# SCube: A Tool for Segregation Discovery

Alessandro Baroni
University of Pisa, Italy
baroni@di.unipi.it

Salvatore Ruggieri
University of Pisa, Italy
ruggieri@di.unipi.it

## ABSTRACT

Segregation is the separation of social groups in the physical or in the online world. Segregation discovery consists of finding contexts of segregation. In the modern digital society, discovering segregation is challenging, due to the large amount and the variety of social data. We present a tool in support of segregation discovery from relational and graph data. The SCube system builds on attributed graph clustering and frequent itemset mining. It offers to the analyst a multi-dimensional segregation data cube for exploratory data analysis. The demonstration first guides the audience through the relevant social science concepts. Then, it focuses on scenarios around case studies of gender occupational segregation. Two real and large datasets about the boards of directors of Italian and Estonian companies will be explored in search of segregation contexts. The architecture of the SCube system and its computational efficiency challenges and solutions are discussed.

## 1 SOCIAL SEGREGATION

Ethical issues in data and knowledge management are gaining momentum in the last few years. In addition to the traditional field of privacy, techniques for data analysis are being designed or enhanced to take into account moral values such as fairness, transparency, accountability, and diversity[1]. We have recentely developed a novel data-driven technique for addressing segregation of social groups through multi-dimensional data analysis [4]. The approach is implemented in the SCube system, which we propose to demonstrate using real case studies.

*Social segregation* refers to the "separation of socially defined groups" [11]. 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 to a national minority, etc. Contact, communication, or interaction among groups are limited by their physical, working or socio-economic distance. This can be observed when dissecting society in organizational units (neighborhoods, schools, job types). Due to the ubiquitous presence and pervasiveness of ICT, segregation is shifting from ancient forms of well explored spatial segregation[2] to novel forms of digital segregation. For instance, it has been warned that the filter bubble generated by personalization of online social networks may foster idelogical segregation [6], opinion polarization [10], and informational segregation. A data-driven technology that enables the assessment of the extent, nature, and trends of social segregation in the offline or online world, is of extreme interest for a wide audience: social scientistics, public policy makers, regulation and control authorities, professional associations, civil rights societies, and investigative journalists. Business decision makers should also care of business practices, particularly automated decision making, that segregate customers and products through stereotypes, because this limits diversity and reduces opportunities of cross-selling. Finally, data scientists and professionals should be aware of the unintended consequences of their models (recommender systems, link suggestion systems, classifiers) on the cohesion of society at large.

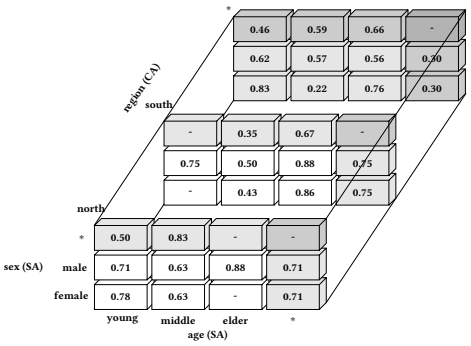



**Figure 1: A segregation data cube with dissimilarity index.**

## 2 SEGREGATION DISCOVERY

From a data analysis perspective, the key problem of assessing social segregation has been investigated so far by hypothesis testing, i.e., by formulating one or more possible contexts of segregation against a certain social group, and then in empirically testing such hypotheses. Such an approach is currently supported by statistical tools, such as the R packages *OasisR*[3] and *seg*[4] [9], or by GIS tools such as the *Geo-Segregation Analyzer*[5] [2]. The formulation of an 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 addition, exploration of multiple hypothesis can be time consuming, since data have to be processed multiple times. Finally, this approach is subject to erroneous conclusions if data is considered at wrong granularity – an instance of the Simpson's paradox.

**Multi-dimensional segregation data cube.** Our approach consists of providing the analysts with a multi-dimensional data cube that can be explored in search of candidate contexts of segregation. An example segregation data cube is shown in Fig.1. Dimensions of the data cube include two types of attributes:

- *segregation attributes* (SA), such as `sex`, `age`, and `ethnicity`, which denote (minority/protected) groups potentially exposed to segregation;
- *context attributes* (CA), such as `region` and `job type`, which denote contexts where segregation may appear.

Metrics of the data cube are chosen among the social science indexes proposed for measuring the degree of segregation of social groups within a society [12]. Here, we recall only one such index, but the SCube system is parametric to the indexes

[1] See e.g., the *Toronto declaration* at www.accessnow.org/toronto-declaration.
[2] See census stats, e.g., www.census.gov/topics/housing/housing-patterns/data.html

[3] cran.r-project.org/package=OasisR
[4] cran.r-project.org/package=seg
[5] geoseganalyzer.ucs.inrs.ca

and it computes 6 of them: dissimilarity, Gini, Information index, Isolation, Interaction, Atkinson. Also, we restrict to binary groups (minority/majority). Let $T$ be the size of the total population under consideration, $0 < M < T$ be the size of a minority group, $T - M$ the size of the rest of society (or majority group) and $P = M/T$ be the overall fraction of the minority group. Assume that there are $n$ organizational units (or simply, units – such as schools, neighboorhoods, job types, etc.), and that for $i \in [1, n]$, $t_i$ is the size of the population in unit $i$, and $m_i$ is the size of the minority group in unit $i$. The *dissimilarity index* $D$ measures the absolute distance between the fractions of minority and majority groups over the units:

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

$D$ ranges over $[0, 1]$, with higher values denoting higher segregation. Dissimilarity is minimum when for all $i \in [1, n]$, $m_i/t_i = M/T$, namely the distribution of the minority group is uniform over units. It is maximum when for all $i \in [1, n]$, either $m_i = t_i$ or $m_i = 0$, namely every unit includes members of only one group (complete segregation). Dissimilarity and other segregation indexes can be interpreted as metrics in a cell of a multi-dimensional cube as follows: set the total population as those individuals that satisfy the CA coordinates of the cell; and, set the minority population as those individuals that satisfy the SA coordinates. For instance, the cube cell in Fig.1 with SA coordinates `sex=female, age=young` and CA coordinates `region=north` contains the dissimilarity index for the population living in the north region and for the minority group of young women. Notice that the number $n$ of organizational units here have to be determined *a-priori*, while the total population and minority groups in each unit depend on the values of cell coordinates. As in standard multi-dimensional modelling [7], the special value "$\star$" allows for considering different granularities of analysis.

**Segregation analysis of tabular data.** We assume in input a relational table with a tuple for every individual in the population, including SA and CA attributes, and with a further attribute `unitID` which denotes the unit an individual belongs to. Unfortunately, segregation indexes are not additive metrics (see [4]). This gives rise to the problem of efficiently computing a data cube for segregation analysis. Our approach is more specialized than generic holistic aggregate computation in datacubes [13]. We resort to frequent closed itemset mining [8]. Data cube coordinates are encoded into itemsets of the form **A**, **B**, where **A** denotes a minority subgroup and **B** denotes a context. Recalling the previous example, **A** =`sex=female, age=young` defines the SA coordinates, and **B** =`region=north` defines the CA coordinates. The *SegregationDataCubeBuilder* algorithm described in [4] fills data cube cells with the value of a segregation index by scanning frequent closed itemsets of the form above. Since relational data is transformed into transaction database for itemset mining, we obtain for free that CA or SA attributes can be multi-valued, e.g., to denote that an individual owns both a house and a car we admit a relation tuple $\sigma$ such that $\sigma$[`owns`] = {`house`, `car`}.

**Segregation analysis of graph data.** While transaction databases are able to cover typical analysis from traditional social science, they are not enough powerful to deal with social network data. We formalize such a case using attributed graphs, where nodes are assigned values on a specified set of attributes. However, in this scenario, there is no *a-priori* defined notion of organizational unit, i.e., the `unitID` attribute assumed in input so far. Some forms of community discovery using graph clustering become necessary in order to determine the organizational units. Clustering attributed graphs consists of partitioning them into disjoint communities of nodes that are both well connected and similar with respect to their attributes [5]. In summary, attributed graph clustering can be used first to partition a social network into communities. At this stage, every node/individual in a community is described by its attributes and the community id, which will be our `unitID` attribute. We have thus reduced the problem to the analysis of relational data, for which the *SegregationDataCubeBuilder* algorithm can be applied.

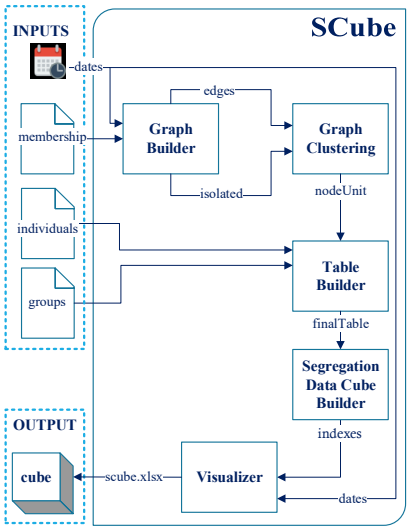



**Figure 2: SCube architecture.**

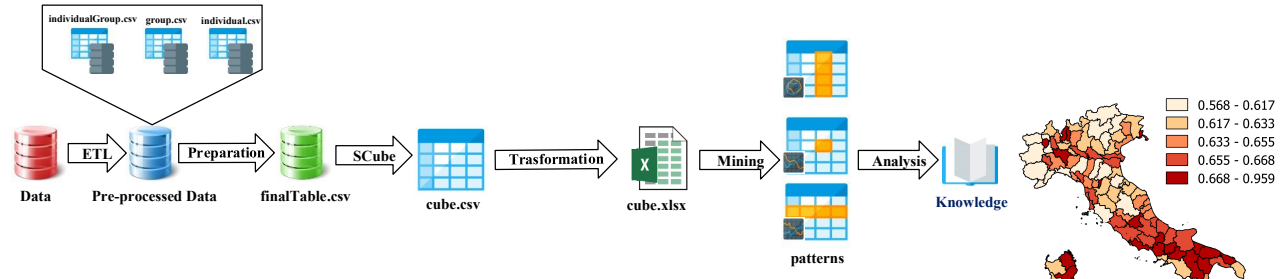
**Segregation analysis of bipartite graphs.** An even more complex scenario is when individuals are not connected among them, e.g., because they are friends, but through a connection with another entity, e.g., because they work in the same company. Here, a form of projection on unipartite graph is needed to reduce to the previous case. For instance, in [4], we adopt a bipartite projection of the bipartite graph of directors and companies to obtain a graph of companies connected by shared directors. Using projection, we have reduced the problem to the previous case, where attributed graph clustering can be adopted to find communities of companies, which then represent the organizational units for segregation analysis.

## 3 SCUBE ARCHITECTURE

The architecture of SCube is shown in Fig. 2. The system is developed in Java, and it relies on a few state-of-the-art libraries[6].

**Inputs.** The user has to provide features for two entities: *individuals* and *groups*. In the reference case studies, individuals are directors and groups are companies. The input `individuals` (a CSV file or a JDBC query) provides for each individual an ID and a number of attribute values, distinguished into segregation attributes (e.g., `gender`, `age`, `birthplace`) and context attributes (e.g., `residence`). A second input `groups` provides for each group an ID and a number of context attributes values (e.g., industrial `sector` of a company and its headquarter `location`). Notice that individuals are subjects to possible segregation, while groups are

[6]EWAH for compressed bitmaps (github.com/lemire/javaewah), Apache POI for OOXML docs (poi.apache.org), Borgelt's FPGrowth for frequent itemset mining (www.borgelt.net), FastUtil for graph storage (fastutil.di.unimi.it).




| *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 |
| ... | ... | ... | ... | ... | ... |

**Figure 3: The process of segregation discovery supported by SCube (left, top), input to SegregationDataCubeBuilder (left, bottom), and an output report on dissimilarity segregation index of the Italian provinces (right).**

not. For this reasons, groups have no SA feature. A third input is `membership`, which includes the edges of the bipartite graph of individuals and groups, i.e., all pairs (`individualID`, `groupID`) for which the individual is related to the group. In our case studies, directors are related to companies they sit in the board of. We also admit that the pairs are labeled with a time interval of validity, thus allowing for temporal analysis of segregation. We have such an information for the Estonian dataset. A fourth input is a list of snapshot dates at which to consider snapshots of the membership relation.

**Modules.** SCube consists of five software modules. *GraphBuilder* projects the bipartite graph of individuals and groups into an unipartite attributed graph, where nodes are groups and an edge connect two groups if they are related by at least one shared individual. In the case studies, nodes are companies, and edges connect companies that share at least one director in their boards. Edges are weighted by the number of shared directors. *GraphBuilder* outputs edges of the projection (`edges`), and nodes that have zero degree (`isolated`). The *GraphClustering* module computes then a clustering of nodes into organizational units (output file `nodeUnit`). Methods for clustering available in SCube include: extraction of connected components (Breadth-First Search), removal of edges from the giant component with weight below a threshold and then extraction of connected components (designed in [4]), and an attributed graph clustering method for very large graphs (SToC algorithm [3]). In our case studies, the result of *GraphClustering* is a partitioning of companies into clusters based on connections among companies determined by shared directors – which can be readily considered a signal of relationships (business, personal, or other) between companies. Clusters represent the organizational units needed for computing segregation indexes. *TableBuilder* joins features of individuals with features of the companies in an organizational unit. This yields a `finalTable` with a row per individual and organizational unit she belongs to. An example is shown in Fig. 3 (left, bottom). This is the input for the *SegregationDataCube* builder module, implementing the algorithm of [4]. Notice that if the data under analysis contains already the assignment of individuals to units, i.e., it is already in the form of `finalTable`, the pre-processing steps of bipartite projection and graph clustering do not need to be performed. The *Visualizer* module transforms the extended datacube in output of SegregationDataCube into a standard OOXML format that can be opened by Microsoft Excel, Libre Office, and other office productivity tools (see Fig. 5). Segregation data cube exploration can be easily interfaced with visualization tools, as in the map overlay in Fig. 3 (right).

**Process, Wizard, and GUI.** The whole process of segregation discovery supported by SCube is shown in Fig. 3 (left, top). To facilitate the adoption of SCube by non-technical users, we have developed two interfaces (see Fig. 4). The first one is a standalone wizard that guides the user throughout all the steps of the process, asking for inputs and parameters when appropriate, and finish launching Microsoft Excel or Libre Office on the output file. Using popular desktop tools as GUI's makes the learning curve of approaching and effectively using SCube more manageable. The second one is a cloud service offered by the SoBigDataLab freely accessible research infrastructure (www.sobigdata.eu/access/virtual), a web front-end comprising a catalogue of data, services, and virtual research environments for big data and social mining research.

## 4 DEMONSTRATION SCENARIO

The demonstration starts with a brief introduction on concepts and methods of segregation measurement [12] and segregation discovery [4]. This provides the audience with the basic definitions for understanding the SCube functionalities. The architecture of SCube is presented next. For interested participants, computational efficiency, algorithmic solutions, and source code internal aspects are discussed. Then, two running case studies in the context of occupational segregation in the boards of company directors [1] are introduced. They are based on a 2012 snaphost of Italian companies (3.6M directors, 2.15M companies), and on a 20-year long dataset of Estonian companies (440K directors, 340K companies). Such anonymized datasets are the largest ever considered in the literature of segregation analysis. We summarize the data pre-processing activities to produce the inputs for SCube.

The demonstration then proceeds by presenting three analysis scenarios based on input data of increasing complexity. In all scenarios, `gender`, `age`, and `birthplace` are used as segregation attributes. The first scenario considers tabular data, where company sector is used as organizational `unitID`, and it is intended

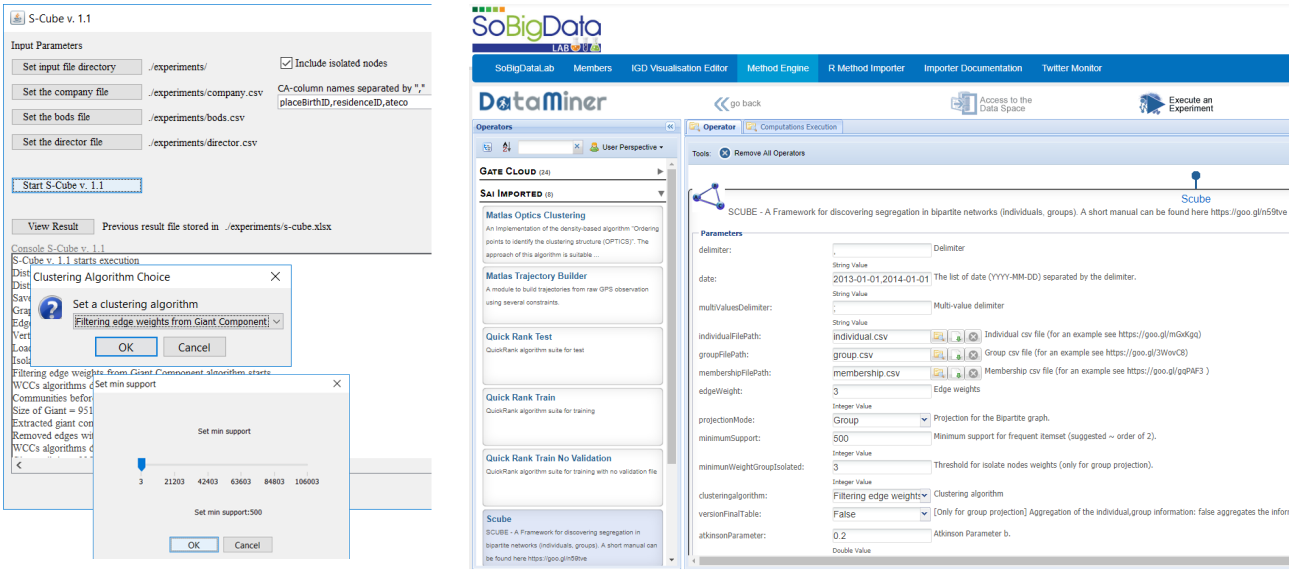
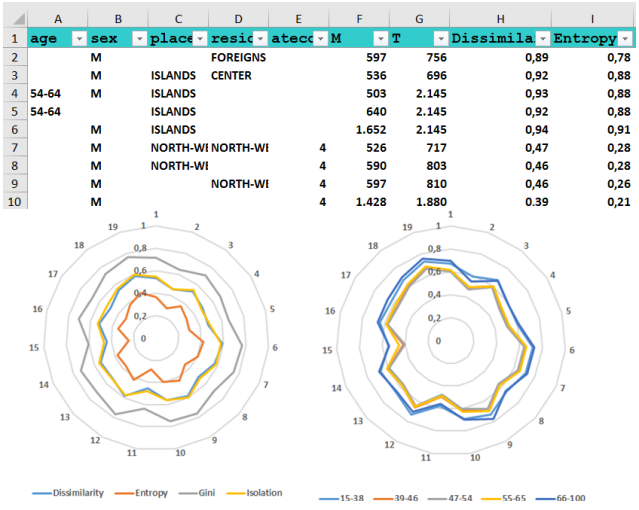



**Figure 4: SCube standalone wizard (left) and SCube method at the SoBigData research infrastructure (right).**



| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| 1 | age | sex | place | resic | atecc | M | T | Dissimila | Entropy |
| 2 | | M | | FOREIGNS | | 597 | 756 | 0,89 | 0,78 |
| 3 | | M | ISLANDS | CENTER | | 536 | 696 | 0,92 | 0,88 |
| 4 | 54-64 | M | ISLANDS | | | 503 | 2.145 | 0,93 | 0,88 |
| 5 | 54-64 | | ISLANDS | | | 640 | 2.145 | 0,92 | 0,88 |
| 6 | | M | ISLANDS | | | 1.652 | 2.145 | 0,94 | 0,91 |
| 7 | | M | NORTH-WI | NORTH-WI | 4 | 526 | 717 | 0,47 | 0,28 |
| 8 | | M | NORTH-WI | | 4 | 590 | 803 | 0,46 | 0,28 |
| 9 | | M | | NORTH-WI | 4 | 597 | 810 | 0,46 | 0,26 |
| 10 | | M | | | 4 | 1.428 | 1.880 | 0.39 | 0,21 |



**Figure 5: Top: sample multidimensional segregation cube. Bottom: radial plot of segregation indexes for directors in each of the 20 Italian company sectors.**

to answer questions such as: how much are women segregated in company sectors? The second scenario considers attributed graph data, where nodes are directors, and edges connect two directors if they belong to a same company board. Here, the organizational units are determined through clustering over attributed graphs. This scenario can answer questions such as: how much are women segregated in communities of connected directors? Finally, the third scenario considers a bipartite attributed graph of directors and companies, as presented throughout the paper. An example of question it can answer is: how much are women segregated in communities of connected companies? For each scenario, the output of SCube is interactively explored using pivot tables and charts. The audience is guided to the discovery of a few actual cases of *a-priori* unknown segregation contexts and to the understanding of which attributes contribute the most to segregation. Moreover, a cross-comparison of the Italian vs Estonian segregation findings will be discussed.

## 5 CONCLUSION

This demonstration illustrates the SCube tool for interactive exploration of social segregation indexes in large and complex data. The audience is made aware of social exclusion issues that can be hidden in data and of the indexes that measure segregation. Real case studies on scenarios of increasing complexity are discussed and explored. Efficiency issues and algorithmic solutions adopted for scaling to large datasets and graphs are detailed.

**Acknowledgements.** This work is partially supported by the European H2020 Program under the funding scheme "INFRAIA-1-2014-2015: Research Infrastructures" grant agreement 654024 "*SoBigData*" (http://www.sobigdata.eu).

## REFERENCES

[1] M. Aluchna and G. Aras, editors. *Women on Corporate Boards*. Routledge, 2018.
[2] P. Apparicio, J. C. Martori, A. L. Pearson, E. Fournier, and D. Apparicio. An open-source software for calculating indices of urban residential segregation. *Social Science Computer Review*, 32(1):117–128, 2014.
[3] A. Baroni, A. Conte, M. Patrignani, and S. Ruggieri. Efficiently clustering very large attributed graphs. In *ASONAM*, pages 369–376. ACM, 2017.
[4] A. Baroni and S. Ruggieri. Segregation discovery in a social network of companies. *J. Intell. Inf. Syst.*, 51(1):71–96, 2018.
[5] C. Bothorel, J. D. Cruz, M. Magnani, and B. Micenková. Clustering attributed graphs: models, measures and methods. *Network Science*, 3(03):408–444, 2015.
[6] S. Flaxman, S. Goel, and J. M. Rao. Filter bubbles, echo chambers, and online news consumption. *Public Opinion Quarterly*, 80:298–320, 2016. Available at SSRN: http://ssrn.com/abstract=2363701.
[7] J. Gray, S. Chaudhuri, A. Bosworth, A. Layman, D. Reichart, M. Venkatrao, F. Pellow, and H. Pirahesh. Data cube: A relational aggregation operator generalizing group-by, cross-tab, and sub totals. *Data Min. Knowl. Discov.*, 1(1):29–53, 1997.
[8] J. Han, H. Cheng, D. Xin, and X. Yan. Frequent pattern mining: Current status and future directions. *Data Min. Knowl. Discov.*, 15(1):55–86, 2007.
[9] S.-Y. Hong, D. O'Sullivan, and Y. Sadahiro. Implementing spatial segregation measures in R. *PLoS ONE*, 9(11):e113767, 2014.
[10] M. Maes and L. Bischofberger. Will the personalization of online social networks foster opinion polarization? *Available at SSRN: http://ssrn.com/abstract=2553436*, 2015.
[11] D. S. Massey. Segregation and the perpetuation of disadvantage. *The Oxford Handbook of the Social Science of Poverty*, pages 369–393, 2016.
[12] D. S. Massey and N. A. Denton. The dimensions of residential segregation. *Social Forces*, 67(2):281–315, 1988.
[13] A. Nandi, C. Yu, P. Bohannon, and R. Ramakrishnan. Data cube materialization and mining over mapreduce. *IEEE Trans. Knowl. Data Eng.*, 24(10):1747–1759, 2012.