Advanced Network Analysis

ERGM Application

Olga Chyzh [www.olgachyzh.com]

Readings

- Steven M. Goodreau, James A. Kitts, and Martina Morris. Birds of a feather, or friend of a friend? using exponential random graph models to investigate adolescent social networks. *Demography*, 46(1):103--125, 2009.

Goodreau et al, 2009

Goal: to identify the determinants of friendship formation that lead to pervasive regularities in friendship structure among adolescent students

Three mechanisms:

- *Sociality*---heterogeneity among individuals in their propensity to establish friendship ties. Individuals with greater sociality have higher degree, although degree may be also influenced by other factors.

- *Selective mixing* is a dyad-level process by which pairs form (or break) relationships based on their combination of individual attributes. *Assortative mixing* is the greater propensity to partner with others having attributes similar to one's own. The resulting pattern---homophily--- is the predominance of within-group ties.

- *Triad Closure* leads to the outcome of transitivity. Possible mechanisms include increased chance for interaction and tendency for structural balance (i.e. a friend of my friend is my friend).
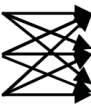
Goodreau et al, 2009

Additional mechanisms:

- Homophily may also be amplified by triad closure if there is already a tendency toward assortative mixing.
- Transitivity may also result from assortative mixing since increasing the likelihood of within category ties enhances the opportunity for completed triangles within categories, especially when groups are small.
- Population composition---the opportunity for partner selection is constrained by the available pool of partners.

Figure 2. Process and Outcome in Social Network Models

Relationships between process and outcome implied by individual analyses

Process		Outcome
Sociality	⟶	Degree distribution
Selective mixing	⟶	Mixing pattern
Triad closure	⟶	Transitivity

Relationships between process and outcome considered here

Process		Outcome
Sociality		Degree distribution
Selective mixing	⤬	Mixing pattern
Triad closure		Transitivity

Notes: Specific forms of selective mixing include assortative mixing and disassortative mixing. Corresponding specific forms of mixing pattern include homophily and heterophily.

Data

- Friendship data from the first wave of Add Health, a sample of more than 90,000 U.S. students in grades 7 through 12, obtained in 1994–1995 through a stratifed sample of schools.
- The questionnaire provided a school roster and asked students to identify their five best male and five best female friends, in order of closeness.
- Students were allowed to nominate friends outside school or missing from the roster, or to stop before nominating five friends of either sex.
- Most students listed fewer friends than the maximum, but for the remainder, there may be some truncation. *When would this be a problem?*

Open Data

```
rm(list=ls())
library(statnet)
data(faux.mesa.high)
mesa <- faux.mesa.high
mesa
```

```
##  Network attributes:
##   vertices = 205
##   directed = FALSE
##   hyper = FALSE
##   loops = FALSE
##   multiple = FALSE
##   bipartite = FALSE
##   total edges= 203
##     missing edges= 0
##     non-missing edges= 203
##
##  Vertex attribute names:
##     Grade Race Sex
##
## No edge attributes
```

Plot Data

```
par(mfrow=c(1,1)) # Back to 1-panel plots
plot(mesa, vertex.col='Grade')
legend('bottomleft',fill=7:12,
       legend=paste('Grade',7:12),cex=0.75)
```

Plot Data

Legend:
- Grade 7
- Grade 8
- Grade 9
- Grade 10
- Grade 11
- Grade 12
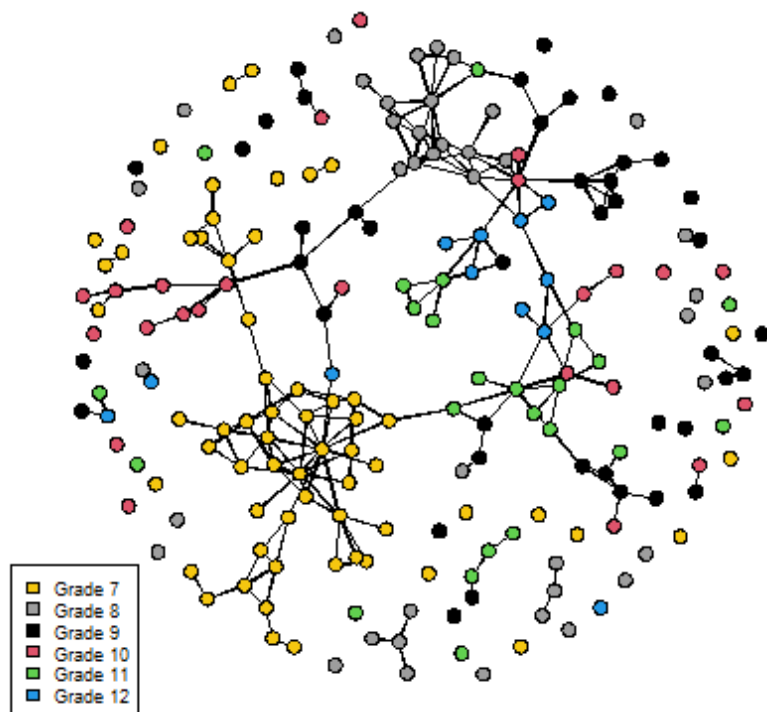
Model Specification

Sociality

Goodreau et al (2009, 111) "infer sociality based on counts of ties observed: s represents the total number of ties, and k_i is the total number of ties for all persons with attribute value i. The s term acts as an intercept [`edges`], and the coefficient for s represents the conditional log-odds of a tie for the reference category (in these models, reference categories are grade 7, white, and male). The k_i terms assume homogeneity within attribute class, allowing each race, sex, and grade to have different mean sociality."

Sociality

Can also use the search function to find the relevant terms.

```
search.ergmTerms(keyword='sociality')
```

```
## Found  8  matching ergm terms:
## b1sociality(nodes=-1)
##   Degree
##
## b1sociality(nodes=-1, form="sum")
##   Degree
##
## b2sociality(nodes=-1)
##   Degree
##
## b2sociality(nodes=-1, form="sum")
##   Degree
##
## receiver(base=1, nodes=-1)
##   Receiver effect
##
## receiver(base=1, nodes=-1, form="sum")
##   Receiver effect
```

Sociality

- None of these match the authors' description, which sounds like they are using the term `nodefactor`.
- Term `nodefactor` adds multiple network statistics to the model, one for each of the unique values of the attribute. Each of these statistics gives the number of times a node with that attribute or those attributes appears in an edge in the network.
- Note: `nodefactor` assumes non-numeric (e.g., character, factor) class, so recode "Grade" as a character.

Sociality

Based on this description, the model so far is specified as:

```
table(mesa %v% 'Race') # Frequencies of race
table(mesa %v% 'Sex') # Frequencies of sex
table(mesa %v% 'Grade') # Frequencies of Grade
mesa %v% "Grade" <-as.character(mesa %v% "Grade")

m1<- ergm(mesa~edges+nodefactor("Race",levels=c("Black","Hisp","Other
          nodefactor("Sex", levels="F")+
          nodefactor("Grade", levels=c("8","9","10","11","12")))
```

Sociality

- Note that we specified levels for each `nodefactor` as to make sure that the reference categories match those of Goodreau et al.
- The description of `nodefactor` in `?ergm.terms` says that including each level is not a good idea (need a reference category, just like with categorical variables in OLS).
- If we get a positive coefficient on any categories within *Race*, *Sex*, and/or *Grade*, we will infer that students with that characteristic are, on average, more social.

Selective Mixing

Two selective mixing dynamics:

1. A homogeneous propensity for assortative mixing across attribute categories ("uniform homophily").
2. A propensity that is specific to individual categories ("differential homophily").

"Statistics are as follows: first, h is the total number of ties between persons in the same attribute category, regardless of category. This uniform homophily is used for sex since there are only three tie types (MM, MF, FF); with main effects included, only one degree of freedom remains. Second, h_i is the total number of ties between persons both in attribute category i. There is one such statistic for each category of the attribute. This differential homophily is used for race and grade 4."

Selective Mixing

Can also use the search function to find the relevant terms.

```
search.ergmTerms(keyword='homophily')
```

```
## Found  13  matching ergm terms:
## b1degrange(from, to=+Inf, by=NULL, homophily=FALSE, levels=NULL)
##  Degree range for the first mode in a bipartite (a.k.a. two-mode) network
##
## b1nodematch(attr, diff=FALSE, keep=NULL, alpha=1, beta=1,)
##  Nodal attribute-based homophily effect for the first mode in a bipartite
##
## b2degrange(from, to=+Inf, by=NULL, homophily=FALSE, levels=NULL)
##  Degree range for the second mode in a bipartite (a.k.a. two-mode) network
##
## b2nodematch(attr, diff=FALSE, keep=NULL, alpha=1, beta=1,)
##  Nodal attribute-based homophily effect for the second mode in a bipartite
##
## degrange(from, to=+Inf, by=NULL, homophily=FALSE, levels=NULL)
##  Degree range
##
## degree(d, by=NULL, homophily=FALSE, levels=NULL)
##  Degree
```

Selective Mixing

Based on this description, we can further specify the model as:

```r
m1<- ergm(mesa~edges+nodefactor("Race",levels=c("Black","Hisp","NatAr
          nodefactor("Sex", levels="F")+
          nodefactor("Grade",  levels=c("8","9","10","11","12"))+
          nodematch("Sex")+nodematch("Race",diff=TRUE)+nodematch("G
          )
```

Note: option `diff` specifies whether there can be differences in probability of friendships between groups.

- Use `absdiff` instead of `nodematch` for continuous variables (e.g. wealth).

Selective Mixing: Interpretation

- To calculate the probability of a tie between two women, set `nodefactor("Sex", levels="F")`=1 and `nodematch("Sex")`=1, and all other variables to the values of interest.

- To calculate the probability of a tie between two men, set `nodefactor("Sex", levels="F")`=0 and `nodematch("Sex")`=1, and all other variables to the values of interest.

- To calculate the probability of a tie between a woman and a man, set `nodefactor("Sex", levels="F")`=1 and `nodematch("Sex")`=0, and all other variables to the values of interest.

Triad Closure

"For the reasons described above, we investigate triad closure using the GWESP statistic. We adopt a value of 0.25 for decay, although results are robust to this choice."

Based on this description, we can further specify the model as:

```
mesa %v% "Grade" <-as.character(mesa %v% "Grade")
m1<- ergm(mesa~edges+nodefactor("Race",levels=c("Black","Hisp","NatAr
           nodefactor("Sex", levels="F")+
           nodefactor("Grade", levels=c("8","9","10","11","12"))+
           nodematch("Sex")+nodematch("Race",diff=TRUE)+nodematch("(
           gwesp(decay=0.25)
           )
par(mfrow = c(3, 2))
mcmc.diagnostics(m1)
```
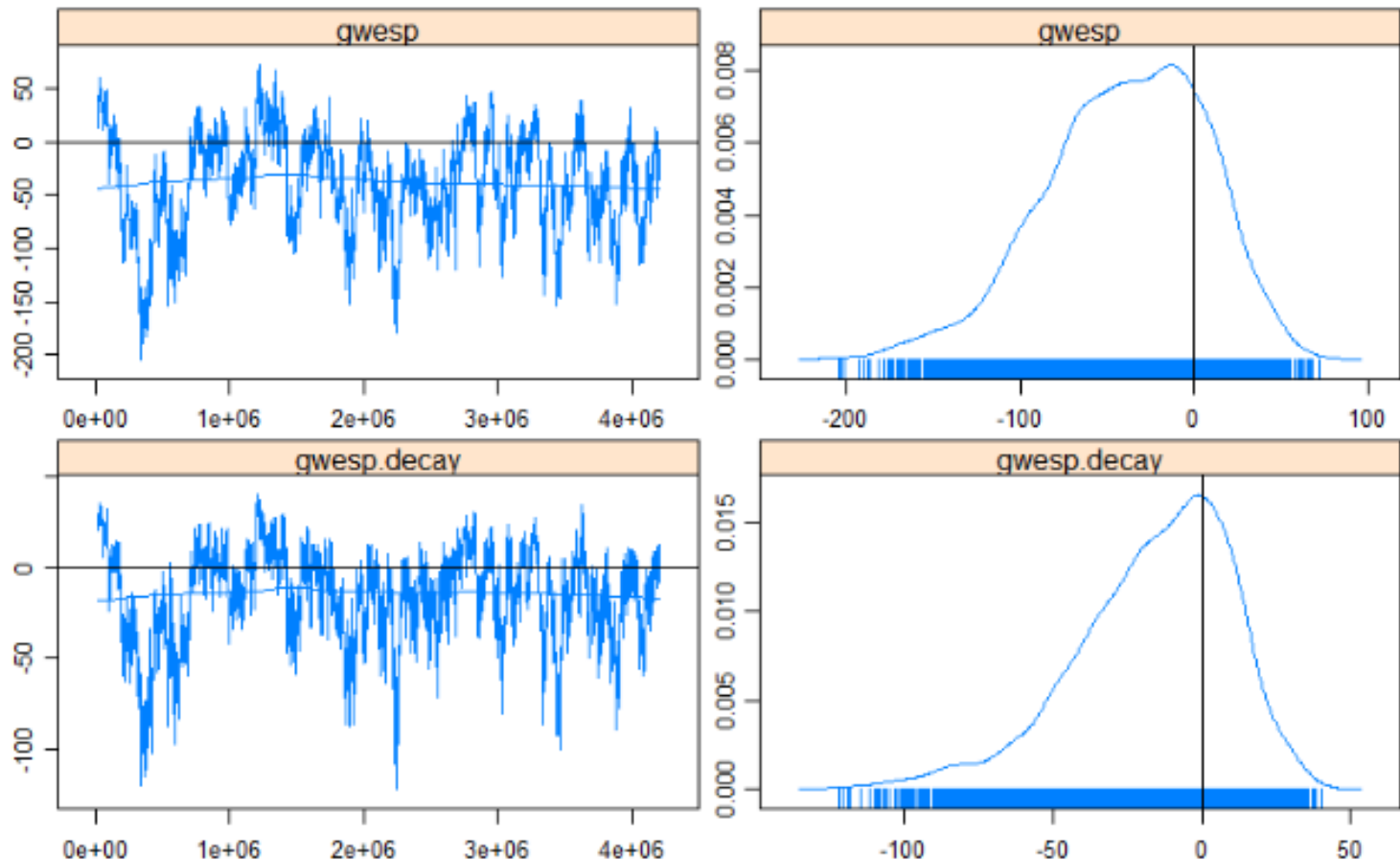
Results

Figure 3.   Coefficients From the Full Model, Plotted Across All 59 Schools
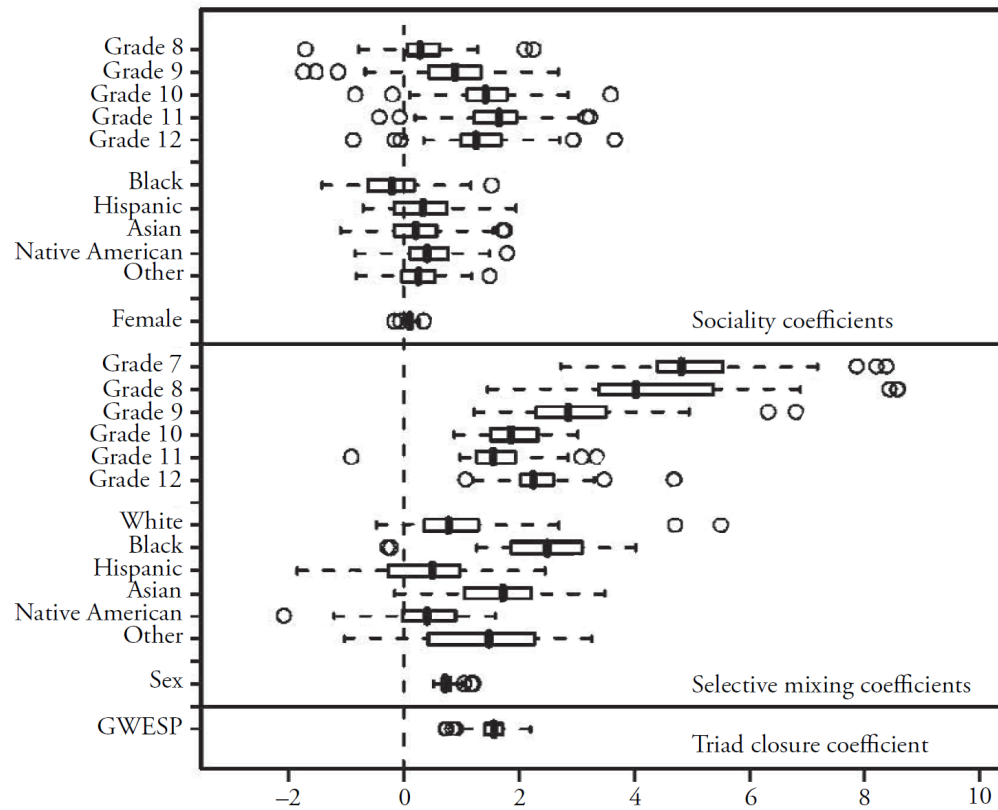
Notes: Boxplots follow the Tukey method. Boxes represent quartiles; whiskers extend to the most extreme data point within 1.5 times the interquartile range from the edge of the box; and points represent outliers.

Our Replication: M1 Diagnostics

Sample statistics

Increase the Number of Simulations.

```r
m2<- ergm(mesa~edges+nodefactor("Race",levels=c("Black","Hisp","NatAm
          nodefactor("Sex", levels="F")+
          nodefactor("Grade", levels=c("8","9","10","11","12"))+
          nodematch("Sex")+nodematch("Race",diff=TRUE)+
          nodematch("Grade",diff=TRUE)+gwesp(decay=0.25),
          control=control.ergm(seed=6886,           MCMC.samplesize=10
          )
par(mfrow = c(3, 2))
mcmc.diagnostics(m2)
```

Results
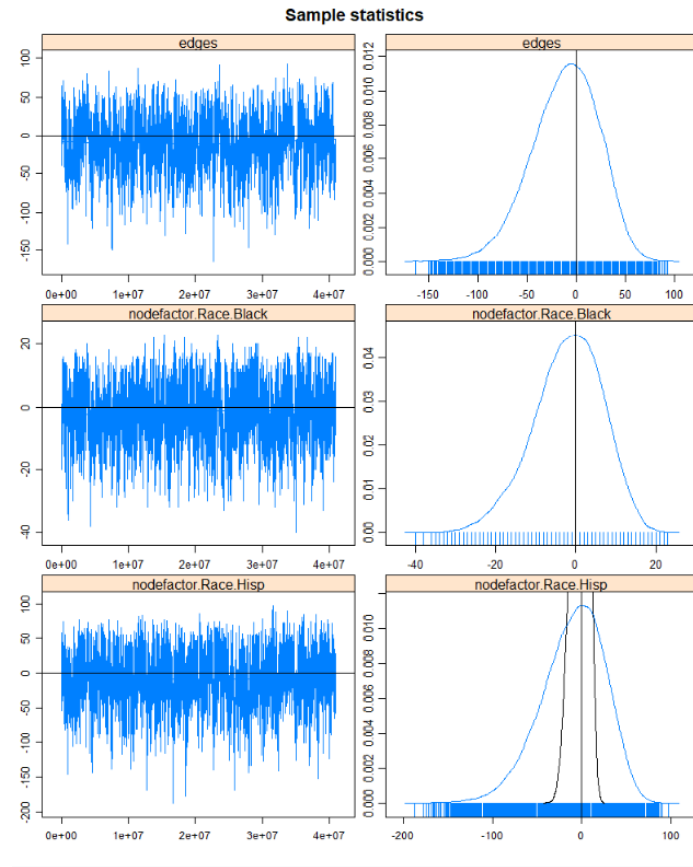
Results

```
##  Results:
##
##                        Estimate Std. Error MCMC % z value
## edges                  -10.22281    1.17910      0  -8.670
## nodefactor.Race.Black    0.61845    0.24268      0   2.548
## nodefactor.Race.Hisp    -0.49867    0.23138      0  -2.155
## nodefactor.Race.NatAm   -0.47808    0.20968      0  -2.280
## nodefactor.Race.Other   -1.54877    0.96497      0  -1.605
## nodefactor.Sex.F         0.13086    0.06715      0   1.949
## nodefactor.Grade.8       1.45656    0.66174      0   2.201
## nodefactor.Grade.9       2.22394    0.62197      0   3.576
## nodefactor.Grade.10      2.55889    0.62167      0   4.116
## nodefactor.Grade.11      2.30632    0.62423      0   3.695
## nodefactor.Grade.12      2.93496    0.62202      0   4.718
## nodematch.Sex            0.52444    0.13175      0   3.980
## nodematch.Race.Black        -Inf    0.00000      0    -Inf
## nodematch.Race.Hisp      0.57111    0.31521      0   1.812
## nodematch.Race.NatAm     1.08712    0.30657      0   3.546
## nodematch.Race.Other        -Inf    0.00000      0    -Inf
## nodematch.Race.White     0.30392    0.62741      0   0.484
## nodematch.Grade.10       1.07180    0.53553      0   2.001
## nodematch.Grade.11       1.83540    0.51002      0   3.599
## nodematch.Grade.12       0.98219    0.57182      0   1.718
```

Results

- Another problem is that several of our parameters are estimated as $-Inf$.

- Check Our Summary Statistics

```
summary(mesa~edges+nodefactor("Race")+
            nodefactor("Sex", levels="F")+
            nodefactor("Grade", levels=c("8","9","10","11","12"))+
            nodematch("Sex")+
            nodematch("Race",diff=TRUE)+
            nodematch("Grade",diff=TRUE)+
            gwesp(decay=0.25)
        )
#Also can:
mixingmatrix(mesa, "Race")
```

Check Our Summary Statistics

```
##                     edges   nodefactor.Race.Hisp
##                       203                    178
## nodefactor.Race.NatAm nodefactor.Race.Other
##                       156                      1
## nodefactor.Race.White     nodefactor.Sex.F
##                        45                    235
##     nodefactor.Grade.8   nodefactor.Grade.9
##                        75                     65
##    nodefactor.Grade.10  nodefactor.Grade.11
##                        36                     49
##    nodefactor.Grade.12        nodematch.Sex
##                        28                    132
##    nodematch.Race.Black  nodematch.Race.Hisp
##                         0                     53
##   nodematch.Race.NatAm  nodematch.Race.Other
##                        46                      0
##   nodematch.Race.White   nodematch.Grade.10
##                         4                      9
##     nodematch.Grade.11   nodematch.Grade.12
##                        17                      6
##      nodematch.Grade.7    nodematch.Grade.8
##                        75                     33
##      nodematch.Grade.9                esp#1
```

Remove the Categories with Empty Cells:

```
m3<- ergm(mesa~edges+nodefactor("Race",levels=c("Black","Hisp","NatAm
          nodefactor("Sex", levels="F")+
          nodefactor("Grade", levels=c("8","9","10","11","12"))+
          nodematch("Sex")+
          nodematch("Race",levels=c("White","Hisp","NatAm"),diff=TR
          nodematch("Grade",diff=TRUE)+
          gwesp(decay=0.25),
          control=control.ergm(seed=6886,          MCMC.samplesize=10
          )
summary(m3)
```

Results

```
Monte Carlo MLE Results:
                        Estimate Std. Error MCMC %  z value Pr(>|z|)
edges                   -10.23242    1.17306      0   -8.723   < 1e-04 ***
nodefactor.Race.Black     0.50708    0.22028      0    2.302  0.021338 *
nodefactor.Race.Hisp     -0.48685    0.22539      0   -2.160  0.030774 *
nodefactor.Race.NatAm    -0.46212    0.20547      0   -2.249  0.024505 *
nodefactor.Race.Other    -1.53903    0.95837      0   -1.606  0.108301
nodefactor.Sex.F          0.13052    0.06733      0    1.938  0.052569 .
nodefactor.Grade.8        1.45427    0.65789      0    2.211  0.027070 *
nodefactor.Grade.9        2.22469    0.62209      0    3.576  0.000349 ***
nodefactor.Grade.10       2.54794    0.62156      0    4.099   < 1e-04 ***
nodefactor.Grade.11       2.30707    0.62273      0    3.705  0.000212 ***
nodefactor.Grade.12       2.91888    0.61877      0    4.717   < 1e-04 ***
nodematch.Sex             0.53114    0.13293      0    3.996   < 1e-04 ***
nodematch.Race.White      0.34575    0.61488      0    0.562  0.573916
nodematch.Race.Hisp       0.56367    0.30562      0    1.844  0.065137 .
nodematch.Race.NatAm      1.06819    0.30388      0    3.515  0.000439 ***
nodematch.Grade.10        1.08583    0.53473      0    2.031  0.042295 *
nodematch.Grade.11        1.83819    0.51002      0    3.604  0.000313 ***
nodematch.Grade.12        1.04165    0.56946      0    1.829  0.067373 .
nodematch.Grade.7         6.00230    1.14645      0    5.236   < 1e-04 ***
nodematch.Grade.8         3.24257    0.64615      0    5.018   < 1e-04 ***
nodematch.Grade.9         1.63799    0.47345      0    3.460  0.000541 ***
gwesp                     1.22453    0.12893      0    9.498   < 1e-04 ***
gwesp.decay               0.50137    0.12661      0    3.960   < 1e-04 ***
---
```

Summary of Results

- Sociality increases by grade.
- Grade-based selective mixing is consistently assortative (i.e., the selective mixing coefficient is positive), but is strongest among 7th graders and declines with seniority.
- The triad closure (GWESP) coefficient is positive.

Interpreting GWESP coefficient

- Two nodes i and j have an edgewise shared partner when they are (1) connected to each other and (2) both i and j are also connected to a third individual k.
- If i and j were also connected to node l, then i and j would have two edgewise shared partners.
- When nodes have edgewise shared partnerships, they form triangles.
- The GWESP term models the tendency for ties that close triangles to be more likely than ties that do not close triangles.
- The GWESP term gradually decreases as pairs of individuals have more existing shared partners.

Interpreting GWESP coefficient

$$\omega = e^{\alpha} \sum_{i=1}^{n-2} (1 - (1 - e^{-\alpha})^i) p_i,$$

where α is the decay parameter, p_i is the number of actor pairs who have exactly i shared edgewise partners, and n is the numebr of nodes in the network.

- The maximum number of edgewise-shared partners for any pair of nodes is $n - 2$.
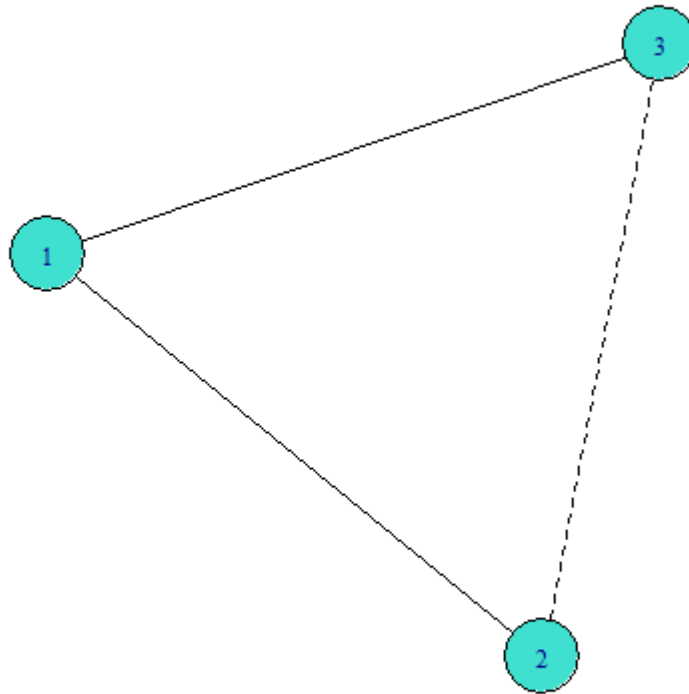
Interpreting GWESP coefficient

Goal: calculate the change in GWESP statistic that will result from adding a particular tie.

Depends on (1) the number of triangles that the tie closes and (2) the existing number of edgewise shared partnerships that the nodes involved in the triangles already belong to.

- Adding a tie that closes no triangles has no effect on GWESP.

- Adding a tie that closes one triangle and no nodes in the group have any existing ESPs will add *three* ties with one edgewise shared partnership (1 and 2 share partner 3, 1 and 3 share partner 2, and 2 and 3 share partner 1:

Interpreting GWESP coefficient

- Consider a GWESP coefficient of 1.8.
- If a tie will close one triangle, and all actor pairs in that triangle currently have no shared partners, the log-odds of the tie are increased by 5.4 (1.8*3), because it would add an edgewise shared partner to three different pairs) relative to an otherwise similar tie that would not close a triangle.
- Can calculate other scenarios, but not very intuitive.