doi:10.5477/cis/reis.177.43

# Resilient Homophily Patterns in Youth Friendship Networks: A Case Study Using a Computer Simulation Experiment

*Patrones de homofilia resilientes en redes de amistad juvenil: estudio de caso mediante un experimento de simulación computacional*

**Francisco Linares Martínez, Francisco J. Miguel Quesada and Mona Kohl**

## Key words

Homophily
- Coleman's Homophily Index
- Friendship Networks
- Resilience
- Agent-Based Simulation
- Computational Sociology

## Abstract

This paper deals with how to recognise if the patterns of homophily found in a social network are resilient to small disturbances that may occur in that network. Data from a survey of students in a secondary school in the Canary Islands were replicated using an agent-based model. The model calculated homophily indexes and their statistical significance and then simulated small alterations in the distribution of links. The results clearly show that some homophily indexes resist these kinds of perturbations and others do not. Evidence suggests that the distribution of individuals across the social network communities is a key factor in explaining why certain patterns of relationships are more resilient than others.

## Palabras clave

Homofilia
- Índice de homofilia de Coleman
- Redes de amistad
- Resiliencia
- Simulación basada en agentes
- Sociología computacional

## Resumen

El presente trabajo aborda la cuestión de cómo conocer si los patrones de homofilia hallados en una red social son resilientes ante pequeñas perturbaciones que pueden producirse en dicha red. Para ello se han replicado con un modelo basado en agentes los datos de una encuesta realizada a los estudiantes de un instituto de enseñanza secundaria de las islas Canarias. Dicho modelo calcula los índices de homofilia y su significatividad estadística para posteriormente proceder a la simulación de pequeñas alteraciones en la distribución de los vínculos. Los resultados muestran claramente que algunos índices de homofilia resisten dichas perturbaciones y otros no. La evidencia hallada apunta a que la distribución de los individuos entre las comunidades que configuran la red es un factor clave que explica que ciertos patrones de relaciones sean más resilientes que otros.

## Citation

Linares Martínez, Francisco; Miguel Quesada, Francisco J. and Kohl, Mona (2022). "Resilient Homophily Patterns in Youth Friendship Networks: A Case Study Using a Computer Simulation Experiment". *Revista Española de Investigaciones Sociológicas*, 177: 43-68. (doi: 10.5477/cis/reis.177.43)

**Francisco Linares Martínez:** Universidad de La Laguna | flinares@ull.es

**Francisco J. Miguel Quesada:** Universitat Autònoma de Barcelona | Miguel.Quesada@uab.cat

**Mona Kohl:** Atos Consulting (Canarias) (México) | mona.kohl@atos.net

## INTRODUCTION[1]

The concept of homophily was coined by Robert K. Merton in his seminal article on[2] friendship relations co-authored with Paul Lazarsfeld (Lazarsfeld and Merton, 1954). Merton justified the need to introduce a new term into the sociological vocabulary by arguing that there was no word in the English language to concisely refer to friendships between people "of the same kind". Today, the most common definition is the one provided in a much-cited literature review article: "the principle that a contact between similar people occurs at a higher rate than among dissimilar people" (McPherson, Smith-Lovin and Cook, 2001: 416). As will be seen below, there is some ambiguity or confusion in the use of the term, despite its apparent clarity, since some academics associate it with a person's preference to hold relationships with similar ones, while others use it to denote an empirical regularity, a pattern of collective behaviour (the frequency of contacts between similar individuals) wich may be the result of various social mechanisms. The second meaning is more closely related to Lazarsfeld and Merton's original use of the term and the most obvious one in McPherson, Smith-Lovin and Cook's definition. It will be the meaning adopted in this paper.

Simply put, there is homophily regarding an attribute in a network of relationships between individuals if the proportion of links between individuals who have that attribute is higher than the proportion of individuals with that attribute in that population. Thus, for instance, if a link between two Catholic individuals is more likely than the proportion of individuals who profess that religion, there are fewer inter-group links and more intra-group links than would be expected in the absence of social mechanisms inducing the structuring of social relations, which implies non-random behavior. The aim of this paper is to study a very specific problem in the identification of this phenomenon, namely, the extent to which the homophily indexes that can be calculated in a network of individuals are resistant to small alterations in their links, given that some links disappear, and new ones are created in the course of social life.

To address this question, a simulation model was built to replicate data from a network of students in a secondary school (coded as IES San Borondón). Once statistically significant homophily indexes were identified, a "virtual" experiment was carried out in which the real links were recursively replaced by randomly chosen links, up to 15% of the total number of homophilic links. As a result of this procedure, all indexes decreased but some nevertheless remained statistically significant, while others were no longer significant. This immediately raises the question of why some homophily patterns are resilient to small disturbances in network structure while others are not. The general hypothesis is that the explanation for why certain patterns of homophily are resilient and others are not rests on how individuals are distributed across the different communities which make up the network as a whole. To our knowledge, no systematic study of this issue exists in the literature.

From a methodological point of view, tackling the problem requires moving through different conceptual levels or planes of reality (individuals, communities, and network) where information is coded into three different databases.

At the most basic level, the units of analysis are the individuals. A questionnaire was used to collect information on their typical attributes of the units of analyisis are the in-

[1] This research has received funding from projects CSO2015-6474-R (MINECO) and PID2019-107589GB-100 (MICIN).

[2] The article was published under both names, but was divided into two parts: the substantive part, authored by Merton, and the methodological part, written by Lazarsfeld.

dividuals (e.g., age, gender, religion, etc.). A name generator was used in the same questionnaire to collect relational information.

The network of friendships was reconstructed using the information obtained from the questionnaire. This is a *small world* network, made up of communities or groups[3] of individuals who, in turn, have some links that *bridge* with other communities. The second database was made up of the properties of the communities, which are supra-individual units.

Finally, since the analysis computed all possible homophily indexes in the network, a third database was constructed that included the indexes and the data associated with each of them (such as: number of individuals, communities to which they belonged, number of links, etc.). This provided an exhaustive description of the patterns of homophilic behaviour observable in the network as a whole.

The study is structured as follows. The next section contains a theoretical review with special emphasis on the explanatory mechanisms of homophily patterns. Methodological details are then described, including a brief outline of the agent-based simulation model (ABM) built specifically to address the research question; its operation is also illustrated. The results of the analysis of three different types of data are then presented: firstly, homophily indexes, identifying those that remained significant from those that did not; secondly, the groups of individuals (communities) that were most frequently associated with the homophily indexes that remained significant; and thirdly, the characteristics of the homophily indexes that remained resilient. The article ends with a discussion of these results and some conclusions.

## INBREEDING HOMOPHILY MECHANISMS

As McPherson, Smith-Lovin and Cook pointed out in the literature review cited in the introduction, empirical research into the phenomenon became significantly popular from the 1970s onwards, largely stimulated by Peter Blau's theory of social structure and by the development of network analysis. However, it has not always been suitably recognised that the concept had a dual status in Blau's (1977) study, which may have led to some confusion in some scholarly works.

On the one hand, Blau used the concept of homophily as a basic assumption concerning individuals' preferences. If individuals *i* and *j* share a quality, they will be interested in creating a friendship bond if given the opportunity. That is, the bond of friendship is explained by that love (*philia*) of equals (*homo*). In this case, homophily is a mechanism that operates at the individual level and helps to explain certain patterns of social relations. This is, for example, the notion that Shalizi and Thomas (2011) used to examine the problem of the distinction between homophily and social contagion.

On the other hand, the term homophily is also used in Blau's theory to mean the proportion of individuals in a category of a given parameter of the social structure who maintain relationships with individuals in the same category[4]. As Blau showed, this proportion would depend on the relative weight of each category in the population as a whole. If one group is larger than another, all other things being equal, its members will have fewer opportunities for heterophilic relationships than those from the smaller group. Therefore, in this second use, the terms homophily/heter-

[3] We use the term group or community interchangeably, leaving the use of the term *cluster* for the statistical technique named *cluster analysis.*

[4] In Peter Blau´s theory, social structure was conceived as an intersection of parameters (nominal or graded) that reflect the relationships that individuals have with each other.

ophily are linked to certain population characteristics related to the degree of social cohesion (Lozares and Verd, 2011), not to a type of individual motivation.

The degree of homophily resulting from the opportunities created by the quantitative distribution of the population is termed baseline homophily. But when the links between individuals with a certain trait exceed the proportion of individuals with that trait in the population, i.e., when such relationships occur more frequently than those offered by the opportunity set, this fact is termed inbreeding homophily. In this case, there must be psycho-social mechanisms that make intra-group relations more frequent than expected and inter-group relations less frequent than expected. In this paper, the term homophily is used to mean the tendency (observed in a network of individuals) for contacts between individuals who have a similar trait or characteristic to occur more frequently than with individuals who are dissimilar in terms of that characteristic, irrespective of the particular mechanism causing such in-breeding.

The mechanisms that produce in-breeding homophily, in turn, belong to two types of families. The first is the family of mechanisms based on the structure of relationships that facilitate the maintenance of social contacts. Such opportunities are provided by *social foci* of interaction (Feld, 1981, 1982) such as organisations, and by the social networks created in everyday life. The second family of mechanisms relates primarily to individual decision-making that involves the creation, maintenance, and dissolution of social ties.

This second type of mechanisms involve psychosocial reinforcement processes. As Lazarsfeld and Merton pointed out in their analysis of *value homophily*, two strangers with similar values are likely to form a bond of friendship if they have the opportunity to meet on a regular basis (this is connected with the meaning of the proverb "Birds of a feather flock together". It is also possible that two individuals may be motivated to change their values precisely as a result of their friendship, in a give-and-take that progressively smooths out initial discrepancies . This second process falls into the category of 'social influence' or *peer influence* (Cohen, 1977). A third possibility is that an individual may acquire one or more traits of another individual with whom they have a bond through some imitation mechanism; in this case influence is not reciprocal, so we call it "social contagion", although the outcome is difficult to distinguish empirically from the previous case. Finally, the preference for similarity or homophilic preference (which, as noted, is often confused with the concept of homophily itself) is possibly the mechanism that operates most frequently in cases of *status homophily*, regarding variables such as gender, ethnicity, educational status, age, etc.

Another question related to the two families of mechanisms briefly presented is which of them is more prominent in explaining the empirically observed patterns of homophily. While no definitive answer can be given to this question in the current state of the art, some evidence suggests that an important part of the explanation lies in structural elements such as the existence of social environments, organisations, etc. that attract individuals with similar characteristics (McPherson and Smith-Lovin, 1987; Moody, 2001; Kossinets and Watts, 2009). Compared to these structural elements which shape individuals' opportunities to find similar others, psychosocial mechanisms operating at the individual level seem to play a relatively minor role in explaining patterns of relationships, although evidence is inconclusive.

## Methodology

The paper is based on data from a previous study (Linares and Kohl, 2017) in

which a questionnaire was distributed to 194 post-compulsory secondary school students in a secondary school in the Canary Islands Region, which we called IES San Borondón[5]. The survey contained three modules: 1) socio-demographic characteristics, 2) leisure and consumption habits, 3) friendship relationships. In this third module, the students were asked to provide the names of up to four "friends with whom you talk about your problems" (Marsden, 1987). They were also asked to give the name of the person with whom they had a romantic relationship, if any. The result is a network with a giant component that included 163 nodes. The characteristics of this network are shown in Table 1.

An initial analysis showed varying degrees of homophily[6] consistent with the findings in the literature (Kandel, 1978; McPherson, Miller and Smith-Lovin, 1986, 1987; Moody, 2001; Shrum, Cheek and MacD. Hunter, 1988), considering variables such as gender, age, smoking/non-smoking, religion, etc. although not all indexes were statistically significant[7]. The value of the 86 calculated homophily indexes and their level of statistical significance are shown in Appendix 1. However, as noted in the introduction, the focus of this paper was not on the statistical significance *per se*, but on the resilience of significant homophily indexes in light of the small disturbances that inevitably occur in a social network due to the dynamics of link creation and dissolution.

To address this issue, an agent-based simulation model (ABM) was built using the Netlogo platform (Wilenski and Rand, 2015)[8]. This model, described in simple terms in Diagram 1, contained a set of modules that sequentially performed the following operations:

1. Importing the survey data of the students from IES San Borondón. This allowed for the *in silico* replication of the real subjects, who were "transformed" into virtual agents with the same attributes and relationships as those reported in the survey.
2. Calculating the Coleman homophily index (CHI) for each of the agents' attributes and its statistical significance.
3. Random rewiring of links (RRL) for each of the agents' attributes, followed by calculating of the new CHI and its statistical significance.

[5] Fieldwork was conducted between 25 February 2015 and 1 March 2015. The questionnaire was distributed in the classrooms to all students present, who accounted for 67% of the total study population.

[6] There are several options for measuring the degree of homophily (Bojanowski and Corten, 2014). We use the Coleman Homophily Index (CHI) (Coleman, 1957), which compares the proportion of actually existing homophilic relationships with what would be expected if the relationships between individuals were established randomly. The CHI ranges between +1 and –1, with the value 0 representing the *baseline homophily* level, which relates exclusively to the relative size of the subpopulations, and therefore means that there is no social mechanism inducing the choices of individuals.

[7] Statistical significance was measured using a test specifically designed for the CHI by Signori and O'Shea (1965), who addressed the problem of finding the probability that a given value of the homophily index, or a larger value, can be obtained assuming that there is no relationship between the attributes of the node from which the link departs (the individual mentioned) and the node which receives it (the individual mentioned); finding the parameters $\mu$ and $\sigma$ which allow the value to be standardised according to a normal distribution. In what follows, the analysis focuses exclusively on the statistically significant indexes.

[8] Appendix 2 provides a brief description of this type of model.

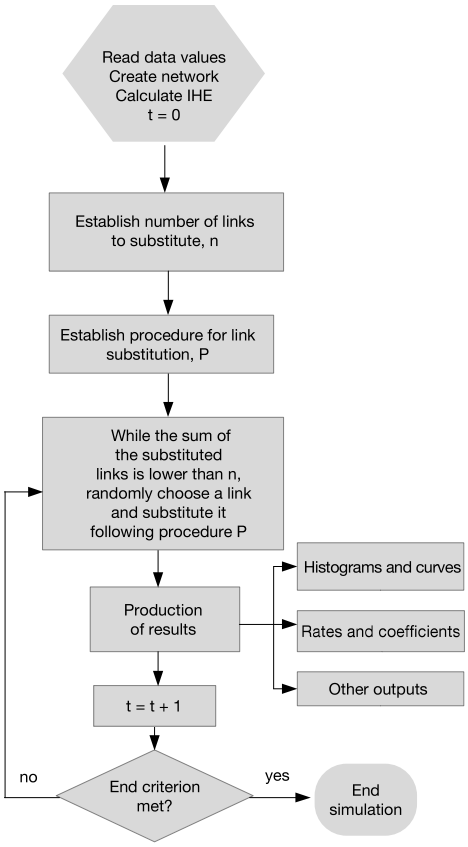
**TABLE 1.** *Basic characteristics of the giant component of the student network from IES San Borondón*

| No. of nodes | No. of links | Average number of links | Maximum number of links | Density | Average geodesic distance | Max geodesic distance | No. of communities (*) |
|---|---|---|---|---|---|---|---|
| 163 | 275 | 3.35 | 9 | 0.021 | 5.92 | 14 | 14 |

(*) Number of communities found using the Clauset-Newman-Moore algorithm.

*Source:* Own elaboration.

**DIAGRAM 1.** *Model flowchart*




*Source:* Own elaboration.

The RRL procedure[9] involved the computer randomly choosing one of the homophilic links and replacing it with a new link; it then immediately recalculated the homophily index and reassessed its statistical significance. This operation was repeated until a number equivalent to 15% of the homophilic links in the selected category were replaced. By substituting one link for another in each case, the basic properties of the network, such as density, mean geodesic distance,

[9] See Appendix 3 for a detailed description of the RRL procedure.

etc., remained virtually identical, since these measures were insensitive to these small disturbances; however, some of the homophily indexes were no longer statistically significant.

This procedure was repeated using different algorithms for link substitution, so 780 cases of artificially manipulated networks were available for each of the attributes under analysis at the end of the simulation. Restricting the analysis to the 36 attributes (shown in tables 3a and 3b, in the next section) whose homophily indexes were positive and statistically significant prior to any manipulation gave a final population of 29,640 cases.

As an example, Tables 2a and 2b show the results of the simulations for the categories "smoker" and "male", respectively. All numbers in the table show the averages of the 780 simulations. The first column shows the percentage of homophilic links replaced (%HLR) related to the total homophilic links of that category. For each of its values (0, 3, 6, 9, 12, 15), the rest of the columns indicate the percentage of homophilic links replaced "related to the total links on the network" (% "total links rewired", TLR), the value of the Coleman homophily index (CHI), the difference between the original value of the CHI and the new value ("drop value", DV), the relative frequency with which the index remained significant with probabilities of 95 % and 99% and, finally, the number of communities in which individuals with the quality of "man" or "smoker" were distributed, a number that remained constant in the simulations, since the model did not recompute the communities in the network.

The information contained in Table 2a shows that smokers were concentrated in six communities and that their initial homophily index had a statistically significant value of 0.329. As the simulation model randomly replaced links, this value dropped to 0.222 once 15% of the homophilic links (which constituted 1.97% of the total links in the network) had been replaced, a value that always remained significant.

**TABLE 2A.** *Simulation results for the "smoker" category*

| %HLR | %TLR | CHI | DV | P95 | P99 | Cs |
|---|---|---|---|---|---|---|
| **0** | 0.00 | 0.329 | 0.000 | 1.00 | 1.00 | 6 |
| **3** | 0.39 | 0.307 | –0.022 | 1.00 | 1.00 | 6 |
| **6** | 0.79 | 0.283 | –0.047 | 1.00 | 1.00 | 6 |
| **9** | 1.18 | 0.260 | –0.059 | 1.00 | 1.00 | 6 |
| **12** | 1.57 | 0.239 | –0.091 | 1.00 | 1.00 | 6 |
| **15** | 1.97 | 0.222 | –0.107 | 1.00 | 1.00 | 6 |

*Notes:* %RHL = percentage of homophilic links replaced out of the total number of homophilic links in that category; %TLR = percentage of homophilic links replaced out of the total number of links in the network; CHI = Coleman homophily index; DV = "drop value" or difference between the original CHI value and the new value; P95 = frequency with which the index remains significant (p-value < 0.05); P99 = frequency with which the index remains significant (p-value < 0.01); Cs= number of communities in which individuals with the "smoker" quality were distributed.

*Source:* Own elaboration.

In the case of Table 2b, for the "male" category, the individuals were distributed into twelve groups and the value of the initial homophily index was 0.331, similar to

that of smokers and equally significant. However, once the random replacement process started, the value dropped to 0.076 and the percentage of times it remained significant at the 0.05 and 0.01 levels was 0.03% and 0.00%, respectively.

**TABLE 2B.** *Simulation results for the "male" category*

| %HLR | %TLR | CHI | DV | P95 | P99 | Cs |
|---|---|---|---|---|---|---|
| **0** | 0.00 | 0.331 | 0.000 | 1.00 | 1.00 | 12 |
| **3** | 1.31 | 0.274 | –0.058 | 1.00 | 1.00 | 12 |
| **6** | 2.49 | 0.223 | –0.108 | 1.00 | 0.98 | 12 |
| **9** | 3.81 | 0.171 | –0.160 | 0.96 | 0.21 | 12 |
| **12** | 4.99 | 0.121 | –0.210 | 0.51 | 0.01 | 12 |
| **15** | 6.30 | 0.076 | –0.255 | 0.03 | 0.00 | 12 |

*Notes:* %RHL = percentage of homophilic links replaced out of the total number of homophilic links in that category; %TLR = percentage of substituted homophilic links replaced out of the total number of links in the network; CHI = Coleman homophily index; DV = "drop value" or difference between the original CHI value and the new value; P95 = frequency with which the index remains significant (p-value < 0.05); P99 = frequency with which the index remains significant (p-value < 0.01); Cs= number of communities in which individuals with the "male" quality were distributed.

*Source:* Own elaboration.

## Results

### Relevance of the numbers of communities in which individuals were distributed

Tables 3a and 3b show the averages of the 780 homophily index values calculated for each attribute after the RRL process had been completed.[10] Table 3a shows those attributes for which the index value remained significant in all simulation rounds (i.e., the probability of the index remaining significant was equal to 1), which therefore corresponded to resilient homophily patterns. Table 3b shows those attributes where the probability of the index remaining significant was less than 1. In this table, a wide variety of possibilities can be seen, ranging from indexes that never remained significant ("Baccalaureate student" (class_der1), "father employed in the public administration" (fathempl1) and "has a partner" (partner1)), to indexes that were very likely to remain significant ("belongs to the NSG music association" (assoctype1), "started a relationship less than five months ago" (startrel1) and "mother working in the education and social services sector" (mothsec3).

The first step in the analysis was to rule out the possibility that the differences between the indexes in Table 3a and those in Table 3b was simply due to the absolute number of homophilic links in each case. Thus, since the random replacement of links necessarily decreases the value of the CHI (due to the fact that new links have a very high probability of not being homophilic links), it could be the case that the number of replaced links and the probability of the CHI remaining significant were negatively associated.

[10] For reasons of space, the variable codes appear in the tables. Their labels can be found in Appendix 1.

**TABLE 3A.** *Resistant homophily indexes*

| Category | CHI(*) |
|---|---|
| assoctype2 | 0.238 |
| class_der2 | 0.467 |
| class1 | 0.278 |
| class2 | 0.059 |
| class3 | 0.364 |
| class4 | 0.313 |
| drugtype1 | 0.158 |
| age16 | 0.282 |
| In-degree_status3 | 0.256 |
| municipality1 | 0.208 |
| municipality2 | 0.271 |
| municipality3 | 0.209 |
| municipality4 | 0.207 |
| municipality5 | 0.195 |
| religion1 | 0.145 |
| gender2 | 0.202 |
| tobacco1 | 0.222 |
| dorm1 | 0.546 |

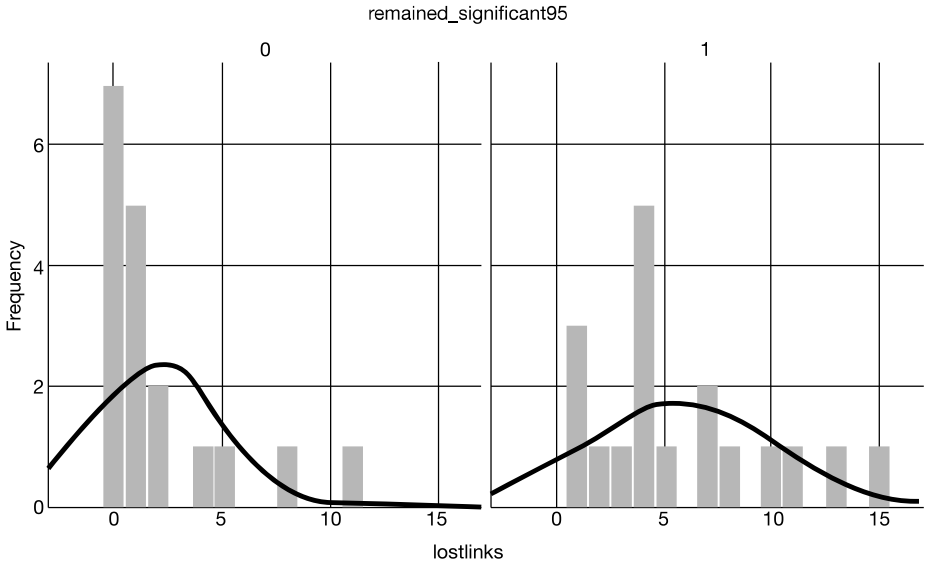
(*) Average CHI after RRL procedure.
*Source:* Own elaboration.

**TABLE 3B.** *Resistant homophily indexes*

| Category | CHI(*) | $fi$ (**) |
|---|---|---|
| assoctype1 | 0.123 | 0.92 |
| class_der1 | –0.010 | 0.00 |
| sporttype0 | –0.050 | 0.01 |
| sporttype1 | 0.022 | 0.01 |
| age18 | 0.067 | 0.57 |
| startrel1 | 0.118 | 0.91 |
| startrel3 | 0.069 | 0.40 |
| readtype1 | 0.081 | 0.44 |
| mothsec3 | 0.127 | 0.91 |
| musictype1 | 0.124 | 0.91 |
| musictype2 | 0.090 | 0.72 |
| fathempl1 | 0.010 | 0.00 |
| fathsec5 | 0.066 | 0.46 |
| rel1 | 0.021 | 0.00 |
| religion0 | 0.090 | 0.48 |
| gender1 | 0.076 | 0.03 |
| Videogames1 | 0.066 | 0.08 |
| dorm2 | 0.315 | 0.84 |

(*) Average CHI after RRL procedure.
(**) Relative frequency of statistically significant index ($p<0.05$).
*Source:* Own elaboration.

**FIGURE 1.** *Distribution of the number of reallocated links*




*Note:* Panel 0 shows the distribution of the number of missing homophilic links in the case of the indexes where "no" remained significant, while panel 1 shows the same distribution for the case of the indexes where "yes" remained significant.

*Source:* Own elaboration.

As can be seen in Figure 1, this was not the case. The right panel (1) shows the histogram of the missing links for the indexes that remained significant after the RRL procedure. It can easily be seen that about half of the indexes withstood the loss of more than six links, up to a maximum of fifteen. The left-hand panel (0) shows that most of the indexes that did not remain significant did not withstand the loss of more than a few links, or even a single link.

**FIGURE 2A.** *Probability that the CHI would remain significant (p<0.05)*

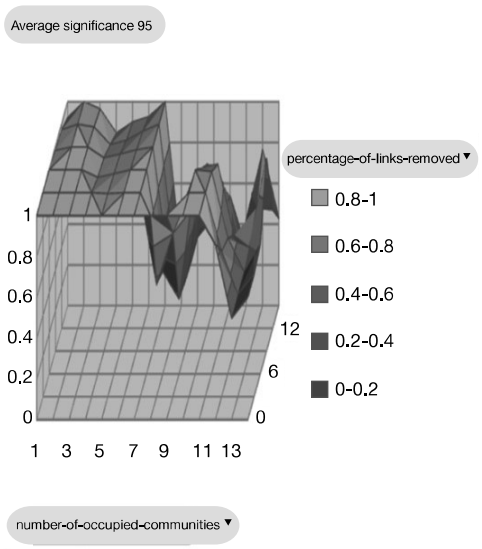
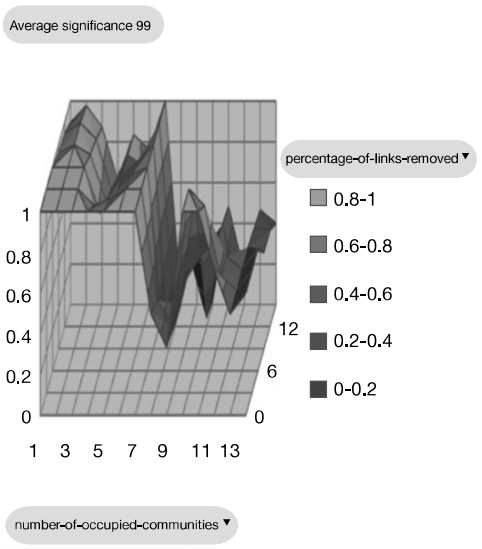



**FIGURE 2B.** *Probability that the CHI would remain significant (p <0.01)*



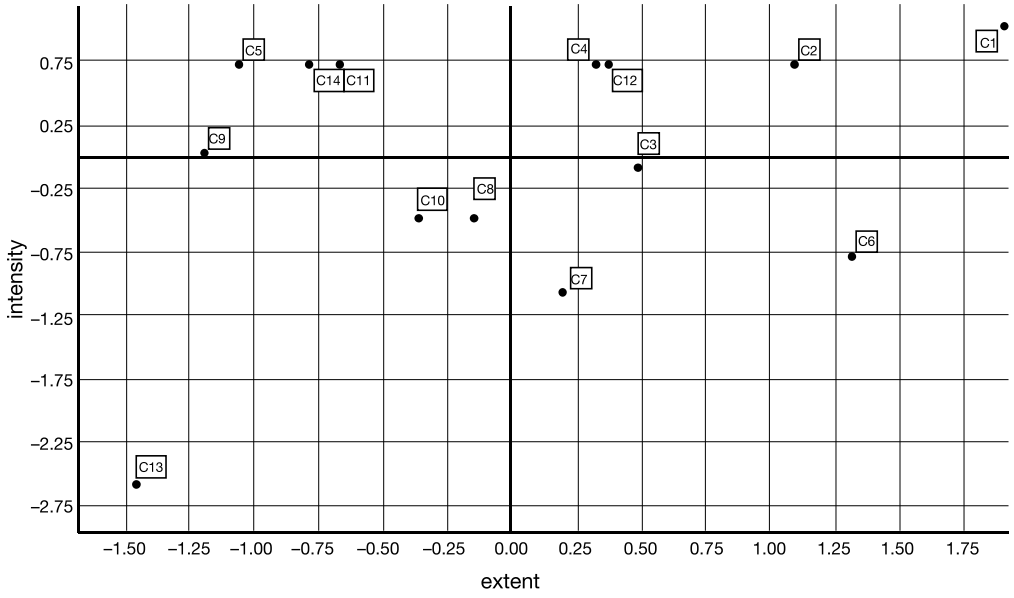

Note: Changes in the probability that the index would remain significant (y-axis) according to the number of communities to which individuals belong (x-axis) and the percentage of homophilic links replaced (z-axis).

*Source:* Own elaboration.

As shown in Figures 2a and 2b, there was a clear relationship between the probability that the CHI would remain significant (y-axis values) with respect to the number of communities[11] among which individuals with the corresponding attribute were distributed (x-axis values), relatively independent of the percentage of links replaced using the RRL procedure (z-axis values). Thus, the probability that the CHI would remain significant decreased very significantly if the number of communities was greater than seven. For example, "men" were distributed into twelve groups and "smokers" into six groups. In the second case, the homophily index was resilient to the RRL procedure, while in the first case the probability of remaining significant started to drop significantly after 9% of the links were replaced.

[11] A community is a subset of the population of individuals in which the density of relationships is higher than in the network as a whole. Community identification algorithms assign each individual in the network to a single group.

**DIAGRAM 2.** *Dispersion of communities by "extent" and "intensity"*




*Source:* Own elaboration.

### Identification of the most frequent communities in the resilient homophily indexes

The evidence shown in the previous section raises the question of whether all communities will be equally important in the "production" of resilient indexes. Two new indexes that we called "extent" and "intensity" were constructed with the purpose of measuring the degree to which each of the fourteen communities identified by the Clauset-Newman-Moore algorithm contributed to resilient or non-resilient CHIs. The term "contribution" relative to a given community $C_i$ was used to denote that at least a fraction of the individuals displaying the attribute whose CHI was being calculated belonged to that community. These indexes were defined as follows:

— The "intensity index" of a community, $IIC_i$, measured the probability that the CHI indexes to which $C_i$ contributed would remain significant.

— The "extent index" of a community, $EIC_i$, measured the proportion of CHI indexes that remained significant where $C_i$ contributed, relative to the proportion of CHI indexes that remained significant where $C_i$ did not contribute[12].

Diagram 2 shows the arrangement of the fourteen communities in a Cartesian space where the horizontal and vertical axes represent their scores for $IIC_i$ and $IEC_i$. Four communities (numbers 1, 2, 4 and 12) scored above average in both dimensions, deviating by approximately 0.75 standard deviations from the mean value for the intensity index, and by 0.25 to 1.80 standard deviations from the mean value of the extent index. Moreover, as can be seen in Table 4, the scores of both indexes were strongly correlated with some community characteristics: there was a strong negative correlation of both indexes

[12] A detailed account of the construction of the extent index can be found in Appendix 3.

with link density (the ratio of existing links to possible links) and a positive correlation with both the average distance and the maximum geodesic distance between two nodes. These correlations suggest that certain topological features of the communities could make it easier for the indexes in which they participate to remain significant.

**TABLE 4.** *Correlations of extent and intensity indexes with the characteristics of the communities*

| | Link density | Average geodetic distance | Maximum geodetic distance |
|---|---|---|---|
| $EIC_i$ | –0.772 | 0.771 | 0.668 |
| $IIC_i$ | –0.614 | 0.592 | 0.573 |

*Source:* Own elaboration.

**TABLE 5.** *Some salient attributes of the homophilic core*

| ATTRIBUTE | Frequency in the homophilic core (N = 59) | Frequency in the giant component (N = 163) |
|---|---|---|
| MUNICIPALITY1 | 20.9% | 52.00% |
| MUNICIPALITY2 | 30.6% | 16.00% |
| GENDER MALE | 52.0% | 56.00% |
| GENDER FEMALE | 48.0% | 44.00% |
| AGE 16 | 30.5% | 24.50% |
| AGE 17 | 45.8% | 53.30% |
| AGE 18 | 13.6% | 14.11% |
| AGE 19 | 6.8% | 3.70% |
| DORMITORY (*) | 61.3% | 31.00% |
| B. SCIENCE (**) | 36.0% | 41.00% |
| MODULES (***) | 24.0% | 17.00% |

(*) Had a place in the student dormitory.

(**) Studying the Science and Technology strand of the Baccalaureate.

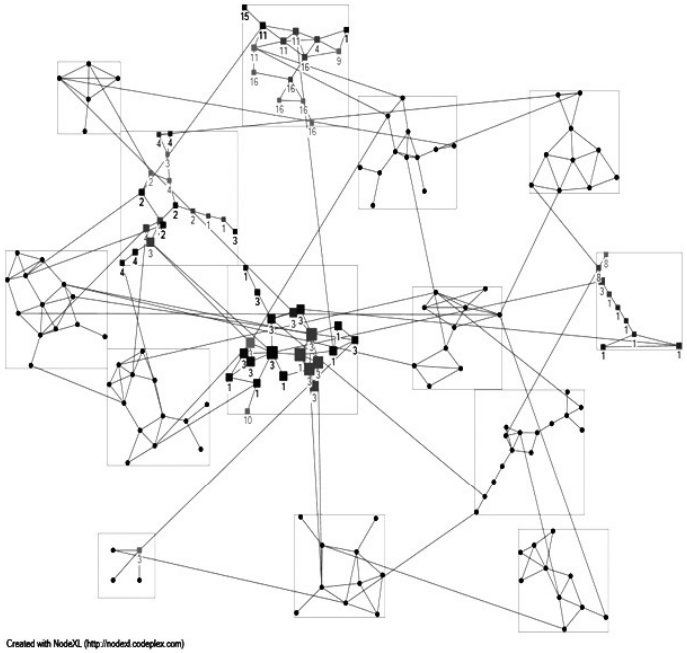
(***) Studying a Vocational Training Module.

*Source:* Own elaboration.

Table 5 shows some differential characteristics of the individuals that formed part of the communities indicated, which were called the "homophilic core", the disposition of which in the giant component of the network is shown in Figure 1a (Figure 1b shows exclusively the nodes belonging to the homophilic core). Of particular importance was the higher proportion of students who had a place in the school dormitory, correlated with a lower proportion of individuals coming from municipality 1 (where IES San Borondón[13] is located), suggesting that part of the resilient homophily patterns were linked to the relational web generated by the cohabitation of young people in the student dormitory.

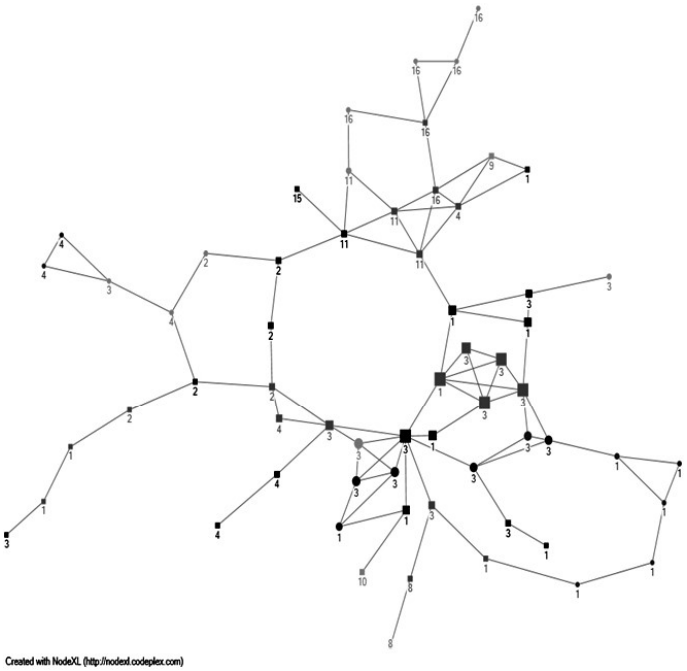
[13] Due to the orographic features of the island, the municipality where IES San Borondón is located had a dormitory for students coming from other municipalities.

**NETWORK 1A.** *Giant component of IES San Borondón*




*Source:* Own elaboration.

**NETWORK 1B.** *Homophilic core*



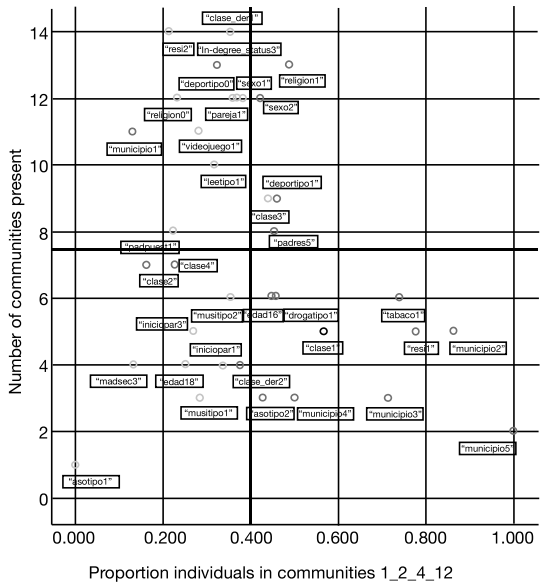

*Source:* Own elaboration.

### Resilience of homophily patterns

The previous sections have shown two findings: (1) when individuals with the attribute under study were concentrated in seven or fewer communities, homophily indexes were more likely to remain statistically significant after the RRL process, and (2) communities 1, 2, 4 and 12 participated in a higher proportion in indexes that remained significant and, at the same time, their participation increased the probability of an index remaining significant.

Figure 3 shows the arrangement of homophily indexes in a Cartesian space defined by the variable "proportion of individuals in communities 1, 2, 4 and 12" (x-axis) and the variable "number of communities present" (y-axis). For example, the point "municipality5" at the bottom right represents the homophily index among students coming from that municipality of San Borondón; these students were distributed into two groups and 100% of them belonged to communities 1, 2, 4 and 12. The point "dorm2" in the upper left corner represents the homophily index for the individuals who did not have a place in the student dormitory, who were distributed into fourteen groups, and only 20% of them belonged to communities 1, 2, 4 and 12. The lines represent the means of the variables, and most of the indexes that remained significant appear in the lower-right quadrant, as expected.

**FIGURE 3.** *Distribution of homophily indexes*



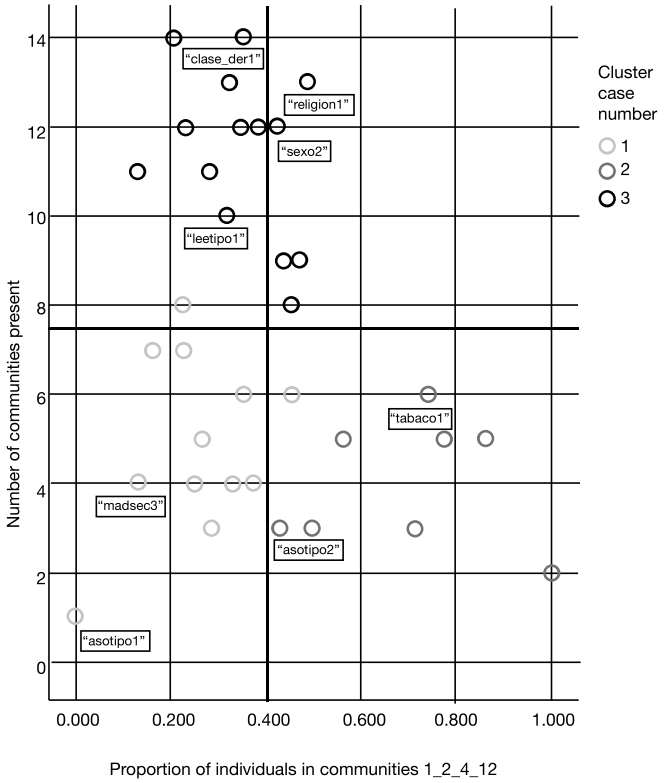

*Source:* Own elaboration.

The resilient indexes can be classified in three main categories. Firstly, indexes that resulted from a focal point for individuals, such as grade's classroom of which they are part, the student dormitory, or the football club. Secondly, indexes that were the result of pre-existing relationships (i.e., all those linked to the municipalities of origin), which clearly illustrated the concept of status homophily. Thirdly, indexes reflecting behaviour susceptible to social contagion, such as tobacco and marijuana use, illus-

trated the concept of value homophily. The full list can be found in Table 3a above.

Figure 4 shows the arrangement of the homophily indexes in the same Cartesian space as above. In this case, the indexes are classified into three clusters identified by the k-means algorithm. It is interesting to note that the algorithm has classified in the same cluster #3 most of the homophily indexes that corresponded to features that were widely distributed in the network (eight or more communities), and that, precisely because of this, individuals belonging to communities 1, 2, 4 and 12 represented a smaller percentage of the total. Examples of these indexes are "class_der1" (individuals who doing their baccalaureate), "religion1" (individuals who define themselves as Catholics), "video game1" (video game users), "readtype1" (readers) or "gender2" (female individuals). The vast majority of the elements in this cluster #3 did not withstand the RRL procedure.

**FIGURE 4.** *Clustering of homophily indexes (k-means algorithm)*




*Source:* Own elaboration.

Cluster #1 consisted of a heterogeneous set of indexes that corresponded to characteristics that were not widely distributed among the population, in which the percentage of individuals belonging to communities 1, 2, 4 and 12 were relatively low. Examples are: "musictype2" (individuals who like pop music), "startrel3" (individuals who have been in a relationship for more than a year), and "mothsec3" (mothers working in the ed-

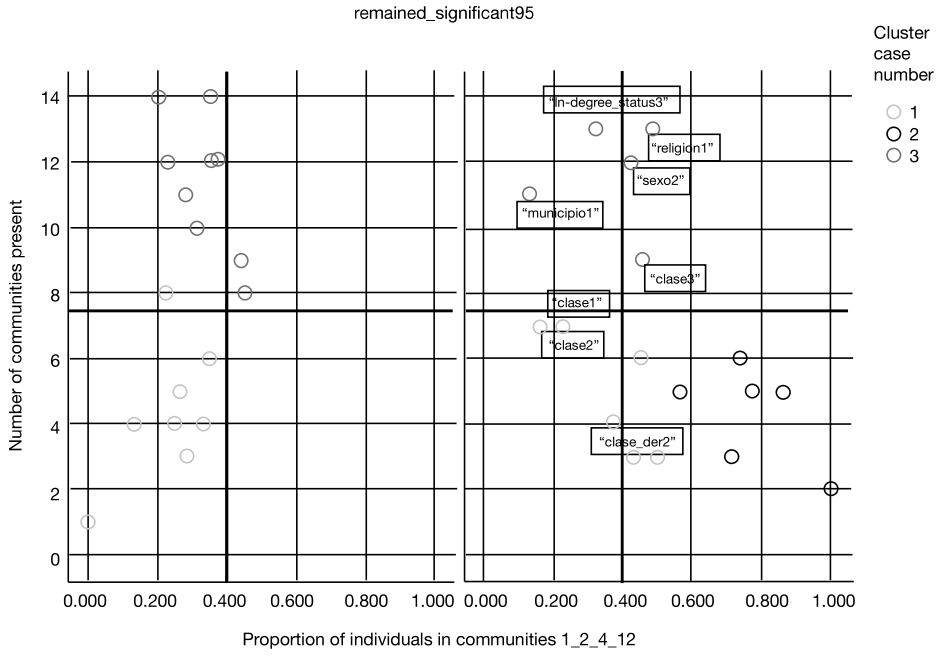
ucation and service sector). Although most of the indexes in this cluster did not withstand the RRL procedure, some of them did.

Finally, all elements of cluster #2 (characteristics not widely distributed but with a high degree of presence of communities 1, 2, 4 and 12) corresponded to indexes that remained significant, such as "tobacco1" (individuals who smoked), "assoctipe2" (individuals in the football club), "class1" (individuals in the first year of the science and technology baccalaureate) or "dorm" (individuals who lived in student accommodation).

To complete the description, Figure 5 shows the information in Figure 4 divided into two panels: the one on the right (1) represents the indexes that withstood the RRL procedure and the one on the left (0) those that did not. This figure clearly shows the "deviant" cases, i.e., the indexes which resisted the RRL procedure, despite not being in the lower-right quadrant.

Looking at the right-hand panel, three cases from cluster #1 are in the lower-left quadrant: "class2" (students in the 2nd year science and technology baccalaureate), "class4" (students in the 2nd year social sciences and humanities baccalaureate ) and "class_der2" (students taking vocational training modules); these are part of the group of resilient homophily indexes related to a given focal point.

**FIGURE 5.** *"Deviant" cases*




*Source:* Own elaboration.

We can also see five indexes classified in cluster #3 in the top half. Of these, "municipality1" and "class3" (individuals in the first year of the social sciences and humanities baccalaureate) belonged to the above-mentioned types of resilient indexes too. The presence in this panel of "in-degree_status3" (individuals mentioned three times

or more) was a result consistent with similar findings in the literature (Maggio and Gari, 2012) regarding the opportunities provided by the structure of social networks: people with more relationships also relate more to each other. Finally, the resilience of "religion1" (Catholic religion) certainly seems to be an anomalous case; as was "gender2" (female individual), whose characteristics were very similar to those of "gender1" (male individual), which, in contrast, was not a resilient index.

## Discussion and conclusions

This paper has addressed the issue of what makes a pattern of homophily robust, so it can therefore be considered a stable characteristic of a network of individuals. A robust or resilient pattern has been understood to be one that produces a homophily index that remains statistically significant even when the network of relationships undergoes a number of disturbances which, for practical purposes, has been set at the random replacement of 15% of the homophilic links.

The evidence suggests that the distribution of individuals across communities in the network is a key factor, albeit in a counter-intuitive sense: there appears to be an inverse relationship between the number of communities in which individuals are distributed and the likelihood that the homophily index remains significant after applying the RRL procedure. Thus, most of the characteristics widely distributed in the population did not give rise to resilient homophily indexes (with some notable exceptions, such as the "female" attribute). This inverse relationship appears to be due to the fact that a few communities (namely, numbers 1, 2, 4 and 12 in our case study) contributed disproportionately to the production of resilient homophily patterns. Paradoxically, this involves that the indexes that remained significant did not provide information on the network as a whole, but on some of the communities within it; and, consequently, the statement "there was homophily in the whole of the network N regarding attribute A" is a generalisation that is not true without further qualification.

The second result is that the most important communities in the production of resilient homophily patterns seem to have specific topological features, such as greater spread and lower density than the average. However, given the small number of communities in the network analysed, this result requires replication of the study with other networks to be confirmed or discarded.

With regard to the mechanisms explaining homophilic patterns, the results are consistent with studies that have underlined the importance of focal points of activity and the opportunity structure of networks, given that most of the resilient indexes corresponded to these cases. A significant number (but not all) of characteristics that can be labelled as status homophily, as well as some characteristics susceptible to social contagion, also showed resilient homophily indexes.

However, most homophily indexes for attributes associated with students' parental status were not robust; a result that is actually consistent with Peter M. Blau's conceptual framework, since when parameters are intersecting, a homophilic association in one parameter necessarily implies a heterophilic association in another. Thus, for example, for the homophily index of students whose parents worked in the service sector to be resilient, the school-year-based association would have to be non-resilient. In other words, homophilic association between individuals in, say, the first year of natural sciences necessar-

ily produced heterophilic associations between individuals with parents of a different occupational status.

The robustness of the results of this study obviously depends on the possibility of replicating them with other populations of individuals and with other algorithms for both community search and link replacement. However, the idea that a pattern of behaviour will be resilient if it is concentrated in certain communities in the network has a theoretical significance that transcends the particular case we have analysed. It is reasonable to suggest that this may be a generalisable result for the 'small world' class of networks. If this were the case, the analysis procedure designed for this research would be useful for a wide number of empirical domains, some belonging to other disciplines such as economics, anthropology or ecology, where the objects of analysis could be expected to be embedded in *small world* networks.

## Bibliography


Aral, Sinan; Muchinik, Lev and Sundarajan, Arun (2009). "Distingusing Influence-Based Contagion from Homophily Driven Diffusion in Dynamic Networks". *PNAS*, 16(51).

Blau, Peter M. (1977). *Inequality and Heterogeneity: A Primitive Theory of Social Structure*. New York: Free Press.

Bojanowski, Michel and Corten, Rense (2014). "Measuring Segregation in Social Networks". *Social Networks*, 39: 14-32.

Cohen, Jere M. (1977). "Sources of Peer Group Heterogeneity". *Sociology of Education*, 50: 227-241.

Coleman, James S. (1957). "Relational Analysis: the Study of Social Organization with Survey Methods". *Human Organization*, 17(4): 28-36.

DiMaggio, Paul and Garip, Filiz (2012). "Networks Effects in Social Inquality". *Annual Review of Sociology*, 38: 93-118.

Feld, Scott (1981). "The Focused Organization of Organizational Ties". *American Journal of Sociology*, 86: 1015-1035.

Feld, Scott (1982). "Structural Determinants of Similarity among Associates". *American Sociological Review*, 47: 797-801.

Kandel, Denise B. (1978). "Homophily, Selection and Socialization in Adolescent Friendship". *American Journal of Sociology*, 84(2): 427-436.

Kossinets, Gueorgi and Duncan, Watts (2009). "Origins of Homophily in an Evolving Social Network". *American Journal of Sociology*, 115(2): 405-50.

Lazarsfeld, Paul F. and Merton, Robert K. (1954). "Friendship as a Social Process: A Substantive and Methodological Analysis". In: Berger, M. (ed.). *Freedom and Control in Modern Society*, pp. 18-66. New York: Van Nostrad.

Linares, Francisco (2018a). *Sociología y teoría social analíticas: la ciencia de las consecuencias inintencionadas de la acción*. Madrid: Alianza Editorial.

Linares, Francisco (2018b). "Agent Based Models and the Science of Unintended Consequences of Social Action"/"Los modelos basados en agentes y la ciencia de las consecuencias inintencionadas de la acción". *Revista Española de Investigaciones Sociológicas*, 162: 21-37.

Linares, Francisco and Kohl, Mona (2017). "Social Networks and Homophily Patterns among Post-Secundary Students in San Borondón". *I Encuentro de Sociología Analítica y Migraciones*. Universidad de A Coruña.

Lozares, Carlos and Verd, Joan M. (2011). "De la homofilia a la cohesión social y viceversa". *Redes – Revista Hispana para el Análisis de Redes*, 20(2): 29-50.

Marsden, Peter V. (1987). "Core Diffusion Networks among Americans". *American Sociological Review*, 52: 122-131.

McPherson, Miller and Smith-Lovin, Lynn (1986). "Sex Segregation in Voluntary Associations". *American Sociological Review*, 51: 61-79.

McPherson, Miller and Smith-Lovin, Lynn (1987). "Homophily in Voluntary Organizations: Status Distance and the Composition of Face-to-Face Groups". *American Sociological Review*, 55: 370-379.

McPherson, Miller; Smith-Lovin, Lynn and Cook, James M. (2001). "Birds of a Feather: Homophily

in Social Networks". *Annual Review of Sociology*, 27: 415-444.

Moody, James (2001). "Race School Integration, and Friendship Segregation in America". *American Journal of Sociology*, 107(3): 679-716.

Shalizi, Cosma R. and Thomas, Andrew C. (2011). "Homophily and Contagion are Generically Confounded in Observational Social Network Studies". *Sociological Methods and Research*, 40(2): 211-239.

Shrum, Wesley; Cheek, Neil H. Jr. and MacD. Hunter, Saundra (1988). "Friendship in School: Gender and Racial Homophily". *Sociology of Education*, 61: 227-239.

Signorile, Vito and O'Shea, Robert M. (1965). "A Test of Significance for the Homophily Index". *American Journal of Sociology,* 70(4): 467-470.

Wilensky, Uri and Rand, William (2015). *An Introduction to Agent-based Modeling*. Cambridge, Massachusetts: The MIT Press.

**RECEPTION:** May 27, 2020
**REVIEW:** November 12, 2020
**ACCEPTANCE:** December 23, 2020

## APPENDIX 1. LIST OF VARIABLES AND CATEGORIES

| Variable | Category | Description | N | CHI |
|---|---|---|---|---|
| ASSOCTYPE | 0 | Does not belong to an association | 117 | 0.128 |
| (membership of an association) | 1 | NSG Music Association | 9 | 0.170** |
| | 2 | Unión Deportiva G | 11 | 0.305** |
| CLASS_DER | 1 | Taking the baccalaureate | 133 | 0.608** |
| (secondary education option) | 2 | Taking vocational training modules | 21 | 0.612** |
| CLASS | 1 | 1st year Science and Technology baccalaureate | 30 | 0.401** |
| (year and type of baccalaureate) | 2 | 2nd year Science and Technology baccalaureate TC | 38 | 0.676** |
| | 3 | 1st year Social Sciences and Humanities baccalaureate | 39 | 0.540** |
| | 4 | 2nd year Social Sciences and Humanities baccalaureate | 26 | 0.438** |
| SPORTTYPE | 0 | Does not do sport | 69 | 0.153** |
| (Does sport) | 1 | Plays football | 37 | 0.164** |
| | 2 | Goes to the gym | 14 | 0.192** |
| DRUGTYPE | 0 | No drug use | 131 | 0.090 |
| (drug use) | 1 | Takes drugs | 29 | 0.240** |
| AGE | 16 | 16 years old | 40 | 0.419** |
| | 17 | 17 years old | 87 | 0.234** |
| | 18 | 18 years old | 23 | 0.127** |
| FREQALC | 0 | No alcohol use | 33 | 0.089 |
| (alcohol use) | 1 | Drinks alcohol only at parties | 122 | 0.398** |
| FREQDRUG | 0 | Never | 131 | 0.090 |
| (frequency of drug use) | 1 | Only uses drugs at parties | 18 | 0.169** |
| | 2 | Drug use also at other times | 12 | 0.195** |
| FREQSMOK | 0 | Non-smoker | 133 | 0.264** |
| (frequency of tobacco use) | 1 | Only smokes at parties | 18 | 0.057 |
| | 2 | Also smokes at other times | 11 | 0.414** |
| GENDER | 1 | Male | 91 | 0.340** |
| | 2 | Female | 72 | 0.544** |
| WEEKENDCURF | 1 | Weekend curfew | 29 | –0.021 |
| (curfew set by parents) | 2 | No weekend curfew | 130 | 0.113 |
| STARTSMOK | 0 | Non-smoker | 133 | 0.264** |
| (curfew set by parents) | 1 | Started before the age of 15 | 11 | 0.212** |
| (started smoking) | 2 | Started at age 15 or older | 11 | 0.016 |

## APPENDIX 1. LIST OF VARIABLES AND CATEGORIES *(CONTINUATION)*

| Variable | Category | Description | N | CHI |
|---|---|---|---|---|
| STARTREL | 1 | Started a relationship less than 5 months ago | 17 | 0.203** |
| (start of the current relationship) | 2 | Started a relationship between 5 months and a year ago | 18 | 0.042 |
| | 3 | Started a relationship over a year ago | 25 | 0.164** |
| IN-DEGREE_STATUS | 1 | Not mentioned or mentioned once in the questionnaire | 57 | –0.730** |
| (number of mentions received in the questionnaire) | 2 | Mentioned twice or three times (2=Avg) in the questionnaire | 67 | 0.046 |
| | 3 | Mentioned more than 3 times in the questionnaire | 39 | 0.352** |
| RDGTYPE | 1 | Is a keen reader | 50 | 0.153** |
| (fond of reading) | 2 | Does not read | 113 | –0.032 |
| MOTHSEC | 1 | Agriculture, livestock, fisheries | 1 | NC |
| (mother's work sector) | 2 | Hospitality and tourism | 33 | 0.078 |
| | 3 | Education and social services | 26 | 0.206** |
| | 4 | Business | 15 | –0.086 |
| | 5 | Construction | 0 | NC |
| | 6 | Health | 11 | 0.156** |
| MOTHJOB | 1 | Public administration employee | 37 | 0.045 |
| (mother's job) | 2 | Company employee | 53 | 0.095* |
| | 3 | Owner of a company or business | 18 | 0.064 |
| MUSICTYPE | 1 | Mentions *reggaeton* music | 18 | 0.211** |
| (preferred type of music) | 2 | Mentions pop music | 28 | 0.206** |
| MUNICIPALITY | 1 | municipality SS | 85 | 0.425** |
| (municipality of origin) | 2 | municipality VG | 26 | 0.443** |
| | 3 | municipality AG | 9 | 0.299** |
| | 4 | municipality HE | 14 | 0.288** |
| | 5 | municipality AL | 15 | 0.270** |
| | 6 | municipality VH | 14 | 0.305** |
| NUMPART | 0 | 0 relationships in the last 18 months (not counting the current one) | 62 | 0.012 |
| (number of past love relationships) | 1 | 1 relationship in the last 18 months (not counting the current one) | 42 | 0.070 |
| | 2 | 2 relationships in the last 18 months (not counting the current one) | 21 | 0.025 |
| | 3 | 3 or more relationships in the last 18 months (not counting the current one) | 16 | –0.250 |

## APPENDIX 1. LIST OF VARIABLES AND CATEGORIES *(CONTINUATION)*

| Variable | Category | Description | N | CHI |
|---|---|---|---|---|
| FATHSEC | 1 | Agriculture, livestock, fisheries | 11 | 0.027 |
| (father's work sector)2 | 2 | Hospitality and tourism | 24 | –0.155 |
| | 3 | Education and social services | 14 | 0.193** |
| | 4 | Business | 7 | 0.019 |
| | 5 | Plumbing, electricity, construction | 32 | 0.176** |
| | 6 | Health | 1 | NC |
| FATHJOB | 1 | Public administration employee | 29 | 0.085* |
| (father's job) | 2 | Company employee | 55 | 0.067 |
| | 3 | Owner of a company or business | 28 | 0.040 |
| PARRELIG (Parents' religious beliefs) | 0 | Parents with no religious beliefs | 24 | –0.033 |
| | 1 | At least one Catholic parent | 122 | –0.012 |
| | 2 | At least one parent who professes a religion other than Catholic | 14 | –1.000** |
| ALLOWTYPE | 1 | Receives a weekly allowance from parents | 31 | –0.018 |
| (allowance received from parents) | 2 | Does not receive a weekly allowance from parents | 132 | –0.026 |
| RELATIONSHIP | 1 | Is currently in a relationship | 63 | 0.139* |
| (current relationship) | 2 | Is not currently in a relationship | 96 | 0.084 |
| STUDENT DORMITORY | 1 | Living in the student dormitory | 51 | 0.706** |
| (living in student accommodation) | 2 | Not living in the student dormitory | 112 | 0.635** |
| Religion | 0 | No religious beliefs | 65 | 0.314** |
| (religious beliefs) | 1 | Has Catholic beliefs | 80 | 0.319** |
| | 2 | Has beliefs other than Catholic | 16 | 0.000** |
| SMOKING | 1 | Smoker | 30 | 0.329** |
| (tobacco use) | 2 | Non-smoker | 133 | 0.429** |
| VIDEOGAME | 1 | Is a fan of video games | 56 | 0.177** |
| (fond of videogames) | 2 | Is not a fan of video games | 108 | 0.267** |

*Notes:* N = number of individuals; CHI = Coleman homophily index; NC = Not Computable; (*) = Significant ($p<0.05$); (**) = Significant ($p<0.01$).

*Source:* Own elaboration.

## Appendix 2. Agent-based models

There is a broad group of techniques for programming models and running simulations. The technique used in this article is agent-based modelling, which is different from other techniques of the same type that are also used in the social sciences, such as system dynamics.

The use of this type of model involves writing a sequence of instructions detailing the variables that characterise the system (in the case of this study, homophily indexes), the characteristics of the agents (the attributes of the real individuals), and the rules by which certain attributes of the agents change (the replacement of links with other agents) and, in turn, the characteristics of the system (the new homophily indexes).

The computer executes the set rules recursively until the end condition of the simulation is met. Each simulation is repeated N times, manipulating various parameters to obtain a 'population' of cases that is sufficiently diverse to perform statistical sensitivity analyses. A more detailed explanation can be found in Linares (2018a and 2018b).

## Appendix 3. Pseudocode of RRL procedure

1. Set the number of links to be replaced, N = 0.15 times total homophilic links between individuals with attribute A.
2. Set the number of substituted links M = 0.
3. As long as M < N, repeat steps 4 to 11.
4. Randomly choose an individual *i* from the set of individuals with homophilic links with respect to attribute A.
5. Randomly choose a link, $v_{i?}$ from the set of homophilic links of *i* with respect to attribute A.
6. Delete $v_{i?}$.
7. According to procedure $P_i$ ($P_i$ belongs to the set of procedures P for the creation of new links), choose an individual *j* (*j* !=*i*) from the set of individuals with homophilic links with respect to attribute A.
8. Create link $v_{ij}$.
9. Calculate the value of the CHI.
10. Calculate the statistical significance of the CHI.
11. Set M = M + 1.

## Appendix 4. Definition and calculation of the "extent" index

For every $C_i$ there is a number $N_i$ of indexes in which it participated and a number $M_i$ in which it did not participate. In turn, as can be seen in Table 6, both $N_i$ and $M_i$ are the result of the sum of the number of indexes that remained significant, $n_1$ or $m_1$, and the number of indexes that did not remain significant after the RRL procedure, $n_2$ or $m_2$. In addition, $n_1 + m_1$ must be the number of indexes that remained significant and $n_2 + m_2$ must be the number of indexes that did not remain significant.

Let $p_i$ be the proportion of indexes that remained significant relative to the total number of indexes in which $C_i$ participates, that is, $n_1 / N_i$, and let $q_i$ be the proportion of indexes that remained significant relative to the total number of indexes in which $C_i$ did not participate, $m_1 / M_i$, then the extension index is given by the following equation:

$$IEC_i = \frac{p_i}{p_i + q_i}$$

**TABLE 6.** *Distribution of homophily indexes*

| | $C_i$ contributed | $C_i$ did not contribute |
|---|---|---|
| Remained significant | $n_1$ | $m_1$ |
| Did not remain significant | $n_2$ | $m_2$ |
| TOTAL | $N_i$ | $M_i$ |

*Source:* Own elaboration.