603     0.2640       0   4.016  < 1e-04 ***
gwesp.fixed.0.7      1.0322     0.1918       0   5.382  < 1e-04 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

     Null Deviance: 923.3  on 666  degrees of freedom
 Residual Deviance: 454.9  on 660  degrees of freedom

AIC: 466.9    BIC: 493.9    (Smaller is better.)
```

```{r}
plogis(coef(W1mod0))
plogis(coef(W2mod0))
plogis(coef(W3mod0))
plogis(coef(W4mod0))
```

```
      edges
0.1876877
      edges
0.1906907
      edges
0.1951952
      edges
0.1996997
```

Conclusion

For all four datasets, null model always the worst one while model with match and gwesp is always the best one. Factors in these models are always non-significant. Match function is a very interesting method to reduce AIC/BIC. Gwesp can reduce the AIC/BIC based on the match function.

VI Appendix

```
---
title: "Network_HW5"
author: "Chenrui Xu"
date: "2021/3/21"
output: html_document
---
```

```{r}
library(igraph)
library(UserNetR)
library(ergm)
library(intergraph)
```

```{r}
data("Coevolve")
Coevolve
```

```{r}
fr_w1 <- Coevolve$fr_w1
class(fr_w1)
fr_w1 <- as.undirected(fr_w1)
fr_w1 <- asNetwork(fr_w1)
```

```r
fr_w2 <- Coevolve$fr_w2
fr_w2 <- as.undirected(fr_w2)
fr_w2 <- asNetwork(fr_w2)

fr_w3 <- Coevolve$fr_w3
fr_w3 <- as.undirected(fr_w3)
fr_w3 <- asNetwork(fr_w3)

fr_w4 <- Coevolve$fr_w4
fr_w4 <- as.undirected(fr_w4)
fr_w4 <- asNetwork(fr_w4)
```

```{r}
fr_w1
```

```{r}
W1mod0 <- ergm(fr_w1 ~ edges, control=control.ergm(seed = 40))
summary(W1mod0)
```

```{r}
W1mod1 <- ergm(fr_w1 ~ edges + nodefactor('gender') + nodefactor('smoke'),
control=control.ergm(seed = 40))
summary(W1mod1)
```

```{r}
W1mod2 <- ergm(fr_w1 ~ edges
              +nodefactor('gender')+nodefactor('smoke')
              + nodematch('gender') +
                nodematch('smoke'), control=control.ergm(seed = 40))
summary(W1mod2)
```

```{r}
W1mod3 <- ergm(fr_w1 ~ edges
              #+ nodefactor('gender') + nodefactor('smoke')
              + nodematch('gender', diff=TRUE)+ nodematch('smoke', diff=TRUE),
              control=control.ergm(seed = 40))
summary(W1mod3)
```

```{r}
W1mod4 <- ergm(fr_w1 ~ edges
```

```
                    +       nodematch('gender',       diff=TRUE)+       gwesp(0.7,       fixed=TRUE),
control=control.ergm(seed = 40))
summary(W1mod4)
```

```{r}
W2mod0 <- ergm(fr_w2 ~ edges, control=control.ergm(seed = 40))
summary(W2mod0)
```

```{r}
W2mod1  <-  ergm(fr_w2  ~  edges  +  nodefactor('gender')  +  nodefactor('smoke'),
control=control.ergm(seed = 40))
summary(W2mod1)
```

```{r}
W2mod2 <- ergm(fr_w2 ~ edges
                + nodematch('gender', diff=TRUE) +
                  nodematch('smoke', diff=TRUE), control=control.ergm(seed = 40))
summary(W2mod2)
```

```{r}
W2mod3 <- ergm(fr_w2 ~ edges + nodematch('gender', diff=TRUE)
                + nodematch('smoke', diff=TRUE)
                + gwesp(0.7, fixed=TRUE), control=control.ergm(seed = 40))
summary(W2mod3)
```

```{r}
W3mod0 <- ergm(fr_w3 ~ edges, control=control.ergm(seed = 40))
summary(W3mod0)
```

```{r}
W3mod1  <-  ergm(fr_w3  ~  edges  +  nodefactor('gender')  +  nodefactor('smoke'),
control=control.ergm(seed = 40))
summary(W3mod1)
```

```{r}
W3mod2 <- ergm(fr_w3 ~ edges + nodematch('gender', diff=TRUE) +
                  nodematch('smoke', diff=TRUE), control=control.ergm(seed = 40))
summary(W3mod2)
```

```{r}
W3mod3 <- ergm(fr_w3 ~ edges + nodematch('gender', diff=TRUE) +
```

```
                    nodematch('smoke',    diff=TRUE)    +    gwesp(0.7,    fixed=TRUE),
control=control.ergm(seed = 40))
summary(W3mod3)
```


```{r}
W4mod0 <- ergm(fr_w4 ~ edges, control=control.ergm(seed = 40))
summary(W4mod0)
```


```{r}
W4mod1  <-  ergm(fr_w4  ~  edges  +  nodefactor('gender')  +  nodefactor('smoke'),
control=control.ergm(seed = 40))
summary(W4mod1)
```


```{r}
W4mod2 <- ergm(fr_w4 ~ edges + nodematch('gender', diff=TRUE) +
                    nodematch('smoke', diff=TRUE), control=control.ergm(seed = 40))
summary(W4mod2)
```


```{r}
W4mod3 <- ergm(fr_w4 ~ edges + nodematch('gender', diff=TRUE) +
                    nodematch('smoke',    diff=TRUE)    +    gwesp(0.7,    fixed=TRUE),
control=control.ergm(seed = 40))
summary(W4mod3)
```


```{r}
plogis(coef(W1mod0))
plogis(coef(W2mod0))
plogis(coef(W3mod0))
plogis(coef(W4mod0))
```
```