Contents lists available at ScienceDirect

Social Networks

journal homepage: www.elsevier.com/locate/socnet

Measuring segregation in social networks

Michał Bojanowski[a,*], Rense Corten[b]

[a] Interdisciplinary Centre for Mathematical and Computational Modelling (ICM), University of Warsaw, Prosta 69, 00-838 Warsaw, Poland
[b] Department of Sociology/ICS, Utrecht Unviersity, Padualaan 14, 3584 CH Utrecht, The Netherlands

ARTICLE INFO

MSC:
91D30

Keywords:
Social networks
Segregation
Homophily
Measurement

ABSTRACT

Network homophily is a pattern in which ties are more likely to exist between nodes similar to each other. It is frequently observed for various types of social relations. At the same time, segregation is often encountered in urban areas as a tendency of families to occupy neighborhoods inhabited by other families similar to them. In this paper we conceptualize both phenomena as in the language of networks of interlinked positions occupied by a population of actors characterized by some node-level attribute. We review existing indexes and approaches to measuring the extent of homophily/segregation in social networks. We pursue an approach of, first, specifying a set of properties that a generic segregation measure might possess, and which, in our view, are relevant in substantial contexts. Second, we check which measures satisfy which properties. The use of measures is illustrated with four empirical examples. Given the particular application and the need for some descriptive measure of segregation, the results presented in this paper can help in selecting an optimal measure for the task at hand. We conclude that the most crucial aspects for the choice of a particular segregation measure include (1) whether the network ties or actors' attributes are assumed to be subject to change, and (2) how one should treat the presence of network isolates.

1. Introduction

In many types of social relations, ties are more likely to form between similar entities than between dissimilar entities. For example, individuals tend to marry others who are similar in terms of age, education, and socio-economic status (Kalmijn, 1998). The discussion of important matters, friendship, and social support also share this feature of *homophily* (see McPherson et al., 2001 for an extensive review of the empirical evidence regarding homophily). We also observe homophily in situations in which individuals affect or influence each other (Erickson, 1988; Cialdini and Goldstein, 2004). For example, people tend to be strongly influenced by others when choosing cultural products (Salganik and Watts, 2009), and friends tend to have similar opinions, especially when the choice of friends is somewhat constrained by the social context (de Klepper et al., 2010).

A related phenomenon, often discussed outside of the social networks literature, is *segregation*. Massey and Denton (1988) defined segregation as "the degree to which two or more groups live separately from one another" in the context of racial segregation of city neighborhoods. The concept of segregation is also applied to the "unequal" distribution of two or more groups of people across different units or social positions. Racial segregation of neighborhoods focuses on the distribution of people belonging to different racial groups across neighborhoods or city blocks constituting the units. In a largely similar fashion, Charles and Grusky (1995) address the way in which groups of men and women are unequally represented in different occupational classes. The literatures on ethnic segregation and gender segregation both emphasize the constraining aspect of segregation as a form of social organization because it places "limits on interactions" (van der Zanden, 1972) and induces a "form of isolation which places limits upon contact, communication, and social relations" (Hunt and Walker, 1974).

While homophily in networks and segregation in neighborhoods or occupations may emerge from very different social processes, the *outcome* in each case is a social structure of interrelated positions occupied by a population of actors consisting of at least two groups. This structure can generally be modeled as a network with the nodes corresponding to the actors and the links corresponding to the relations between the actors. For example, school children from different ethnic groups in a newly assembled class start to form friendships with one another. The typical

[*] Corresponding author. Tel.: +48 22 8749 427.
E-mail addresses: m.bojanowski@uw.edu.pl (M. Bojanowski), r.corten@uu.nl (R. Corten).
URLs: http://www.bojanorama.pl (M. Bojanowski), http://www.rensecorten.org (R. Corten).

http://dx.doi.org/10.1016/j.socnet.2014.04.001
0378-8733/© 2014 Elsevier B.V. All rights reserved.

outcome of preferential friendship formation processes is a highly homophilous network in which the nodes correspond to children and the links to friendship (Moody, 2001). As another example, consider families of different ethnicities moving to a neighborhood. The neighborhood consists of heterogeneously placed dwellings. In this context, the locations of the dwellings and their spatial proximities can be represented as a network in which the nodes correspond to dwellings and the edges link the dwellings that are adjacent to each other. If the dwellings become occupied by the families, each node of the graph is characterized by the ethnic group of the resident family. Therefore, the outcome is again a network with a node-level attribute designating the *groups* of the nodes. As a third example, consider a contagion-like process, in which some trait spreads in a social network. Also in this case, the outcome is a social structure of types of actors, in this case "infected" and non-infected.

In different literatures, researchers have identified various dimensions of segregation (see Massey and Denton, 1988, for a complete discussion and proposed measures). For the purpose of this paper, we take the view of segregation as (the lack of) *exposure*: the extent to which groups are exposed to one another by occupying nearby positions. This aspect of segregation is intrinsically *relational*, which brings us very close to the social network literature.

Like other types of segregation, segregation in social networks may emerge from different types of processes. Probably the more familiar process is one in which we have a population of actors with fixed attributes (say, gender or ethnicity), among whom a social network is formed. The extent to which actors connect to others with the same attributes generates a level of segregation. However, the same result may also be obtained if the network is fixed and instead the attribute (say, adoption of a certain behavior) change. Segregation then emerges if this attribute tends to cluster in certain parts of the network.

It is important to stress that, while the two processes may substantively be very different, the end result is equivalent from a measurement perspective, namely a situation in which actors with certain attributes are to some degree connected to others with the same attributes. Yet, as will become clear in our analysis, the distinction between the two processes has implications for the interpretation of a given measure of segregation.

A frequent goal of empirical investigations is to compare specific outcomes across different groups, settings, or time points. For example, one could compare different year groups, schools, or classes with respect to the level of friendship segregation (Moody, 2001). In other settings, one might want to compare different districts of a city, or several cities, in terms of the ethnic residential segregation of neighborhoods (Freeman and Sunshine, 1970). Performing such comparisons necessitates the *measurement* of the *level of segregation* in the given network.

Various measures and approaches have been proposed in the network literature. Although these measures are intended for describing the same phenomenon, they originate from different literatures, follow different logics, and are typically proposed without referencing one another. Thus, it is possible for different measures to lead to different conclusions in the same situation. To our knowledge, no systematic overview of the available measures exists. In this paper, we provide a systematic overview of existing segregation measures and highlight the similarities and differences between those measures, with the goal of enabling the researchers to choose the right measure for their respective purposes.

The somewhat dissatisfying state of affairs concerning the measurement of network segregation may be attributed to the same causes that Duncan and Duncan (1955) identified in the realm of segregation measurement (in the stratification sense) in the 50s, namely "naive operationalism" and "[arbitrarily] matching some convenient numerical procedure with the verbal concept of segregation". What is needed is a measurement theory to enable the careful theoretical grounding of segregation measurement.

One particular strategy for building this theoretical basis is the *axiomatic method*. The axiomatic method starts by positing a set of basic properties, or axioms, that a generic measure should possess. In the deductive steps that follow, the goal is to derive classes of measures that logically result from different combinations of the proposed axioms. In the ideal case, the ultimate goal is to arrive at collections of axioms that pin down a single measure of a concept at hand. In other words, given a certain collection of axioms, there exists one and only one measure that simultaneously satisfies all of them.

The axiomatic method has been fruitfully applied in the social sciences. Examples include such diverse domains as utility measurement (Suppes and Winet, 1955), measurement of inequality (Schwartz and Winship, 1980; Cowell and Kuga, 1981; Chakravarty, 1999), income mobility (Cowell, 1985), numerous problems in social choice theory such as the axiomatization of the simple majority rule (May, 1952) or various implications of the assumptions about measurability and comparability of individual utility functions (for example, d'Aspremont and Gevers, 1977, 1985). Regarding segregation, much of the progress in the social stratification research on segregation has been made through the employment of an axiomatic approach (or its associated elements) in the work of James and Tauber (1985), in the later work by Reardon and Firebaugh (2002a) and others (e.g., Egan et al., 1998; Massey and Denton, 1998; Grannis, 2002; Reardon and Firebaugh, 2002b), and recently in work by Alonso-Villar and del Río (2010).

While we believe that a full axiomatization of segregation in social networks is desirable, in this paper we opt for a more practical approach by providing a systematic overview of existing approaches to measuring segregation. We do so by considering a set of atomic properties that a generic segregation measure might possess, and that we believe have practical consequences for research. We then compare existing measures of segregation against this set of properties. Although we do not provide definite results in the form of axiomatizations, we believe that what follows provides an attractive perspective on the problem. The results we obtained should enable researchers to choose an appropriate measure in a particular substantive research context.[1]

In the following section, we define the notation that will be used in the remainder of the paper. In Section 3, we formulate the properties that will guide our analyses of existing segregation measures. Then, the main part of the paper is devoted to an overview and analysis of nine existing segregation measures (Section 4). For each measure, we provide a brief explanation and verify the extent to which the measure conforms to the properties formulated in Section 3. We then demonstrate the use of the measures we discussed by way of a number of empirical examples in Section 5. In the concluding Section 6, we summarize the results of this endeavor and discuss the implications of the results on the practical use of the measures reviewed.

2. Definitions and notation

We introduce the necessary notation and basic definitions that will be used throughout the paper. The notation is loosely based on the standards proposed by Wasserman and Faust (1994).

Network nodes The set of nodes is denoted by $\mathcal{N} = \{1, \ldots, i, \ldots, N\}$.

[1] Instead of "reviewing" the measures, a truly axiomatic method would be to combine the axioms and arrive at some parametrized class(es) of measures. That, however, is beyond the scope of this paper.

Groups Nodes of the network are assigned to groups (for example, based on ethnicity). Grouping implies a partition of the set of nodes into exhaustive and mutually exclusive subsets. The set of K groups may be denoted as: $\mathcal{G} = \{G_1, \ldots, G_k, \ldots, G_K\}$ where G_k is a generic kth group that is a subset of $\mathcal{N}$. In the remainder of the paper, a simpler notation will be used wherever it does not introduce ambiguity. Groups will be referred to with the index k, i.e., group 1, group 2, and group k. The letters h and g will also be used to refer to generic groups.

The partition of nodes into groups can be formalized as a *type vector*:

$$\mathbf{t} = [t_1, \cdots, t_i, \cdots, t_N] \quad \text{where} \quad t_i \in \{1, \ldots, K\}. \tag{2.1}$$

The values in the type vector assign the nodes to the groups, with the value at the ith position designating the group number to which node i belongs. The set of all possible type vectors is denoted as $\mathcal{T}$.

The numerous properties of graphs and group distributions are stated using linear algebra notation. Another way of representing group membership of the nodes is with a *type indicator vector* for group k:

$$\mathbf{v}_k = [v_1, \cdots, v_i, \cdots, v_N] \quad \text{where} \quad v_i \in \{0, 1\}, \tag{2.2}$$

derived from type vector $\mathbf{t}$ such that the value of v_i is 1 if node i belongs to group k. Formally,

$$v_i = \begin{cases} 1 & \text{if } t_i = k \\ 0 & \text{if } t_i \neq k \end{cases}. \tag{2.3}$$

Additionally, it is convenient for some computations to use a matrix that combines the type indicator vectors for all groups columnwise. We call this matrix *type indicator matrix*, which is defined as follows: for a given type vector $\mathbf{t}$ of length N describing the membership in K groups a type indicator matrix T is a matrix with N rows and K columns with entries that are either 0 or 1. $T_{ik} = 1$ if node i is a member of group k and zero otherwise. Consequently, the kth column of T is equivalent to $\mathbf{v}_k$ – the type indicator vector for group k.

It is important to realize that all three representations – the partition, the type vectors and the type indicator matrix – are equivalent in that they contain the same information about the group membership of the nodes.

Network ties Following the sociometric tradition of Wasserman and Faust (1994) the network is defined by a binary, irreflexive, and (a)symmetric[2] relation R defined over $\mathcal{N} \times \mathcal{N}$. This relation implies a squared adjacency matrix $X = [x_{ij}]_{N \times N}$ such that

$$iRj \quad \leftrightarrow \quad X_{N\times N} \begin{cases} x_{ij} = 1 & & \text{in the directed case,} \\ x_{ij} = 1 \quad \leftrightarrow \quad x_{ji} = 1 & & \text{in the undirected case.} \end{cases} \tag{2.4}$$

In specific contexts, and when noted, we will use other capital letters such as Y, Z to represent graph adjacency matrices. By $\mathcal{X}$, we denote a set of all possible network matrices.

Degree of a node The degree of a node i is denoted with η_i, that is $\eta_i = \sum_{j=1}^{N} x_{ij}$.

Sizes of groups The number of nodes in group G_k is denoted by n_k.

Mixing matrix We define the mixing matrix as a three-dimensional distribution of all the dyads (pairs of actors) based on three characteristics:

1. The group to which the first node in the dyad belongs;
2. The group to which the second node in the dyad belongs;
3. Whether the two nodes in the dyad are connected in the analyzed network.

Formally, for a network with adjacency matrix X and type vector $\mathbf{t}$, the mixing matrix $M = [m_{ghy}]_{K\times K\times 2}$ is defined as

$$m_{gh1} = \sum_{i\in G_g} \sum_{j\in G_h} x_{ij}, \tag{2.5}$$

$$m_{gh0} = \sum_{i\in G_g} \sum_{j\in G_h} (1 - x_{ij}). \tag{2.6}$$

The "contact layer" of the mixing matrix, m_{gh1}, summarizes the pattern of existing ties in the network in terms of the group memberships of the nodes. The "non-contact layer", m_{gh0}, provides supplementary and analogous information about disconnected dyads (Koehly et al., 2004).

The values of m_{gh1} can be conveniently calculated based on the adjacency matrix X and a type indicator matrix T with[3]

$$m_{gh1} = T^T XT. \tag{2.7}$$

Additionally, with the + sign we denote summation over a particular subscript when dealing with marginal distributions of the mixing matrix. For example:

$$m_{gh+} = \sum_{y=1}^{2} m_{ghy}, \quad m_{+hy} = \sum_{g=1}^{K} m_{ghy},$$
$$m_{++y} = \sum_{g=1}^{K} \sum_{h=1}^{K} m_{ghy}. \tag{2.8}$$

Segregation indices A generic index of segregation on the network level, $S(\cdot)$, is a function that maps every

[2] The network may be directed or undirected.

[3] The symbol ^{T} denotes matrix transposition.

Table 1
Summary of notation.

Symbol	Meaning
N	Number of nodes
$\mathcal{N}$	Set (population) of nodes: $\mathcal{N} = \{1, \ldots, i, \ldots, N\}$
X	$N \times N$ network adjacency matrix
x_{ij}	Element of X
$\mathcal{X}$	Set of all possible networks of size N
η_i	Degree of node i
$\mathcal{G}$	Set of all groups: $\mathcal{G} = \{G_1, \ldots, G_g, \ldots, G_K\}$
n_g	Number of nodes in group G_g
$\mathbf{t}$	Type vector $\mathbf{t} = [t_1, \ldots, t_i, \ldots, t_N]$
t_i	Element of $\mathbf{t}$, $t_i \in \{1, \ldots, g, \ldots, K\}$
$\mathbf{v}$	Type indicator vector
T	Type indicator matrix
M	Mixing matrix $[m_{ghy}]_{K \times K \times 2}$
m_{ghy}	Element of M: number of dyads between nodes from group G_g with nodes in group G_h, $y = 1$ dyad is connected, $y = 0$ if it is not connected
S	Generic network-level segregation index, $S : \mathcal{X} \times \mathcal{T} \mapsto \mathbb{R}$
S^g	Generic group-level segregation index, $S^g : \mathcal{X} \times \mathcal{T} \times \mathcal{G} \mapsto \mathbb{R}$
S_i	Generic node-level segregation index, $S_i : \mathcal{X} \times \mathcal{T} \times \mathcal{N} \mapsto \mathbb{R}$

network matrix and a type vector to a real number:

$$S : \mathcal{X} \times \mathcal{T} \mapsto \mathbb{R}. \tag{2.9}$$

Some of the indices reviewed next in Section 4 are not defined on the network level but on the lower levels, assigning segregation scores to groups or even individual nodes. Moreover, some of them can be conveniently aggregated from lower levels to higher levels (e.g., from the group level to the network level) or disaggregated from higher levels to lower levels (e.g., from the group level to the node level). The group-level segregation index can be defined as a function that assigns a segregation score to every group in each combination of network and type vector, i.e.,

$$S^g : \mathcal{X} \times \mathcal{T} \times \mathcal{G} \mapsto \mathbb{R}. \tag{2.10}$$

Analogously, the node-level segregation score is a function that assigns a segregation score to every node in each combination of network and type vector:

$$S_i : \mathcal{X} \times \mathcal{T} \times \mathcal{N} \mapsto \mathbb{R}. \tag{2.11}$$

The notation is summarized in Table 1.

3. Some properties for segregation measures

For a systematic comparison of various measures, it is useful to establish a common benchmark, or a frame of reference, to allow the positioning of the different measures. One possibility is to establish such a benchmark based on empirical data. Specifically, the measures can be applied to the sets of data and the between-measures correlations can be examined. This kind of approach was taken by Fagiolo et al. (2007). Another possibility is to use a "theoretical" benchmark by formulating a set of properties that capture different aspects of the network structure that are relevant in the context of segregation. The latter possibility is also the starting point of an axiomatic approach, as described in Section 1. Each measure can then be evaluated by stating the properties satisfied and violated. In this section, we propose such a set of basic properties.

Obviously, there is a certain arbitrariness to our choice of properties below. Why these properties and not others? In our analysis, it is important to justify each property selected and clarify the specific role that each property plays. In particular, whereas these properties serve as useful reference points for evaluating various segregation measures, we do not intend to make claims about normativity for any one of the properties. In other words, we will not argue about the properties that an "ideal" segregation measure *should* satisfy. We believe that such ideals are specific to the particular question at hand. For example, certain details of a network formation process that brings about the segregation, or other types of phenomena affected by the segregation.

Instead, the properties that we define in our analysis serve to highlight *differences* between the measures in terms of their application and interpretation in research. Thus, we focus on properties that strongly differentiate between the measures and do not discuss those on which the measures do not strongly differ, even if those properties would be considered crucial for any measure of segregation. In particular, all the measures discussed here satisfy some general properties that can be consider "fundamental" for any measure of segregation in networks. Given a network with nodes assigned to groups, the value of the measures will not decrease when ties are added *within* groups and will not increase when ties are added *between* groups.[4]

In the context of this paper, the properties serve as a tool for evaluating the instruments for measuring segregation. The selected properties

1. capture substantive intuitions related to the concepts of segregation or homophily in social networks and
2. are expected to differentiate between the various existing measures of segregation (i.e., satisfied by some measures by not by others).

3.1. *Null models*

The first set of properties refers to whether a measure *embeds* a *null model*, that is, whether a measure is defined in such a way that the types of situations for which the measure would return the value 0 (i.e., no segregation) are clearly specified. As we will see, not all measures clearly specify the conditions under which they identify no segregation, and if they do, those conditions differ significantly between measures. This leads to important differences in interpretation.

Specification of network null models usually follows from certain hypotheses about network structure and its relation to the distribution of the node-level attribute in question. Such specification can consist of the following two mutually non-exclusive elements:

1. Formulating a set of constraints on the network structure that are implied by the hypothesis.
2. Formulating a probability distribution over all possible networks assuming that the hypothesis is true.

For example a simple null model is a random network with a given size and density. The constraints, as in (1), consist of fixing the size and density at specified values. Formulating the probability distribution, as in (2), would require supplementing the constraints with specification of the sampling mechanism, e.g., independent sampling of ties. In some cases, most notably most of non-trivial ERGMs, specifying the constraints (1) is infeasible, but it is still possible to simulate draws and approximate the probability distribution (2).

[4] There are, however some special cases: it can be shown that not all measures respond monotonically to the addition of ties within or between groups. Those results are not presented here but are available on request.

With respect to segregation measures and embedding, if we consider network model as a set of constraints imposed on the network structure, a segregation measure embeds a given model as null model if satisfying the associated constraints is a sufficient condition for the measure to return the value of 0. At the same time, if we consider the model as a probability distribution, then a segregation measure embeds the model as a null model if the expected value of the segregation index using that distribution is 0. In this article we follow the approach (1). The models formulated below apply to directed networks. However, it should be noted that this approach is easily extended to undirected networks as well.

The simplest model corresponds to a random network with a fixed density.

Definition 1 *(Random network model: RN).* Under the random network model conditional probability of tie existence is independent of group membership of ego and alter. This is equivalent to the requirement that odds for tie existence for all combinations of group memberships of ego and alter are equal to one another:

$$\frac{m_{111}}{m_{110}} = \frac{m_{gh1}}{m_{gh0}} \tag{3.1}$$

for all $g = 2, \ldots, K$ and $h = 2, \ldots, K$.

This model is equivalent to the Erdös and Rényi (1959) random graph model $G(n, M)$.

One of the implications of the RN model is that all nodes have the same expected degree. One can relax this assumption by allowing groups to have different expected degrees. We call the associated model the "marginal effects" model, defined as:

Definition 2 *(Marginal effects network model: ME).* Under the marginal effects network model conditional tie probability depends on group membership of ego and alter, but there is no interaction effect. This model is equivalent to the requirement that all conditional odds ratios are equal to one another, e.g., conditioning on the *y* margin:

$$\frac{m_{111}m_{gh1}}{m_{g11}m_{1h1}} = \frac{m_{11y}m_{ghy}}{m_{g1y}m_{1hy}} \tag{3.2}$$

for all $g = 2, \ldots, K$, $h = 2, \ldots, K$, and $y = 0, \ldots, 1$. Analogous formulas hold if we condition on g or h margin.

Another important model focuses only on the contact layer of the mixing matrix and corresponds to a situation of *proportionate mixing* (Nold, 1980):

Definition 3 *(Proportionate mixing network model: PM).* Under the proportionate mixing network model group memberships of ego and alter in all connected dyads are stochastically independent. This translates to the requirement that all odds-ratios in the contact layer of the mixing matrix are equal to 1:

$$\frac{m_{111}m_{gh1}}{m_{g11}m_{1h1}} = 1 \tag{3.3}$$

for all $g = 2, \ldots, K$, $h = 2, \ldots, K$.

A related model extends this restriction to also hold in the non-contact layer of the mixing matrix.

Definition 4 *(Conditional independence model: CI).* Under the conditional independence network model group memberships of ego and alter are conditionally independent given the status of the dyad (connected or disconnected). This requirement is equivalent to the restriction that all odds-ratios in the contact and non-contact layer of the mixing matrix are equal to 1:

$$\frac{m_{111}m_{gh1}}{m_{g11}m_{1h1}} = \frac{m_{110}m_{gh0}}{m_{g10}m_{1h0}} = 1 \tag{3.4}$$

for all $g = 2, \ldots, K$, $h = 2, \ldots, K$.

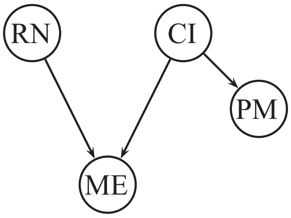


Fig. 1. Relations between the models defined in Section 3.1. Two models are connected with an arrow if the restrictions implied by the sender model imply the restrictions of the receiver model. Equivalently, the sender model is nested in the receiver model.

The models proposed above are related. First, the restrictions defining the RN model include the restrictions defining the ME model: equality of conditional odds as specified in (3.1) implies the equality of conditional odds-ratios as specified in (3.2). In that sense we can say that the RN model is nested in the ME model. Second, the restrictions defining the CI model include the restrictions of the ME model: CI model assumes, as ME does, the equality of conditional odds-ratios. Consequently, the CI model is also nested in the ME model. Third, the restrictions defining the CI model include the restrictions of the PM model: both models assume independence in the contact layer of the mixing matrix, but CI also assumes independence in the non-contact layer. The CI model is nested in the PM model. Fourth, although the PM model and the CI model are formally distinct, they are approximately equivalent for practical purposes when applied to empirical data (cf. Koehly et al., 2004). The relations between the models that we describe above are elements of a fuller hierarchy of models that could be fit to a three-dimensional mixing matrix. We present that hierarchy in Appendix A (Fig. 1).

The above relations between the models have consequences for the behavior of segregation measures. If a certain segregation measure is always 0 for networks consistent with the ME model then it will be also 0 for networks consistent with the models nested in the ME model (so RN and CI models). However, the converse is not generally true, e.g., there might be measures that always return 0 for networks consistent with RN model but return a value different from 0 for networks consistent with ME. When discussing segregation measures, we are interested to find the least restrictive null model, so, as in the previous paragraph, if a measure assumes value of 0 for both ME model and RN model, we would conclude that it embeds the ME model. When reviewing the measures we also discuss the implication in the other direction, namely, whether the fact that a measure assumes value of 0 implies that the network is consistent with a particular null model.

3.2. Other properties

Consider the way in which the number of isolates in a network affects segregation. On the one hand, one could argue that disconnected actors in a network should not play any role in segregation as they do not contribute any "relational" information. Because disconnected actors do not have any ties to anyone, it is impossible to state the extent to which they lead to segregation. However, on the other hand, disconnected actors may represent *opportunities* for creating ties. For example, one could argue that adding isolated actors from minority groups creates many opportunities for integration in the form of between-group ties. As such, the sensitivity of a measure to isolated nodes is especially relevant in a research context in which one studies networks among *growing* populations. More generally, as will become clear in the analysis, the sensitivity of a measure to isolated nodes is indicative for the extent to which it is sensitive to changes in opportunities for creating ties, and by extension, for its applicability to dynamic social networks.

We convey these considerations in the following property:

Property 1 (Insensitivity to adding isolates: ISO). *Let network X be defined on N nodes with an associated type vector* $\mathbf{t}$*. Construct a network Y defined for N + 1 nodes with an associated type vector* $\mathbf{u}$ *by adding an isolate to X. Consequently:*

1. *Networks X and Y are identical for all the nodes other than* $(N+1)$*th:* $\forall p, q \in \mathcal{N}\, y_{pq} = x_{pq}$.
2. *The* $(N+1)$*th node does not have any links in network Y:* $\sum_{p=1}^{N} y_{p\ N+1} = \sum_{q=1}^{N} y_{N+1\ q} = 0$.
3. *Group membership of all the nodes is identical in both networks:* $\forall k \in \mathcal{N} \quad t_k = u_k$.

Network segregation index S is insensitive to isolates if and only if

$S(X, \mathbf{t}) = S(X, \mathbf{u})$.

In other words, adding isolates to the network does not affect the segregation level.

As we will see in the following sections, some of the measures will not satisfy this property in various ways. These departures are investigated measure-by-measure in Section 4 and summarized in Section 6.

The final property refers to the network as a whole. Often, social networks consist of two or more *components* or disconnected parts. In such cases, each component can be perceived as a subnetwork and characterized by a separate segregation score while maintaining a focus on the network as a whole requires a network-level segregation score. The following property relates to how a network measure behaves if several components are studied as a single network.

Property 2 (Symmetry: SYM). *Define two identical networks X and Y and some type vector* $\mathbf{t}$*. Network segregation index S satisfies symmetry if and only if*

$S(X, \mathbf{t}) = S(Y, \mathbf{t}) = S(Z, \mathbf{z})$,

where the network Z is constructed by considering X and Y together as a single network, namely: $Z = [z_{pq}]_{2N \times 2N}$ *such that*

- $\forall p, q \in \{1, \ldots, N\} \quad z_{pq} = x_{pq}$,
- $\forall p, q \in \{N+1, \ldots, 2N\} \quad z_{pq} = y_{pq}$,
- *otherwise* $z_{pq} = 0$.

This property resembles the *population principle* in the axiomatizations of social inequality measures (Foster, 1983). The population principle postulates that social inequality should be identical when replicating a population under study, and allows for a "per capita" interpretation of inequality that is easily comparable populations. For measures of segregation in social networks, this property is relevant, for example, when one studies networks that are subsets of a larger, population, say, classes within a school. The symmetry property implies that if two classes in the school have the same measured level of segregation, the measured level of segregation of the two classes combined would be the same. Again, whether this is desirable or not depends on the research context and research questions. Our goal here is to clarify how different measures behave in this regard.

Parallel to the properties specified above, it is also important to determine the level on which a measure assigns the segregation scores, as we discussed in our notation section. Some measures provide only a group-level score. Others may specify an additional rule by which group-level segregation scores can be aggregated to produce network-level score. Other segregation measures can be conveniently aggregated or disaggregated across all three levels (node, group, or network). Thus, we will discuss the level on which a measure can be applied as an additional dimension that differentiates between the segregation indices.

4. Existing approaches

In this section we review a set of prominent measures and approaches to measuring segregation in social networks and examine the extent to which these approaches satisfy the properties specified in the preceding section. The measures that are applicable to both directed and undirected networks are first examined, followed by an examination of specialized indices.

4.1. The E-I index

A particularly simple approach to quantifying segregation in networks is the *E-I index*, which was originally proposed by Krackhardt and Stern (1988) and is implemented as a measure for homophily in the popular social network analysis package UCINET (Borgatti et al., 2002). Despite the fact that it was meant to measure homophily, its directionality is opposite to that of most of the other measures discussed here. That is: low values indicate relatively little connections between groups, while high values indicate relatively many connections between groups. The E-I index is defined simply as the difference between between-group ties and within-group ties, divided by the total number of ties for normalization. That is:

$$S_{\mathrm{EI}} = \frac{\sum_{g=1}^{K} \sum_{h \neq g} m_{gh1} - \sum_{g=1}^{K} m_{gg1}}{\sum_{g=1}^{K} \sum_{h=1}^{K} m_{gh1}} \tag{4.1}$$

This measure takes the value +1 if all ties in the network are between groups, and −1 if all ties are within groups.

Because the directionality of S_{EI} is opposite to that of all the other measures discussed here, we apply our analysis to $-S_{\mathrm{EI}}$ rather than to S_{EI} to keep our terminology consistent.

To verify whether S_{EI} embeds one of the null-models we defined above, we check whether any of the restrictions specified in Section 3.1 imply that $S_{\mathrm{EI}} = 0$. By definition, $S_{\mathrm{EI}} = 0$ if the number of between-group ties equals the number of within-group ties. It is easily verified that the random network model (Eq. (3.1)) does not imply that $S_{\mathrm{EI}} = 0$. Consider an undirected network with two groups. The number of expected within-group ties under the random network model is proportional to $(n_1(n_1-1))/2 + (n_2(n_2-1))/2$, while the expected number of between-group ties is proportional to $n_1 n_2$. While these quantities will converge with increasing group size if $n_1 = n_2$, they are not generally equal. In particular, in the case of unequal group sizes ($n_1 \neq n_2$), the number of within- and between-group ties will be unequal. Hence, S_{EI} does not embed the random network model. This conclusion is not altered if we allow for the groups to have different degrees, that is, under the marginal effects network model.

It is also easily verified that the restriction of the proportionate mixing network model (Eq. (3.3)), namely that all odds ratios in the contact layer of the mixing matrix are equal to 1, does not imply $S_{\mathrm{EI}} = 0$. To see this, simply consider the situation for two groups, where it is immediately clear that $(m_{111} + m_{221}) = (m_{121} + m_{211})$, that is, $S_{\mathrm{EI}} = 0$, does not imply that $(m_{111} m_{221}) = (m_{121} m_{211})$, that is, the odds ratio equals 1, and vice versa. The same holds for the conditional independence model. Thus, we can conclude that S_{EI} does not embed any of the network models considered in this paper as a null model.

Conversely, we can also observe that because S_{EI} uses only information from the contact layer of the mixing matrix, the restrictions of the random network model (Eq. (3.1)), the marginal effects network model (Eq. (3.2)), and the conditional independence model

(Eq. (3.4)) are not necessarily satisfied if $S_{\text{EI}} = 0$, because these models also rely on the non-contact layer of the mixing matrix.

Because $-S_{\text{EI}}$ is defined exclusively in terms of numbers of ties, adding nodes without any ties will not change the measure. That is, $-S_{\text{EI}}$ is insensitive to adding isolates, and thus satisfies ISO.

To verify if $-S_{\text{EI}}$ satisfies the Symmetry property we consider how the measure would change if the network to which it is applied is duplicated. Notice that as the network is duplicated, both the numerator and the denominator of (4.1) are doubled, and hence S_{EI} does not change and therefore satisfies Symmetry. Another way to view this latter result is to note that S_{EI} does not take into account that as the network is duplicated, the number of possibilities for between-group ties increases disproportionately as compared to the possibility of within-group ties. Krackhardt and Stern (1988, p. 128) notice this issue when they write that

> In an organization of any reasonable size it would be difficult to find an E-I index of less than zero because [the potential number of between-group ties] outnumbers [the potential number of within-group ties] so strongly.

but apparently do not consider this to be problematic.

4.2. The Assortativity Coefficient

In the context of analyzing mixing patterns in networks of sexual contacts and marriage matching, Newman and colleagues (Newman, 2003; Newman and Girvan, 2002) proposed the *Assortativity Coefficient*, which is presented here using our notation.

The Assortativity Coefficient can be formulated using the mixing matrix M. It summarizes the contact layer by evaluating the relative "weight" of the diagonal; the more likely it is for actors to be connected within groups, the larger the numbers in the diagonal cells of the contact layer of the mixing matrix.

Given the mixing matrix M, a matrix of proportions is defined as $p_{gh} = m_{gh1}/m_{gh+}$. The Assortativity Coefficient is defined as:

$$S_{\text{Newman}} = \frac{\sum_{g=1}^{K} p_{gg} - \sum_{g=1}^{K} p_{g+}p_{+g}}{1 - \sum_{g=1}^{K} p_{g+}p_{+g}}. \tag{4.2}$$

S_{Newman} is, in fact, equivalent to *Cohen's Kappa* applied to the contact layer of the mixing matrix (Cohen, 1960), although this does seem to have gone unnoticed by Newman (2003). Cohen's Kappa is a classical psychometric measure of agreement on nominal variables (Reynolds, 1977, Section 2.7.1)

The index S_{Newman} reaches its maximum of 1 for "perfect assortative mixing" when all the ties are within-group and the diagonal entries sum up to 1. The minimum value of the index for "perfect disassortativity" depends on the relative number of ties in each group and is equal to

$$\min_{\mathcal{G}}(S_{\text{Newman}}) = \frac{-\sum_{g} p_{g+}p_{+g}}{1 - \sum_{g} p_{g+}p_{+g}}. \tag{4.3}$$

S_{Newman} does not necessarily take the value -1 for perfectly integrated networks. Newman (2003) defends this apparent asymmetry with the following argument:

> a perfectly disassortative network is normally closer to a randomly mixed network than is a perfectly assortative network. (. . .) random mixing will most often pair unlike vertices, so that the network appears to be mostly disassortative. Therefore, it is appropriate that the value $[S_{\text{Newman}}] = 0$ for the random network should be closer to that for the perfectly disassortative network than for the perfectly assortative one.

Because, like the E-I index discussed in the previous section, S_{Newman} uses only information from the contact layer of the mixing matrix, we can conclude that it does not embed the random network model, the marginal effects model, or the conditional independence model as a null model. The index assumes the value of 0 when $p_{gh} = p_{g+}p_{+h}$, that is, group memberships in connected dyads are stochastically independent, which is equivalent to proportionate mixing as specified in Eq. 3.3. Conversely, however, $S_{\text{Newman}}(X, t) = 0$ does not imply that X is consistent with proportionate mixing. The reason is that S_{Newman} only takes the values at the diagonal of the mixing matrix into account. Thus, for a mixing matrix larger than 2 × 2, even if the values on the diagonal are the expected values according to proportionate mixing, off-diagonal values may still deviate. In other words, we may say that while S_{Newman} successfully "detects" networks consistent with proportionate mixing, observing that $S_{\text{Newman}} = 0$ is no guarantee that the network is consistent with proportionate mixing for more than two groups.

Because the non-contact layer of the mixing matrix does not enter S_{Newman}, adding isolates to the network does not change the value of the index. Thus, the property of ISO is satisfied. Lastly, the distribution p_{gh} does not change when the number of existing ties and nodes is duplicated (the relative frequencies of ties linking different groups stay the same), thus satisfying the property of SYM.

4.3. Gupta, Anderson, and May's Q

Gupta et al. (1989) analyzed the effects of mixing patterns of sexual contacts on the spread of the HIV epidemic. The measure of "within-group mixing" in the population is designed for undirected networks and based on the contact layer of the mixing matrix. Define f_{gh} as the proportion of ties of actors in group g to actors in group h:

$$f_{gh} = \frac{m_{gh1}}{m_{g+1}}. \tag{4.4}$$

The proposed index (denoted by Q in Gupta et al., 1989, Eq. (8)), is defined as:

$$S_{\text{GAM}} = \frac{\sum_{g=1}^{K} \lambda_g - 1}{K - 1} = \frac{\sum_{g=1}^{K} f_{gg} - 1}{K - 1}, \tag{4.5}$$

where λ_g are the eigenvalues of the matrix $[f_{gh}]$ and f_{gg} are the diagonal entries of the matrix.[5]

The measure captures "assortativeness" by varying between $-1/(K-1)$ for the maximal dissassortativity (integration) and 1 for maximal assortativity (segregation). The measure yields a value of zero if the sum of diagonal entries of the f_{gh} matrix is equal to 1. A sufficient condition for that to happen is when the conditional probabilities that both connected actors belong to the same group g given that at least one of them belongs to g are proportional to the overall activity level of actors in group g (i.e., proportion of ties in which at least one actor from g participates). This requirement is equivalent to the PM model from Section 3.1. However, the converse is not generally true: $S_{\text{GAM}} = 0$ is not a sufficient condition for the studied network to be consistent with the PM model.

Because the index is based on the contact layer of the mixing matrix, it is insensitive to the number isolates. In this way, ISO is also satisfied.

Finally, the property SYM is satisfied because duplicating the network does not affect the contact layer of the mixing matrix.

[5] For some reason Gupta et al. (1989) failed to recognize that the sum of the eigenvalues of a square matrix of real or complex numbers is equal to its trace, i.e., the sum of diagonal entries (Harville, 1997, Ch. 21.6, Eq. (6.2)).

4.4. *Odds-ratio for within-group ties*

We noted already in Section 3 that ignoring opportunities for creating ties may be undesirable under certain circumstances. That is, it is possible to examine the proportion of existing between-group ties to the number of all possible dyads instead of focusing only on the existing ties (connected dyads). A simple approach based on the research on occupational segregation (Charles and Grusky, 1995), was employed by Moody (2001). The approach is to calculate the odds ratio for tie existence versus non-existence for within-group dyads and between-group dyads. In this paper, we call it the Odds-Ratio for Within-Group ties (ORWG). To calculate ORWG, one can use the information from the mixing matrix *M*. The odds ratio is equal to:

$$\begin{aligned} S_{\mathrm{ORWG}} &= \frac{\sum_{g=1}^{K} m_{gg1} \Big/ \sum_{g=1}^{K} m_{gg0}}{\sum_{g=1}^{K} \sum_{h \neq g} m_{gh1} \Big/ \sum_{g=1}^{K} \sum_{h \neq g} m_{gh0}} \\ &= \frac{\sum_{g=1}^{K} m_{gg1} \sum_{g=1}^{K} \sum_{h \neq g} m_{gh0}}{\sum_{g=1}^{K} m_{gg0} \sum_{g=1}^{K} \sum_{h \neq g} m_{gh1}}. \end{aligned} \qquad (4.6)$$

If S_{ORWG} equals 1, we would conclude that between- and within-group ties are equally likely in the analyzed network when group sizes are taken into account, therefore there is no segregation. The more likely it is to observe within-group ties, the closer the value of the index approaches infinity. Thus, larger values indicate higher segregation levels. Conversely, in the case of integration (as opposed to segregation), the more likely it is to observe between-group ties, the closer the value of the index approaches 0. Taking a logarithm of this index makes the values distribute symmetrically around zero and vary between plus and minus infinity.

A sufficient condition for S_{ORWG} to be equal to 1 (or to 0 on the log scale) is when the network is consistent with the RN model. However, $S_{\mathrm{ORWG}} = 0$ is not a guarantee that the studied network is consistent with any network models described in Section 3.1.

Adding isolates changes the opportunities for creating ties. In the context of S_{ORWG} adding isolates affects the values in the non-contact layer of the mixing matrix m_{gh0}. With a network of size *N* with two groups of sizes n_1 and n_2, we have

$$S_{\mathrm{ORWG}} = \frac{(m_{111} + m_{221}) m_{120}}{(m_{110} + m_{220}) m_{121}}, \qquad (4.7)$$

where the number of all dyads within group 1 is $m_{110} + m_{111} = n_1(n_1 - 1)/2 \approx n_1^2/2$, and the number of all dyads between groups 1 and 2 is $m_{120} - m_{121} = (n_1 n_2)/2$. The value of S_{ORWG} is proportional to the ratio $m_{120}/(m_{110} + m_{220})$. If expressed in terms of the group sizes this ratio is equal to $(n_1 n_2)/(n_1^2 + n_2^2)$. The value of this ratio increases with n_1 when $n_1 < n_2$ and decreases when $n_1 > n_2$. Therefore, adding isolates to the minority group *increases* segregation, which means that the property ISO is not satisfied.

Duplicating the network affects all the components of the measure. Denoting the mixing matrix resulting from duplication as m' we get:

$$\begin{aligned} m'_{111} &= 2m_{111} & m'_{221} &= 2m_{221} \\ m'_{121} &= 2m_{121} & m'_{120} &\approx \frac{1}{2} 2n_1 2n_2 = 2n_1 n_2 \\ m'_{110} &\approx \frac{1}{2}(2n_1)^2 = 2n_1^2 & m'_{220} &\approx \frac{1}{2}(2n_2)^2 = 2n_2^2 \end{aligned}$$

Now, if we compute the index, we obtain:

$$\begin{aligned} S'_{\mathrm{ORWG}} &= \frac{2(m_{111} + m_{221}) 2n_1 \cdot 2n_2}{(2n_1)^2 + (2n_2)^2 \cdot 2m_{121}} = \frac{2(m_{111} + m_{221}) \cdot 4n_1 n_2}{4(n_1^2 + n_2^2) \cdot 2m_{121}} \\ &= \frac{(m_{111} + m_{221}) \cdot n_1 n_2}{(n_1^2 + n_2^2) \cdot m_{121}} = S_{\mathrm{ORWG}}. \end{aligned} \qquad (4.8)$$

Therefore, the value does not change, satisfying the property of SYM.

4.5. *ERGM and other log-linear models for networks*

Another way of capturing dependence between tie existence and nodal attributes is offered by a log-linear approach to network modeling. These models include the Exponential Random Graph Models (ERGM, for example, Snijders et al., 2006) and other conditional (Morris, 1991; Koehly et al., 2004) and unconditional (Fienberg and Wasserman, 1981) log-linear models. These families of models offer much flexibility in terms of specification. Here, we will focus on the models that capture the effect of nodal attributes on the probability of the network tie existence. In the rest of the discussion, we will further assume *conditional tie independence* (Frank, 1988), which states that the probabilities of network ties are independent given the attributes of the nodes. One of the implications of conditional tie independence is that all the nodes with the same attributes are assumed to be homogeneous (exchangeable).

Given the arguments above, it is sufficient to consider the network in the form of a three-dimensional mixing matrix $M = [m_{ghy}]_{K \times K \times 2}$ as defined in Section 2. Two types of models are considered:

1. Conditional Log-Linear Models for the contact layer of the mixing matrix (m_{gh1}).
2. Logit models for the full mixing matrix, which are special cases of ERGM.

4.5.1. *Conditional Log Linear Models*

A general log-linear model for a two-dimensional contact layer of the mixing matrix models the logarithm of quantities m_{gh1} as a linear function of marginal and interaction effects.[6] We will consider the following models, taken from Koehly et al. (2004):

$$\log m_{gh1} = \mu + \lambda_g^A + \lambda_h^B + \lambda_{gh}^{UHOM} \begin{cases} \lambda_{gh}^{UHOM} = \lambda^{UHOM} & g = h \\ \lambda_{gh}^{UHOM} = 0 & g \neq h \end{cases} \qquad (4.9)$$

$$\log m_{gh1} = \mu + \lambda_g^A + \lambda_h^B + \lambda_{gh}^{DHOM} \begin{cases} \lambda_{gh}^{DHOM} = \lambda_g^{DHOM} & g = h \\ \lambda_{gh}^{DHOM} = 0 & g \neq h \end{cases} \qquad (4.10)$$

$$\log m_{gh1} = \mu + \lambda_g^A + \lambda_h^B + \lambda_{gh}^{AB} \qquad (4.11)$$

where *UHOM* and *DHOM* stand for *uniform* and *differential* homophily effects. The main effects λ_g^A and λ_h^B capture the tendency for the groups to initiate and accept ties. The interaction effects λ_{gh}^{AB}, λ_g^{DHOM}, and λ^{UHOM} are of our primary concern given that they measure the degree of over- and under-representation of certain types of ties compared to the independence model, which contains only the main effects. For these models to be identified, additional

[6] For a general introduction to log-linear models see, Goodman (1978, 1996), Agresti (2002).

restrictions are placed on λs, such that their appropriate sums are 0:

$$\sum_g \lambda_g^A = \sum_h \lambda_h^B = \sum_g \sum_h \lambda_{gh}^{AB} = \sum_g \sum_h \lambda_{gh}^{UHOM} = \sum_g \sum_h \lambda_{gh}^{DHOM} = 0 \qquad (4.12)$$

Alternatively, the values of λs for one of the levels of the variables is fixed at 0 by setting it as a reference category (Koehly et al., 2004; Agresti, 2002).

Model (4.11) is a saturated model in which the interaction terms λ_{gh}^{AB} capture all possible deviations from the independence model, reproducing the observed matrix. Models (4.9) and (4.10) impose additional restrictions on the interaction terms to allow the measurement of homophily. Model (4.9) is the *uniform homophily model* that distinguishes only between within- and between-group ties. The parameter λ^{UHOM} measures the extent to which the ties connect actors within the same group rather than actors from different groups. These deviations are assumed to be the same for all groups, i.e., the model assumes that all groups manifest the same degree of homophily. Model (4.10) relaxes this last assumption and is called the *differential homophily model*. In this model, the groups can be characterized with group-specific homophily effects as captured by the parameters λ_g^{DHOM}. We will treat λ^{UHOM} and λ_g^{DHOM} as measures of segregation comparable to other measures discussed in this paper. Both measures vary between plus and minus infinity and take a value of zero whenever the independence model holds. The independence situation corresponds to "proportional mixing" (PM model in Section 3.1), in which the relative numbers of ties between the groups are proportional to the group activity levels.

Adding isolates to the network does not affect λ^{UHOM} as it does not modify the entries in the contact layer of the mixing matrix, thus satisfying ISO. Finally, the measure satisfies SYM because the merging identical networks keeps the mixing matrix of proportions constant. Doubling the frequency counts is absorbed by the constant μ of the log-linear model.

Parameters λ_g^{DHOM} in the differential homophily model can capture in-breeding tendencies for the different groups separately and serve as network segregation measures on the group level. The properties ISO and SYM are both satisfied for the same reasons as for λ^{UHOM}.

It is important to note that for networks with only two groups, the differential homophily model is not identified for undirected network, and for directed networks it is a re-parametrization of the saturated model (4.11) (also refer to the last paragraph of this section).

4.5.2. Exponential Random Graph Models

The log-linear models described above can be perceived as models for conditional probabilities $P(i \in G_1 \wedge j \in G_2 | x_{ij} = 1)$, that is, the probability that connected individuals belong to groups G_1 and G_2. Alternatively, one could consider modeling the conditional probabilities of tie existence given group membership, i.e., $P(x_{ij} = 1 | i \in G_1 \wedge j \in G_2)$. In logit form for the mixing matrix the models take the form:

$$\log \left(\frac{m_{gh1}}{m_{gh0}} \right) = \alpha + \beta_g^A + \beta_h^B + \beta_{gh}^{AB}. \qquad (4.13)$$

This model is a special case of ERGM and is limited to effects related to actor attributes, namely, the main effects β_g^A and β_h^B and the interaction effects β_{gh}^{AB}. Constraints similar to those presented in (4.12) apply to this model as well. The interaction effects β_{gh}^{AB} in this model are in the form of log odds ratios, with the odds for tie existence compared depending on the group membership of ego and alter. For a more complete overview of Exponential Random Graph Models, consult the rich and growing literature that includes Holland and Leinhardt (1981), Frank and Strauss (1986), Wasserman and Pattison (1996), Robins et al. (2001a,b) and Snijders et al. (2006).

As in the previous section, we will consider two restricted versions of the model (4.13):

$$\log \left(\frac{m_{gh1}}{m_{gh0}} \right) = \alpha + \beta_g^A + \beta_h^B + \beta_{gh}^{UHOM} \quad \begin{cases} \beta_{gh}^{UHOM} = \beta^{UHOM} & g = h \\ \beta_{gh}^{UHOM} = 0 & g \neq h \end{cases} \qquad (4.14)$$

$$\log \left(\frac{m_{gh1}}{m_{gh0}} \right) = \mu + \beta_g^A + \beta_h^B + \beta_{gh}^{DHOM} \quad \begin{cases} \beta_{gh}^{DHOM} = \beta_g^{DHOM} & g = h \\ \beta_{gh}^{DHOM} = 0 & g \neq h \end{cases} \qquad (4.15)$$

Model (4.14) is a model of *uniform homophily* as the parameter β^{UHOM} measures the tendency for ties to be formed within groups, assuming that this tendency is the same for all the groups. Model (4.15) relaxes the assumption and allows for group-specific in-breeding levels as captured by β_g^{DHOM}. If the parameters β^{UHOM} and β_g^{DHOM} are equal to 0 in Eqs. (4.14) and (4.15), the models reduce to marginal effects model, which is equivalent to ME model in Section 3.1.

Parameter β^{UHOM} is a log odds ratio measuring the relative likelihood for network ties to exist in dyads between nodes that belong to the same group rather than to different groups. Segregation as measured by β^{UHOM} is ceteris paribus differential "popularity" and "sociality" of the groups, which are measured by the main effects β_g^A and β_h^B. It is worth noting that if both β_g^A and β_h^B parameters are simultaneously 0, then β^{UHOM} is equal to S_{ORWG}. Because of that, as in the case of S_{ORWG}, the property of ISO is not satisfied whereas the property of SYM is satisfied for β^{UHOM}, and this also holds for the differential homophily effects β_g^{DHOM}.

4.5.3. CLLM versus ERGM: a brief comparison

Conditional Log Linear Models and Exponential Random Graph Models are closely related (Koehly et al., 2004). CLLMs model the contact layer of the mixing matrix, accounting for the joint probability of group memberships of ego and alter nodes that are conditional on the existence of the tie: $P(i \in G_1 \wedge j \in G_2 | X_{ij} = 1)$. ERGMs model the conditional probability of tie existence given the group memberships of the participating nodes: $P(X_{ij} = 1 | i \in G_1 \wedge j \in G_2)$. These two probabilities are related through the Bayes' Rule:

$$P(i \in G_1 \wedge j \in G_2 | X_{ij} = 1) = \frac{P(X_{ij} = 1 | i \in G_1 \wedge j \in G_2) \cdot P(i \in G_1 \wedge j \in G_2)}{P(X_{ij} = 1)} \qquad (4.16)$$

We refer the reader to the original paper by Koehly et al. (2004) for details and implications (see also Robins et al., 2001a,b).

As a final remark to close the section on both CLLMs and ERGMs, it is important to note the following fact. For both CLLMs and ERGMs, when a network has only two groups, the independence models result from setting the interaction terms in the saturated model to 0 have only one degree of freedom. In this way, the uniform homophily models are simply a re-parametrization of the saturated model, with both models producing the same fitted values for the mixing matrix. The differential homophily model implies parameterizing the interaction term with the number of parameters equal to the number of groups, and is thus not identified in the case of two groups.

4.6. Freeman's segregation index

In its original formulation, Freeman's segregation measure (Freeman, 1978b) applies to undirected networks defined for *two* groups. 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. Freeman characterized segregation as follows:

> (...) segregation could be thought of as restriction on social network ties between members of two distinguishable "kinds" of people. Thus, segregation was seen as a systematic – as opposed to random – social arrangement that reflects limitations on the access of different classes of people to one another. (Freeman, 1978a)

Formally, we have an undirected network X consisting of two groups of nodes G_1 and G_2. The observed proportion of between-group ties is equal to:[7]

$$p = \frac{m_{121}}{m_{++1}}. \tag{4.17}$$

The expected proportion of between-group ties in the random graph is given by

$$\pi = \frac{m_{12+}}{m_{+++}} = \frac{2n_1n_2}{N(N-1)}. \tag{4.18}$$

where n_1 and n_2 are the sizes of groups G_1 and G_2 respectively. Given these two quantities Freeman's segregation index is equal to

$$\begin{aligned} S_{\text{Freeman}} &= \frac{\pi - p}{\pi} = 1 - \frac{p}{\pi} = 1 - \frac{2N(N-1)m_{121}}{n_1n_2m_{++1}} \\ &= 1 - \underbrace{\frac{m_{121}}{n_1n_2}}_{(1)} \times \left(\underbrace{\frac{2m_{++1}}{N(N-1)}}_{(2)} \right)^{-1}. \end{aligned} \tag{4.19}$$

It is worth noting that the two highlighted terms in the last transformation can be substantively interpreted. The first term is equivalent to the "density" of between-group ties: the proportion of existing between-group ties out of all possible between group ties. The second term is the density of the whole graph.

Before we analyze this index in more detail from the perspective of the proposed properties it is worth making one observation. The index as defined in (4.19) can take both positive and negative values. The negative values correspond to the networks for which the proportion of between-group ties is higher than would have been expected by chance. Freeman originally proposed to truncate the index at 0 by assuming $S_{\text{Freeman}} = 0$ if $p > \pi$.

This truncation deficiency of Freeman's index was criticized by Mitchell (1978). In a response Freeman (1978a) proposed an alternative measure for *integration* (the opposite of segregation). With a similar structure as the segregation index, the integration measure captures the features of networks in which the number of between-group ties is larger than what would have been expected if tie formation were random. For this type of networks the original segregation index is assigns a value of 0. Freeman proposed an index of *integration* for $p > \pi$ of the form

$$S_{\text{FreemanI}} = \frac{p - \pi}{p_{\max} - p}, \tag{4.20}$$

where π and p refer to the expected and observed proportion of between-group ties, and $p_{\max}$ is the maximal proportion of between-group ties.

The measure varies between 0 and 1, taking the value 1 whenever all the ties that exist in the given network are between-group (perfect integration) and taking the value 0 whenever the proportion of between-group ties is equal to or less than the proportion expected in the case of random tie formation. This measure suffers from the same problems as the original segregation index. Realizing this shortcoming, Freeman advocates the use of both segregation and integration measures together until a unified solution is found.

Now we examine the performance of Freeman's original segregation index with respect to the proposed properties.

With respect to the models defined in Section 3.1, a sufficient condition for S_{Freeman} to be equal to zero is when the network is random, i.e., corresponds to the RN model.

Turning to ISO, adding an isolate to the network affects the opportunities for making ties (π) only, with the value of p remaining constant. Intuitively, the effect of adding the isolate depends on relative group sizes. The segregation would increase or decrease depending on whether the isolate that is being added belongs to the majority or minority group. For two groups of sizes n_1 and n_2 adding an isolate from group 1 will *decrease* segregation as long as $n_1 < n_2 - 1$. In practice, this means that adding isolates belonging to the majority group *decreases* segregation, whereas adding isolates belonging to the minority group *increases* it.[8] The ISO property is not satisfied.

Freeman's index does not satisfy SYM either. When merging two identical networks the value of p (4.17) stays the same but the value of π decreases. Consequently, the value of the index also *decreases*.[9]

Although Freeman's original idea for the measure was limited to only two groups it is possible to extend it to an arbitrary number of groups (e.g., K). When generalized for use with more groups, the formula for the observed number of between-group ties (4.17) stays almost identical:

$$p = \frac{\sum_{g,h:g \neq h} m_{gh1}}{m_{++1}}. \tag{4.21}$$

Formula (4.18) requires a slightly more substantial modification for the expected number of cross-group ties π. The numerator of π can be stated as a sum of products of group sizes leading to

$$\pi = \frac{2\sum_{k=1}^{K-1}\sum_{l=k+1}^{K} n_k n_l}{N(N-1)}, \tag{4.22}$$

which, upon algebraic procedures, can be represented as a function of the difference between the squared sum and the sum of squares of the group sizes:

$$\pi = \frac{\left(\sum_{k=1}^{K} n_k\right)^2 - \sum_{k=1}^{K} n_k^2}{N(N-1)}, \tag{4.23}$$

[7] Assuming that the entries in the lower triangle of the mixing matrix for undirected networks are all equal to 0.

[8] Appendix B contains additional details.

[9] As long as there are fewer between-group ties than expected by chance, the truncation to 0 does not apply.

which leads to a *generalized Freeman's index* equal to:

$$S_{\text{Freeman}} = 1 - \frac{pN(N-1)}{\left(\sum_{k=1}^{K} n_k\right)^2 - \sum_{k=1}^{K} n_k^2}. \tag{4.24}$$

The full derivation is shown in Appendix B.

4.7. *The Spectral Segregation Index*

The Spectral Segregation Index (SSI) (Echenique and Fryer, 2007) was developed for measuring the extent of residential segregation, such as race segregation of neighborhoods in a city. It is also applicable in other contexts. In its original form, the network underlying the computation of SSI represents residential areas and their spatial proximities, so it is undirected. However, this can be substituted with other types of undirected, and possibly weighted, relations.

Although defined on the group level, the SSI can be easily decomposed to the node level giving segregation values for individual nodes. It can also be aggregated flexibly to provide segregation scores for network components and for the network as a whole.

The basis for this measure is a normalized adjacency matrix $R = [r_{ij}]_{N\times N}$, which is formed from the original network by normalizing the rows so that they sum up to 1. This creates the possibility for applying SSI in contexts in which network ties have a certain weight or value attribute. In the case of individuals embedded in a social network, this attribute might be a time constraint, for example, the value of r_{ij} could be the proportion of time that i spends with j. For spatially located neighborhoods r_{ij} could be the ratio of the length of the border between i and j to the total circumference of the neighborhood i. In case of binary adjacency matrices, all the entries are simply equal to $1/\eta_i$. Additionally, for every group G_g, the measure defines a matrix B_g which is a sub-matrix of R that contains only the nodes belonging to group G_g. In other words, the matrix B_g contains only within-group interactions for group G_g. Echenique and Fryer (2007) define the SSI axiomatically. Here we discuss the most important characteristics of SSI and refer the reader to the original publication for detailed discussion.

A distinctive feature of the SSI is that it can be decomposed to node-level values. Thus, in this context, one can speak of segregation levels of individual nodes. The amount of segregation of higher-level entities such as components or the whole network can be calculated by averaging of the values on the node level. The basic assumption is that the individual level segregation of node i belonging to group G_g, s_i^g should be equal to the average segregation levels of same-group neighbors of i. Formally,

$$s_i^g = \frac{1}{S^g_{C_i}} \sum_j r_{ij} s_j^g \tag{4.25}$$

where $S^g_{C_i}$ is the average level of segregation in a connected component C_i of within-group interactions specified by B to which i belongs (see Linearity axiom in Echenique and Fryer, 2007).

At the level of connected components of within-group interactions (i.e., connected components of B matrices defined above), SSI is equal to the largest eigenvalue of that matrix. Individual-level SSIs are calculated by distributing the component level value across individuals in proportion to the values in the corresponding eigenvector. As an example, take an actor i who is a member of group G_1 in the (normalized) network R. Create a sub-matrix B by selecting only the actors that belong to group G_1, the group of i. Then, extract from B a sub-matrix C_i corresponding to the connected component of B to which actor i belongs. These are all the nodes belonging to group G_1 which are reachable from i by traversing only within-group ties. The value of SSI for that component is equal to the largest eigenvalue λ of the matrix C_i. The level of individual-level segregation is calculated by distributing the value of λ using the corresponding eigenvector l

$$S_{\text{SSI}}(i) = \frac{l_i}{\bar{l}}\lambda, \tag{4.26}$$

where $\bar{l}$ is the mean of the values in the eigenvector l.

Even though a single closed-form formula for the SSI is not available, it is still possible to infer the performance of SSI with respect to the properties proposed in Section 3 based on the original axioms proposed by Echenique and Fryer (2007). The value of SSI on the network level is 0 if and only if all ties are between groups. From that perspective, SSI is not related to any of the network models proposed in Section 3.1.

Turning to the effects of adding isolates to the network, the individual segregation of an isolate is 0 by definition (Echenique and Fryer, 2007, Section V.B.). Adding isolates to the network always *decreases* the network-level average (unless it is already 0). Therefore, ISO is not satisfied.

The values of the SSI are calculated based on the connected components of the within-group interaction networks. Accordingly, the individual SSIs and the component-level average will be identical for two identical components. Because all the component- and network-level SSIs are simple averages of individual level quantities, the SSI satisfies the Symmetry property.

4.8. *The Segregation Matrix Index*

This measure, proposed by Fershtman (1997), is designed for directed graphs. It is based on the mixing matrix M. The original version assumes only two groups of nodes, but it is straightforward to generalize the measure to an arbitrary number of groups. We start with the version for two groups. Given a network mixing matrix m_{ghy} and two groups G_1 and G_2, define the following quantities:

$$d_{11} = \frac{m_{111}}{m_{11+}}, \tag{4.27}$$

$$d_{12} = \frac{m_{121}}{m_{12+}}. \tag{4.28}$$

The value of d_{11} is the density of the ties within group G_1, and the value of d_{12} is the density of the ties between the groups G_1 and G_2. The tendency for the group to have segregative ties is the ratio of these two densities:

$$R(G_1) = \frac{d_{11}}{d_{12}}, \tag{4.29}$$

$$R(G_2) = \frac{d_{22}}{d_{21}}. \tag{4.30}$$

The value of $R(\cdot)$ ranges from 0 to ∞, but can be normalized to a quantity that varies between -1 and 1:

$$S^1_{\text{SMI}} = \frac{R(G_1)-1}{R(G_1)+1} = \frac{d_{11}-d_{12}}{d_{11}+d_{12}} \quad \text{for group } G_1, \tag{4.31}$$

$$S^2_{\text{SMI}} = \frac{R(G_2)-1}{R(G_2)+1} = \frac{d_{22}-d_{12}}{d_{22}+d_{12}} \quad \text{for group } G_2. \tag{4.32}$$

The index is called the *Segregation Matrix Index*. It is defined at the group level and is computed for each group separately. The original publication by Fershtman (1997) does not suggest any way to compute a network-level segregation score.

The Segregation Matrix Index can be extended for us with an arbitrary number of groups by reformulating the densities in Eqs. (4.27) and (4.28) to take into account the ties to other groups. The multi-group Segregation Matrix Index takes the following form:

$$w_g = \frac{m_{gg1}}{m_{gg+}} \quad \text{(density of within-group ties)}, \tag{4.33}$$

$$b_g = \frac{m_{g+1} - m_{gg1}}{m_{g++} - m_{gg+}} \quad \text{(density of between-group ties)}, \tag{4.34}$$

$$S^g_{\text{SMI}} = \frac{w_g - b_g}{w_g + b_g}. \tag{4.35}$$

See Appendix C for the detailed derivation.

To see that S_{SMI} embeds the marginal effects model as its null model, consider that $S_{\text{SMI}} = 0$ for group g if and only if the density of within-group ties equals the density of between-group ties. As fershtman (1997, Section 2) shows, this is equivalent to stating that given the total degree of group g, the numbers of between-group and within-group ties are proportional to the group sizes of group g and the respective other group(s), which is equivalent to the marginal effects network model.

The effect of adding isolates to the network on the value of S^g_{SMI} depends on the group membership of the added isolate. If it is added to group G_g, then the index always increases. However, if it is added to any group other than G_g, the index decreases. Therefore, the ISO property is not satisfied. See Appendix C.2 for a complete demonstration.

Symmetry is not satisfied given that doubling the network always decreases the value of $R(\cdot)$. See Appendix C.3 for further details.

4.9. Coleman's Homophily Index

Coleman (1958) defines a segregation measure for *directed* networks. In its original formulation, this measure was defined for each subgroup in a population. We first explain this group-wise formulation and propose a network-level version. Let m_{gg1} denote the number of ties *within* group G_g. The expected number of ties within the gth group in a random network is then

$$m^*_{gg1} = \sum_{i \in G_g} \eta_i \frac{n_g - 1}{N - 1}, \tag{4.36}$$

where η_i is the *out-degree* of actor *i*.

The fraction $\pi_g = (n_g - 1)/(N - 1)$ is the probability for a node to choose a node from the same group if the choice is random.[10] The segregation index $S^g_{Coleman}$ for group G_g is established to represent the propensity of an individual to create a tie to someone from the same group (i.e., the extent of *homophily*), as opposed to choosing randomly. The index is constructed as

$$S^g_{\text{Coleman}} = \begin{cases} \dfrac{m_{gg1} - m^*_{gg1}}{\sum_{i \in G_g} \eta_i - m^*_{gg1}} & \text{if} \quad m_{gg1} \geq m^*_{gg1}, \\ \dfrac{m_{gg1} - m^*_{gg1}}{m^*_{gg1}} & \text{if} \quad m_{gg1} < m^*_{gg1}. \end{cases} \tag{4.37}$$

Eq. (4.37) provides an index that varies between −1 (perfectly avoiding one's own group) and 1 (perfect segregation). The index assumes the value 0 if and only if the expected number of within-group ties under random choice is exactly equal to the observed number of within-group ties, given the total degree of a group. As such, S_{Coleman} embeds the marginal effects network model as its null model.

By taking the random network as a baseline for comparison, Coleman (1958) follows the same logic as (Freeman (1978b), see Section 4.6). The major conceptual difference between the two indices is that Freeman's measure is intended for undirected networks whereas Coleman's measure is intended for directed networks. Although there are no technical reasons not to apply S_{Coleman} to undirected networks, there are *conceptual* objections. In the case of Coleman's index, it is assumed that individuals' choices of ties are *independent*. This is typically not the case in undirected networks in which, for example, the consent of both individuals is required to create an undirected relationship. Given this process, the expected number of ties within groups (Eqs. (4.36) and (4.38)) may be different (indeed, the procedure applied in Freeman (1978b) would be more appropriate; see Section 4.6). Thus, we recommend caution when applying S_{Coleman} to undirected networks.

The index can be generalized to provide a measure at the network level that is more comparable to the other indices discussed in this paper. Let $\omega = \sum_g m_{gg1}$ denote the total number of (directed) ties within the same group. The expectation ω^* given that actors choose partners randomly is equal to

$$\omega^* = \sum_g \sum_{i \in G_g} \eta_i \frac{n_g - 1}{N - 1}, \tag{4.38}$$

and the segregation index S_{Coleman} on the network level is given by

$$S_{\text{Coleman}} = \begin{cases} \dfrac{\omega - \omega^*}{\sum_{i=1}^{N} \eta_i - \omega^*} & \text{if} \quad \omega \geq \omega^*, \\ \dfrac{\omega - \omega^*}{\omega^*} & \text{if} \quad \omega < \omega^*. \end{cases} \tag{4.39}$$

Whether the addition of an isolated node leads to an increase (as both of the above examples) or decrease of the index value depends on the group membership of the node and the distribution of ties between the groups. Thus the property ISO is not satisfied.

Finally, the property Symmetry is not satisfied either. If we duplicate the network X to obtain network Z, we have $\pi(X)_g = (n_g - 1)/(N - 1)$ and $\pi(Z)_g = (2n_g - 1)/(2N - 1)$. Clearly, $\pi(Z)_g > \pi(X)_g$, and thus $S_{\text{Coleman}}(Z) < S_{\text{Coleman}}(X)$. Note, however, that this difference is due to the term −1 in the numerator and denominator in (4.38). For large Ns, $\pi(Z)_g \approx \pi(X)_g$ and $S_{\text{Coleman}}(Z) \approx S_{\text{Coleman}}(X)$.

5. Empirical examples

In this section we provide examples of applying the measures reviewed in Section 4 to four network datasets. These four datasets (shown on Fig. 2) come from three sources:

White's kinship network Data from White (1975) used as an example by Freeman (1978b). The undirected network, illustrates communication links between kinship positions (upper left on Fig. 2). A link is absent if communication between the two positions is ever restricted in any of 219 societies analyzed by white. There are two groups of vertices corresponding to gender.

Galesburg friendship network Directed friendship network among physicians from Columbia University Drug Study (Coleman et al., 1966) and it is shown in the upper right of Fig. 2. The two groups correspond to physicians who did or did not adopt prescribing a new drug. The data was retrieved from the Pajek collection of datasets accompanying the book by de Nooy et al. (2005).[11]

[10] Coleman proposes that π can be conveniently approximated by $\pi_g \approx (n_g)/(N)$ for large N.

[11] http://vlado.fmf.uni-lj.si/pub/networks/data/esna/Galesburg2.htm.

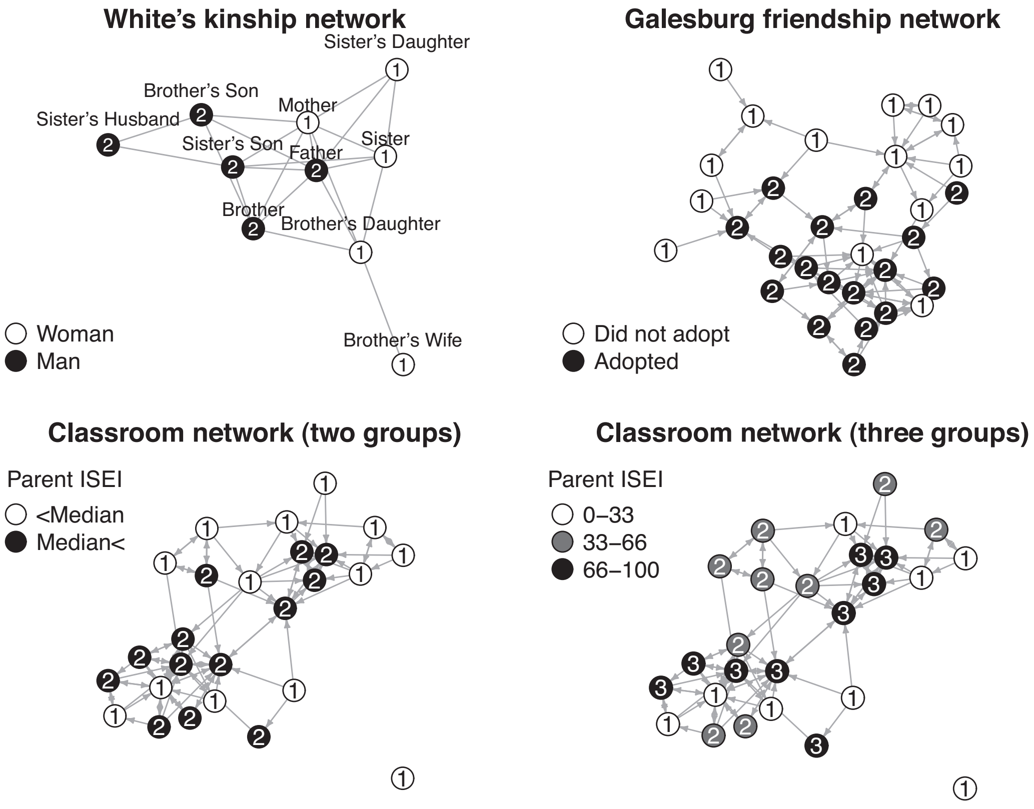


Fig. 2. Networks used as examples.

Classroom network A directed classroom network based on a study by (Polish) Educational Research Institute[12] (Educational Research Institute, 2012) based on answers to a question "With whom would you like to play with?" Children are assigned to groups based on an average International Socio-Economic Index score of their parents. We consider to versions of this dataset. In the first version (lower left of Fig. 2), there are two groups created by splitting ISEI on the median. In the second version (lower right of Fig. 2), there are three groups based on terciles of ISEI.

For every network dataset we have calculated all segregation measures that are applicable given the directionality of the network and the number of groups. Table 2 presents the results.

Let us start with measures based on the contact layer of the mixing matrix: the E-I index, the Assortativity Coefficient, the Gupta–Anderson–May measure, and CLLMs. These are also measures that embed proportionate mixing model (PM, Definition 3). Given those common properties we would expect these measures to behave in a similar way. For Galesburg network all measures lead toward the same conclusion that the network is mildly segregated (e.g., Gupta–Anderson–May is 0.350 with maximum of 1). For the remaining three datasets the results are not that consistent. The two-group classroom network does not seem to be segregated at all according to the E-I index, the Assortativity Coefficient, and the uniform homophily effect of CLLM, with the first two being small and negative, and the third small and positive. For White's network and the three-group classroom network different measures provide even more different results. In particular, White's network is mildly segregated according to the E-I index and the Assortativity Coefficient, but Gupta–Anderson–May and CLLM are negative, suggesting integration. For three-group classroom network, the E-I index suggests slight segregation while the Assortativity coefficient and Gupta–Anderson–May are negative, but rather close to 0. Differential homophily effects in CLLM suggest that groups 2 and 3 are to a similar extent more segregated than group 1.

Let us move on to segregation measures that are based on the full mixing matrix: the odds-ratio for within-group ties (ORWG), ERGMs, Freeman's index, the Segregation Matrix Index, and Coleman's index. According to the results in Table 2 these measures seem to behave more consistently, but not without exceptions. ORWG and uniform homophily effect in ERGM suggest the same ordering of the four network datasets in terms of segregation: the two-group classroom network is the least segregated, followed by the three-group classroom network, the Galesburg network, and White's network as the most segregated. The Segregation Matrix Index and Coleman's index provide very similar results for Galesburg network and two-group classroom network: both suggest that group 1 is less segregated than group 2. Classroom network with three groups allow us to compare differential homophily effects in ERGM and the Coleman index. Both measures suggest the same ordering of the groups in terms of segregation. For group 1 the measures are negative meaning that actors belonging to that group have much more ties to actors in different groups than ties to actors within group 1. Group 3 is the most the most segregated. Group 2 is somewhere in between groups 1 and 2 in terms of segregation level: ERGM effect is positive suggesting segregation while Coleman index is negative suggesting integration.

To conclude discussing the examples we would like to point out a few cases of how different measures could lead to different

[12] http://www.ibe.edu.pl.

Table 2
Reviewed segregation measures applied to four example networks datasets. Cells with '–' correspond to situations, in which a particular measure is not defined (e.g., it is not applicable for directed or undirected network or it is not defined on a particular level of analysis).

Measure	Level[c]	Network			
		White	Galesburg	Classroom (2 groups)	Classroom (3 groups)
E-I[a]	Network	−0.182	−0.385	−0.023	0.295
Assortativity Coefficient	Network	0.180	0.350	−0.025	−0.017
Gupta–Anderson–May	Network	−0.163	0.350	−0.182	−0.113
Odds-ratio WG ties	Network	3.302	2.542	1.156	1.265
CLLM: uniform homophily	Network	−0.328	0.752	0.032	0.087
CLLM: differential homophily	Group 2[b]	–	–	–	1.533
CLLM: differential homophily	Group 3[b]	–	–	–	1.625
ERGM: uniform homophily	Network	1.202	0.891	0.133	0.242
ERGM: differential homophily	Group 1	–	–	–	−1.058
ERGM: differential homophily	Group 2	–	–	–	0.621
ERGM: differential homophily	Group 3	–	–	–	0.769
Freeman	Network	0.264	–	–	–
SSI	Network	0.631	–	–	–
SSI	Group 1	0.606	–	–	–
SSI	Group 2	0.657	–	–	–
Segregation Matrix Index	Group 1	–	0.325	−0.361	–
Segregation Matrix Index	Group 2	–	0.448	0.072	–
Coleman	Group 1	–	0.294	−0.370	−0.571
Coleman	Group 2	–	0.464	0.444	−0.242
Coleman	Group 3	–	–	–	0.510

[a] The E-I index has a reverse scale ($S_{\text{E-I}} = 1$ for full integration) as compared to other measures. Here we report it on the original scale.
[b] Differential homophily CLLMs for classroom network with three groups treat group 1 as a reference category.
[c] Whether a measure describe segregation of the network as a whole or of a particular group (possibly relative to other groups). Group numbers correspond to numbers on vertices in networks shown on Fig. 2. Please note that the two classroom networks differ in how the vertices are assigned to groups.

Table 3
Performance of segregation measures with respect to the proposed properties.

Measure	Properties			Level[c]			Network type[d]	Scale
	Null model[a]	ISO[b](→)	SYM[b](→)	Network	Group	Node		
E-I index[e] (S. 4.1)	–	→	→	+	–	–	D/U	$[-1;1]$
Assortativity Coefficient (S. 4.2)	PM	→	→	+	–	–	D/U	$\left[-\frac{\sum_g p_{g+}p_{+g}}{1-\sum_g p_{g+}p_{+g}};1\right]$
Gupta–Anderson–May (S. 4.3)	PM	→	→	+	–	–	D/U	$[-\frac{1}{G-1};1]$
Odds-ratio WG ties (S. 4.4)	RN	↕	→	+	–	–	D/U	$(0;\infty)$
CLLM: uniform homophily (S. 4.5.1)	PM	→	→	+	–	–	D/U	$(-\infty;\infty)$
CLLM: differential homophily (S. 4.5.1)	PM	→	→	–	+	–	D/U	$(-\infty;\infty)$
ERGM: uniform homophily (S. 4.5.2)	ME	↕	→	+	–	–	D/U	$(-\infty;\infty)$
ERGM: differential homophily (S. 4.5.2)	ME	↕	→	–	+	–	D/U	$(-\infty;\infty)$
Freeman (S. 4.6)	RN	↕	↘	+	–	–	U	$[0;1]$
SSI (S. 4.7)	–	↘	→	+	+	+	U	$[0;\infty)$
Segregation Matrix Index (S. 4.8)	ME	↕	↘	–	+	–	D	$[-1;1]$
Coleman (S. 4.9)	ME	↕	↘	+	+	–	D	$[-1;1]$

[a] Null models defined in Section 3.1: PM: proportionate mixing network model; ME: marginal effects network model; RN: random network model.
[b] Properties: ISO: insensitivity to adding isolates, page 13; SYM: symmetry, page 14; micro-modification effects: ↗: value of the measure always increases, ↘: value of the measure always decreases, →: value of the measure does not change, ↕: value of the measure increases, decreases or stays the same depending on other features of the network.
[c] Level: + / −: the measure does or does not provide segregation scores on the given level of analysis.
[d] Network type: U: undirected, D: directed, D/U: directed or undirected
[e] Because the E-I index is operationalized in the opposite direction than the other measures, results reported here apply to $-S_{EI}$ rather than S_{EI}.

conclusions. Perhaps the most extreme case is the White's network. For that network measures embedding the proportional mixing null model, notably the Gupta–Anderson–May index and CLLM, but also the E-I index suggest that it is integrated. At the same time, measures based on a full mixing matrix, including ORWG, ERGM, and Freeman's index, suggest segregation. A similar inconsistency, but to a lesser extent, can be observed for the three-group classroom network. Measures embedding proportional mixing suggest low integration, while ORWG and ERGM suggest segregation.

In the case of classroom data, it is rather clear that any segregation/integration is a result of tie formation process rather then contagion. Given that, measures embedding the marginal effects – or random network null models seem more appropriate, which would lead us to conclude that it is rather segregated. In the case of White's network, we should note that is rather specific because it represents relations between kinship positions, not between natural actors. Nevertheless, the tie formation perspective seem more appropriate too.

6. Conclusions and discussion

Upon reviewing the set of segregation measures, we now turn to summarizing the main results. Table 3 lists the 11 network segregation measures evaluated in this paper.

As shown in the column "Network type", most of the measures are applicable in the context of both directed and the undirected networks. Only Freeman's index and the Spectral Segregation Index are designed specifically for undirected networks. Simultaneously, only the Segregation Matrix Index and Coleman's index are designed specifically for directed networks.[13]

The measures are designed for various levels of analysis (columns "Level" in Table 3), though most of them yield a network-level scalar value. Any other heterogeneity between the groups, such as different network activity levels, are incorporated into this scalar quantity. For measures on the group level, the choices include the Spectral Segregation Index, Coleman's index, the Segregation Matrix Index, and the differential homophily effects of CLLMs and ERGMs. Of all the measures the Spectral Segregation Index is the most versatile because it is defined on the individual node level and can be flexibly aggregated for the higher levels of groups, components, or the whole network.

The analyzed indices differ in terms of their measurement unit and zero point (see column "Scale" in Table 3). Having a well-defined unit and zero point facilitates the interpretation of the results. For all the measures apart from the E-I index and SSI the zero point corresponds to a particular null model that a measure embeds (see column "Null model"). This is indicated in the first column of Table 3. For the measures based on the contact layer of the mixing matrix, the zero point corresponds to *proportionate mixing* (Definition 3), referring to stochastical independence of ego and alter attributes. This holds for the Assortativity Coefficient, the Gupta–Anderson–May index, and homophily effects in CLLMs. At the same time, other measures take into account the number of disconnected dyads. The Odds Ratio for Within-Group ties (ORWG), Coleman's index, or homophily effects in ERGM assume a value of 0 for networks consistent with the Marginal Effects model (Definition 2), i.e., when conditional probability of tie existence given the attributes of the actors depends only on the relative number of ties associated with each group. Freeman's segregation index is guaranteed to assume value of 0 if the network is consistent with the Random Network model (Definition 1). Consequently, it may quantify networks as to some extent segregated, even if they are consistent with the ME model. The Spectral Segregation Index behaves differently and takes the value of 0 for networks that contain only between-group ties.

The unit of measurement depends on the normalization procedure used in each measure. Some measures (Freeman's index, the Assortativity Coefficient, Gupta–Anderson–May's index, Coleman's index, and the Segregation Matrix Index) are scaled such that the maximum of 1 is reached for full segregation (i.e., when only within-group ties are present in the network). The E-I index has a reverse scale, such that full segregation corresponds to a value of −1. For these measures, the particular value indicates how far an observed network is located from the case of full segregation and the other relevant extreme. The effects in CLLMs and ERGMs vary between plus and minus infinity but can, nevertheless, be interpreted in the same way as coefficients in logistic regression or log-linear models for contingency tables because the values often correspond to certain log odds ratios in the mixing matrix. For example, the uniform homophily effect refers to the extent to which ties are more likely to exist within groups than between groups.

The SSI is a special case in this context, taking on only non-negative values with no maximum. However, the interpretation is implied by the homogeneity property. As an example, if the SSI of a given network is equal to 0.6, then in this network on average everybody devotes 60% of their ties to others from the same group. However, when SSI exceeds 1, interpretation becomes more unclear.

6.1. *Measures versus properties*

The rest of Table 3 summarizes the performance of the measures reviewed in Section 4 with respect to the properties defined in Section 3. Recall that each of the proposed properties implied a network modification operation that, in turn, can (and often does) change the value of the segregation measure. Arrow symbols are used to depict graphically the direction of this change (row) when the network is subjected to a particular micro-modification (column). Explanations of the arrows are provided in the bottom of Table 3, and also included in the table header next to the names of the properties. Serving as only a reference, each arrow indicates the "expected" direction of the effect as formulated in Section 3 based on informal considerations about segregation measurement.

Much more interesting results are obtained from the investigation of the behavior of the segregation indices when the isolates are included into the network (property ISO). All measures based on the contact layer of the mixing matrix (i.e., the E-I index, the Assortativity Coefficient, Gupta–Anderson–May's index, and homophily effects in CLLMs) are insensitive to isolates and thus satisfy ISO. The Spectral Segregation Index is the only measure that decreases (unless it is already 0) whenever isolates are added to the network. The rest of the measures decrease or increase depending on the existing group sizes.

The property of Symmetry is satisfied by most of the analyzed indices. The segregation level of the combined baseline networks is the same as the level when each baseline network is considered individually. However, there are three exceptions: Freeman's index, Coleman's index, and the Spectral Segregation Index. Freeman's index and Coleman's index both decrease because duplicating the network slightly changes the opportunity structure for creating ties. Freeman's index decreases, because the expected fraction of between-group ties decreases. Coleman's index decreases, because the probability of randomly choosing a within-group network partner increases. Finally, the Segregation Matrix Index decreases if the network is doubled because the doubling decreases the ratio of densities of within- and between-group ties.

6.2. *Network ties and nodal attributes*

The crucial point differentiating among the reviewed measures is the question of whether a network should be considered from one of two perspectives:

1. The configuration of network ties can be treated as fixed while the group attributes of the nodes are dynamic.
2. The observed network is a product of some tie formation process between actors possessing more or less stable attributes.

In case 1, the central questions pertain to explaining the observed pattern of group memberships given the existing network ties. From this perspective, the proper approach is to concentrate on the conditional probability distribution of group memberships of egos and alters given the existing ties. This is equivalent to the contact layer of the mixing matrix, The E-I index, the Assortativity Coefficient, Gupta–Anderson–May, and homophily effects in CLLMs follow that approach. In terms of the analyzed properties, these measures are insensitive to the number and group membership of the isolates.

In case 2 we are more interested in capturing the segregative character of the formed ties, as opposed to ties that *were not* formed. This perspective focuses on the conditional probability of

[13] We restricted our review to measures for unweighted networks, but some of the measures discussed may be extended to weighted networks as well. In particular, if the weights are counts (e.g., number of relational events within a dyad) then all the measures based on a mixing matrix can be readily applied as this information is naturally aggregated into a mixing matrix. Additionally, the SSI can be applied to weighted networks if the weight on a tie w_{ij} can be interpreted as "exposure" of i to j.

tie existence given the configuration of attributes of the actors involved. Freeman's index, the odds-ratio for within-group ties, the Segregation Matrix Index, homophily effects in ERGMs, and Coleman's measure all follow that approach. Not coincidentally, the distinction between the two approaches is mostly aligned with the different null models — the measures associated with case 1 all rely on the proportionate mixing null model, while those associated with case 2 rely on either the random network model or the marginal effects network model.

Classifying the Spectral Segregation Index in that context is not straightforward. On the one hand, the SSI is sensitive to the number of isolates in the network (property ISO), which, at first sight, appears to run counter its primary design purpose of studying spatial segregation. On the other hand, the network-level SSI decreases when adding isolates to the network simply because the node-level SSI for any isolate is 0 by definition. The values for the remaining nodes remain the same, which implies that the measure is component-separable and satisfies the Symmetry property, which is a desirable property for category 1.

The above distinction implies that the choice for a segregation measure in a given research context should be motivated by theoretical notions about the underlying process driving segregation. Choosing a measure insensitive to network dynamics while the network *is* (potentially) dynamic might lead to misleading results. As an example, consider a case where one compares patterns of friendship formation between two groups of people, and want to measure network segregation by gender within each group. Suppose that the networks in the two groups are identical, except that one of the groups includes a number of men without any friends. The presence of such isolates is relevant for any conclusion about gender-driven friendship choices, yet the measures that are based on the contact layer of the mixing matrix alone will not detect a difference in segregation between the two groups. Thus, in any context in which segregation is studied as resulting from network dynamics, it is crucial that a segregation measure is used that is not only sensitive to the pattern of existing ties, but also to the pattern of ties that are *not* there.

In summary, we propose that the choice of an appropriate segregation measure in a given research context should be primarily motivated by theoretical ideas about the process driving segregation. Only after that should more practical considerations be taken into account, such as the directionality of ties, the number of groups, and the preferred scale be taken into account.

6.3. Concluding remarks

The aim of this paper was to systematically compare and categorize the existing measures of network segregation/homophily. We compared a set of existing measures (Section 4) against a set of properties (Section 3) that are relevant to the issue of segregation in networks. There are two main areas for future work.

The first area concerns the axiomatic characterization of the measures. Axiomatic characterizations clarify all the assumptions that are built into the indices and, thus, facilitate the comparison across measures. Such stringency would greatly contribute to the social networks literature. However, out of all the measures, only the SSI was derived in this way.

The second area concerns the need to bridge the gap between data-driven empirical research and substantive theoretical models (see Granovetter, 1979). All the segregation indices reviewed in this paper (with a few exceptions mentioned below) were created to provide statistical descriptions of network data. However, a crucial element that is still missing is a clear behavioral interpretation of the measures, which would establish a firmer link to theoretical models. This necessity was explicated by coleman:

> Every good measure of purported tendency is based on an underlying model. The model shows, in effect, how this tendency operated to produce observed result. Thus, once one knows the model, he can work backward from the observed result to obtain a measure of the size of the tendency which supposedly produced it. (Coleman, 1958)

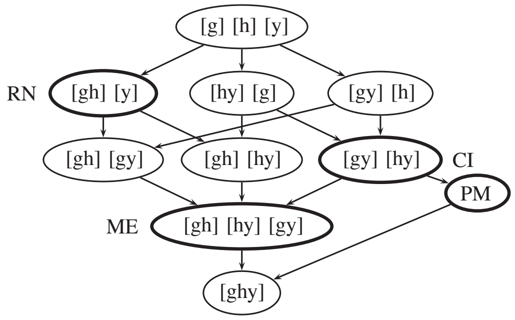
Following the comment, Coleman shows how his index can be derived from a simple probabilistic model of the way in which individual actors choose network partners. An alternative model related to Coleman's index has been recently proposed by Currarini et al. (2010). Analyses in a similar spirit have also been proposed in the context of network centrality measures (see, e.g., Ballester et al., 2006). We do hope to see additional research along these lines.

Acknowledgements

This research benefited from the support from Utrecht University 2004 High Potential Grant *Dynamics of Cooperation, Networks, and Institutions*. We are grateful to two anonymous reviewers for providing suggestions to improve the initial version of the manuscript. We thank Bastian Westbrock for joint discussions that inspired us to write this article. In addition, we would like to thank Vincent Buskens, Stephanie Rosenkranz, and Werner Raub, as well as participants of the ISCORE seminars at the ICS at Utrecht University for useful comments and suggestions. The authors are responsible for any remaining imperfections of the presented research.

Appendix A. Hierarchy of null models for mixing matrices

The diagram below presents a hierarchy of models for a mixing matrix m_{ghy}. Models marked with thicker ovals correspond to labeled null models from Section 3.1.



Each oval corresponds to a model for mixing matrix m_{ghy}. Symbols within an oval represent sets of marginal distributions of m_{ghy} (each in square brackets) that are sufficient statistics for model. For example, the model [g][h][y] in the top oval is a three-way independence model, in which all three variables are stochastically independent. The three marginals are sufficient statistics for that model because to compute the expected values $\widehat{m}_{ghy}$ we would multiply the corresponding from the marginal distributions m_{g++}, m_{+h+}, and m_{++y}.

Arrows between models show nesting relations. The arrow-sender model is nested in the arrow-receiver model. See for example works of Agresti (2002) and Goodman (1978, 1996) for a more comprehensive discussions on hierarchical log-linear models.

Appendix B. Additional derivations related to Freeman's segregation index

B.1. Multiple group variant

To derive the multiple-group variant of Freeman's index, it is sufficient to focus on the formula for the expected number of between-group ties π. The proportion of the between-group ties in a random graph is equivalent to the ratio of the number of between-group ties in a full network to the number of all possible ties. In the case of two groups, the number of all possible between-group ties is equal to $n_1 n_2$. We can rewrite it as:

$$\begin{aligned} n_1 n_2 &= \frac{1}{2}(2n_1 n_2 + n_1^2 + n_2^2 - n_1^2 - n_2^2) \\ &= \frac{1}{2}[(n_1+n_2)^2 - n_1^2 - n_2^2] \\ &= \frac{\left(\sum_{k=1}^{2} m_k\right)^2 - \sum_{k=1}^{2} m_k^2}{2}. \end{aligned} \tag{B.1}$$

In the general case for K groups, the number of possible between-group ties is equal to

$$\sum_{k=1}^{K-1}\sum_{l=k+1}^{K} n_k n_l, \tag{B.2}$$

which in turn, using the identity

$$(a_1 + a_2 + \cdots + a_n)^2 = \sum_{i=1}^{n} a_i^2 + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} a_i a_j, \tag{B.3}$$

can be rewritten as

$$\sum_{k=1}^{K-1}\sum_{l=k+1}^{K} n_k n_l = \frac{1}{2}\left[\left(\sum_{k=1}^{K} n_k\right)^2 - \sum_{k=1}^{K} n_k^2\right]. \tag{B.4}$$

Consequently, the expected proportion of between-group ties in a network with K groups is equal to

$$\pi = \frac{\left(\sum_{k=1}^{K} n_k\right)^2 - \sum_{k=1}^{K} n_k^2}{N(N-1)}, \tag{B.5}$$

and Freeman's segregation index to

$$S_{\text{Freeman}} = 1 - \frac{pN(N-1)}{\left(\sum_{k=1}^{K} n_k\right)^2 - \sum_{k=1}^{K} n_k^2}. \tag{B.6}$$

B.2. Effect of adding an isolate

The effect of adding an isolate on the value of S_{Freeman} can be shown in the following way. Formally, given a network X, we create a network X' by adding an isolate belonging to group 1. Then, we show that

$$S_{\text{Freeman}}(X) > S_{\text{Freeman}}(X') \quad \leftrightarrow \quad n_1 > n_2 - 1 \tag{B.7}$$

Adding isolates to the network affects only the value of π in (4.18). Therefore, we proceed with

$$\begin{aligned} &\pi > \pi' \\ &\frac{2n_1 n_2}{(n_1+n_2)(n_1+n_2-1)} > \frac{2(n_1+1)n_2}{(n_1+1+n_2)(n_1+n_2)} \\ &n_1(n_1+n_2+1) > (n_1+1)(n_1+n_2-1) \\ &n_1 > n_2 - 1 \end{aligned}$$

Consequently, Freeman's segregation index *decreases* in n_1 if and only if n_1 is greater than $n_2 - 1$. In practical terms, this means that adding isolates to the majority group *decreases* segregation.

B.3. Symmetry

Merging two identical networks (as in the Symmetry property) also affects only π. Before merging, each network is characterized by some $\pi_0 = \frac{2n_1 n_2}{N(N-1)}$. After duplicating and merging, the new value, π_1, is equal to

$$\pi_1 = \frac{8n_1 n_2}{2N(2N-1)} = \frac{4n_1 n_2}{N(2N-1)}. \tag{B.8}$$

The ratio of the two is equal to

$$\frac{\pi_1}{\pi_0} = \frac{4n_1 n_2}{N(2N-1)} \times \frac{N(N-1)}{2n_1 n_2} = \frac{N-1}{2N-1}, \tag{B.9}$$

which, for positive a N, is always strictly increasing and bounded within the interval $[0; 0.5)$. Consequently, as the ratio is smaller than 1 for any N, duplicating the network always *decreases* the segregation. This relationship holds independently of relative group sizes.

Appendix C. Additional derivations related to segregation matrix index

C.1. Multiple group variant

To generalize the original version of the SMI index to multiple groups, we first generalize the densities from Eqs. (4.27) and (4.28) to

$$w_g = \frac{m_{gg1}}{m_{gg+}} \quad \text{(density of within-group ties)} \tag{C.1}$$

$$b_g = \frac{m_{g+1} - m_{gg1}}{m_{g++} - m_{gg+}} \quad \text{(density of between-group ties).} \tag{C.2}$$

Next, the formula for R becomes

$$R(G_g) = \frac{w_g}{b_g}, \tag{C.3}$$

which allows us to define the multi-group segregation matrix index for group g as

$$S_{\text{SMI}}^{g} = \frac{R(G_g) - 1}{R(G_g) + 1} = \frac{w_g - b_g}{w_g + b_g}, \tag{C.4}$$

which is identical to Eq. (4.35).

C.2. Effect of adding an isolate

To show the effect of adding isolates, it is sufficient to focus on $R(\cdot)$, as S_{SMI} is a monotonic transformation of $R(\cdot)$.

We start by showing that adding isolates to groups other than G_g increases the value of S_{SMI}^{g}. First, notice that adding isolates to group $h \neq g$ affects $R(\cdot)$ only through b_g and m_{g++}. Consquently, increasing m_{g++} will decrease the value of b_g and *increase* the value of $R(\cdot)$, which increases S_{SMI}^{g}.

Demonstrating that S_{SMI}^{g} will always decrease when adding isolates to group G_g is slightly more complicated as it affects both m_{gg+} and m_{g++}. Let R be equal to (C.3) calculated for a group G_g in the given network X. Let R' be equal to $R(\cdot)$ computed for network Y which results from adding an isolate belonging to group G_g to network X. Substituting formulas for w_g and b_g into (C.3) yields the following:

$$R = \frac{m_{gg1}(N - 2n_g + 1)}{(n_g - 1)(m_{g+1} - m_{gg1}}, \tag{C.5}$$

$$R' = \frac{m_{gg1}(N - 2n_g - 1)}{n_g(m_{g+1} - m_{gg1})}. \tag{C.6}$$

Now, we need to show that $R' - R$ is negative for all $n_g \geq 1$. The difference becomes:

$$R' - R = \frac{m_{gg1}(N - 2n_g - 1) - \frac{n_g}{n_g - 1} m_{gg1}(N - 2n_g + 1)}{n_g(m_{g+1} - m_{gg1})}. \tag{C.7}$$

The denominator is always positive whenever the group G_g is not fully segregated in the network X. Thus, we can focus on the numerator, which, after factoring out m_{gg1}, becomes

$$N - 2n_g - 1 - \frac{n_g}{n_g - 1}(N - 2n_g + 1). \tag{C.8}$$

Multiplying by $(n_g - 1)$ preserves the sign, so we obtain

$$(n_g - 1)(N - 2n_g - 1) - n_g(N - 2n_g + 1) = -2n_g - N + 1 < 0, \tag{C.9}$$

i.e., the measure always decreases for $n_g > 0$ and $N > 0$.

C.3. Symmetry

To see why the Symmetry property is not satisfied let, we take R to be the value of $R(\cdot)$ for network X and R' to be the value of $R(\cdot)$ for a network Y that results from combining X and its copy as a single network. To verify the sign of the difference $R' - R$, it is worth noting, in (C.5), that its value does not depend on m_{gg1} nor m_{g+1}. Thus we have

$$R' - R \sim \frac{N - 2n_g + 1}{n_g - 1} - \frac{2N - 4n_g + 1}{2n_g - 1}, \tag{C.10}$$

With some algebra, it can be shown that its sign depends only on the sign of $n_g - N$, which is always negative given our assumption that $N > n_g \geq 1$. Consequently, S_{SMI} always decreases when the analyzed network is doubled.

References

Agresti, A., 2002. Categorical Data Analysis, 2nd edition. John Wiley & Sons, New York.

Alonso-Villar, O., del Río, C., 2010. Local versus overall segregation measures. Math. Soc. Sci. 60, 30–38.

Ballester, C., Calvó-Armengol, A., Zenou, Y., 2006. Who's who in networks. Wanted: the key player. Econometrica 74 (5), 1403–1417.

Borgatti, S.P., Everett, M.G., Freeman, L.C., 2002. Ucinet 6.

Chakravarty, S.R., 1999. Measuring inequality: the axiomatic approach. In: Silber, J. (Ed.), Handbook of Income Inequality Measurement. Kulwer Academic Press, Boston, pp. 163–186.

Charles, M., Grusky, D.B., January 1995. Models for describing the underlying structure of sex segregation. Am. J. Sociol. 100 (4), 931–971.

Cialdini, R.B., Goldstein, N.J., 2004. Social influence: compliance and conformity. Annu. Rev. Psychol. 55, 591–621.

Cohen, J., 1960. A coefficient of agreement for nominal scales. Educ. Psychol. Meas. 20, 37–46.

Coleman, J.S., 1958. Relational analysis: the study of social organizations with survey methods. Hum. Org. 17, 28–36.

Coleman, J.S., Katz, E., Menzel, H., 1966. Medical Innovation: A Diffusion Study. Bobbs-Merrill, Indianapolis.

Cowell, F.A., 1985. Measures of distributional change: an axiomatic approach. Rev. Econ. Stud. 52 (1), 135–151.

Cowell, F.A., Kuga, K., 1981. Inequality measurement: an axiomatic approach. Eur. Econ. Rev. 15, 287–305.

Currarini, S., Jackson, M.O., Pin, P., 2010. Identifying the roles of race-based choice and chance in high school friendship network formation. PNAS 107 (11), 4857–4861.

d'Aspremont, C., Gevers, L., 1977. Equity and informational basis of collective choice. Rev. Econ. Stud. 44 (2), 199–209.

d'Aspremont, C., Gevers, L., 1985. Axioms for social welfare orderings. In: Hurwicz, L., Schmeidler, D., Sonnenschein, H. (Eds.), Social Goals and Social Organization: Essays in Memory of Elisha Pazner. Cambridge University Press, Cambridge, pp. 19–76.

Duncan, O.D., Duncan, B., 1955. A methodological analysis of segregation indexes. Am. Sociol. Rev. 20, 210–217.

Echenique, F., Fryer Jr., R.G., May 2007. A measure of segregation based on social interactions. Quart. J. Econ. 122 (2), 441–485.

Educational Research Institute, 2012. School Effectiveness Study. http://eduentuzjasci.pl/en/oprojekcie-2/502-the-project.html

Egan, K.L., Anderton, D.L., Weber, E., 1998. Relative spatial concentration among minorities: addressing errors in measurement. Soc. Forces 76 (3), 1115–1121.

Erdös, P., Rényi, A., 1959. On random graphs. Publ. Math. 6, 290–297.

Erickson, B., 1988. The relational basis of attitudes. In: Wellman, B., Berkowitz, S.D. (Eds.), Social Structures: A Network Approach. Cambridge University Press, New York, pp. 99–121.

Fagiolo, G., Valente, M., Vriend, N.J., 2007. Segregation in networks. J. Econ. Behav. Org. 64, 316–336.

Fershtman, M., 1997. Cohesive group segregation detection in a social network by the Segregation Matrix Index. Soc. Netw. 19, 193–207.

Fienberg, S.E., Wasserman, S.S.,1981. Categorical data analysis of single sociometric relations. In: Sociological Methodology 1981. Jossey Bass, San Francisco, pp. 156–192, Ch. 4.

Foster, J.E., 1983. An axiomatic characterization of the theil measure of income inequality. J. Econ. Theory 31, 105–121.

Frank, O., 1988. Random sampling and social netoworks: a survey of various approaches. Math. Inf. Sci. Hum. 104, 19–33.

Frank, O., Strauss, D., 1986. Markov graphs. J. Am. Stat. Assoc. 81 (395), 832–842.

Freeman, L.C., 1978a. On measuring systematic integration. Connections 2 (1), 9–12.

Freeman, L.C., 1978b. Segregation in social networks. Sociol. Methods Res. 6 (4), 411–429.

Freeman, L.C., Sunshine, M.H., 1970. Patterns of Residential Segregation. Schenkman, Cambridge.

Goodman, L.A., 1978. A modified multiple regression approach to the analysis of dichotomous variables. In: Magidson, J. (Ed.), Analyzing Qualitative/Categorical Data. Log-linear Models and Latent Structure Analysis. Addison-Wesley Publishing Company, New York, pp. 7–27.

Goodman, L.A., 1996. A single general method for the analysis of cross-classified data: reconciliation and synthesis of some methods of pearson, yule, and fischer, and also some methods of correspondance analysis and association analysis. J. Am. Stat. Assoc. 91 (433), 408–428.

Grannis, R., 2002. Discussion: segregation indices and their functional inputs. Sociol. Methodol. 32, 69–84.

Granovetter, M.S., 1979. The theory-gap in social network analysis. In: Holland, P.W., Leinhardt, S. (Eds.), Perspectives on Social Network Research. Academic Press, New York, pp. 501–518.

Gupta, S., Anderson, R.M., May, R.M., 1989. Networks of sexual contacts: implications for the pattern of HIV. AIDS 3, 807–817.

Harville, D.A., 1997. Matrix Algebra From a Statistician's Perspective. Springer, New York.

Holland, P.W., Leinhardt, S., 1981. An exponential family of probability distributions for directed graphs. J. Am. Stat. Assoc. 76 (373), 33–50.

Hunt, C.L., Walker, L., 1974. Ethnic Dynamics: Patterns of Inter-Group Relations in Various Societies. Dorsey, Homewood.

James, D.R., Tauber, K.E., 1985. Measures of segregation. Sociol. Methodol. 15, 1–32.

Kalmijn, M., 1998. Intermarriage and homogamy: causes, patterns, trends. Annu. Rev. Sociol. 24, 395–421.

de Klepper, M., Sleebos, E., van de Bunt, G., Agneessens, F., 2010. Similarity in friendship networks: selection or influence? The effect of constraining contexts and non-visible individual attributes. Soc. Netw. 32 (1), 82–90.

Koehly, L.M., Goodreau, S.M., Morris, M., 2004. Exponential family models for sampled and census network data. Sociol. Methodol. 34, 241–270.

Krackhardt, D., Stern, R.N., 1988. Informal networks and organizational crises: an experimental simulation. Soc. Psychol. Quart. 51 (2), 123–140.

Massey, D.S., Denton, N.A., 1988. The dimensions of residential segregation. Soc. Forces 67, 281–315.

Massey, D.S., Denton, N.A., 1998. The elusive quest for the perfect index of concentration: reply to Egan, Anderton and Weber. Soc. Forces 76 (3), 1123–1133.

May, K.O., 1952. A set of independent necessary and sufficient conditions for simple majority decision. Econometrica 20 (4), 680–684.

McPherson, M., Smith-Lovin, L., Cook, J.M., 2001. Birds of a feather: homophily in social networks. Annu. Rev. Sociol. 27, 415–444.

Mitchell, J.C., 1978. On Freeman's segregation index: an alternative. Connections 2 (1), 9–12.

Moody, J., 2001. Race, school integration, and friendship segregation in America. Am. J. Sociol. 107 (3), 679–716.

Morris, M., 1991. A log-linear modeling framework for selective mixing. Math. Biosci. 107, 349–377.

Newman, M.E.J., 2003. Mixing patterns in networks. Phys. Rev. A 67 (2), 1050–2947.

Newman, M.E.J., Girvan, M., 2002. Mixing Patterns and Community Structure in Networks, arXiv:cond-mat/0210146 v1.

Nold, A., 1980. Heterogeneity in disease-transmission modelling. Math. Biosci. 52 (3-4), 227–240.

de Nooy, W., Mrvar, A., Batagelj, V., 2005. Exploratory Network Analysis with Pajek. Cambridge University Press, Cambridge, UK.

Reardon, S.F., Firebaugh, G., 2002a. Measures of multigroup segregation. Sociol. Methodol. 32, 33–67.

Reardon, S.F., Firebaugh, G., 2002b. Response: segregation and social distance – a generalized approach to segregation measurement. Sociol. Methodol. 32, 85–101.

Reynolds, H.T., 1977. The Analysis of Cross-Classifications. The Free Press, New York.

Robins, G., Elliot, P., Pattison, P., 2001a. Network models for social selection processes. Soc. Netw. 23, 1–30.

Robins, G., Pattison, P., Elliot, P., 2001b. Network models for social influence processes. Psychometrika 66 (2), 161–190.

Salganik, M., Watts, D.J., 2009. Social influence: the puzzling nature of success in cultural markets. In: Hedström, P., Bearman, P. (Eds.), Oxford Handbook of Analytic Sociology. Sage, Oxford, pp. 315–341, Ch. 14.

Schwartz, J., Winship, C., 1980. The welfare approach to measuring inequality. Sociol. Methodol. 11, 1–36.

Snijders, T.A.B., Pattison, P.E., Robbins, G.L., Handcock, M.S., 2006. New specifications for Exponential Random Graph Models. Sociol. Methodol. 36 (1), 99–153.

Suppes, P., Winet, M., 1955. An axiomatization of utility based on the notion of utility differences. Manage. Sci. 1 (3/4), 259–270.

Wasserman, S., Faust, K., 1994. Social Network Analysis. Cambridge University Press, Cambridge.

Wasserman, S., Pattison, P., 1996. Logit models and logistic regressions for social networks: I. An introduction to markov graphs and p^*. Psychometrika 61 (3), 401–425.

White, D.R., 1975. Communicative Avoidance in Social Networks. University of California, Irvine (mimeo).

van der Zanden, J., 1972. American Minority Relations. Ronald Press, New York.