SCube: A Tool for Segregation Discovery

Alessandro Baroni, Salvatore Ruggieri

Department of Computer Science, University of Pisa, Italy
baroni@di.unipi.it, ruggieri@di.unipi.it

Abstract—Segregation discovery consists of finding contexts of segregation of social groups distributed across units or communities. The SCube system implements an approach for segregation discovery on top of frequent itemset mining, by offering to the analyst a multi-dimensional (segregation) data cube for exploratory data analysis. The demonstration first guides the audience through the social science concepts, then describes the architecture of the SCube system and how it is able to deal with both relational data and with attributed graph data modeling social networks. Finally, we go through two running case studies in the context of occupational segregation in the boards of company directors using a 2012 snapshot of the Italian companies, and a 20-year long dataset of Estonian companies.

I. SOCIAL SEGREGATION

Ethical issues in data and knowledge engineering 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. See e.g., the resources listed in https://fatconference.org. We have recently developed a data-driven technique for addressing segregation of social groups through multi-dimensional data analysis [1]. The approach is implemented in the SCube system, which we propose to demonstrate.

Social segregation refers to the “separation of socially defined groups” [2]. 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. [3]. 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 spatial segregation to modern cultural segregation. For instance, it has been warned that the filter bubble generated by personalization of online social networks may foster ideological segregation [4], opinion polarization [5], and informational segregation [6].

A technology that enables to assess the extent, nature, and trends of social segregation is of extreme interest for a wide audience: social scientists, public policy makers, regulation and control authorities professional associations, civil rights societies, and investigative journalists. Business decision makers should also care of business practices, or even automated decisions and recommendation, that segregate customers and users into stereotypes.

Fig. 1: A segregation data cube.
are \( n \) organizational units (or simply, units – such as schools, neighborhoods, job types, etc.), 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 \( i \). The **dissimilarity index** \( D \) is the weighted mean absolute deviation of every unit’s minority proportion from the global minority proportion:

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

The factor \( 2 \cdot P \cdot (1 - P) \) normalizes the index in the range \([0, 1]\). Since \( D \) measures dispersion of minorities over the units, higher values of the index mean higher segregation. Dissimilarity index (and other segregation indexes) can be used as a metric in a data cube by setting: the total population as those individuals that satisfies the CA coordinates; and, the minority population as those individuals that satisfy the SA coordinates. For instance, a cube cell with SA coordinates \( \text{sex}=\text{female}, \text{age}=\text{young} \) and CA coordinates \( \text{region}=\text{north} \) contains the segregation index for the population living in the north region and for the minority group of young women. Notice that the \( n \) organizational units here have to be determined *a-priori*, but the total population and minority groups in each unit change from one cell to another.

Unfortunately, segregation indexes are not additive metrics (see [1] for some mathematical properties of the indexes). This gives rise to the problem of efficiently computing a data cube for segregation analysis. Our approach resorts to frequent itemset. We assume in input a relational table with a tuple for every individual in the population, with SA and CA attributes, and with a further attribute \text{unitID} which denotes the unit an individual belongs to.\(^2\) We code data cube coordinates as itemsets of the form \( \text{A}, \text{B} \), where \( \text{A} \) denotes a minority subgroup and \( \text{B} \) denotes a context. Recalling the previous example, the itemset \( \text{sex}=\text{female}, \text{age}=\text{young}, \text{region}=\text{north} \) codes in \( \text{A} =\text{sex}=\text{female}, \text{age}=\text{young} \) the SA coordinates, and in \( \text{B} =\text{region}=\text{north} \) the CA coordinates. The **SegregationDataCubeBuilder** algorithm described in [1] 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 transactions for itemset mining, we obtain for free that CA or SA attributes can be multi-valued, e.g., to denote that an individual’s property include both house and cars we admit a tuple to have \{house, car\} as values for the properties attribute. While transactional data is able to cover typical analysis from traditional social science, it is not enough to deal with modern scenarios, such as analysing social network data. Such a case is challenging because there is no *a-priori* defined notion of organizational unit. Some forms of community discovery using graph clustering become necessary in order to determine the organizational units. This can be formalized using attributed graphs, where nodes are assigned values on a specified set of attributes. Clustering attributed graphs consists of partitioning them into disjoint communities of nodes that are both well connected and similar with respect to their attributes \[^{8}\].

In summary, attributed graph clustering can be used first to partition a social network into communities, and then the **SegregationDataCubeBuilder** algorithm can be applied to construct a data cube for exploratory data analysis.

An even more complex scenario is when individuals are not connected directly, e.g., because they are friends, but through a connection with another entity, e.g., because they work in the same company. Here, a second-level form of clustering is needed. Companies must be clustered first, in order to find communities of “similar” companies, and then such communities become the organizational units over which to calculate segregation indexes. We tackle such a case in [1] by means of a bipartite projection \[^{9}\] of the bipartite graph of individuals and companies. After projecting over companies, attributed graph clustering is adopted to find community of companies, and finally the **SegregationDataCubeBuilder** algorithm is applied to build a data cube for individuals, where organizational units are the clusters of companies.

III. The SCube System

The SCube system supports an analyst in discovering contexts of social segregation in a dataset or in a social network. The overall architecture of SCube is shown in Fig. 2. The system is developed in Java 8, and it relies on a number of libraries.\[^{10}\] Throughout the demo, we will present SCube functionalities through two running case studies in the context

\[^{1}\]EWAH for compressed bitmaps (github.com/lemire/javaewah), Apache POI for OXXML docs (poi.apache.org), Borgelt’s FPgrowth for frequent itemset mining (www.borgelt.net), FastUtil for graph storage (fastutil.di.unimi.it).

Fig. 2: SCube architecture.
of occupational segregation in the boards of company directors using a 2012 snapshot of the Italian companies (3.6M directors, 2.15M companies), and a 20-year long dataset of Estonian companies (440K directors, 340K companies).

**Inputs.** The user has to provide features for two entities: *individuals* and *groups*. In the reference case study, individuals are directors and groups are companies. A CSV file *individuals.csv* in input to SCube provides for each individual an ID and a number of attribute values, distinguished into segregation attributes (gender, age, birthplace) and context attributes (residence). A second CSV file *groups.csv* provides for each group an ID and a number of context attributes values (industrial sector of a company and its headquarter location). A third input file is *membership.csv*, which includes the edges of the bipartite graph of individuals and groups, i.e., all pairs (individual ID, group ID) for which the individual is related to the group. In our case study, directors are related to companies they sit in the board of. We also admit that the pairs are extended to include a time interval of validity, thus allowing for temporal analysis of segregation. In the Estonian dataset, we have such an information. For this reason, a fourth input is a list of 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 study, 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.csv*), and nodes that have zero degree (*isolated.csv*). These files are passed in input to the *GraphClustering* module, which computes a clustering of nodes into organizational units (output file *nodeUnit.csv*). 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 (devised in [1]), and an attributed graph clustering method for very large graphs (SToC algorithm [10]). For our case study, the result of *GraphClustering* is a partitioning of companies into clusters of companies connected via 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 final *Table.csv* 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 algorithm proposed in [1]. Notice that if the data under analysis contains already the assignment of individuals to units, i.e., it is already in the form of final *Table.csv*, the pre-processing steps of bipartite projection and graph clustering do not need to be performed. The *SegregationDataCube* algorithm extracts and post-process closed frequent itemsets to compute a multi-dimensional data cube. Dimensions of the data cube include SA and CA attributes. Cells in the data cube include 6 segregation indexes calculated for the minority population specified by the SA coordinates, and for the reference population specified by the CA coordinates. The final module of SCube, *Visualizer*, transforms the 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 Wizard that guides the user throughout all the steps, asking for input files and parameters when appropriate, and finish launching of Microsoft Excel or Libre Office.
on the output file. Using popular desktop tools as GUI for data analysis makes the learning curve of approaching and effectively using SCube more manageable. The second one is a method offered by the SoBigDataLab 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.

IV. DEMONSTRATION SCENARIO

The demonstration starts with a brief introduction on concepts and methods of segregation measurement [7] and segregation discovery [1]. This provides the audience with the basic definitions for understanding the SCube functionalities. The system architecture of SCube is presented next. The demonstration then proceeds by presenting three analysis scenarios based on input data of increasing complexity: relational data, attributed graphs, bipartite attributed graphs, and bipartite attribute graphs with temporal annotation. For each scenario, a sample dataset is analyzed by presenting summaries of the data, and the usage of SCube wizard and SoBigData modules with their parameters explained. Case studies include the Italian and Estonian datasets of company boards. The output of SCube is finally explored using spreadsheet’s pivot tables and charts. A few cases of a-priori unknown segregation are outlined as result of exploratory search on the segregation data cubes.

Acknowledgment. Work partly supported by the European Community’s H2020 Program INFRAIA-1-2014-2015: Research Infrastructures, grant agreement #654024 SoBigData: Social Mining & Big Data Ecosystem (www.sobigdata.eu).

REFERENCES

[1] A. Baroni and S. Ruggieri, “Segregation discovery in a social network of companies,” Journal of Intelligent Information Systems, Sep 2017.
[2] D. S. Massey, “Segregation and the perpetuation of disadvantage,” The Oxford Handbook of the Social Science of Poverty, pp. 369–393, 2016.
[3] A. Romei and S. Ruggieri, “A multidisciplinary survey on discrimination analysis,” The Knowledge Engineering Review, vol. 29, no. 5, pp. 582–638, 2014.
[4] S. Flaxman, S. Goel, and J. M. Rao, “Filter bubbles, echo chambers, and online news consumption,” Public Opinion Quarterly, vol. 80, pp. 298–320, 2016, available at SSRN: http://ssrn.com/abstract=2363701
[5] M. Maes and L. Bischofberger, “Will the personalization of online social networks foster opinion polarization?” Available at SSRN: http://ssrn.com/abstract=2553436, 2015.
[6] A. Chakraborty, M. Ali, S. Ghosh, N. Ganguly, and K. P. Gummadi, “On quantifying knowledge segregation in society,” CoRR, vol. abs/1708.00670, 2017.
[7] D. S. Massey and N. A. Denton, “The dimensions of residential segregation,” Social Forces, vol. 67, no. 2, pp. 281–315, 1988.
[8] C. Bothorel, J. D. Cruz, M. Magnani, and B. Menciková, “Clustering attributed graphs: models, measures and methods,” Network Science, vol. 3, no. 03, pp. 408–444, 2015.
[9] T. Zhou, J. Ren, M. Medo, and Y.-C. Zhang, “Bipartite network projection and personal recommendation,” Physical Review E, vol. 76, no. 4, 2007.
[10] A. Baroni, A. Conte, M. Patrignani, and S. Ruggieri, “Efficiently clustering very large attributed graphs,” in Proc. of Int. Conf. on Advances in Social Networks Analysis and Mining (ASONAM 2017). IEEE, 2017.