Journal of Intelligent Information Systems manuscript No.
(will be inserted by the editor)

Segregation discovery in a social network of companies*

Alessandro Baroni · Salvatore Ruggieri

Received: date / Accepted: date

Abstract We introduce a framework for the data-driven analysis of social segregation of minority groups, and challenge it on a complex scenario. The framework builds on quantitative measures of segregation, called segregation indexes, proposed in the social science literature. The segregation discovery problem is introduced, which consists of searching sub-groups of population and minorities for which a segregation index is above a minimum threshold. A search algorithm is devised that solves the segregation problem by computing a multi-dimensional data cube that can be explored by the analyst. The machinery underlying the search algorithm relies on frequent itemset mining concepts and tools. The framework is challenged on a cases study in the context of company networks. We analyse segregation on the grounds of sex and age for directors in the boards of the Italian companies. The network includes 2.15M companies and 3.63M directors.

Keywords Segregation discovery, segregation indexes, frequent itemset mining, network of company board directors.

1 Introduction

The term *social segregation* refers to the "*separation of socially defined groups*" [39]. People are partitioned into two or more groups on the grounds of personal or cultural traits that can foster discrimination, such as gender, age, ethnicity, income, skin color, language, religion, political opinion, membership of a national minority, etc. [54]. Contact, communication, or interaction among groups are limited by their physical, working or socio-economic distance. Members of a group are often observed to cluster together when dissecting the society into organizational units (neighborhoods, schools, job types).

* A preliminary version of the results of this paper appeared in [5].

A. Baroni (🖂) · S. Ruggieri
Dipartimento di Informatica, Università di Pisa, Largo B. Pontecorvo 3, 56127 Pisa, Italy
email: {baroni,ruggieri}@di.unipi.it

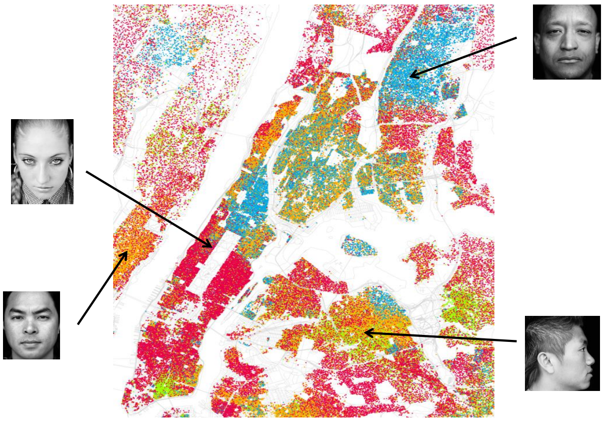

Fig. 1: Racial spatial segregation in New York City, based on Census 2000 data [22]. One dot for each 500 residents. Red dots are Whites, blue dots are Blacks, green dots are Asian, orange dots are Hispanic, and yellow dots are other races.

Early studies on residential segregation trace back to 1930's [20]. In this context, social groups are set apart in neighborhoods where they live in, in schools they attend to, or in companies they work at. As sharply pointed out in Figure 1, racial segregation (a.k.a. residential segregation on the grounds of race) very often emerges in most cities characterized by ethnic diversity. Schelling's segregation model [14, 57] shows that there is a natural tendency to spatial segregation, as a collective phenomenon, even if each individual is relatively tolerant – in his famous abstract simulation model, Nobel laureate Schelling assumed that a person changes residence only if less than 30% of the neighbors are of his/her own race.

Recently, [41] argued that segregation is shifting from ancient forms on the grounds of racial, ethnic and gender traits to modern socio-economic and cultural segregation on the basis of income, job position, and political-religious opinions. An earlier comparison of ideological segregation of the American electorate online and offline is offered in [27]. The paper found that segregation in news consumption is higher online than offline, but significantly lower than the segregation of face-to-face interactions with neighbors, co-workers, or family members. More recently, it has been warned that the filter bubble generated by personalization of online social networks may foster segregation [23], opinion polarization [38], and lack of consensus between different social groups. People are only reinforced in what they already believe and lack exposure to alternative viewpoints and information [4,48]. Polarization in social media may also lead to unfriending peers who expressed different opinions [29]. Consequently, online social network users are sometimes led to self-censorship acts [17] for fear of public opinion on personal thoughts.

The problem of assessing the presence, extent, nature, and trends of social segregation has been investigated so far by hypothesis testing. Hypothesis formulation, however, can be non-trivial and biased. In this paper, we will consider the social segregation problem from a data analysis perspective. We present theory, tools, and examples based on data mining and network science, for data-driven *segregation discovery*. We assume in input a dataset which records the characteristics of a population of individuals, including minority groups, distributed over a number of organizational units. The approach searches for sub-populations and social groups where *a-priori* unknown segregation is quantitatively prominent. Segregation is measured through evenness and exposure segregation indexes well-known in the social science literature. The approach allows for a deeper understanding of segregation phenomena through the design of analytical processes that proactively support policy makers and control authorities in discovering and in anticipating potential segregation problems. We demonstrate the applicability of the proposed methodology in a complex scenario, reflecting the risks of modern segregation in occupational social networks. The scenario considers glass-ceiling barriers for women in accessing boards of company directors. We challenge the proposed framework on the analysis of the real and large network of Italian companies.

The rest of the paper is organized as follows. Section 2 introduces segregation indexes. Section 3 defines the segregation discovery problem, and devises an algorithmic solution which provides the analyst with a multi-dimensional data cube for exploratory analysis. Section 4 deals with the case study of occupational segregation in networks of companies. Section 5 discusses related work. Finally, we summarize paper contribution and open problems for future work.

2 Segregation indexes

A segregation index provides a quantitative measure of the degree of segregation of social groups (e.g., Blacks, Whites, Hispanics, etc.) distributed among units of social organization (e.g., schools, neighborhoods, jobs, etc.). In this paper, we restrict to consider binary indexes, which assume a partitioning of the population into two groups, say majority and minority (but could be men/women, native/immigrant, White/NonWhite, etc.). Several indexes have been proposed in the literature. The surveys [21, 34] represent the earliest attempts to categorize them. Afterward, [40] provided a shared classification with reference to five key dimensions: evenness, exposure, concentration, centralization, and clustering. In this paper, we will consider evenness and exposure indexes. The other three classes of indexes are specifically concerned with spatial notions of segregation. Concentration indexes measure the relative amount of physical space occupied by social groups in an urban area. Centralization indexes measure the degree to which a group is spatially located near the center of an urban area. Clustering indexes measure the degree to which group members live disproportionately in contiguous areas.

Let T be size of the total population, $0 < M < T$ be the size of the minority group, and $P = M/T$ be the overall fraction of the minority group. Assume that there are n organizational units (or simply, units), and that for $i \in [1, n]$, t_i is the size of the population in unit i, m_i is the size of the minority group in unit i, and $p_i = m_i/t_i$ is the fraction of the minority population in unit i.

2.1 Evenness indexes

Evenness indexes measure the difference in the distributions of social groups among organizational units. The indexes mostly used in the social science literature include dissimilarity, information index, and Gini. The *dissimilarity index D* is the weighted mean absolute deviation of every unit's minority proportion from the global minority proportion:

$$D = \frac{1}{2 \cdot P \cdot (1-P)} \sum_{i=1}^{n} \frac{t_i}{T} \cdot |p_i - P|$$

The normalization factor $2 \cdot P \cdot (1-P)$ is to obtain an index in the range $[0,1]$. Since D measures dispersion of minorities over the units, higher values of the index mean higher segregation. Dissimilarity is minimum when for all $i \in [1,n]$, $p_i = P$, namely the distribution of the minority group is uniform over units. It is maximum when for all $i \in [1,n]$, either $p_i = 1$ or $p_i = 0$, namely every unit includes members of only one group (complete segregation).

Example 1 With basic algebra, it is readily checked that an equivalent definition of dissimilarity is:

$$D = \frac{1}{2} \sum_{i=1}^{n} \left| \frac{m_i}{M} - \frac{t_i - m_i}{T - M} \right|$$

D measures how different are the distributions of percentages of total minorities and of total majorities in the units. When $n = 2$, the formula boils down to:

$$D = \left| \frac{m_1}{M} - \frac{t_1 - m_1}{T - M} \right| \qquad (1)$$

The second widely adopted index is the *information index*, also known as the *Theil index* in social sciences [44] and normalized mutual information in machine learning [42]. Let the population entropy be $E = -P \cdot \log P - (1-P) \cdot \log(1-P)$, and the entropy of unit i be $E_i = -p_i \cdot \log p_i - (1-p_i) \cdot \log(1-p_i)$. The information index is the weighted mean fractional deviation of every unit's entropy from the population entropy:

$$H = \sum_{i=1}^{n} \frac{t_i}{T} \cdot \frac{(E - E_i)}{E}$$

Information index ranges in $[0,1]$. Since it denotes a relative reduction in uncertainty in the distribution of groups after considering units, higher values mean higher segregation of groups over the units. Information index reaches the minimum when all the units respect the global entropy (full integration), and the maximum when every unit contains only one group (complete segregation).

The third evenness measure is the *Gini index*, defined as the mean absolute difference between minority proportions weighted across all pairs of units, and normalized to the maximum weighted mean difference. In formula:

$$G = \frac{1}{2 \cdot T^2 \cdot P \cdot (1-P)} \cdot \sum_{i=1}^{n} \sum_{j=1}^{n} t_i \cdot t_j \cdot |p_i - p_j| \qquad (2)$$

Here $\sum_{i=1}^{n} \sum_{j=1}^{n} t_i \cdot t_j \cdot |p_i - p_j|$ is the weighted mean absolute difference. The normalization factor is obtained by maximizing such a value. The definition of the Gini index stems from econometrics, where it is used as a measure of the inequality of income distribution [26].[1] In our context, it measures the inequality of the majority group distribution among units. The Gini index ranges in $[0, 1]$ with higher values denoting higher segregation. The maximum and minimum values are reached in the same cases of the dissimilarity index.

An equivalent formulation of the Gini index (see [21, 59]) can be stated under the assumption that $p_1, \ldots, p_n$ are in descending order:

$$G = \frac{1}{M \cdot (T - M)} \cdot \sum_{i=1}^{n} (X_{i-1} \cdot Y_i - X_i \cdot Y_{i-1}) \tag{3}$$

where, for $i \in [1, n]$:

$$X_i = \sum_{j=1}^{i} m_i \qquad Y_i = \sum_{j=1}^{i} (t_i - m_i)$$

the X_i's (resp., Y_i's) are the cumulative sums of minority (resp., majority) population in units $1, \ldots, i$. Formulation (3) easily derives from the geometric interpretation of the Gini index (see footnote 1). From a computational perspective, it allows for computing G in $O(n \cdot log\ n)$, whilst formula (2) requires $O(n^2)$. This will be particularly relevant in our case study, where the number n of units will be in the order of millions.

2.2 Exposure indexes

Exposure indexes measure the degree of potential contact, or possibility of interaction, between members of social groups. The most used measure of exposure is the *isolation index* [9], defined as the likelihood that a member of the minority group is exposed to another member of the same group in a unit. For a unit i, this can be estimated as the product of the likelihood that a member of the minority group is in the unit (m_i/M) by the likelihood that she is exposed to another minority member in the unit $(m_i/t_i$, or $p_i)$ – assuming that the two events are independent. In formula:

$$I = \frac{1}{M} \cdot \sum_{i=1}^{n} m_i \cdot p_i$$

The right hand-side formula can be read as the minority-weighted average of minority proportions in units. The isolation index ranges over $[P, 1]$, with higher values denoting higher segregation. The minimum value is reached when for $i \in [1, n]$, $p_i = P$, namely the distribution of the minority group is uniform over the units. The maximum value is reached when there is only one $k \in [1, n]$ such that

[1] The geometric interpretation of the Gini index is provided in the space $[0, 1] \times [0, 1]$. The Lorenz curve plots the cumulative fraction of minority against the cumulative fraction of majority. Formally, assume that $p_1, \ldots, p_n$ are in descending order. The Lorenz curve $f()$ is the piece-wise linear function such that $f(0) = 0$, $f(1) = 1$, and, for $i \in [1, n]$, $f(\hat{X}_i) = \hat{Y}_i$ where $\hat{X}_i$ is the cumulative fraction of the minority group up to unit i, and $\hat{Y}_i$ is the cumulative fraction of the majority group up to unit i. The diagonal represents the perfect equality of distribution of majority vs minority population. The Gini index is twice the area between the Lorenz curve and the diagonal. See [21, 59] for details.

$m_k = t_k = M$, namely there is a unit containing all minority members and no majority member.

A dual measure is the *interaction index*, which is the likelihood that a member of the minority group is exposed to a member of the majority group in a unit. By reasoning as above, this leads to the formula:

$$\mathit{Int} = \frac{1}{M} \cdot \sum_{i=1}^{n} m_i \cdot (1 - p_i)$$

It clearly holds that $I + \mathit{Int} = 1$. Hence, lower values denote higher segregation. A more general definition of interaction index occurs when more than two groups are considered in the analysis, so that the exposure of the minority group to one of the other groups is worth to be considered [40].

2.3 Some properties of segregation indexes

There is a long standing debate in social sciences about which mathematical properties segregation indexes are expected to have. For instance, [32] lists seven general properties that indexes measuring occupational segregation should have for allowing comparison over longitudinal studies. Despite pros and cons of adopting a specific index, strong correlation among evenness indexes (D, H, and G) has been observed in practice by empirical analyses [40]. The following are some useful mathematical properties and differences among the evenness and exposure indexes introduced earlier.

(**P1**) All indexes are insensitive to units i with $t_i = 0$, i.e., by adding such "empty" units the value of an index does not change.

(**P2**) I and *Int* are insensitive to units i with $m_i = 0$, whilst D, H, and G are not. By adding units with only majority members, the likelihood of interaction among minority members does not change. The distributions of populations over units, instead, do change.

(**P3**) D, H and G are symmetric, i.e., by inverting the minority and majority groups the index remains unchanged, whilst I and *Int* are not. Intuitively, distance between minority and majority distributions is a symmetric concept, whilst likelihood of contact depends on the groups being considered.

(**P4**) I is anti-monotonic w.r.t. the addition of a majority member, i.e., by adding one majority member to a unit the value of I decrease. Moreover, *Int* is monotonic, and D, H, and G are neither monotonic nor anti-monotonic.

(**P5**) D, H and G are subject to the *Simpson's paradox*. E.g., for the dissimilarity index, there may exist datasets X and Y such that:

$$D_{X \cup Y} > D_X \quad \text{and} \quad D_{X \cup\, Y} > D_Y$$

where D_X, D_Y and $D_{X \cup Y}$ are the dissimilarity indexes for datasets X, Y and $X \cup Y$ respectively.

Let us explain the last two properties in detail. The following example shows what stated in (**P4**) for the dissimilarity index.

Example 2 Assume $n = 2$, with $m_1 = m_2 = 1$, $t_1 = 2$, and $t_2 = 4$. We have $M = 2$, $T = 6$. Using (1), it turns out $D = 1/2 - 1/4 = 0.25$.

Consider adding one majority member to unit 1 (the most segregated because $p_1 = 0.5$, $p_2 = 0.25$ and $P = 0.33$). We have $T = 7$ then $D = 1/2 - 2/5 = 0.1$. Thus, the dissimilarity index has decreased. Consider now instead adding the majority member to unit 2 (the most integrated). Again, we have $T = 7$. But now $D = 1/2 - 1/5 = 0.3$. The dissimilarity index has now increased.

Simpson's paradox is a well-known case in presence of ratios and differences of distributions, in which a trend appears in different groups of data but disappears or reverses when these groups are combined [49]. In our context, this occurs when segregation appears in combined dataset $X \cup Y$, but disappear when looking separately at X and Y, or vice-versa.

Example 3 Assume two university departments ($n = 2$). Faculty (X) and administrative staff (Y) are employed in each department. Assume the numbers on the left hand side of the following table:

	m_1	t_1	m_2	t_2	M	T	D	H	G
X	99	100	1	2	100	102	0.490	0.2903	0.490
Y	1	10	9	100	10	110	0.010	0.0002	0.010
$X \cup Y$	100	110	10	102	110	212	0.811	0.8354	0.811

Using formula (1), $D_{X \cup Y} = 100/110 - 10/102 = 0.811$, i.e., segregation at university level appears to be high. However, when considering separately faculty and administrative staff, we have $D_X = 99/100 - 1/2 = 0.49$ and $D_Y = 1/10 - 9/100 = 0.01$, i.e., segregation is much lower in both sub-groups. Similar conclusions can be drawn for H and G, as shown in the right hand side of the table above.

The previous example is a contrived one. Actually, the reversed effect (segregation in combined dataset $X \cup Y$ lower than in X and Y separately) is more likely to be observed, as we will show later on.

2.4 An extension of segregation indexes

So far, we assumed that individuals are *partitioned* among the units of analysis. Each individual belongs to one and only one unit. The size of the overall population is then the sum of the population in each unit, and similarly for the size of the minority population. This assumption readily holds in cases of residential segregation. The case study that we will present later on, however, breaks such an assumption, since individuals (company directors) may belong to more than one unit (group of companies). This situation resembles the case of segregation analysis in e.g., sports club where players may be associated with more than one team at a time [25,37] or in movie productions, where actors may play in movies of more than one producer [58]. We conservatively extend then the previously introduced definitions of segregation indexes by considering every instance of an individual in an unit as a distinct one. In practice, this turns out to revise the definition of T and M as follows:

$$T = \sum_{i=1}^{n} t_i \qquad M = \sum_{i=1}^{n} m_i$$

namely, the size of the total population is *by definition* the sum of the sizes of the unit populations, and similarly for the minority population.

3 Segregation discovery

Traditional data analysis approaches from social sciences typically rely on formulating an hypothesis, i.e., a possible context of segregation against a certain social group, and then in empirically testing such an hypothesis – see, e.g., [45]. For instance, a suspect case of segregation of female students in high schools from NYC is studied first by collecting data on gender of high school students in NYC (reference population), and then by computing and analysing segregation indexes over female students (minority group). The formulation of the hypothesis, however, is not straightforward, and it is potentially biased by the expectations of the data analyst of finding segregation in a certain context. In this process, one may overlook cases where segregation is present but undetected.

Example 4 By property **(P4)**, segregation can result undetected when the analyst targets an actually segregated minority but considering a reference population that is too small/large. Similarly, by property **(P5)**, segregation is undetected if analysed at a wrong granularity level, as shown in Example 3. However, an analyst does not typically know *a-priori* which granularity is the most appropriate one.

We propose a data-driven approach, which complements hypothesis testing, by driving the search (the “discovery”) of contexts and social groups where *a-priori* unknown segregation factors are quantitatively prominent. Recall the previous example on school segregation. The analyst has to collect data on gender and other possible segregation attributes such as age and race of students, and on location, school type, annual fees and other context attributes that may distinguish conditions of segregation. Although no segregation may be apparent in the overall data, it may turn out that for a specific combination of context attributes (e.g., high schools located in a particular area), a specific minority group denoted by a combination of segregation attributes (e.g., black female students) is at risk of segregation. We quantify such a risk through a reference segregation index, and assume that a value of the index above a given threshold denotes a situation worth for further scrutiny – what legal scholars call a *prima-facie* evidence. We call the problem of discovering a-priori unknown minority groups and reference populations for which segregation indexes are above a given threshold, the segregation discovery problem.

3.1 Notation and itemset mining

Let us recall notation and concepts from itemset mining [30], which will serve to define the search space of segregation discovery. Let $\mathcal{R}$ be a relational table (or, simply, a table or a dataset). Tuples σ in the table will denote individuals, and attribute values will denote information about individuals and organizational units they belong to. We assume that every attribute A has a discrete domain $dom(A)$ of values. Continuous attributes can be considered after discretization into bins. We denote by $\sigma(A)$ the value of the tuple σ on attribute A, as in e.g., $\sigma(\texttt{sex}) = \texttt{female}$.

An *A-item* is a term $A = v$, where $v \in dom(A)$. An *itemset* $\mathbf{X}$ is a set of items. As usual in the literature, we write $\mathbf{X}, \mathbf{Y}$ for $\mathbf{X} \cup \mathbf{Y}$. A tuple σ from $\mathcal{R}$ *supports* $\mathbf{X}$ if for every $A = v$ in $\mathbf{X}$, we have $v \in \sigma(A)$. The *cover* of $\mathbf{X}$ is the set of all tuples that support $\mathbf{X}$: $cover_{\mathcal{R}}(\mathbf{X}) = \{\sigma \in \mathcal{R} \mid \sigma \text{ supports } \mathbf{X}\}$. We omit the subscript $\mathcal{R}$ if it is clear from the context. Intuitively, covers will denote sets of individuals sharing the characteristics stated by the itemset. The (absolute) *support* of $\mathbf{X}$ is the size of its cover, namely $supp(\mathbf{X}) = |cover(\mathbf{X})|$. $\mathbf{X}$ is a *frequent* itemset if $supp(\mathbf{X}) \geq minsupp$, where *minsupp* is a given threshold. $\mathbf{X}$ is *closed* if there is no $\mathbf{Y} \supset \mathbf{X}$ with $cover(\mathbf{Y}) = cover(\mathbf{X})$ or, equivalently, with $supp(\mathbf{Y}) = supp(\mathbf{X})$. A closed itemset is a representative member of the class of equivalence of itemsets with a same cover [6]. Thus, the groups of individuals denoted by the cover of closed itemsets are non-overlapping, i.e., restricting to closed itemsets means pruning duplicate groups from the space of all covers of itemsets.

Example 5 Consider the dataset in Figure 2 (left). The cover of the itemset `sex=female, age=young` is the set of young women in the dataset, which consists of only one tuple. Its support is then 1. The itemset is not closed, since the superset `sex=female, age=young, region=north` has the same cover/support.

3.2 The segregation discovery problem

We introduce here the segregation discovery problem. Let $\mathcal{R}$ be an input relational table. We assume that attributes are partitioned into three groups. First, *segregation attributes* (SA), such as `sex`, `age`, and `race`, which denote minority groups potentially exposed to segregation. Second, *context attributes* (CA), such as `region` and `job type`, which denote contexts where segregation may appear. Third, an attribute `unitID`, which is an ID of the unit the tuple/individual belongs to. We write $\mathbf{A}, \mathbf{B}$ to denote an itemset where $\mathbf{A}$ includes only SA-items, and $\mathbf{B}$ includes only CA-items. We call $\mathbf{A}$ an *SA-itemset*, and $\mathbf{B}$ a *CA-itemset.*

Example 6 (Ctd.) For the itemset `sex=female, age=young, region=north`, it turns out $\mathbf{A}$ =`sex=female, age=young` and $\mathbf{B}$ =`region=north`. In this example, the minority group is the set of young women, and the majority population is all the rest, i.e., men or middle aged or elder. This is one specific pair of population and minority group. Other pairs could be considered in the segregation analysis. Actually, one would like to consider all possible such pairs.

We are now in the position to extend the notation of segregation indexes to itemsets $\mathbf{A}, \mathbf{B}$ where the reference population is the cover of $\mathbf{B}$, and the reference minority group is the cover of $\mathbf{A}, \mathbf{B}$. Recall that $0 < M < T$ is assumed by segregation index definitions.

Definition 1 Let $s()$ be a segregation index. For an itemset $\mathbf{A}, \mathbf{B}$ such that $0 < supp(\mathbf{A}, \mathbf{B}) < supp(\mathbf{B})$, we denote by $s(\mathbf{A}, \mathbf{B})$ the segregation index calculated for the population in $cover(\mathbf{B})$ considering as minority population those in $cover(\mathbf{A}, \mathbf{B})$.

For instance, $D(\mathbf{A}, \mathbf{B})$ is the dissimilarity index for minority $\mathbf{A}$ and population $\mathbf{B}$. Notice that $supp(\mathbf{A}, \mathbf{B}) < supp(\mathbf{B})$ implies that $\mathbf{A}$ cannot be the empty itemset.

Example 7 (Ctd.) $D(\texttt{sex=female, age=young, region=north})$ is then the dissimilarity index of segregation of young women among the units in north region. With reference to the dataset in Figure 2 (left), we have $T = 10$ (the support of $\texttt{region=north}$) and $M = 1$ (the support of $\texttt{sex=female, age=young, region=north}$). Populations in the units amount at $t_1 = 2$, $t_2 = 3$, $t_3 = 2$, $t_4 = 3$, $t_5 = 0$. Minority distribution in the units is $m_1 = m_3 = m_4 = m_5 = 0$ and $m_2 = 1$. By definition of dissimilarity, $D(\texttt{sex=female, age=young, region=north}) = 10^2/(2 \cdot 9) \cdot (2/10 \cdot 1/10 + 3/10 \cdot (1/3 - 1/10) + 2/10 \cdot 1/10 + 3/10 \cdot 1/10) = 7/9 \approx 0.78$.

Notice that all parameters needed to compute D can be defined using the itemset notation. In particular, $T = \mathit{supp}(\mathbf{B})$, $M = \mathit{supp}(\mathbf{A}, \mathbf{B})$, and, for a unit i: $t_i = \mathit{supp}(\mathbf{B}, \texttt{unit=}i)$, and $m_i = \mathit{supp}(\mathbf{A}, \mathbf{B}, \texttt{unit=}i)$.

Let us introduce now the problem of segregation discovery.

Definition 2 Let $s()$ be a segregation index, and α a fixed threshold.

Let $\mathbf{A}, \mathbf{B}$ be an itemset such that $0 < \mathit{supp}(\mathbf{A}, \mathbf{B}) < \mathit{supp}(\mathbf{B})$.

We say that $\mathbf{A}, \mathbf{B}$ is α-integrative w.r.t. $s()$ if $s(\mathbf{A}, \mathbf{B}) \leq \alpha$. Otherwise, $\mathbf{A}, \mathbf{B}$ is α-segregative. The problem of segregation discovery consists of computing the set of α-segregative itemsets.

Intuitively, we are interested in searching the space of itemsets for $\mathbf{A}, \mathbf{B}$ denoting a minority sub-group ($\mathbf{A}$) and a context ($\mathbf{B}$) where the segregation index $s(\mathbf{A}, \mathbf{B})$ is above the α threshold. Notice that we assume that higher values of $s()$ denote higher segregation, which is the case for all introduced indexes except for *Int*. For such an index, the bound in Definition 2 becomes $s(\mathbf{A}, \mathbf{B}) \geq \alpha$.

3.3 A data cube for exploratory analysis of segregation

The problem of segregation discovery can be readily formulated as the one of computing iceberg multi-dimensional data cubes [31]. Let $A_1, \ldots, A_k$ be the collection of segregation and context attributes. They can be considered dimensions of a multi-dimensional array. The i^{th} dimension is named as the attribute A_i, and it takes values in $\mathit{dom}(A_i) \cup \{\star\}$. The coordinate $v \in \mathit{dom}(A_i)$ denotes the itemset $A_i = v$. The coordinate $\star$ denotes the empty itemset (absence of an A_i-item). A multi-dimensional data cube is an array mapping dimension coordinates to values of a measure, which, in our case, is a segregation index $s()$:

$$d[A_1 = v_1, \ldots, A_k = v_k] = s(\cup_{i=1}^{k} \{A_i = v_i \mid v_i \neq \star\})$$

With a little abuse of notation, we write items instead of coordinate values in the index positions of the array $d[]$. For an itemset $\mathbf{A}, \mathbf{B}$ whose support is non-zero and is lower than the support of $\mathbf{B}$, if $d[\mathbf{A}, \mathbf{B}] = s(\mathbf{A}, \mathbf{B}) > \alpha$ then $\mathbf{A}, \mathbf{B}$ is α-segregative. The subset of cells in a data cube whose value is higher than a minimum threshold is called an iceberg data cube. Thus, the problem of segregation discovery is equivalent to computing the iceberg data cube of $d[]$. However, since the number of cells in a data cube grows exponentially with the number of dimensions, a practical additional requirement is to impose also a minimum support threshold. Thus, we aim at computing $s(\mathbf{A}, \mathbf{B})$ only if $\mathbf{A}, \mathbf{B}$ is frequent, namely $\mathit{supp}(\mathbf{A}, \mathbf{B}) \geq \mathit{minsupp}$. Finally, we also aim at considering itemsets $\mathbf{A}, \mathbf{B}$ that denote no duplicate pair

SA		CA	
sex	**age**	**region**	**unitID**
male	young	north	1
male	young	north	1
male	middle	south	2
male	middle	north	2
female	middle	north	2
female	young	north	2
male	middle	north	3
female	middle	north	3
male	elder	south	3
male	middle	south	3
female	elder	south	4
female	elder	south	4
female	middle	south	4
male	elder	north	4
male	young	north	4
male	elder	north	4
male	middle	south	5
male	young	south	5
female	middle	south	5

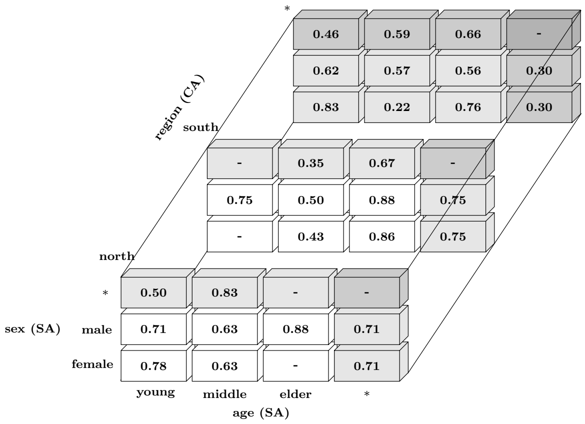

Fig. 2: Input table (left) and segregation data cube with the D index (right).

of reference population *cover*($\mathbf{B}$) and of minority group *cover*($\mathbf{A}, \mathbf{B}$). Distinct reference populations can be considered by enumerating $\mathbf{B}$'s that are frequent and closed (in $\mathcal{R}$). For a fixed $\mathbf{B}$, then the distinct minority groups can be achieved by enumerating $\mathbf{A}$'s that are frequent and closed in the dataset *cover*($\mathbf{B}$).

Example 8 Reconsider Ex. 5. The itemsets $\mathbf{A}_1, \mathbf{B}_1$ = `sex=female, age=young` and $\mathbf{A}_2, \mathbf{B}_2$ = `sex=female, age=young, region=north` have the same cover, i.e., they denote the same minority population. However, the former considers as reference population the whole dataset ($\mathbf{B}_1$ is empty), while the latter considers people from the north region ($\mathbf{B}_2$ is `region=north`). By property **(P4)** stated in Sect. 2.3, the dissimilarity index of $\mathbf{A}_1, \mathbf{B}_1$ can be lower, equal or higher than the one of $\mathbf{A}_2, \mathbf{B}_2$. In Ex. 7, we have seen that $D(\mathbf{A}_2, \mathbf{B}_2) \approx 0.78$. By doing the calculations, it turns out that $D(\mathbf{A}_1, \mathbf{B}_1) \approx 0.83$.

We introduce next the definition of segregation data cube.

Definition 3 Let $s()$ be a segregation index, and *minsupp* > 0 a fixed threshold. A segregation data cube is a multi-dimensional data cube such that:

$$d[\mathbf{A}, \mathbf{B}] = \begin{cases} s(\mathbf{A}, \mathbf{B}) \text{ if } \mathit{minsupp} \leq \mathit{supp}(\mathbf{A}, \mathbf{B}) < \mathit{supp}(\mathbf{B}) \text{ and} \\ \qquad\qquad \mathbf{B} \text{ is closed and } \mathbf{A} \text{ is closed in } \mathit{cover}(\mathbf{B}) \\ \text{“}-\text{”} \quad \text{otherwise} \end{cases}$$

Recalling that $M = \mathit{supp}(\mathbf{A}, \mathbf{B})$ and $T = \mathit{supp}(\mathbf{B})$, we have that the value of $d[\mathbf{A}, \mathbf{B}]$ is undefined, which is written as “-”, if:

- the minority group under analysis is smaller than a minimum threshold ($M < \mathit{minsupp}$);
- or, there is no majority member ($M = T$);
- or, the reference population is analysed in another cell ($\mathbf{B}$ is not closed);
- or, the the minority group within the reference population is analysed in another cell ($\mathbf{A}$ is not closed w.r.t. the dataset *cover*($\mathbf{B}$)).

Algorithm 1: Segregation data cube computation.

```
Input: relational table $\mathcal{R}$ with context attributes (CA), segregation attributes (SA),
       and unit attribute unitID with a total of $n$ units. minsupp threshold.
Output: segregation data cube $d[]$.
foreach B CA-itemset frequent and closed do
    $T = supp(\mathbf{B})$
    foreach $i \in [1, n]$ do
        $t_i = supp(\mathbf{B}, \texttt{unit=}i)$
    end
    foreach A SA-itemset frequent and closed w.r.t. $cover(\mathbf{B})$ do
        $M = supp(\mathbf{A}, \mathbf{B})$
        if $M < T$ then
            foreach $i \in [1, n]$ with $t_i > 0$ do
                $m_i = supp(\mathbf{A}, \mathbf{B}, \texttt{unit=}i)$
            end
            $X = Y = 0$
            $sum = 0$
            foreach $i \in [1, n]$ with $t_i > 0$ order by $m_i/t_i$ desc do
                $X$ += $m_i$
                $Y$ += $t_i$ - $m_i$
                sum += $f_s(m_i, t_i, X, Y, M, T)$
            end
            $d[\mathbf{A}, \mathbf{B}] = g_s(sum, M, T)$
        end
    end
end
```

Example 9 Figure 2 (right) shows the segregation data cube $d[]$ for the input dataset at its left. *minsupp* is set to 1, and dissimilarity is set as segregation index. Several facts are worth to be pointed out:

- dissimilarity is symmetric **(P3)**. In fact, for $X \in dom(\texttt{region}) \cup \{\star\}$, we have $d[\texttt{sex=female, age=}\star\texttt{, region=}X] = d[\texttt{sex=male, age=}\star\texttt{, region=}X]$;
- dissimilarity is neither monotonic nor anti-monotonic **(P4)**. In Ex. 8, we showed $d[\texttt{sex=female, age=young}] > d[\texttt{sex=female, age=young, region=north}]$. From the data cube, we can also see that $d[\texttt{sex=female, age=elder}] < d[\texttt{sex=female, age=elder, region=south}]$.
- the Simpson's paradox holds, in a reversed form than **(P5)**, namely dissimilarity of combined context $X \cup Y$ is lower than dissimilarities in X and Y. For instance, dissimilarity for females in general (0.30) is lower than in the north region (0.71) and in the south region (0.75) separately. Why does it happen? The proportion of females in general ($P = 6/19$) is close to the proportions of females in the north ($P_n = 3/10$) and in the south ($P_s = 3/9$) regions. So, it is the proportion of female in units that must be severely affected. In fact, consider unit 2. The proportion of females in general is $2/4 = 0.50$, which is distant from P only 0.17. The proportion in the north is 2/3, which is distant from $P_n \approx 0.36$, and in the south it is 0/1, which is distant from $P_s \approx 0.33$.

3.4 Computing segregation data cubes

Algorithm 1 provides a solution to the problem of computing a segregation data cube. The input is a relation $\mathcal{R}$ with context and segregation attributes, and a

Index	$f_s(m_i, t_i, X, Y, M, T)$	$g_s(sum, M, T)$
Dissimilarity (D)	$t_i \cdot \lvert \frac{m_i}{t_i} - \frac{M}{T} \rvert$	$\frac{T \cdot sum}{2 \cdot M \cdot (T-M)}$
Gini (G)	$(X - m_i) \cdot Y - X \cdot (Y - t_i + m_i)$	$\frac{sum}{M \cdot (T-M)}$
Information index (H)	$t_i \cdot \mathcal{E}(\frac{m_i}{t_i})$	$1 - \frac{sum}{T \cdot \mathcal{E}(\frac{M}{T})}$
Isolation (I)	$m_i \cdot \frac{m_i}{t_i}$	$\frac{sum}{M}$
Interaction (Int)	$m_i \cdot \left(1 - \frac{m_i}{t_i}\right)$	$\frac{sum}{M}$

Table 1: Function $f_s()$ and $g_s()$. Here, $\mathcal{E}(p) = -p \cdot \log p - (1-p) \cdot \log(1-p)$.

unit attribute, with n units. The output is the segregation data cube for a fixed segregation index $s()$.

Basically, the outer loop is over the set of frequent closed CA-itemsets $\mathbf{B}$. Enumeration of this set can be achieved through state of the art algorithms for closed itemset mining [30]. Our implementation adopts the system provided in [11].

For a given $\mathbf{B}$, we first compute the size of the reference population (line 2), and the size of unit populations $t_1, \ldots, t_n$ (lines 3–5). Support counting is performed by the function *supp*() (lines 2,4). A possible way of implementing *supp*() is through the construction of an FP-tree, a compressed representation of a dataset used for frequent itemset mining [30]. Our implementation, instead, relies on storing all CA and SA attributes of $\mathcal{R}$ in memory as compressed bitmaps, and the `unitID` attribute as an array that maps a tuple position into the unit ID of that tuple. We adopt the Enhanced WAH compression library [35], which relies on word-alignment to provide a good trade-off between space occupation and running time. *cover*($\mathbf{B}$) is explicitly computed by efficient bitmap and's operations. Values $t_1, \ldots, t_n$ are computed by iterating over such cover and incrementing counters based on the unit ID of the tuples in the cover. This approach is more efficient than iterating over units (line 3), when $n \gg supp(\mathbf{B})$.

The inner loop iterates over SA-itemsets $\mathbf{A}$ that are frequent and closed w.r.t. the reference population, namely *cover*($\mathbf{B}$). Again, enumeration of these itemsets can be achieved through closed itemset mining. However, since the number of SA-attributes is typically small (due to the difficulty of collecting sensitive data), our implementation adopts a simpler approach. We compute and store all frequent SA-itemsets $\mathcal{C}$ in memory. For a given $\mathbf{B}$, we first compute the support of itemsets in $\mathcal{C}$ w.r.t. *cover*($\mathbf{B}$). Obviously, itemsets that are not frequent in the whole dataset cannot be frequent in a subset of it. Then, we order itemsets in $\mathcal{C}$ lexicographically based on support and number of items. Finally, we filter out those infrequent (support lower than *minsupp*) and non-closed (support equal to an itemset including one additional item).

In the inner loop, we first check that $M < T$ (line 8) to meet the assumptions of segregation indexes. Then, we proceed with computing m_i's only for those units which are non-empty, i.e., such that $t_i > 0$. This optimization is possible by property **(P1)**. Next, we accumulate the results of a function $f_s()$ over each non-empty unit, and finally pass it to the normalization function $g_s()$. The intermediate functions f_s and g_s depend on the segregation index $s()$ under consideration. Table 1 shows their definitions for the indexed introduced in Section 2. The formulation (3) of the Gini index requires sorting units based on descending proportion of minority (line 14), and to compute cumulative sums of minority and majority population in units (lines 15,16). These computations are not strictly necessary for the other indexes. Finally, in the case of indexes I and Int, property **(P2)** can be exploited to speed up the loop at line 14 by restricting to non-zero m_i's.

3.5 Computational complexity

Let us discuss here the computational complexity of Algorithm 1.

We start with time complexity. An upper bound to the number of outer and inner iterations is given by the number of frequent itemsets $\mathbf{A}, \mathbf{B}$. In the worst case, this is $O(\pi)$, where $\pi = \prod_A |dom(A)|$, with A ranging over context and segregation attributes. The loops calculating t_i's and m_i's (lines 3–5 and 9–11) require at most $k-1$ bitmap and's operations, where k is the total number of context and segregation attributes, and a scan of the `unitID` attribute. This is in the worst case $O(k \cdot |\mathcal{R}|)$. Finally, the loop at lines 14–18 requires $O(n \cdot log\ n)$ for sorting and $O(n)$ for computing an index – in fact, all calculations in Table 1 require constant time. In summary, the worst-case time complexity of Algorithm 1 is $O(\pi \cdot (k \cdot |\mathcal{R}| + n \cdot log\ n))$, namely it is linear in the size of the relation $\mathcal{R}$ and in the number of units, but exponential in the number of attributes of the relation. Since π is an upper bound to the number of itemsets to iterate over, the performances of the algorithm will be inversely proportional to the the minimum support threshold. We will present actual performances on a large dataset in Sect. 4.

Consider now space complexity. Let $\delta = \sum_A |dom(A)|$ be the sum of the sizes of domains of context and segregation attributes. Space complexity is $\Theta(\delta \cdot |\mathcal{R}|)$. Recall, however, that the dataset is stored in memory using (compressed) bitmaps.

Finally, one could consider whether Apriori-like optimizations could be used to directly compute the iceberg segregation data cube (without first computing the whole segregation data cube), namely only the cells with segregation index higher than a given threshold α. Unfortunately, this is not possible due to properties **(P4)** and reversed **(P5)**: $D(\mathbf{A}, \mathbf{B})$ is not necessarily greater than or equal to $D(\mathbf{A}, (\mathbf{B}, B_1))$, where B_1 is an item not in $\mathbf{B}$.

3.6 Multi-valued attributes

In our case study, we will make use of multi-valued attributes in the input relation $\mathcal{R}$. Normally, a tuple σ maps an attribute A to a single value $\sigma(A)$ in its domain, i.e., $\sigma(A) \in dom(A)$. For multi-valued attributes, we admit instead $\sigma(A) \subseteq dom(A)$, as in e.g., $\sigma(\texttt{owns}) = \{\texttt{house}, \texttt{car}\}$. The frequent itemset mining framework allows for a smooth generalization of our approach to include multi-valued attributes. In fact, the input dataset $\mathcal{R}$ can be seen as a transaction database obtained by mapping a tuple σ in $\mathcal{R}$ into a transaction:

$$\{A_1 = \sigma(A_1), \ldots, A_k = \sigma(A_k), \texttt{unitID} = \sigma(\texttt{unitID})\}$$

where $A_1, \ldots, A_k$ are the context and segregation attributes in $\mathcal{R}$. For a multi-valued attribute A_i, the mapping $A_i = \sigma(A_i)$ will become $\cup_{v \in \sigma(A_i)} \{A_i = v\}$.

Example 10 A tuple σ such that $\sigma(\texttt{owns}) = \{\texttt{house}, \texttt{car}\}$ is mapped to a transaction including:

$$\{\texttt{owns} = \texttt{house}, \texttt{owns} = \texttt{car}\}$$

Since our segregation framework builds on the notion of support, which readily applies to the mapped transaction database, it smoothly extends to multi-valued attributes. In particular, the coordinates of a segregation data cube may now include multiple items over a same attributes, e.g., as in $d[\texttt{sex = female}, \texttt{owns=house}, \texttt{owns=car}]$.

4 Case Study

In this section, we challenge the framework for segregation discovery in a complex scenario with a real and large dataset. Our case study targets segregation of minority groups (youngsters, seniors, females) in the boards of companies. The social segregation question we intend to study is: *which minority groups are segregated in the boards of companies and for which type of companies?* A possible answer may lead to the discovery that, e.g., for IT companies, females in a certain age-range appears frequently together in boards and rarely with members of the majority group (men or individuals in other age-ranges).

The case study is challenging in several respects. First, gender segregation in the labour market is a socially relevant problem, with several causes, implications, and policy issues. Case studies, such as [10], have highlighted gender employment segregation in many contexts (university professors, doctors, financial professional, IT technicians, cleaners, retail sector workers, police). Data analysis, however, has been typically conducted at national level, without (the possibility of) drilling-down the investigation in specific sub-sectors. Our explorative approach will allow for achieving this. Second, the data under analysis will consists of a network of relationships (between companies) with no a-priori defined notion of organizational unit. Thus, we will face the problem of how to cluster companies, and their directors, into units for the calculation of segregation indexes.

In the following, we first introduce the notion of social network of companies, then we report basic facts on the case study of the network of Italian companies, and finally challenge the segregation discovery framework on such a case study.

4.1 Social networks of companies

A *director* is a person appointed to serve on the board of a company. The *board of directors* (BoD) is a body of elected or appointed members who jointly oversee the activities of the company. The *presence* of a director is the number of BoDs the director belongs to. If presence is two or higher, the director is called an interlocking director [43, 53]. As an example, the board of a controlled company typically includes directors from the board of the controlling company. Top level managers can be appointed in the board of a company as a means to consolidate partnership with other companies, or to share their expertise and vision. Other reasons for multiple presence include political influence, friendship, kinship, and so on. The presence of a same director in the boards of two companies can then

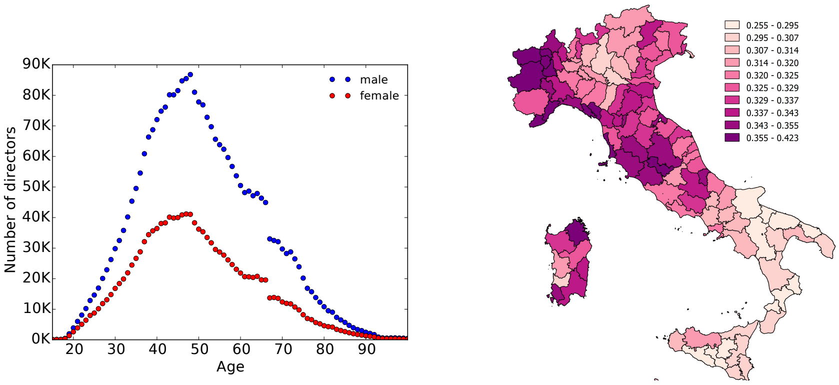

Fig. 3: Director distributions: by gender/age (left), by residence province (right).

be considered a signal of relationships (business, personal, or other) between the two companies [8]. Under this "social tie" assumption, we model a social network of companies by linking those companies that share at least one director.

Formally, let $\mathcal{N} = \{1, \ldots, N\}$ be a set of company IDs, and for $i \in \mathcal{N}$, let $BoD(i) \subseteq \mathcal{D}$ be the board of directors of company i, where $\mathcal{D} = \{1, \ldots, D\}$ is the set of directors IDs. A *social network of companies* is a weighted undirected graph $\mathcal{G} = \langle \mathcal{N}, \mathcal{E} \rangle$ where a weighted edge (i, j, w_{ij}) is in $\mathcal{E} \subseteq \mathcal{N} \times \mathcal{N} \times \mathbb{N}^+$ iff $i \neq j$ and $w_{ij} = |BoD(i) \cap BoD(j)| > 0$, i.e., if companies i and j share at least one director. Intuitively, w_{ij} is a measure of the strength of ties between the boards of directors of i and j. We denote by L the number of edges, i.e., $L = |\mathcal{E}|$. The degree of a node i is the number of edges connecting i to other nodes. A connected component is a maximally connected subgraph of $\mathcal{G}$.

4.2 The social network of Italian companies

The Italian Business Register records information on all Italian companies and directors. The register is managed by the Italian Chamber of Commerce. Data are keep up-to-date by the companies themselves, since the register is recognized by the law as the official source of information about companies. We had a unique access to a complete snapshot of the registry regarding the year 2012. Data on companies stored in the register include legal and financial information.

Data on directors include gender, birth year and city, and city of residence. The age distribution of directors is shown in Figure 3 (left). The plot sadly highlights the glass-ceiling reality for women, who suffer from a under-proportional representativeness in top-level job positions. The plot also shows a net reduction of the number of directors around the age that gives the option for retiring. Figure 3 (right) shows the percentage of female directors over the province[2] of residence. Values range from a minimum of 25% in the historically more depressed regions in the south of Italy to a maximum of 43.5% in the more developed regions.

[2] Level 3 of the Nomenclature of Territorial Units for Statistics (NUTS) [33].

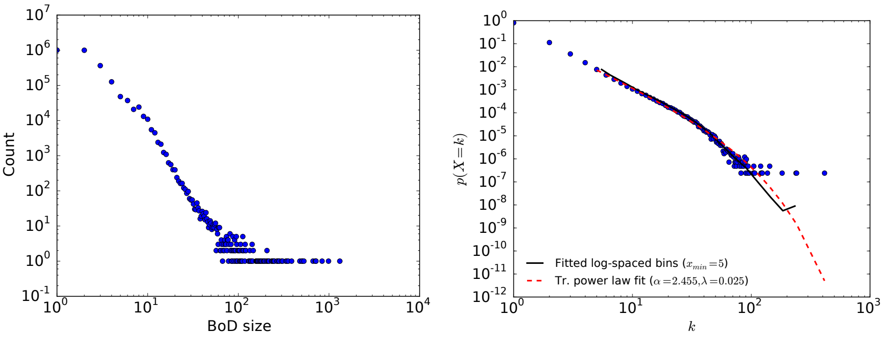

Fig. 4: Distributions: BoD size (left), director presence (right).

A company can be structured as a sole proprietorship, a partnership, a corporation, or other national forms. For corporations, the BoD is elected by shareholders, while for a partnership the BoD includes all partners. We filtered out sole proprietorships, since this type of business does not exist separately from its owner. Similarly, we do not consider companies with only one director who is not shared with any other company.

The social network resulting after preprocessing and filtering raw data includes $N \simeq 2.15 \cdot 10^6$ companies/nodes, and $D \simeq 3.63 \cdot 10^6$ directors. The network has $L \simeq 6.75 \times 10^6$ edges. About $631 \cdot 10^3$ nodes are isolated, i.e., their degree is 0. This amounts at 29.3 % of the total number of nodes, and it is quite representative of the Italian scenario, where tiny/family businesses are widespread. Figure 4 reports the distributions of BoD size and director presence. Distributions are heavily tailed (notice the log-log scale), but only for director presence there is a good fit by a truncated powerlaw[3]. A few directors appear in hundreds of boards, with one appearing in as many as 404 boards. We investigated the reasons of such impressively high numbers, and found two explanations. First, when a company is winding-up because of bankruptcy, an official receiver is appointed by the court as an interim receiver and manager of the company. Such directors are independent experts appointed in many boards and for a possibly long period. Second, there are groups of companies with a pyramidal structure of management and control [1, 55] which share the same directors in their boards. An outlier case that we found consists of a clique of 108 companies having the same person as their unique director. In order to reduce the impact of the two special cases above on the density of the social network of companies, we removed from the set of directors the 0.01% with the highest presence.

4.3 Segregation discovery input

We aim at exploiting the segregation discovery framework of Section 3 to the case study of the social network of Italian companies. The dataset under analysis will

[3] We adopted the methods and software from [2] for fitting heavily tailed distributions. The distribution with the best loglikelihood ratio is selected among power laws, truncated power laws, exponentials, stretched exponentials, and log-normals.

segregation atts			*context atts*		
gender	**age**	**birthplace**	**residence**	**sector**	**unitID**
`M`	`15-38`	`foreign`	`north`	{`education`}	1
`F`	`39-46`	`south`	`south`	{`electricity`, `transports`}	2
`M`	`55-65`	`north`	`south`	{`agriculture`}	1
...	...	...	...	...	...

Table 2: Sample input for segregation discovery.

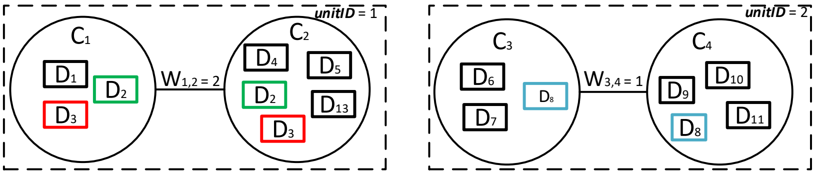

Fig. 5: Sample social network of companies.

have the form of the relation shown in Table 2. A tuple in the dataset regards a director. Segregation attributes include: gender, age, and birth place. Age values are discretized into 5 equal-frequency bins (`15-38`, `39-46`, `47-54`, `55-65`, `66-100`). Birth place can be one region of Italy (`islands`, `south`, `center`, `north-east`, and `north-west`) or any foreign country (`foreign`). This classification corresponds to Level 1 of the Nomenclature of Territorial Units for Statistics (NUTS) [33]. We will also consider finer-grained levels, such as provinces. Context attributes include residence of the director, and the sectors of companies the director seats in the board of. Residence has the same grain of the birth place attribute. Sectors are classified into a number of categories (`agriculture`, `education`, `transportation`, etc.) defined by the Italian institute of statistics. Notice that the sector attribute is multi-valued, since interlocking directors may seat in boards of companies belonging to different industry sectors.

In this section, we discuss two issues that challenge the framework of Section 3, and devise solutions for tackling them.

Segregation index definitions assume a partitioning of individuals into units of social organization (schools, neighborhoods, communities). *The first challenge* in the context of social networks of companies is then to define how such units are defined. Intuitively, a unit is a set of companies within which directors can get in contact, either directly (because they belong to a same BoD) or indirectly (e.g., through an interlocking director connecting two BoDs). Our approach is to consider a structural decomposition of the social network graph into groups of companies, i.e., sub-graphs, each one representing a unit. A natural candidate is to consider the decomposition based on connected components (CCs).

Example 11 Figure 5 shows a sample social network of companies. There are 4 companies (C_1–C_4) and 11 directors (D_1–D_{11}). Edges connect C_1 and C_2 (interlocking directors are D_2 and D_3) and C_3 and C_4 (D_8 is the only interlocking director). There are 2 CCs, which are then the organizational units to be considered.

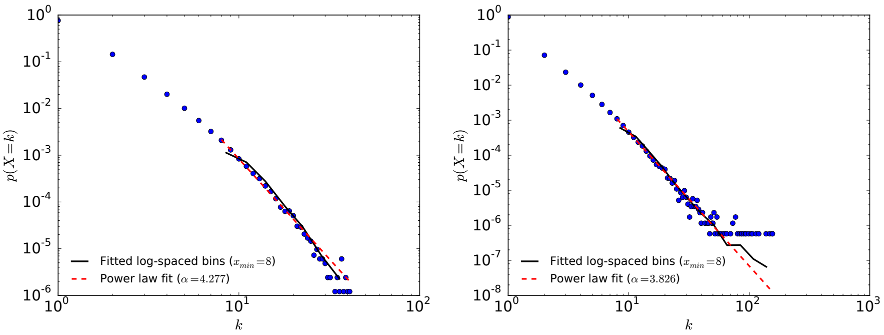

Fig. 6: Distribution of size of CCs before (left, without the giant component) and after (right) splitting the giant component by removing the edges with weight ≤ 3.

The distribution of the size of CCs in the social network of Italian companies is shown in Figure 6 (left). It is fitted by a power law distribution. In addition to the isolated nodes, there are $196 \cdot 10^3$ other CCs with size in the range [2-99], and one giant component consisting of $947 \cdot 10^3$ nodes (not shown in the figure). The total number of CCs is $827 \cdot 10^3$. The giant component accounts for more than 40% of the total number of nodes. This may prevent the discovery of some segregation conditions, which may be hidden in units which are finer-grained than the giant component. We argue then that the giant component needs to be further split. Observe that our assumption that interlocking directors represent signals of relationships between two companies does not account for the strength of such signals. We exploit this intuition to split the giant component into components by removing edges in it that represent “weaker ties”. Recall that the weight of an edge between nodes i and j is $w_{ij} = |BoD(i) \cap BoD(j)|$, i.e., the number of shared directors. We remove edges from the giant component whose weight is lower or equal than a threshold. The selected threshold ($w_{ij} \leq 3$) is the lowest that leads to no giant component. The resulting distribution of CCs, shown in Figure 6 (right), is fitted by a power law with exponent close the the original distribution without the giant component, shown in Figure 6 (left). The total number of CCs is now $\simeq 1.74 \cdot 10^6$. They are the organizational units considered by segregation indexes. In other words, the value of the `unitID` attribute in Table 2 is the ID of the CC a director appears in.

The *second challenge* in our case study originates from the splitting of the giant component. In fact, a side effect is that in the resulting network an interlocking director may appear in two or more units, if the companies the director is in the board of belong to distinct CCs. This situation was accounted for in the extension of the segregation indexes reported in Section 2.4. Therefore, we will consider the multiple occurrences of a director in distinct CCs as distinct individuals, and then as separate tuples in Table 2. Gender, age, birth place, and residence attribute values will be same for each tuple. The sector attribute will be the set of industry sectors of only the companies in the specific CC in which the individual appears as a director.

	H	*G*	*I*	*Int*
D	0.968	0.977	0.280	-0.280
H		0.937	0.364	-0.364
G			0.333	-0.333
I				-1

Table 3: Pearson's correlation between pairs of indexes.

4.4 Segregation discovery findings

The dataset resulting from data preparation consists of $4.88 \cdot 10^6$ tuples. We executed Algorithm 1 on such dataset, setting $minsupp = 100$, which means considering minority groups of at least 100 directors. This section presents a few exploratory analyses over the segregation data cube produced by the algorithm. The advantage of providing a segregation data cube is that an analyst is free to explore sub-cubes of interest defined along any combination of context and segregation attributes values. Our implementation outputs the segregation data cube into a spreadsheet with a Pivot table for multi-dimensional exploration.

Indexes correlation. As a first analysis, we test correlation among the segregation indexes over all 125,318 cells of the segregation data cube. Table 3 shows the Pearson's correlation coefficient between pairs of indexes. The evenness indexes (D, H, and G) are strongly correlated, as already observed in other empirical analyses [40]. There is instead low correlation between evenness indexes and exposure indexes, since they measure different aspects of segregation. Finally, isolation I and Interaction *Int* are obviously negatively correlated because $I + Int = 1$.

Gender segregation by province. The second analysis consists of mapping segregation indexes over the province of residence of directors. For example, D(`gender=F, residence=Pisa`) is the dissimilarity index for the reference population of directors with residence in Pisa province and for the minority group of female directors. A visual representation of dissimilarity and isolation indexes for all Italian provinces is shown in Figure 7. Provinces in the south of Italy have the highest dissimilarity, followed by center provinces, islands, and the north provinces. Contrasting this with the distribution of female directors (see Figure 3 right), it is worth noting how provinces in the center of Italy have a relatively high percentage of female directors, who however result to be more segregated than e.g., in the provinces of islands and of north-east. Isolation follows a similar pattern, except that the south provinces are less isolated than the center provinces. This means that female directors have more chances of getting in contact with male directors in the south compared to the center of Italy. This can be explained by observing that the percentage of female directors in the south is lower than in the center provinces (see Figure 3 right).

Gender segregation by company sector. Similarly to the previous analysis, Figure 8 (left) maps segregation indexes over the sectors of companies. All indexes show a common pattern. Sectors with the highest segregation are: 6 (constructions), 12 (real estate), and 9 (accommodation and food). Sectors with the lowest segregation are: 11 (finance and insurance), 2 (mining), and 5 (water supply).

Foreigner segregation by company sector. A variant of the previous analysis is to consider as minority group the foreigner directors. Actually, we do not have the

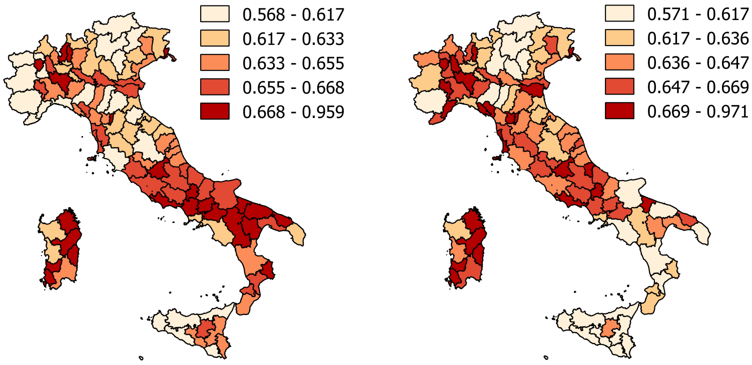

Fig. 7: Dissimilarity (left) and Isolation (right) index for Italian provinces' population and for female director minority group.

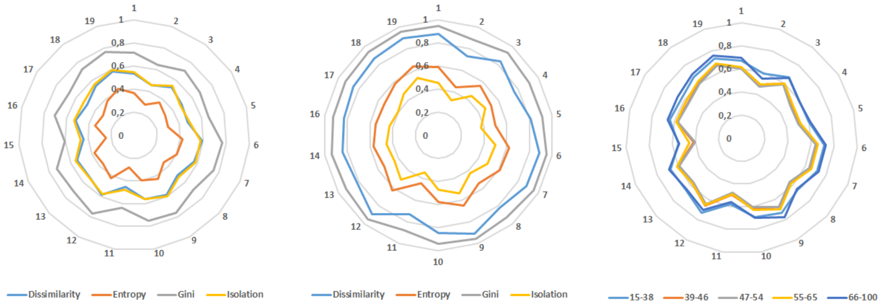

Fig. 8: Segregation indexes over the 19 company sectors for minority groups of: female directors (left), foreign directors (center), and age-band directors (right, only dissimilarity index).

exact information whether a director is Italian or not, but only whether she/he was born in Italy or abroad. Figure 8 (center) maps segregation indexes for directors born abroad over the sectors of companies. Values of evenness indexes are higher than for female segregation at the left hand side plot of the same figure. Foreigners experience an even worse distribution than females among companies in each sector. Immigration studies such as [15] have previously highlighted forms of spatial segregation of foreign workers in Italian cities. Values of isolation are instead lower for foreigners compared to females. Again, this is due to the smaller number of foreigners, hence to lower chances of getting in contact among them in BoDs.

Age-band segregation by company sector. Another variant is to consider whether there is segregation of directors of a specific age-band. Figure 8 (right) shows the dissimilarity index for various age-band groups. Youngsters and elderly directors experience higher dissimilarity values than middle-aged directors across all company sectors. Such values are in between dissimilarity for female and for foreigner directors. Notice that all age-bands have medium-to-high dissimilarity indexes, which means that directors tends to distribute oddly with regard to their ages.

minsupp	Cube size	Elapsed (s)
400	36,945	312
100	125,318	572
25	389,743	1051

Table 4: Cube size and running time of Algorithm 1. Input size: $4.88 \cdot 10^6$ tuples.

Top segregated groups. As a final investigation, we consider looking at the cells in the segregation cube that have the highest values of segregation indexes and, at the same time, a significant size of the minority group. Segregation occurs often when considering the birth place as minority ground. For instance, the cube cell:

`birthplace=Center, sector=9` $(M = 97,173\ T = 506,786\ P = 19.2\%\ D = 0.91)$

shows that directors born in the center regions do not integrate with directors born in other regions in the industry sector 9 (accommodation and food). This is quite reasonable for a very localized sector. For sector 11 (financial and insurance services), a lower value of the index can be observed:

`birthplace=Center, sector=11` $(M = 25,550\ T = 142,672\ P = 17.9\%\ D = 0.79)$

The sub-group of women aged 66 or more appears segregated in several contexts. For instance:

`age=66+, sex=F, residence=South` $(M = 22,912\ T = 817,073\ P = 2.8\%\ D = 0.956)$

shows residential segregation, and:

`age=66+, sex=F, sector=19` $(M = 6,995\ T = 166,827v\ P = 4.2\%\ D = 0.90)$

shows occupational segregation in the industry sector 9 (other services). Segregation indexes for women in sector 1 (agriculture) is not high. For instance, in the south regions we have:

`sex=F, sector=1, residence=South` $(M = 10,604\ T = 38,767\ P = 27.4\%\ D = 0.57)$

and similarly for other regions. This deviates from previous empirical studies that observed segregation in agriculture when considering the whole workforce [16].

4.5 On the efficiency of the approach

Let us finally discuss the efficiency of the Algorithm 1 on the large input dataset. Table 4 reports the size of the segregation data cube and the running time of the algorithm at the variation of the minimum support threshold. The running times refer to the total computation of all of the five segregation indexes considered in this paper. Our implementation is almost entirely in Java 8, with only frequent closed itemset extraction using a C program [11]. The test machine was a commodity PC with Intel Core i5-2410@2.30GHz with 16 Gb of RAM and Windows 10 OS. The running times show that the implementation is fast and scalable to small minimum support thresholds. The size of the segregation cube grows exponentially with lower supports, and this is intuitive since the number of cells depends on the number of frequent closed itemsets. However, the running time grows less than linearly with the size of the segregation data cube. Since frequent closed itemsets with lower support have, by definition, a smaller cover to iterate over (using compressed bitmaps as described in Section 3.4), lowering the minimum support leads to lowering the average time per itemset. Overall, the moderate running time for the large input show that our approach is efficient in pratice. It is worth noting that no multi-core programming was adopted which could futher speedup the approach. Multi-core computing could be exploited for two purposes. First, for frequent closed itemset mining [46], which, however, is only a small percentage of the total running time – less than 10%. Second, for a data parallel execution of the main loop of Algorithm 1. Such a parallelization would be trivial and highly scalable, since there is no dependency between iterations.

5 Related work

This paper is the first to look at segregation from a knowledge discovery perspective. Our approach relies on frequent itemset mining, which is a well-established research area with solid theory [30] and efficient tools [28]. Preliminary results appeared in a conference version of this paper [5]. The extension reported here is significant, and it covers: the Gini and interaction indexes and the characterization of properties of indexes (Section 2), the restriction to closed itemsets in segregation discovery and index computation (Section 3), and a deeper analysis of the case study (Section 4).

The case study presented in this paper targets gender occupational segregation, a relevant social problem with deep roots [10,24]. More specifically, we considered segregation in top company positions such as BoDs. This is a new topic, which adds to related research on social and economic studies of the glass-ceiling effect for women representation[4] in BoDs [13], of wage gap for top positions [3], and of power-concentration in the hands of a small number of directors [19]. Our topic is closely linked to the analysis of benefits of demographic diversity in BoDs [12,47,52].

Another related strand of research concerns the decision dynamics of the corporate boards. [7,8] study the network characteristics of a bipartite graph of directors and companies linked by board membership. The aim is to understand whether the graph structure influences the overall set of strategies and decisions of boards. Bipartite projection [60] over directors consists of a graph with a node for every director, and a link between directors appearing in a same board. For such networks, a high level of homophily has been observed [56]. Bipartite projection over companies consists of our network of companies, namely nodes are companies and edges link two companies that share at least one director. In empirical analysis, such networks have been observed to exhibit a small-world effect [18,36,53], and to include a giant component [51]. The network of our case study (see Sec. 4.3 for details) is at least two orders of magnitude larger than the ones considered by the cited papers.

[4] See also en.wikipedia.org/wiki/Gender_representation_on_corporate_boards_of_directors.

6 Conclusions

We have introduced a knowledge discovery perspective on segregation data analysis by formulating the problem of segregation discovery. This is modelled as a search problem in the space of combinations of reference populations and minority groups. The search is driven by quantitative measures, called segregation indexes, which are taken from the social science literature. Our solution provides an algorithm for constructing a segregation data cube, i.e. a multi-dimensional data cube, for exploratory (OLAP) data analysis. Only cells with distinct population-minority groups are filled, and for which minority size is greater or equal than a minimum threshold. Theory and tools from frequent itemset mining are adopted in the design and implementation of the solution. The approach is challenged on a complex and intriguing case study, concerning segregation of board directors in networks of companies. Here, there is no a-priori defined notion of organizational unit. Thus, we faced the original problem of how to cluster companies, and their directors, into units for the calculation of segregation indexes. The case study is discussed in deep to provide a guidance on the steps necessary for data preparation and cube exploration. The efficiency of the proposed segregation data cube algorithm has been demonstrated on the large input dataset of the case study.

While our approach provides a powerful exploratory tool for segregation analysis, several issues remain open for future investigation. Let us mention two relevant ones. First, a higher layer of analysis on top of our approach must be devised to solve the Simpson's paradox in a given domain of analysis. The problem of choosing the right level of aggregation at which considering segregation indexes can be solved by adopting causal graphs or simulation methods as shown in [50]. Second, segregation discovery is half way towards the more challenging objective of *segregation-aware* data mining and social network analysis. The objective here is the development of *responsible* predictive models, such as link prediction and group recommendation, that, *by design*, can provide quantitative guarantees on the impact of their recommendations over social integration values. As an ethical requirement, such recommender systems should promote suggestions that combat emergent segregation and polarization of social groups, increase exposure to diverse social groups, and improve social ties and cohesion in general.

References

1. Almeida, H.V., Wolfenzon, D.: A theory of pyramidal ownership and family business groups. The Journal of Finance **61**(6), 2637–2680 (2006)
2. Alstott, J., Bullmore, E., Plenz, D.: powerlaw: A Python package for analysis of heavy-tailed distributions. PLoS ONE **9**(1), e85777 (2014)
3. Atkinson, A.B., Piketty, T., Saez, E.: Top incomes in the long run of history. Journal of Economic Literature **1**(49), 3–71 (2011)
4. Bakshy, E., Messing, S., Adamic, L.A.: Exposure to ideologically diverse news and opinion on Facebook. Science **348**(6239), 1130–1132 (2015)
5. Baroni, A., Ruggieri, S.: Segregation discovery in a social network of companies. In: Advances in Intelligent Data Analysis XIV, *LNCS*, vol. 9385, pp. 37–48. Springer (2015)
6. Bastide, Y., Taouil, R., Pasquier, N., Stumme, G., Lakhal, L.: Mining frequent patterns with counting inference. SIGKDD Explorations **2**(2), 66–75 (2000)
7. Battiston, S., Bonabeau, E., Weisbuch, G.: Decision making dynamics in corporate boards. Physica A: Statistical Mechanics and its Applications **322**, 567–582 (2003)

8. Battiston, S., Catanzaro, M.: Statistical properties of corporate board and director networks. The European Physical Journal B **38**(2), 345–352 (2004)
9. Bell, W.: A probability model for the measurement of ecological segregation. Social Forces **32**(4), 357–364 (1954)
10. Bettio, F., Verashchagina, A.: Gender segregation in the labour market: Root causes, implications and policy responses in the EU. Publications Office of the European Union (2009)
11. Borgelt, C.: Frequent item set mining. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **2**(6), 437–456 (2012). www.borgelt.net
12. Burke, R.J.: Company size, board size and numbers of women corporate directors. In: Women on corporate boards of directors, pp. 157–167. Springer (2000)
13. Burke, R.J., Mattis, M.C.: Women on corporate boards of directors: International challenges and opportunities, vol. 14. Springer Science & Business Media (2013)
14. Clark, W.A.V.: Residential preferences and neighborhood racial segregation: A test of the Schelling segregation model. Demography **28**(1), 1–19 (1991)
15. Cristaldi, F.: Immigrazione e territorio: lo spazio con/diviso. Pàtron (2012)
16. Croppenstedt, A., Goldstein, M., Rosas, N.: Gender and agriculture: Inefficiencies, segregation, and low productivity traps. World Bank Research Observer **28**, 79–109 (2013)
17. Das, S., Kramer, A.D.I.: Self-censorship on Facebook. In: Proc. of the Int. Conference on Weblogs and Social Media (ICWSM 2013). The AAAI Press (2013)
18. Davis, G.F., Yoo, M., Baker, W.E.: The small world of the American corporate elite, 1982-2001. Strategic organization **1**(3), 301–326 (2003)
19. Demb, A., Neubauer, F.F.: The corporate board: Confronting the paradoxes. Long range planning **25**(3), 9–20 (1992)
20. Denton, N.A., Massey, D.S.: Residential segregation of Blacks, Hispanics, and Asians by socioeconomic status and generation. Social Science Quarterly **69**(4), 797–817 (1988)
21. Duncan, O.D., Duncan, B.: A methodological analysis of segregation indexes. American Sociological Review **20**(2), 210–217 (1955)
22. Fischer, E.: Distribution of race and ethnicity in US major cities (2011). Published on line at www.flickr.com/photos/walkingsf under Creative Commons licence, CC BY-SA 2.0
23. Flaxman, S., Goel, S., Rao, J.M.: Ideological segregation and the effects of social media on news consumption. Available at SSRN: http://ssrn.com/abstract=2363701 (2013)
24. Flückiger, Y., Silber, J.: The measurement of segregation in the labor force. Springer Science & Business Media (1999)
25. Frey, J.H., Eitzen, D.S.: Sport and society. Annual Review of Sociology **17**, 503–522 (1991)
26. Gastwirth, J.L.: A general definition of the Lorenz curve. Econometrica: Journal of the Econometric Society **39**(6), 1037–1039 (1971)
27. Gentzkow, M., Shapiro, J.M.: Ideological segregation online and offline. Quarterly Journal of Economics **126**(4), 1799–1839 (2011)
28. Goethals, B.: Frequent itemset mining implementations repository (2010). http://fimi.cs.helsinki.fi
29. Grevet, C.: Being nice on the internet: Designing for the coexistence of diverse opinions online. Ph.D. thesis, Georgia Institute of Technology (2016)
30. Han, J., Cheng, H., Xin, D., Yan, X.: Frequent pattern mining: Current status and future directions. Data Mining and Knowledge Discovery **15**(1), 55–86 (2007)
31. Han, J., Kamber, M., Pei, J.: Data Mining: Concepts and Techniques, 3rd edn. Morgan Kaufmann Publishers Inc. (2011)
32. Hutchens, R.M.: Segregation curves, Lorenz curves, and inequality in the distribution of people across occupations. Mathematical Social Sciences **21**(1), 31–51 (1991)
33. International Organization for Standardization: ISO 3166-1:2013 International standard for country codes and codes for their subdivisions (2013)
34. James, D.R., Tauber, K.E.: Measures of segregation. Sociological Methodology **13**, 1–32 (1985)
35. Kaser, O., Lemire, D.: Compressed bitmap indexes: Beyond unions and intersections. Software: Practice and Experience **46**, 167–198 (2016). https://github.com/lemire/javaewah
36. Kogut, B., Walker, G.: The small world of Germany and the durability of national networks. American Sociological Review **66**, 317–335 (2001)
37. Loy, J.W., Elvogue, J.F.: Racial segregation in American sport. International Review for the Sociology of Sport **5**(1), 5–24 (1970)
38. Maes, M., Bischofberger, L.: Will the personalization of online social networks foster opinion polarization? Available at SSRN: http://ssrn.com/abstract=2553436 (2015)

39. Massey, D.S.: Segregation and the perpetuation of disadvantage. The Oxford Handbook of the Social Science of Poverty pp. 369–393 (2016)
40. Massey, D.S., Denton, N.A.: The dimensions of residential segregation. Social Forces **67**(2), 281–315 (1988)
41. Massey, D.S., Rothwell, J., Domina, T.: The changing bases of segregation in the United States. Annals of the American Academy of Political and Social Science **626**, 74–90 (2009)
42. Mitchell, T.: Machine Learning. The Mc-Graw-Hill Companies, Inc. (1997)
43. Mizruchi, M.S.: What do interlocks do? An analysis, critique, and assessment of research on interlocking directorates. Annual Review of Sociology **22**(1), 271–298 (1996)
44. Mora, R., Ruiz-Castillo, J.: Entropy-based segregation indices. Sociological Methodology **41**, 159–194 (2011)
45. Musterd, S.: Social and ethnic segregation in Europe: Levels, causes, and effects. Journal of Urban Affairs **27**(3), 331–348 (2005)
46. Negrevergne, B., Termier, A., Rousset, M.C., Méhaut, J.F.: Para Miner: a generic pattern mining algorithm for multi-core architectures. Data Mining and Knowledge Discovery **28**(3), 593–633 (2014)
47. Ooi, C.A., Hooy, C.W., Som, A.P.M.: Diversity in human and social capital: Empirical evidence from Asian tourism firms in corporate board composition. Tourism Management **48**, 139 – 153 (2015)
48. Pariser, E.: The Filter Bubble: What the Internet is hiding from you. Penguin UK (2011)
49. Pearl, J.: Causality: Models, Reasoning, and Inference, 2 edn. Cambridge University Press, New York, USA (2009)
50. Pearl, J.: Comment: Understanding simpsons paradox. The American Statistician **68**(1), 8–13 (2014)
51. Piccardi, C., Calatroni, L., Bertoni, F.: Communities in Italian corporate networks. Physica A: Statistical Mechanics and its Applications **389**(22), 5247–5258 (2010)
52. Randøy, T., Thomsen, S., Oxelheim, L.: A nordic perspective on corporate board diversity. Tech. Rep. 0.5428 (2006)
53. Robins, G., Alexander, M.: Small worlds among interlocking directors: Network structure and distance in bipartite graphs. Computational & Mathematical Organization Theory **10**(1), 69–94 (2004)
54. Romei, A., Ruggieri, S.: A multidisciplinary survey on discrimination analysis. The Knowledge Engineering Review **29**(5), 582–638 (2014)
55. Romei, A., Ruggieri, S., Turini, F.: The layered structure of company share networks. In: Proc. of the IEEE Int. Conference on Data Science and Advanced Analytics (DSAA 2015), pp. 1–10. IEEE Computer Society (2015)
56. Sankowska, A., Siudak, D.: The small world phenomenon and assortative mixing in Polish corporate board and director networks. Physica A: Statistical Mechanics and its Applications **443**, 309–315 (2016)
57. Schelling, T.C.: Dynamic models of segregation. Journal of Mathematical Sociology **1**(2), 143–186 (1971)
58. Smith, S.L., Choueiti, M.: Black characters in popular film: Is the key to diversifying cinematic content held in the hand of the black director. Annenberg School for Communication & Journalism. Retrieved March **12**, 2013 (2011)
59. Xu, K.: How has the literature on Gini's index evolved in the past 80 years? Economics working paper, Dalhousie University (2003). Available at SSRN: http://ssrn.com/abstract=423200
60. Zhou, T., Ren, J., Medo, M., Zhang, Y.C.: Bipartite network projection and personal recommendation. Physical Review E **76**(4), 046115 (2007)