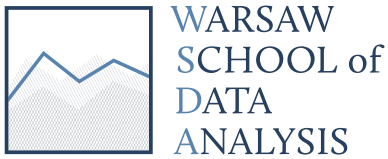
Further Topics in Social Network Analysis

Homophily and segregation

Dominik Batorski Michał Bojanowski Bartosz Chroł Kamil Filipek

Interdisciplinary Centre for Mathematical and Computational Modelling, University of Warsaw



Contents

HUMAN CAPITAL NATIONAL COHESION STRATEGY

UNIVERSITY OF WARSAW

ICS RUG / UU / RU

EUROPEAN UNION EUROPEAN SOCIAL FUND

Project co-financed by the European Union under the European Social Fund

Syntetic definition of homophily has been offered by McPherson, Smith-Lovin and Cook "Homophily is a principle that a contact between similar people occurs at a higher rate that among dissimilar people. (...) Homophily implies that distance in terms of social characteristics translates into network distance, the number of relationships through which a piece of information must travel to connect two individuals" (McPherson, Smith-Lovin, and Cook 2001, 416). Homophilic relations are based on shared characteristics e.g. values, knowledge, skills, beliefs, wealth, social status, geographic closure, ethnicity etc. If we consider a social network that is made of two types of nodes, the density of connections should be higherer between similar nodes. A related concept is "network segregation" (Freeman 1978; Bojanowski and Corten 2014).

The difference between concepts of "homophily" and "segregation" is subtle: by segregation we usually mean a property of network structure while "homophily" is an individual-level propensity to form social network ties with similar others. While homophily usually leads to segregation, lack of homophily can lead to segregation too, as Schelling (1971) famously demonstrated.

In this tutorial we present examples of descriptive tools to analyze homophily/segregation in social networks. We focus on methods that take into account a single node-level variable which is nominal. Consequently this variable exhaustively divides all the nodes into a set of mutually exclusive *groups*.

Data and functions presented here are available in package "isnar" (Bojanowski 2015).

1 Mixing matrix

The probably most important tool for analyzing homophily/segregation patterns in social networks is the *mixing matrix*. Mixing matrix is a three-dimensional distribution of all the dyads in the network understudy which are crossclassified according to the following dimensions:

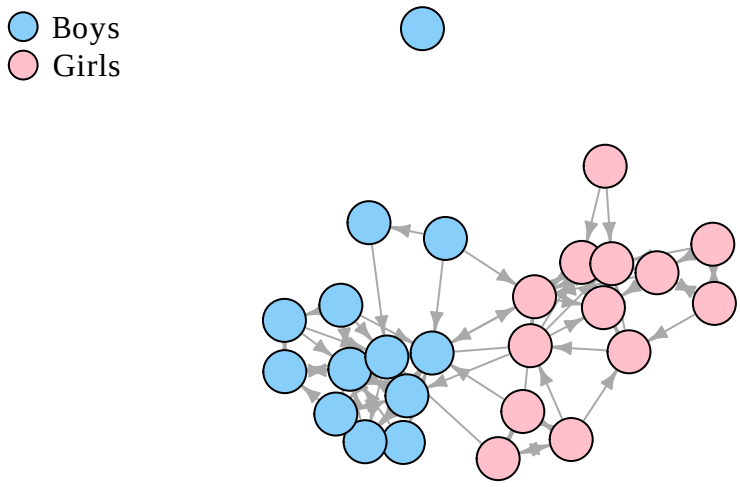
1. Group membership of ego
2. Group membership of alter
3. Whether the dyad is connected or not.

Technically, let us have a network of size N represented with a adjacency matrix $X = [x_{ij}]_{N \times N}$ and a node level variable $t = [t_i]_N$ such that $t_i \in \{1, ..., K\}$. In other words, the variable represents the memebership of each node in one of the K groups. The mixing matrix $M(X, t)$ is an array $M = [m_{ghy}]_{K \times K \times 2}$ where $g, h \in \{1, ..., K\}$ and $y \in \{0, 1\}$. In other words the value m_{gh0} is the number of *disconnected* dyads between nodes belonging to groups g and h, and m_{gh1} is the number of *connected* dyads in between nodes belonging to groups g and h.

Recall the classroom network from package "isnar" built from "would like to play with" nominations:

```
library(isnar)
data(IBE121)
playnet <- delete.edges(IBE121, E(IBE121)[question != "play"])
plot(playnet, vertex.label=NA,
     vertex.color=ifelse(V(playnet)$female, "pink", "lightskyblue"))
legend("topleft", pch=21, pt.bg=c("lightskyblue", "pink"),
      legend=c("Boys", "Girls"), bty="n", pt.cex=2 )
```

HUMAN CAPITAL NATIONAL COHESION STRATEGY
UNIVERSITY OF WARSAW
ICS RUG / UU / RU
EUROPEAN UNION EUROPEAN SOCIAL FUND
Project co-financed by the European Union under the European Social Fund



We can create the mixing matrix using `mixingm` function from package "isnar":

```
m <- mixingm(playnet, "female", full=TRUE)
m
```

```
## , , tie = FALSE
##
##         alter
## ego      FALSE TRUE
##   FALSE    116  167
##   TRUE     164  115
##
## , , tie = TRUE
##
##         alter
## ego      FALSE TRUE
##   FALSE     40    2
##   TRUE       5   41
```

As we can see, there are alltogether $40 + 41 = 81$ homophilous ties that connect children of the same sex, and only $5 + 2 = 7$ ties connecting children of opposite sex.

Let us extend the mixing matrix notation to accomodate marginal distributions of the mixing matrix by denoting by subscript $+$ summation over corresponding dimension. For example:

$$m_{gh+} = \sum_{y=0}^{1} m_{ghy}$$

$$m_{++y} = \sum_{g=1}^{K} \sum_{h=1}^{K} m_{ghy}$$

... and so on.

We can analyze homophily by analyzing the mixing matrix using tools that are used to analyze any cross-classification. For example, we may calculate conditional probabilities of tie existence m_{ghy}/m_{gh+}:

```
round( prop.table(m, c(1,2)) * 100, 1)
```

```
## , , tie = FALSE
##
##        alter
## ego     FALSE TRUE
##   FALSE  74.4 98.8
##   TRUE   97.0 73.7
##
## , , tie = TRUE
##
##        alter
## ego     FALSE TRUE
##   FALSE  25.6  1.2
##   TRUE    3.0 26.3
```

or summarize the connected dyads by analyzing the *contact layer* of the mixing matrix (i.e. m_{gh1}) by calculating conditional probabilities of nominations m_{gh1}/m_{g+1}:

```
round( prop.table(m[,,2], 1 ) * 100, 1)
```

```
##        alter
## ego     FALSE TRUE
##   FALSE  95.2  4.8
##   TRUE   10.9 89.1
```

This tells us, for example, that boys consists 95% of all nominations made by boys, and 11% nominations made by girls.

Most of the existing indexes of homophily/segregation can be derived as functions of the mixing matrix. We provide some examples below.

2 Assortativity coefficient

The assortativity coefficient (Newman 2003; Newman and Girvan 2004) summarizes the contact layer of the mixing matrix by evaluating the relative "weight" of the values on the diagonal. The more likely it is for nodes to be connected within groups, the larger the values in the cells on the diagonal of the contact layer of the mixing matrix.

If we use $p_{gh} = m_{gh1}/m_{gh+}$ as joint probabilities in the contact layer of the mixing matrix, we can write assortativity coefficient as

$$\text{Assortativity} = \frac{\sum_{g=1}^{K} p_{gg} - \sum_{g=1}^{K} p_{g+}p_{+g}}{1 - \sum_{g=1}^{K} p_{g+}p_{+g}}$$

The index returns 1 for perfeclty segregated networks (all ties exist within groups). It returns 0 for random networks. The minimum value of the index depends on average degrees of nodes in each group (see Bojanowski and Corten 2014 for details).

Assortativity coefficient can be computed with `assort` from package "isnar". `assort` expects an igraph object and the name of vertex attribute defining the groups. In our classroom network the assortativity with respect to gender is:

```
assort(playnet, "female")
```

```
## [1] 0.8408885
```

3 Freeman segregation index

Freeman's segregation index (Freeman 1978) is applicable to undirected networks and two groups of nodes. The basic idea behind this measure is to compare the proportion of between-group ties in the observed network with a benchmark representing null segregation. Freeman proposed a baseline proportion of between-group ties expected to exist in a purely random graph with group sizes and density identical to the observed network. As the number of between-group ties in the observed network increases, segregation decreases.

Based on the mixing matrix the observed proportion of between-group ties is equal to

$$p = \frac{m_{121}}{m_{++1}}$$

The expected proportion of between-group ties in the random graph is equal to the proportion of between-group dyads in the total number of dyads in the network (ignoring their connected/disconnected state):

$$\pi = \frac{m_{12+}}{m_{+++}}$$

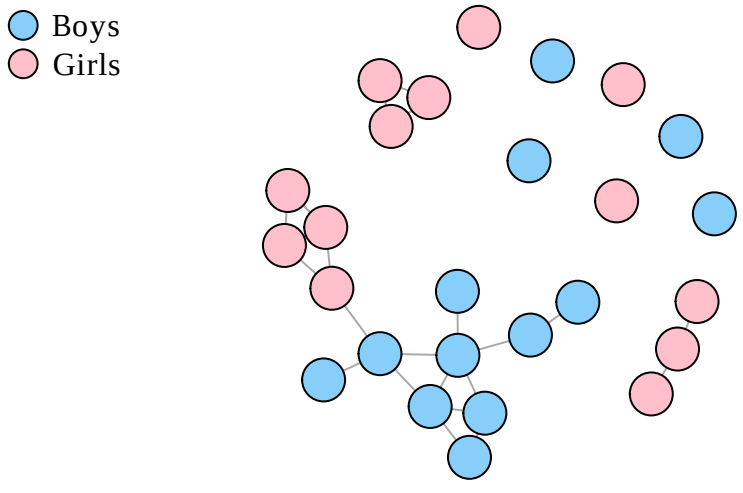
Given these two quantities Freeman's segregation index is equal to

$$\text{Freeman} = \frac{\pi - p}{\pi} = 1 - \frac{p}{\pi}$$

Freeman index varies between 0 and 1. It returns 0 for a random network and returns 1 for a perfectly segregated network.

To try it out in practice let us create an undirected network based on classroom data such that there is an edge only in reciprocated dyads. In other words, we are considering only reciprocated "play with" nominations

```
z <- as.undirected(playnet, mode="mutual")
plot(z, vertex.label=NA,
     vertex.color=ifelse(V(z)$female, "pink", "lightskyblue"))
legend("topleft", pch=21, pt.bg=c("lightskyblue", "pink"),
      legend=c("Boys", "Girls"), bty="n", pt.cex=2 )
```



For this network the gender segregation as measured by Freeman's index is:

```
freeman(z, "female")
```

```
## [1] 0.9125874
```

which is very close to 1 (perfect segregation). Indeed, we see only a single reciprocated nomination between a boy and a girl.

4 Coleman's homophily index

Coleman (1958) homophily index is defined for *directed* networks. By design it assigns a segregation scores for each group in the network, unlike previously described measures which assign a single value for the whole network. In other words, Coleman's index is a group-level measure.

The index captures the tendency for members of a particular group to nominate others from the same group. The tendency is evaluated vis a vis the situation in which the nominations would be random. The expected number of ties within group g if choices were made at random is:

$$m^*_{gg1} = \sum_{i:t_i=g} \eta_i \frac{n_g - 1}{N - 1}$$

where η_i is the out-degree of node i, n_g is the number of nodes in group g, and N is the size of the network.

Coleman's homophily index for group g is then equal to:

$$\text{Coleman}_g = \begin{cases} \frac{m_{gg1} - m^*_{gg1}}{\sum_{i:t_i=g} \eta_i - m^*_{gg1}} & \text{if} \quad m_{gg1} \geq m^*_{gg1} \\ \frac{m_{gg1} - m^*_{gg1}}{m^*_{gg1}} & \text{if} \quad m_{gg1} < m^*_{gg1} \end{cases}$$

Coleman's index returns 0 if and only if expected proportion of within-group ties in group g is equal to the proportion of nodes from group g in the total number of nodes (excluding ego). It is equal to 1 if all nominations of group g nodes belong to group g. It is equal to -1 if all nominations of group g nodes do not belong to group g.

For the `playnet` network Coleman's indexes for boys and girls are:

```
coleman(playnet, "female")
```

```
##     FALSE      TRUE
## 0.9084249 0.7909699
```

As we can see both values are close to 1 (high segregation). The value for boys is larger as girls are slightly more likely to nominate boys than boys are to nominate girls.

5 See also

- The article of Bojanowski and Corten (2014) contains a more detailed comparative analysis of segregation measures presented in this tutorial as well as other measures.
- Homophily effects can be expressed in Exponential Random Graph Model framework. See ERGM section of this tutorial for more details.

HUMAN CAPITAL NATIONAL COHESION STRATEGY
UNIVERSITY OF WARSAW
ICS RUG / UU / RU
EUROPEAN UNION EUROPEAN SOCIAL FUND
Project co-financed by the European Union under the European Social Fund

References

Bojanowski, Michał. 2015. *Isnar: Introduction to Social Network Analysis with R*. https://github.com/mbojan/isnar.

Bojanowski, Michał, and Rense Corten. 2014. "Measuring Segregation in Social Networks." *Social Networks* 39: 14–32.

Coleman, James S. 1958. "Relational Analysis: The Study of Social Organizations with Survey Methods." *Human Organization* 17: 28–36.

Freeman, Linton C. 1978. "Segregation in Social Networks." *Sociological Methods & Research* 6 (4): 411–29.

McPherson, Miller, Lynn Smith-Lovin, and James M Cook. 2001. "Birds of a Feather: Homophily in Social Networks." *Annual Review of Sociology*. JSTOR, 415–44.

Newman, Mark E. J. 2003. "Mixing Patterns in Networks." *Physical Review A* 67 (2): 1050–2947.

Newman, Mark E. J., and M. Girvan. 2004. "Finding and Evaluating Community Structure in Networks." *Phys. Rev. E* 69 (2). American Physical Society: 026113. http://link.aps.org/doi/10.1103/PhysRevE.69.026113.

Schelling, Thomas C. 1971. "Dynamic Models of Segregation." *Journal of Mathematical Sociology* 1 (2): 143–86.

HUMAN CAPITAL NATIONAL COHESION STRATEGY
UNIVERSITY OF WARSAW
ICS RUG / UU / RU
EUROPEAN UNION EUROPEAN SOCIAL FUND

Project co-financed by the European Union under the European Social Fund