Paper being replicated: DiMaggio, Paul and Filiz Garip. 2011. "How Network Externalities Can Exacerbate Intergroup Inequality." *American Journal of Sociology* 116(6): 1887-1933.

Author: Diego F. Leal (www.diegoleal.info)

1. Generating a population of ("realistic") agents:

1.1. According to Dimaggio and Garip (2001: 1900-1901):

> To ensure that the distribution of parameters and associations among parameters are realistic, we base our agents on 2,257 African-American and white respondents to the 2002 General Social Survey (GSS). The survey included items on network size, race, education, and income.

Therefore, GSS 2002 data was downloaded in Stata format from: http://www.thearda.com/Archive/Files/Downloads/GSS2002_DL2.asp

1.2. After inspecting the GSS data set, I found the GSS variables that correspond to those employed by Dimmagio and Garip:

Variable	GSS Name	Var. Name in this Document	Total	Missing Values
Race	race	race/RACE	2765	188
Education (years)	educ	educ/EDUC	2765	12
Degree	degree	degree/DEGREE	2765	5
Income category	INCOME98	fIncome/FINCOME	2765	302
# friend	numcntct	numcntct/NUMCNTCT	2765	54
# close friends	numprobs	numprobs/NUMPROBS	2765	82

1.3. I dropped all missing values as per lines 19 to 24 e of Roberto's Rscript (titled "DiMaggio_Garip_2011_model.R"). (

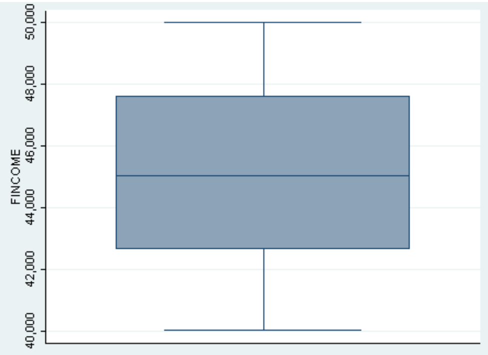
WARNING: After dropping all missing values, there are 2241 agents (observations), 16 less than the number of agents reported by DiMaggio and Garip (2011: 1900). See lines 15-29 of my dofile ("replication_dimaggio.do") or the logfile "logfileDiMaggio.smcl." Despite multiple attempts to understand this difference, I have not been able to understand why the authors have 16 agents more than myself.

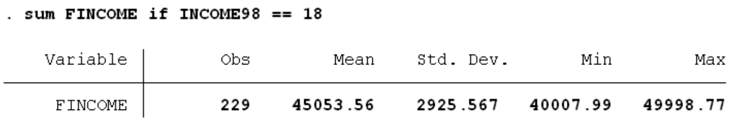
1.4. Generating the variable family income. According to DiMaggio and Garip (2011: 1901, footnote 9):

> GSS reports income as a series of ranges: we treat income as uniformly distributed within each interval and randomly assign individuals to points in their distribution

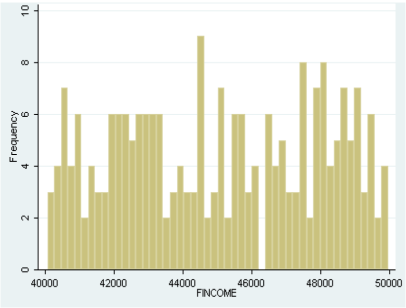
After randomly and uniformly assigning incomes to agents in a given bracket, I checked results which did in fact follow a uniform distribution within each interval/bracket. The distribution shown below corresponds to agents' earnings within income (FINCOME) bracket # 18 ($40,000 - $49,000), the bracket with the highest number of data points. As can be seen, values seem to be randomly and uniformly distributed:

```
. graph box FINCOME if INCOME98 ==18
```





```
. sum FINCOME if INCOME98 == 18

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
     FINCOME |       229    45053.56    2925.567   40007.99   49998.77
```



WARNING: The authors do not explicitly mention how they handle the first income bracket (< $1,000), so I also generated random and uniformly distributed incomes for this bracket. I am not sure how "realistic" these bracket is.

1.5. Following the procedure in lines 52-64 of "generate_population_gss.R," the US census distribution for households in 2007 was used for individuals who reported a family income greater than $110,000. It was necessary to do this because:

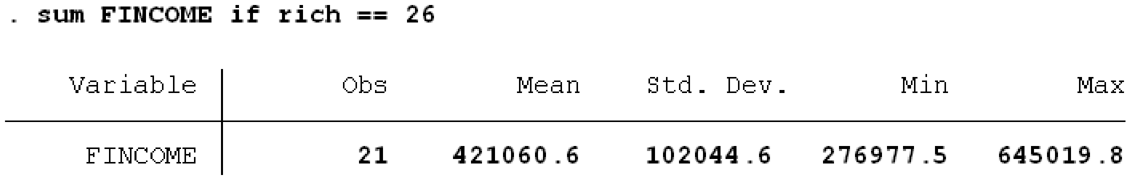
> Individuals who reported family incomes of $110,000 or more (about 10%) were randomly allocated to incomes up to $650,000 on the basis of CPS data on actual income distributions in that range. (CPS is top coded at $250,000, but the mean income in the top-income category is reported to be $450,000. We assume the range in the top category to be [$250,000, $650,000], producing a mean of $450,000 if income is uniformly distributed.) (DiMaggio and Garip 2011: 1901, footnote 9)

I took the US census data from (See "households_income_greater_than_100000.xlsx" the get the whole table): https://www.census.gov/hhes/www/cpstables/macro/032008/hhinc/new06_000.htm

After obtaining the data, I calculated the number of individuals that needed to be assigned to each one of the four new income brackets (see table below). It is important to remember that the number of individuals in the GSS dataset who reported earning $110,000 or more is **212**. As mentioned by the authors, these 212 individuals represent about 10% (9.42% according to my data, to be more precise) of the total number of individuals in the GSS data set. These 212 individuals were distributed across the four new income brackets as follows:

Income bracket according to CPS 2007	**Households CPS 2007**	**Proportion per Bracket According to CPS 2007**	**number of agents per new income bracket GSS**
$100,000 to $149,999	14,214	0.60	127
$150,000 to $199,999	5,115	0.22	47
$200,000 to $249,999	2,012	0.08	17
$250,000 to $650,000	2,245	0.10	21
Total	23,586	1	212

To check that the procedure used to assign incomes to these (rich) agents was correct, I tabulated a statistical summary of FINCOME for the agents in the last bracket ($250,000 - $650,000). As mentioned by DiMaggio and Garip (2011: 1901, footnote 9), the mean income in the last bracket should be close to $450,000.



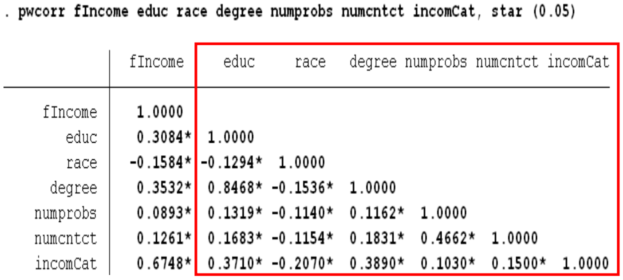
```
. sum FINCOME if rich == 26

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     FINCOME |        21    421060.6    102044.6   276977.5   645019.8
```

1.6. I generated two variables to group income data: incomeGroupMean and IncomeGroup. In order to do this, I divided the income data using 33rd and 67th percentiles cut-off points. See Rscprit lines 76 to 78. The variable incomeGroupMean is based on fIncome percentiles and incomeGroup is based of fIncomeMean percentiles.

1.7. I generated one variable to regroup degree data: educationGroup. See Roberto’s Rscript line 92. Education Groups: (0) Less than HS; (1) HS; (2) Any college, no BA; (3) BA or more.

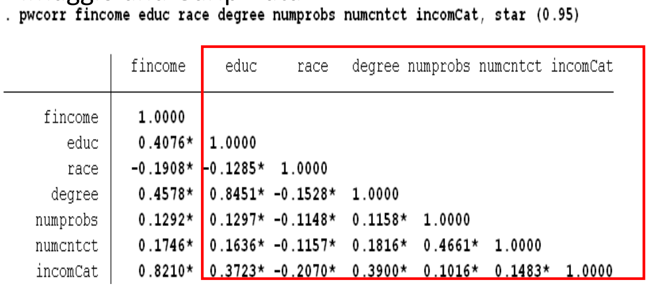
1.8. After saving the new data set, called “dataDiMaggio.csv” I run a correlation between the variables contained in the matrix I produced. I did the same with the original GSS data used by DiMaggio and Garip (“gss02.mat”). A comparison between these two data sets allowed me to determine whether the extra 16 agents that DiMaggio and Garip report affect the basic relationships between the variables. Fortunately, this does not seem to be the case because each of the variables I left unchanged as described above(educ, race, degree, numprons, numcntct, and inccat) have very similar correlations in the data set I produced vis-à-vis DiMaggio and Garip’s.

Replication Data



```
. pwcorr fIncome educ race degree numprobs numcntct incomCat, star (0.05)

             |  fIncome     educ     race   degree numprobs numcntct incomCat
-------------+---------------------------------------------------------------
     fIncome |   1.0000
        educ |   0.3084*  1.0000
        race |  -0.1584* -0.1294*  1.0000
      degree |   0.3532*  0.8468* -0.1536*  1.0000
    numprobs |   0.0893*  0.1319* -0.1140*  0.1162*  1.0000
    numcntct |   0.1261*  0.1683* -0.1154*  0.1831*  0.4662*  1.0000
    incomCat |   0.6748*  0.3710* -0.2070*  0.3890*  0.1030*  0.1500*  1.0000
```

DiMaggio and Garip Data



```
. pwcorr fincome educ race degree numprobs numcntct incomCat, star (0.95)

             |  fincome     educ     race   degree numprobs numcntct incomCat
-------------+---------------------------------------------------------------
     fincome |   1.0000
        educ |   0.4076*  1.0000
        race |  -0.1908* -0.1285*  1.0000
      degree |   0.4578*  0.8451* -0.1528*  1.0000
    numprobs |   0.1292*  0.1297* -0.1148*  0.1158*  1.0000
    numcntct |   0.1746*  0.1636* -0.1157*  0.1816*  0.4661*  1.0000
    incomCat |   0.8210*  0.3723* -0.2070*  0.3900*  0.1016*  0.1483*  1.0000
```

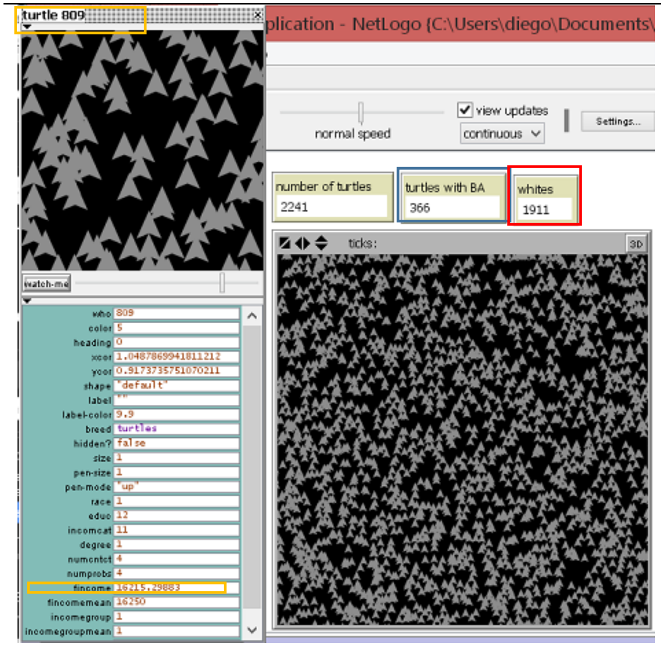
2. Reading-in the data into NetLogo

2.1. In order to read-in the data in NetLogo, I first created as many turtle-specific variables as there were variables in the GSS data set I developed in the first part of this memo. These included the following 11 variables: race, educ, incomCat, degree, numcntct, numprobs, fIncome, fIncomeMean, incomeGroup, incomeGroupMean, educationGroup.

2.2. I developed a procedure called “setup” in Netlogo to read-in the data. This procedure opens the GSS data in .txt format (the actual file is called “dataToImportNetLogo.txt”). Data are read-in using a while loop that stops after the entire file/data set is read. This is achieved using a NetLogo built-in boolean reporter called “file-at-end?” Within the while loop, I created a local variable called “agent-features.” After doing this, I read the first line of the data set (i.e. the data corresponding to the first respondent in the GSS data) using the NetLogo reporter “file-read-line.” This allowed local variable agent-features to store 11 different items corresponding to the values of the 11 different variables reported by the first respondent in the original GSS data set. Then, I created one turtle and matched the 11 items stored in agent-features to the corresponding 11 turtle-specific variables created in point 2.1 of this memo. I then asked the turtle to choose a place in the world at random. This procedure was repeated as many times as lines (i.e. respondents) are in the GSS data. Finally, the while loop was closed and the file (“dataToImportNetLogo.txt”) is closed too.

2.3. To make sure that the data were correctly read-in, I ran a couple of basic tests. I created three reporters: the total number of turtles, the total number of whites, and the total number of turtles with a B.A. All the numbers matched the GSS data. I also chose one turtle at random, in this case turtle number 809, and made sure its values matched the corresponding values for the observation (i.e. participant) with the same id number in the GSS data.

Turtles in Netlogo



☑ view updates

continuous

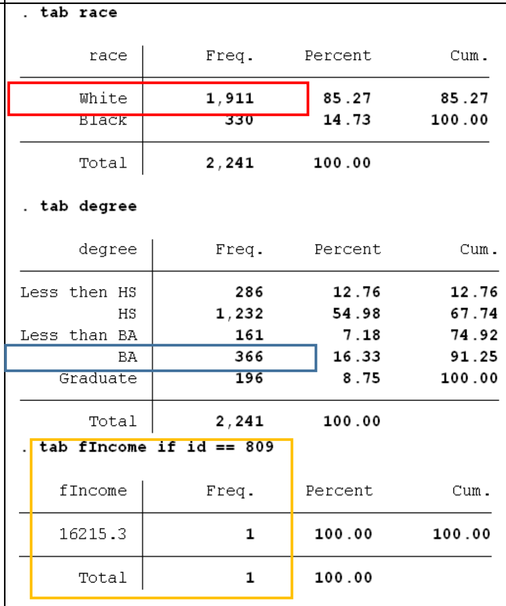
number of turtles: 2241

turtles with BA: 366

whites: 1911

who: 809
color: 5
heading: 0
xcor: 1.0487869941811212
ycor: 0.9173735751070211
shape: "default"
label: ""
label-color: 9.9
breed: turtles
hidden?: false
size: 1
pen-size: 1
pen-mode: "up"
race: 1
educ: 12
incomcat: 11
degree: 1
numcntct: 4
numprobs: 4
fincome: 16215.29883
fincomemean: 16250
incomegroup: 1
incomegroupmean: 1

Respondents in GSS



```
. tab race

       race |      Freq.     Percent        Cum.
------------+-----------------------------------
      White |      1,911       85.27       85.27
      Black |        330       14.73      100.00
------------+-----------------------------------
      Total |      2,241      100.00

. tab degree

     degree |      Freq.     Percent        Cum.
------------+-----------------------------------
Less then HS |        286       12.76       12.76
         HS |      1,232       54.98       67.74
Less than BA |        161        7.18       74.92
         BA |        366       16.33       91.25
   Graduate |        196        8.75      100.00
------------+-----------------------------------
      Total |      2,241      100.00
. tab fIncome if id == 809

    fIncome |      Freq.     Percent        Cum.
------------+-----------------------------------
    16215.3 |          1      100.00      100.00
------------+-----------------------------------
      Total |          1      100.00
```