Segregation Discovery in a Social Network of Companies

Alessandro Baroni and Salvatore Ruggieri

Dipartimento di Informatica, Università di Pisa
Largo B. Pontecorvo 3, 56127 Pisa, Italy
{baroni,ruggieri}@di.unipi.it

Abstract. We introduce a framework for a data-driven analysis of segregation of minority groups in social networks, 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 consists of searching sub-graphs and sub-groups for which a reference segregation index is above a minimum threshold. A search algorithm is devised that solves the segregation problem. The framework is challenged on the analysis of segregation of social groups in the boards of directors of the real and large network of Italian companies connected through shared directors.

1 Introduction

Social networking services record our connections to friends, colleagues, collaborators. The analysis of those digital traces can create new comprehensive pictures of individual and group behaviour, through the discovery of patterns and models, with the potential to transform the understanding of our lives, organizations, and societies. In this paper, we will consider the social problem of group *segregation* in social networks [8], which is an unjustified separation or distance in social environments (physical, working, or on-line) of individuals on the basis of any physical or cultural trait. We present theory and tools, based on data mining and network science, for data-driven *segregation discovery*, with two main goals. First, we aim at providing 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. Second, we aim at studying the applicability of proposed methodology in a complex scenario through the analysis of segregation of minority groups in the network of Italian companies linked through shared directors in their boards.

The paper is structured as follows. Section 2 provides an overview of segregation indexes from the social science literature. Section 3 introduces the problem of segregation discovery and provides a solution using concepts from itemset mining. Section 4 challenges the solution on the network of Italian companies by tackling a few issues arising from the case study. Section 5 concludes and presents directions for future work.

Fig. 1. Racial spatial segregation in New York City, based on Census 2000 data [7]. 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.

2 Segregation Indexes

2.1 On the Notion of Segregation

The term *segregation* refers to restrictions on the access of people to each other. 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. Contact, communication, or interaction among groups are limited by their physical, working or socio-economic distance. Members of a group tend to cluster together when dissecting the society into organizational units (neighborhoods, schools, job types).

In spatial segregation, 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 Fig. 1, racial segregation very often emerges in most cities characterized by ethnic diversity. Schelling's segregation model [19] 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, Schelling assumed that a person changes residence only if less than 30% of the neighbors are of his/her own race.

Recently, [13] 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. For instance, it has been warned that the personalization of online social networks may foster segregation and lack of consensus between different social groups, because people are only reinforced in what they already believe and lack exposure to alternative viewpoints and information [16] or because they are led to self-censorship acts [6] for fear of public opinion on personal thoughts.

2.2 Segregation Indexes

A segregation index provides a quantitative measure of the degree of segregation of social groups (e.g., Blacks and Whites) among units of social organization (e.g., schools). Many indexes have been proposed in the literature. [10] represents the earliest attempt to categorize them. Afterward, the survey [12] provided a shared classification with reference to five key dimensions: evenness, exposure, concentration, centralization, and clustering. We restrict ourselves to binary indexes, which assume a partitioning of people into two groups, say majority and minority (but could be Blacks/Whites, women/men, etc.). Let T be size of the total population, M be the size of the minority group, and $P = M/T$ be the overall fraction of the minority group. Assume that there are n units, and that for $i \in [1, n]$, t_i is the population in unit i, m_i is the minority population in unit i, and $p_i = m_i/t_i$ is the fraction of the minority group in unit i.

Evenness indexes. *Evenness* indexes measure the difference in the distributions of social groups among the units. They are widely adopted in the social science literature on segregation. The mostly referenced indexes are dissimilarity and entropy. The *dissimilarity index* D is the weighted mean absolute deviation of every unit's minority proportion from the global minority proportion:

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

The normalization factor $2 \cdot P \cdot (1 - P)$ is to obtain values 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).

A second widely adopted index is the *information index*, also called the *Theil index* [15] in social science, and normalized mutual information in machine learning. 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:

$$H = \sum_{i=1}^{n} \frac{t_i}{T} \cdot \frac{(E - E_i)}{E} \tag{2}$$

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 all units contain only one group (complete segregation).

Exposure indexes. *Exposure* indexes measure the degree of potential contact, or possibility of interaction, between members of different groups.

The most used measure of exposure is the *isolation index* [4], 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:

$$I = \frac{1}{M} \cdot \sum_{i=1}^{n} m_i \cdot p_i \tag{3}$$

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 in the same conditions of the previous two indexes. The differences between indexes are the following: (a) H and I are insensitive to units i where $m_i = 0$, whilst D is not; (b) D and H are symmetric, i.e., by inverting the minority and majority groups the index remains unchanged, whilst I is not.

Other indexes. The other three classes of indexes are specifically concerned with spatial notions of segregation. Concentration measures the relative amount of physical space occupied by social groups in an urban area. Centralization measures the degree to which a group is spatially located near the center of an urban area. Finally, clustering measures the degree to which group members live disproportionately in contiguous areas. We refer the reader to [12] for details.

3 Segregation Discovery

Traditional data analysis approaches from social science 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. For instance, a suspect case of segregation of black female students in high schools from NYC is studied first by collecting data on race and gender of all high school students in NYC (reference population), and then by computing and analysing segregation indexes over black females (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. 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. The analyst has to collect data on gender, age, race of students (called segregation attributes), and on city location, school type, and annual fees (called context attributes). 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 NYC), a specific minority group denoted by a combination of segregation attributes (e.g., black women) 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.

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*. The problem statement will be formalized using notation and concepts from itemset mining [9]. This allows for re-using methods and tools from this widely investigated research area. In particular, itemsets 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 in the table denote individuals, and attribute values denote information about individuals and units they belong to. Attributes are partitioned into *segregation attributes* (SA), such as `sex`, `age`, and `race`, which denote minority groups potentially exposed to segregation; *context attributes* (CA) attributes, such as `city` and `job type`, which denote contexts where segregation may appear; and an attribute `unit`, which is an ID of the unit the tuple/individual belongs to. For a discrete attribute A, an *A-item* is a term $A = v$, where $v \in dom(A)$, the domain of A. We assume that continuous attributes are discretized into bins [11]. 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 $\sigma[A] = v$, where $\sigma[A]$ is the value of the attribute A in the tuple σ. The *cover* of $\mathbf{X}$ is the set of 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. E.g., $cover$(`sex=female, age=[20-29]`) is the set of women in their 20s included in the dataset. The (absolute) *support* of $\mathbf{X}$ is the size of its cover, namely $supp(\mathbf{X}) = |cover(\mathbf{X})|$.

We write $\mathbf{A}, \mathbf{B}$ to denote an itemset where $\mathbf{A}$ is non-empty and it includes only SA-items, and $\mathbf{B}$ includes only CA-items. We call $\mathbf{A}$ a non-empty SA-itemset, and $\mathbf{B}$ a CA-itemset. We are now in the position to extend the notation of the segregation indexes to a reference population, which is the cover of $\mathbf{B}$, and to a reference minority group, which is the cover of $\mathbf{A}$.

Definition 1. *Let $s()$ be a segregation index. For an itemset $\mathbf{A}, \mathbf{B}$ 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 $cover(\mathbf{A}, \mathbf{B})$.*

As an example, $D(\mathbf{A}, \mathbf{B})$ is the dissimilarity index, where $T = supp(\mathbf{B})$, $M = supp(\mathbf{A}, \mathbf{B})$, $t_i = supp(\mathbf{B}, \texttt{unit=}i)$, and $m_i = supp(\mathbf{A}, \mathbf{B}, \texttt{unit=}i)$. Reconsidering the example above, we would fix $\mathbf{A}$ as `race=black,sex=female` and $\mathbf{B}$ as `city=NYC`. $D($(`race=black,sex=female`), `city=NYC`$)$ is then the dissimilarity index of segregation of black females in the high schools of NYC.

We introduce now the problem of segregation discovery.

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

An itemset $\mathbf{A}, \mathbf{B}$ is α-integrative w.r.t. $s()$ if $cover(\mathbf{B}) = \emptyset$ or $s(\mathbf{A}, \mathbf{B}) \leq \alpha$. 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 finding itemsets $\mathbf{A}, \mathbf{B}$ denoting a minority sub-group (non-empty $\mathbf{A}$) and a non-trivial context ($\mathbf{B}$ with non-empty cover) where the segregation index $s()$ is above the α threshold.

Input: relational table $\mathcal{R}$ with context attributes (CA), segregation attributes (SA), unit attribute `unit` with a total of n units.
Output: segregation index values $s(\mathbf{A}, \mathbf{B})$.

```
1  foreach B CA-itemset do
2  |  T = 0;
3  |  foreach i ∈ [1, n] do
4  |  |  t_i = supp(B, unit=i);
5  |  |  T += t_i
6  |  end
7  |  foreach A non-empty SA-itemset do
8  |  |  M = 0;
9  |  |  foreach i ∈ [1, n] with t_i > 0 do
10 |  |  |  m_i = supp(A, B, unit=i);
11 |  |  |  M += m_i
12 |  |  end
13 |  |  sum = 0
14 |  |  foreach i ∈ [1, n] with t_i > 0 do
15 |  |  |  sum += f_s(m_i, t_i, M, T)
16 |  |  end
17 |  |  s(A, B) = g_s(sum, M, T)
18 |  end
19 end
```

Algorithm 1: Segregation index computation.

Algorithm 1 is a solution to the problem of computing $s(\mathbf{A}, \mathbf{B})$ for a segregation index $s()$ and all itemsets $\mathbf{A}, \mathbf{B}$. It can readily solve the segregation problem by filtering itemsets whose index is lower or equal than the threshold α. We assumes that the support counting function *supp*() is available. We implemented it by storing the subset of $\mathcal{R}$ at each unit as an array of bitmaps, one bitmap per each CA and SA item. Position i of a bitmap is set to 1 iff the i^{th} tuple of the unit supports the item associated to the bitmap. Support counting consists then of bitmap and-operations. An alternative way of implementing *supp*() is through the construction of an FP-tree, a compressed representation of a dataset used for frequent itemset mining [9]. The outer loop (lines 1-19) of the algorithm iterates over all CA-itemsets $\mathbf{B}$. For each of them, the total population size T is calculated at lines 3–6. The inner loop (lines 7–18) iterates over all non-empty SA-itemsets $\mathbf{A}$. First, the size M of the minority is calculated at lines 9–12. We accumulate the results of a function $f_s()$ over each unit, and then pass it to the normalization function $g_s()$. These two functions depend on the segregation index $s()$. For the information index, we observe that $H = 1-(\sum_{i=1}^{n} t_i \cdot E_i)/(T \cdot E)$. Hence, $f_s(m_i, t_i, M, T) = t_i \cdot E_i$ and $g_s(sum, M, T) = 1 - sum/(T \cdot E)$, where E_i and E are clearly calculable from m_i, t_i and from M, T respectively.

Let $\delta = \sum_A |dom(A)|$ be the sum of the sizes of domains of context and segregation attributes, and $\pi = \prod_A |dom(A)|$ be their product. Algorithm 1 has worst-case time complexity of $O(\pi|\mathcal{R}|)$. Our bitmap-based implementation has space complexity of $\Theta(\delta|\mathcal{R}|)$. We will present actual performances on a large

dataset later on. Notice that Apriori-like optimizations in the index calculations are not possible since D and H are not anti-monotonic, and I is monotonic only w.r.t $\mathbf{A}$ – i.e., $I(\mathbf{A} \cup \mathbf{A}', \mathbf{B}) \leq I(\mathbf{A}, \mathbf{B})$.

4 Segregation Discovery in Social Networks of Companies

We will challenge the framework for segregation discovery in a complex scenario with a real and large dataset. We are interested in studying 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 appear frequently together in boards and rarely with members of the majority group (men or individuals in other age-ranges). In the following, we first introduce the notion of social network of companies, then report some basic facts on the running 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

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 a *bridge director*. As an example, the board of a controlled company typically includes directors from the board of the controlling company. Other reasons for multiple presence include partnership consolidation, collusion, cooptation, monitoring, political influence, friendship, kinship, etc. The presence of a same director in the boards of two companies (*interlocking directorate*) can then be considered a signal of business, personal, or other forms of relationship and information exchange between the two companies [14]. Under this "social tie" assumption, we model a social network of companies by linking those companies that share at least one director [3].

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) is in $\mathcal{E} \subseteq \mathcal{N} \times \mathcal{N} \times \mathbb{R}$ iff $w = |BoD(i) \cap BoD(j)| > 0$, i.e., if companies i and j share at least one director. Intuitively, w is a measure of the strength of ties between the boards of directors of i and j. We write $e_{ij} = 1$ if $(i, j, w) \in \mathcal{E}$, and $e_{ij} = 0$ otherwise. We denote by L the number of edges, i.e., $L = |\mathcal{E}|$, and by k_i the degree of node i, i.e., $k_i = \sum_{j=1}^{N} e_{ij}$. A node is called *isolated* if its degree is 0. A connected component (CC) is a maximally connected subgraph of $\mathcal{G}$.

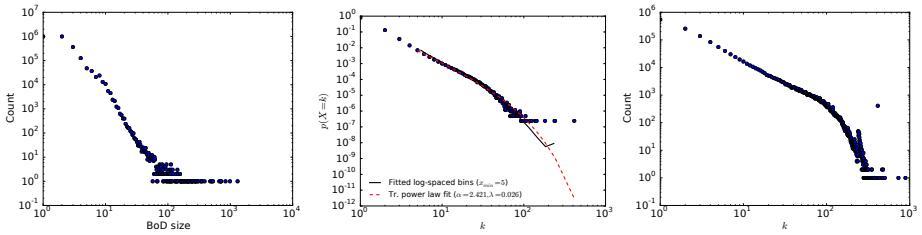

Fig. 2. Distrib.: BoD size (left), director presence (center), node degree (right).

4.2 The Social Network of Italian Companies

The Italian Business Register records information on all Italian companies and directors. We had a unique access to a complete 2012 snapshot of the registry. A company can be structured as 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.

There is a total of $N \simeq 2.2 \cdot 10^6$ registered companies, and $D \simeq 3.7 \cdot 10^6$ directors. The network has $L \simeq 5.9 \times 10^6$ edges. Around $0.7 \cdot 10^6$ nodes are isolated (i.e., degree is 0). This amounts at 35.2 % of the total number of nodes, and it is quite representative of the Italian scenario, where tiny/family businesses are widespread. Fig. 2 reports the distributions of BoD size, director presence, and node degree. Distributions are heavily tailed (notice the log-log plot), but only for director presence there is a good fit by a truncated powerlaw (we used the software from [2]). A few directors appear in hundreds of boards (one of them appears 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 [1] sharing the same directors. An example is the outlier in Fig. 2 (right), representing a clique of 250 companies having a same person as the unique director in their boards. 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. The age distribution of directors is shown in Fig. 3 (left). The plot sadly highlights the glass-ceiling reality for women, who suffer from a under-proportional representativeness in top-level job positions.

4.3 Segregation Discovery

We aim at exploiting the segregation discovery framework and algorithm of Section 3 to the case study of the social network of Italian companies. The dataset under analysis will contain one tuple for each director. Available segregation attributes include: gender, age (discretized into 5 equal-frequency bins). Context

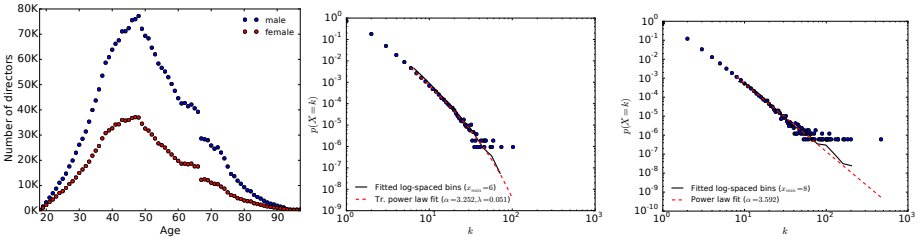

Fig. 3. Left: age distribution. Center: distribution of size of CCs before (center, without the giant component) and after (right) splitting the giant component.

attributes include the company sector (the top level of a hierarchical classification used by the Italian official statistics institute), with 21 possible values, and the region of residence of the director (north-east, north-west, center, south, islands, abroad). In this section, we discuss three issues that challenge the framework of Section 3, and devise solutions for overcoming 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 a bridge 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). The distribution of the size of CCs, shown in Fig. 3 (center), is fitted by a truncated powerlaw. In addition to the isolated nodes, there are $251 \cdot 10^3$ other CCs with size in the range [2-99], and one giant component consisting of $642 \cdot 10^3$ nodes (not shown in the figure). The number of directors in the giant component amounts at 20% of the total. This means that the giant component weights 20% in the calculation of dissimilarity and information gain segregation indexes (for the isolation index, the weight depends also on the size of minority m_i). This may prevent segregation from being discovered, because the giant component may hide segregated finer-grained units within it. We claim that the giant component need to be further split. Observe that our assumption that bridge 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 $w = |BoD(i) \cap BoD(j)|$, 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 ($w \leq 3$) is the lowest that leads to no giant component. The resulting distribution of CCs, shown in Fig. 3 (right), is fitted by a powerlaw

with exponent close the the original distribution without the giant component. The total number of CCs is now $\simeq 1.6 \cdot 10^6$.

The *second challenge* in segregation discovery originates by the splitting of the giant component. In fact, a side effect of *any* splitting is that in the resulting network a bridge director may appear in two or more units. This is not accounted for in the framework of Section 3, which assumes that an individual belongs to only one unit. We will consider multiple instances of bridge directors in different units as distinct individuals. With reference to the notation of Sect. 2.2, we revise the definitions of the size of population T and minority group M by setting $T = \sum_i^n t_i$ and $M = \sum_i^n m_i$, i.e., by counting every *occurrence* of an individual in any unit, not every individual. Algorithm 1 remains unchanged because it already computes T and M as above.

The *third challenge* is motivated by the need of including characteristics of companies among the context attributes, so that segregation, e.g., in the subnetwork of IT companies, can be discovered. However, bridge directors may appear in BoDs of companies with different characteristics. How do we model this in our framework? We use multi-valued attributes, by admitting that, for an attribute A and a tuple σ, $\sigma[A] \subseteq dom(A)$ (instead of simply, $\sigma[A] \in dom(A)$). As an example, the industry sector of a director is defined as the set of industry sectors of companies where the director appears, e.g., $\sigma[\texttt{sector}] = \{$ `IT`, `Banks` $\}$. Our framework can be extended to admit multi-valued tuples by simply extending the notion of support as follows: a tuple σ *supports* $\mathbf{X}$ if for every $A = v$ in $\mathbf{X}$, we have $v \in \sigma[A]$ if A is multi-valued, and $\sigma[A] = v$ otherwise. On the implementation side, this extension does not require drastic changes. The support counting method has to be initialized with a set of transaction items $A = v_1, \ldots, A = v_k$ for $\{v_1, \ldots, v_k\} = \sigma[A]$ instead of simply with $A = v$ for $v = \sigma[A]$. In our bitmap based implementation, for a multi-valued attribute A, a tuple σ will lead to set to 1 all the bitmaps of the values in $\sigma[A]$.

4.4 Segregation Discovery: Findings

The dataset processed as described in the previous section consists of $4.6 \cdot 10^6$ tuples, 2 context attributes (residence, sector), 2 segregation attributes (age, sex), and the `unit` attribute. We have applied Algorithm 1 on the dataset to calculate the D, H, and I segregation indexes. The total running time of the algorithm was of 110 seconds, on a commodity PC with Intel Core i5-2410@2.30GHz with 16 Gb of RAM, Windows 7 OS, and Java 8 as programming language.

The affordable running time allows for more advanced data analysis than the one stated by the definition of segregation discovery, namely selecting/ranking itemsets $\mathbf{A}, \mathbf{B}$ whose index is above a given threshold. We are in the position of providing the segregation analyst with a data cube of indexes for exploratory analysis in the style of OLAP cubes. Here, indexes play the role of metrics, and context and segregation attributes play the role of dimensions. Also, constraints on the sizes T (resp., M) of the population (resp., minority group) can be provided to guide the analysis.

Let us present here three real cases. By setting a minimum $M \geq 10^3$, the itemset with the highest dissimilarity index:

`sector='agriculture', age='<=38', sex='F'` ($D = 0.916, H = 0.605, I = 0.431$)

regards the population of directors of in the agriculture sector, with women up to 38 years old as minority population. Segregation in agriculture is a well-known phenomenon. Excluding such a sector, the highest information index is for:

`residence='abroad', age='>=53'` ($D = 0.75$, $H = 0.675$, $I = 0.805$)

the population of directors with residence abroad, and for the minority of directors with age of 53 years or more. Finally, excluding foreign directors, the highest isolation index is for:

`sector='electricity', sex='M'` ($D = 0.625$, $H = 0.411$, $I = 0.907$)

directors of companies producing or supplying electric power or gas, with minority population the male directors. In this case, segregation of males means they have 90.7% of likelihood of getting in contact with other males in their board or through bridge directors.

5 Conclusions and Future Work

We have formulated the problem of segregation discovery in social networks, devised a solution that provides the data analyst with a data cube of segregation indexes for exploratory analysis, and challenged the approach on a complex scenario with a real and large dataset regarding segregation in boards of directors.

Several issues remain open for future investigation.

First, relations with research streams that appear closely linked must be explored. One related field is community discovery in attributed graphs [5], where graph clustering algorithms exploit both the structural dimension of the social graphs as well as a compositional dimension represented by features of nodes. Another related field is discrimination discovery [18], where the objective is to search for contexts with a disproportionate distribution of socially sensitive decisions (granting of a loan, admission to school, hiring, etc.) among social groups.

Second, the proposed framework need to be further validated, e.g., on whether it is able to cover more complex segregation index definitions and application scenarios, and on whether Algorithm 1 scales to a large number of attributes. The impact of different segregation indexes on the top segregative itemsets should also be evaluated, as done in [17] for discrimination indexes. The final objective will be a complete framework and working system for OLAP analysis of segregation in social networks.

Finally, we argue that 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 *segregation-aware* data analysis and data mining models that, by design, can provide a guarantee about the impact of computer-supported decisions (e.g., link predictions, group recommendation) on individuals and social groups, about the possibilities of interaction between them, and about the increase of social cohesion of society at large.

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 (2004)
3. Battiston, S., Catanzaro, M.: Statistical properties of corporate board and director networks. The European Physical Journal B 38(2), 345–352 (2004)
4. Bell, W.: A probability model for the measurement of ecological segregation. Social Forces pp. 357–364 (1954)
5. Bothorel, C., Cruz, J.D., Magnani, M., Micenková, B.: Clustering attributed graphs: Models, measures and methods. Network Science FirstView, 1–37 (2015)
6. 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)
7. Fischer, E.: Distribution of race and ethnicity in US major cities (2011), published on line at http://www.flickr.com/photos/walkingsf under Creative Commons licence, CC BY-SA 2.0
8. Freeman, L.C.: A set of measures of centrality based on betweenness. Sociometry pp. 35–41 (1977)
9. 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)
10. James, D.R., Tauber, K.E.: Measures of segregation. Sociological Methodology 13, 1–32 (1985)
11. Liu, H., Hussain, F., Tan, C.L., Dash, M.: Discretization: An enabling technique. Data mining and knowledge discovery 6(4), 393–423 (2002)
12. Massey, D.S., Denton, N.A.: The dimensions of residential segregation. Social forces 67(2), 281–315 (1988)
13. 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)
14. 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)
15. Mora, R., Ruiz-Castillo, J.: Entropy-based segregation indices. Sociological Methodology 41, 159–194 (2011)
16. Pariser, E.: The Filter Bubble: What the Internet is hiding from you. Penguin UK (2011)
17. Pedreschi, D., Ruggieri, S., Turini, F.: A study of top-k measures for discrimination discovery. Proc. of ACM Int. Symposium on Applied Computing (SAC 2012). pp. 126–131. ACM (2012)
18. Romei, A., Ruggieri, S.: A multidisciplinary survey on discrimination analysis. The Knowledge Engineering Review 29(5), 582–638 (2014)
19. Schelling, T.C.: Dynamic models of segregation. Journal of mathematical sociology 1(2), 143–186 (1971)