A hybrid approach with agent-based simulation and clustering for sociograms

Iván García-Magariño[a], Carlos Medrano[b], Andrés S. Lombas[c], Angel Barrasa[c]

[a]*Department of Computer Science and Engineering of Systems, University of Zaragoza*
[b]*Department of Electronics Engineering and Communications, University of Zaragoza*
[c]*Department of Psychology and Sociology, University of Zaragoza*

Abstract

In the last years, some features of sociograms have proven to be strongly related to the performance of groups. However, the prediction of sociograms according to the features of individuals is still an open issue. In particular, the current approach presents a hybrid approach between agent-based simulation and clustering for simulating sociograms according to the psychological features of their members. This approach performs the clustering extracting certain types of individuals regarding their psychological characteristics, from training data. New people can then be associated to one of the types in order to run a sociogram simulation. This approach has been implemented with the tool called CLUS-SOCI (an agent-based and CLUStering tool for simulating SOCIograms). The current approach has been experienced with real data from four different secondary schools, with 38 real sociograms involving 714 students. Two thirds of these data were used for training the tool, while the remaining third was used for validating it. In the validation data, the resulting simulated sociograms were similar to the real ones in terms of cohesion, coherence of reciprocal relations and intensity, according to the binomial test with the correction of Bonferroni.

Key words: agent-based simulation, agent-based social simulation, multi-agent system, social relation, sociogram

Email addresses: ivangmg@unizar.es (Iván García-Magariño), ctmedra@unizar.es (Carlos Medrano), slombas@unizar.es (Andrés S. Lombas), abarrasa@unizar.es (Angel Barrasa)

1. Introduction

Establishing positive social relationships is thought to be a crucial aspect for well-being. Thus, for example, the famous Maslow's Hierarchy of Human Needs [34] includes "the love needs", described as "hunger for affectionate relations with people in general, namely, for a place in their group", as fundamental need for achieving self-realization. A more modern theory based on basic psychological needs, Self-Determination Theory [43], proposes the need of social contact, referred as "relatedness need", as one of the nutrients that is essential for ongoing personal growth, integrity and psychological health.

A frequent instrument used to measure social relationships is the sociometric techniques developed by Moreno [35]. Sociometric techniques identify types of relationships among peers, formulated in terms of attraction and rejection. It is acknowledged that the degree of acceptance and rejection among peers at school is key to psychosocial adjustment and academic success in adolescence [6] and other stages such as childhood [26]. Popular adolescents, who are accepted by the majority of their peers, are characterized for their interpersonal abilities, their empathy for others, and their willingness to cooperate non-aggressively [10]. In addition, peer acceptance is positively associated to social prominence [51]. On the contrary, rejected adolescents are perceived as unpleasant and are less liked [6]. Besides, they show more psychosomatic symptoms and are affected by more psychiatric disorders [29]. They have more conflictual relationships with other classmates and teachers, being more frequently involved in disruptive and aggressive behaviors that lead to the violation of institutional rules [37]. As a matter of fact, conflictual relationships have been proposed as the cause of being rejected by their peers [22].

Sociograms represent the social relations within a group. In sociograms, the cohesion and structures of relations have been previously proven to be related with the performance of the group, as stated for example in [24]. Specifically in education, Yu et al. [49] improve academic performance of student groups by considering student profiles and different levels of academic performance when making the groups. Thus, sociograms of student groups considering student profiles may be relevant when taking academic performance into account. In addition, simulation is useful for predicting emergent behaviors of groups of individuals [45]. In this manner, in some cases one can apply or prepare the necessary actions beforehand. Nevertheless, to the best of authors' knowledge, there is not any simulator about

student sociograms. In this context, the current work presents a novel hybrid approach with an Agent-based Simulator (ABS) of sociograms that uses clustering for classifying students.

The current work selects developing an ABS, because in the literature [33] this kind of simulators has been proven to be especially useful when simulating environments with several autonomous individuals that make their decisions and establish their social relations. The current approach selects clustering as the mechanism for classifying students, as clustering has already proven to be useful when classifying students or other features related to them, as one can observe in [46].

The current work presents a tool that implements the present approach. This tool is called CLUS-SOCI (an agent-based and CLUStering tool for simulating SOCIograms). CLUS-SOCI is prepared to be trained with certain known input data, so that then it can be applied to simulate certain unknown data. CLUS-SOCI is mainly composed of three modules. The first module has been named the clustering module. This module receives input from training data, i.e. existing real and known sociograms of certain individuals with specific psychological features. This module is responsible for clustering the existing individuals in certain clusters (also referred as types from this point forward) according to their psychological features, and relating these clusters with certain social behavioral information. Secondly, the management module allows practitioners (1) to load and learn these types, (2) to classify any individual into one of the learned types according to their psychological features, (3) show and measure any sociogram, and (4) to invoke the ABS of the remaining module. The third module is the ABS for simulating sociograms for a group of individuals classified with the corresponding previous step.

The current approach and its CLUS-SOCI tool have been experienced with student sociograms from four different secondary schools of the Aragón region of Spain. The data involve 714 students with their psychological features extracted with the corresponding tests. These students constitute 38 groups with their corresponding sociograms. Approximately, two thirds of these data were used as training data, and the remaining data were used for validating the current approach. The results showed that the simulated sociograms are similar to the real ones for some characteristics according to the corresponding statistical test.

The current work extends our previous work about the FTS-SOCI tool [20]. The most relevant improvements are (a) the clustering process for

obtaining the student types in the training phase, (b) the classification of students according to their psychological features, and (c) the enhancement of experiments, based on 38 real sociograms instead of two, and considering four sociometrics instead of two.

The remaining of the article is organized as follows. The next section analyzes the related works highlighting the gaps of the literature that are covered with the current approach. Section 3 presents some psychological instruments as background of the presented approach. Section 4 introduces the current hybrid approach describing its modules. Section 5 presents the experimentation including the training and the validation phases. Finally, section 6 depicts the conclusions of the current work, and mentions some future lines of research.

2. Related work

The current work presents an ABS approach for simulating sociograms based on the psychological features of the individuals. It applies a clustering technique for detecting individual profiles, which are the base for training the ABS. Hence, this work organizes the discussion of related works in three main blocks: (1) sociology simulations, (2) simulations with clustering, (3) Multi-Agent Systems (MASs) with clustering, and (4) social network algorithms. Since the current work is implemented with an ABS, most analyzed works fall into this category.

2.1. Sociology simulations

ABSs have been widely applied to simulate sociology experiments. To begin with, Macy and Willer [33] present a survey about the modeling of ABSs for being applied in sociology. They present the main features that make MASs especially useful for simulating sociology aspects. The construction of the behaviors is performed with a bottom-up approach, in which the agents simulate individual behaviors within a society. These systems make it possible to simulate emergent behaviors from the interactions of individuals. They also mention some aspects such us collaboration between peers and whether these trust each other. Their survey introduces several works that relate ABS modeling with sociology.

Moreover, Andrei et al. [2] present an ABS for simulating the tax compliance and evasion according to certain social networks. In their simulations, they show how the network structures can influence the propagation of tax

evasive behaviors. In addition, Serban et al. [44] apply agent-based modeling and simulation for analyzing the relationship between (1) cognitive ability, extroversion and self efficacy, and (2) leadership emergence. Their results advocate that this relationship is reduced when using virtuality in teams. Prenkert and Følgesvold [39] present an ABS in order to describe the influence of sharing some focal resource and market, by several business networks. These networks are assumed to be inter-related but with different internal organizations. They simulate the repercussion of different network structures on the international business relations.

Nonetheless, these works do not apply clustering techniques for explicitly simulating sociograms.

2.2. Simulations with clustering

Clustering techniques have been applied in ABSs. Specifically, Serrano et al. [45] present an ABS for simulating voting in political elections, based on the social choice theory. This ABS allowed them to support their hypothesis about that the combination of social choice theory and ambient intelligence systems can improve users' satisfaction when accessing shared resources. Their simulations were executed by means of their VoteSim tool, and were compared with certain metrics that they propose. They applied cluster analysis techniques to improve social welfare in the experiments. In particular, they estimate each user's vote as classifying their last known vote into a cluster of known votes.

Moreover, Raberto et al. [40] analyze the behavior on financial markets with clustering techniques. Their clustering approach took the relation of traders and the market volatility into account for conforming clusters of traders. They developed an ABS for financial markets, in which the agents play roles (i.e. behavioral types) that correspond to the obtained clusters. Their simulations show the stylized fact of volatility clustering that is common in real-world markets. In addition, Becu et al. [5] present an ABS that simulated the rural stakeholders as well as the individual decisions of farmers. This ABS considers the existence of the resources such as the distribution of the hydrological resources. They analyzed clusters of households to obtain a realistic scenario. They experienced their ABS by comparing their simulated results with real scenarios in the north of Thailand.

Furthermore, Elhabyan and Yagoub [16] present a technique for clustering and routing Wireless Sensor Networks (WSNs). They introduce an algorithm with two lineal programming formulations that are based on particle swarm

optimization. They perform 50 simulations with homogeneous and heterogeneous WSN models, and their proposed approach performs better than some well-known clustered-based sensor network protocols.

By contrast, all these works do not apply ABSs and clustering for simulating sociograms.

2.3. MASs with clustering

Clustering has been previously applied in the MASs literature. In particular, Dos Santos and Bazzan [15] applied a clustering algorithm based on swarm intelligence for classifying agents. This classification of agents in clusters allowed a MAS to improve the performance in the RoboCup Rescue scenario where tasks with different features must be assigned to agents with different capabilities. Thus, this work applied clustering for classifying agents to improve the performance in assigning tasks to these.

Ayala-Cabrera et al. [3] present the combination of a MAS with a clustering technique to analyze subsoil characteristics based on the output files of the ground penetrating radars. Their tool provides information that supports decision-making in the management of water supplies. In addition, López-Ortega and Rosales [32] provide a MAS for assisting decision-making. This MAS applies fuzzy clustering to group individual evaluations. The fuzzy clustering obtains relevant information such as the largest group of evaluations around a centroid value. Then, it applies the analytical hierarchy process to reach a final decision. They present a case-study about selecting a new robotic manipulator for a manufacturing department.

The MAS architecture of Garruzzo and Rosaci [21] applies a clustering technique based on the HIerarchical SEmantic NEgotiation (HISENE) protocol. Their architecture allows practitioners to form groups of agents. Their novelty was the analysis of the semantic component when grouping agents. For example, their architecture takes into consideration the context in which a term is used in an agent ontology. Their results show the effectiveness, efficiency and scalability of their approach.

Conversely, these works do not use the clustering for grouping individuals according to their psychological features concerning their social relations. In addition, these works are not specifically aimed at simulating or explicitly modeling sociograms.

2.4. Social network algorithms

There are several works that use algorithms in the same context as the current work. To begin with, some works present algorithms that use clustering related to social networks. For instance, Di et al. [14] present a social-network based algorithm for organizing the water distribution. Firstly it performs a clustering process. Then, it applies partitioning to determine the configuration of gate valves. Their algorithm pursues to construct an efficient water distribution network. In addition, Johnson et al. [28] analyze the social networks by means of the dot product model. This model assumes that two individuals are connected when they have similar opinions and attributes. They measure diversity and clustering in social networks. Nevertheless, these works do not simulate social networks considering certain common sociogram aspects such as the evolution of relationships considering certain trends, the influence of the relationship durations, the repercussion of the group size, and the high proportion of reciprocal relationships.

Moreover, there are some algorithms that predict the establishment of relationships in social networks. For example, Ahmed and Chen [1] present an algorithm for predicting relationships in uncertain social networks. Their algorithm is based on transforming the problem in uncertain networks to a random walk in a deterministic network. Their approach considers temporal and global topological information. Their approach obtains more accurate results than other similar methods. Bliss et al. [9] introduce an evolutionary algorithm for predicting relationships in dynamic social networks. They apply a covariance matrix adaptation evolution strategy with linear combination of some neighborhoods. They apply their approach in the Twitter reciprocal reply networks for the experimentation. Furthermore, Xu et al. [47] propose an algorithm for predicting friendship relationships in mobile social networks. Their algorithm is mainly based on the locations and times where the users check in. The main improvement of the current approach over all these works is the consideration of the psychological features of individuals for simulating and predicting network relationships.

2.5. Discussion

On the whole, there are several works that use ABSs for simulating sociology aspects. There are simulators (including ABSs) that apply clustering techniques. There are also MASs that benefit from clustering techniques. In addition, some algorithms analyze social networks. However, to the best of authors' knowledge, there is not any work that applies a hybrid approach

with an ABS and a clustering technique for simulating sociograms based on the psychological features. This gap of the literature is covered with the current approach, which is introduced in the next sections.

3. Psychological Instruments

The current approach analyzes the psychological features of people in order to classify them in certain types (i.e. clusters). This work chooses the psychological features that are most related with selections and rejections within a group, so that the ABS can properly simulate these.

Research has shown that social acceptance and rejection is associated to aggression. Rejected adolescents are characterized by high levels of aggression [7], whereas popular (i.e., accepted) adolescents by low levels [37]. Nonetheless, more recent studies indicate that although some aggressors are rejected by their peers, others are accepted [25]. Like aggression, victimization is also related to social status. Data obtained through peer-report informed that victimization is associated with more peer rejection and less peer acceptance [51], although some victims are popular among their peers [25]. Aggression and victimization are not only related to social status, but also they are related to each other. Specifically, victimization is associated with more relational aggression [50]. Therefore, the current approach selects three types of relational aggression and the relation victimization as psychological features for classifying people with clustering.

In particular, the Relational Aggression Scale was assessed by means of the Aggressive Behavior Questionnaire of Little et al. [31]. This scale consists of 12 items and evaluates three types of relational aggression: pure (e.g., "I am a person who does not allow others to come into my group of friends"), reactive (e.g., "When somebody gets angry with me, I tell my friends to avoid contact with that person"), and instrumental (e.g., "To get what I want, I either treat others with indifference or stop talking to them"). The answers were expressed on a seven-point Likert scale, ranging from one ("Do not agree at all") to seven ("Agree completely"). The reliability of the subscales (Cronbach's alpha) in this study was of .84, .79 and .90 respectively.

To measure Relational Victimization Scale, the current approach applies the Multidimensional Victimization Scale by Mynard and Joseph [36]. The Relational Victimization Scale uses ten items to measure the exposure to behaviors aimed at damaging relationships or one's social reputation, such as exclusion, manipulation, and rumor-spreading [11]. An example item is "A

classmate has told the others not to have anything to do with me". Responses are rated on a scale of one (never) to seven (always). Cronbach's alpha of this scale was .94 in this research.

The sociometric test designed in this study consisted of two questions to respectively measure peer acceptance ("Who would you choose as team mate in the class?") and peer rejection ("Who would you NOT choose as team mate in the class?"). Each person replies with a list of peers for each question, and sociograms can graphically represent these responses.

Based on the answers given by the participants or their representation (i.e. the sociogram), it is possible to measure the following sociometric indices: Cohesion (the degree of mutual acceptance between members of a group; IAg), Dissociation (the degree of mutual rejections between members of a group; IDg), Coherence (the relation between reciprocal acceptance and the number of total acceptance elections received by the members of a group; ICg) and Group Intensity (the degree of acceptance and rejection elections received by the members of a group; IIg). These indices were measured for each group from the selections and rejections of participants, according to the definitions and formulas of these indices provided by Barrasa and Gil [4].

4. Hybrid approach with agent-based simulation and clustering for sociograms

The current approach combines agent-based simulation with clustering for simulating sociograms based on the psychological features of the individuals in certain groups. The current approach is implemented with the CLUS-SOCI tool. The first subsection of this section introduces an overview of CLUS-SOCI and its modules, while their remaining subsections respectively describe the most relevant ones of these.

4.1. Overview of CLUS-SOCI

CLUS-SOCI is composed of several modules, which are (1) the clustering module, (2) the management module (3) and the ABS module. Figure 1 illustrates the main functioning of CLUS-SOCI with its main components.

CLUS-SOCI is built on the basis that the researcher has a large set of existing sociograms of people with their psychological features. In this manner, CLUS-SOCI can be trained with these data. In particular, the clustering module is responsible for extracting the most relevant types of people according to the psychological features indicated in section 3. This module performs

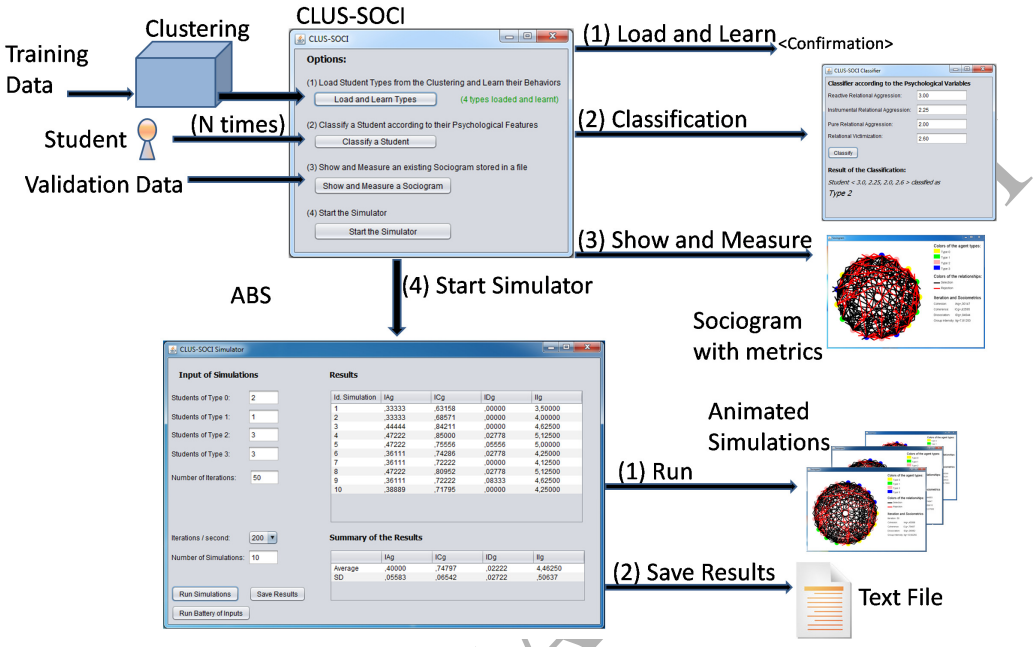


Figure 1: Overview of CLUS-SOCI.

a clustering process, in which the output is (a) several types of individuals (i.e. clusters) and (b) a formula for classifying any individual into one of these types according to their psychological features. Besides the clustering, this module also extracts some relevant information about some ratios of different kinds of relationships between the individuals of each possible pair of individual types.

The second module provides a Graphical User Interface (GUI), so that users can manage CLUS-SOCI and perform the following operations:

1. *Load and learn types*: This operation loads the types that are provided by the clustering module. CLUS-SOCI learns the social behaviors of different types, and incorporates these behaviors within the simulator.
2. *Classify a person*: This operation automatically classifies any person within one of the types loaded in the previous operation. This classification is based on their psychological features.
3. *Show and measure a sociogram*: This operation retrieves a sociogram represented in a file with a specific format, and graphically shows this sociogram. This operation also measures the corresponding sociomet-

rics. This operation has been useful in the validation phase, to graphically compare an existing sociogram with the corresponding simulated one.

4. *Start the simulator*: This operation starts running the ABS, so users can perform simulations. The ABS is configured with the loaded types and their corresponding behaviors. This ABS is implemented within the next module.

The ABS module simulates the social status of a group of people. Users introduce the number of students of each type, according to the classification performed in the previous module, and other input parameters. The ABS can perform several simulations with the same parameters. The evolution of each simulation is represented with a graphical animation of the corresponding sociogram. The results are presented in the GUI and can be saved into a text file.

The people of groups are referred as students in some cases from this point forward, as CLUS-SOCI was adapted to be mainly experienced in groups of students.

4.2. Clustering module of CLUS-SOCI

The goal of the clustering module is to detect types of students based on their psychological characteristics. The psychological characteristics generally influence the establishment of relationships between them, as previously discussed in section 3. By reducing the information to a few prototypes (cluster centers), each student will be classified to be of a certain type, in agreement with previous studies on sociograms [4, 41]. Thus the main input of this module is a set of students' identification numbers (IDs) with their psychological variables, obtained through standardized questionnaires. The input is in a simple text file format for training, in which there is one row per student. The main output of the module is the label of each student and, more important, the cluster information that allows classifying a new student.

In order to select the algorithm for clustering, several characteristics were taken into account. The input vector dimension is only four, which is not very high, and all the dimensions are similar in range. Students with close psychological variables should be in the same group and hyperellipsoid shaped clusters are expected. Thus, the current work ruled out many algorithms that are especially suited for clustering in high dimension spaces and for fitting

complex manifolds. Therefore, appropriate candidates were classical general purpose algorithms like K-means or a Gaussian Mixture Model (GMM) [27, 48]. Clusters found by GMM are more flexible in terms of cluster shape, so that GMM was finally selected. The underlying assumption of GMM is that samples are drawn from one of several distributions. This assumption agrees with the idea of the existence of several psychological types. In addition, GMM allows classifying a new sample once the model is built, which is a requirement in our system.

Selecting the number of clusters is not an easy task and there is no golden rule [27]. In this work, this number was chosen with the Bayesian Information Criteria (BIC), which has some advantages. Firstly, it gives an unambiguous number, as far as a minimum can be found with respect to the number of clusters. Secondly, it is not a heuristic approach, but it is based on a model selection approach [27]. It is an approximation to the model evidence [8] that takes into account how the model fits the data but including a term that penalizes the number of parameters in the model. Its use with GMM was proposed in [42].

All the programs were done in Python using the Scikit-learn library for Machine Learning [38]. In GMM, the probability of an input vector $\mathbf{X}$ is given by a sum over all clusters:

$$p(\mathbf{X}) = \sum_{c \in C} \pi_c \mathcal{N}(\mathbf{X}|\mathbf{m}_c, \mathbf{\Lambda}_c^{-1}) \tag{1}$$

where C is the set of clusters, π_c is the mixture weight of cluster c, $\mathbf{m}_c$ is the mean of cluster c, $\mathbf{\Lambda}_c$ is the precision matrix of cluster c (inverse of covariance) and $\mathcal{N}$ indicates the normal distribution.

Once the model is built with a training set, for a given new input $\mathbf{Y}$, the posterior probability that it belongs to cluster c, a kind of mark $r(\mathbf{Y}, c)$, is given in log-scale by (removing unnecessary constants):

$$r(\mathbf{Y}, c) = log\, p(c|\mathbf{Y}) = log(\pi_c) + 0.5\, log(|\mathbf{\Lambda}_c|) - 0.5\, (\mathbf{Y} - \mathbf{m}_c)^T \mathbf{\Lambda}_c (\mathbf{Y} - \mathbf{m}_c) \tag{2}$$

where $||$ denotes the determinant and T the transpose. Thus, the new input is assigned to the label of the cluster for which $r(\mathbf{Y}, c)$ is maximum.

In order to let the classification module classify a new input, the following values are stored in a text file:

- The number of clusters "NC".

- The number of psychological variables "NV".
- $\mathbf{m}_c = \{m_{c,1}, m_{c,2}, ..., m_{c,NV}\}$: List of the means of the psychological variables for each particular c cluster.
- b_c: The mark of each c cluster that is calculated beforehand, defined as $b_c = log(\pi_c) + 0.5\, log(|\mathbf{\Lambda}_c|)$. Note that this value does not depend on the input vector.
- $\mathbf{\Lambda}_c$: Precision matrix of each c cluster.

In this manner, the management module can load the GMM model when loading the student types, and the classifier can use it for allowing users to classify students with a GUI. In the current approach, the input vector is composed of the four psychological variables that were introduced in section 3.

The clustering module also processes a set of sociograms included in the training set. The purpose of this is to find the ratios of relationships between each possible pair of types. These are calculated as the number of actual relationships of a given kind (i.e. between students of a particular pair of types) divided by the number of possible relationships of this kind. In a group of students, this number of possible relationships is obtained by multiplying the number of students of the first type and the number of students of the second type. The module obtains as well how probable is that a relationship changes between two students for each possible pair of types, by comparing the relation of students in two different times. This value is necessary as later the tool simulates the evolution of sociograms through time. The details of the calculations are included later in this section.

The input sociograms are stored in text format as matrices indexed by students' IDs. The value (ID_1,ID_2) is one if student ID_1 selected student ID_2 as friend, or zero otherwise. A similar format applies for the rejection relations. Given a set of sociograms and the labels of students (obtained after the clustering), the clustering module outputs the values Q_{xy}, W_{xy} and Z_{xy} for selection, rejection and time evolution of relationships respectively. In order to illustrate the calculation of the values used in the learning process of CLUS-SOCI, firstly some basic notations are defined as follows:

- $C = \{t_0, t_1, t_2, ..., t_N\}$: the set of Clusters (i.e. the student types), where $t_0, t_1, t_2, ..., t_N$ are the corresponding types.

- G: the set of the Groups of students used in the training process.
- T_{xg}: the set of students that are classified as t_x type within a specific g group according to the clustering, in which $t_x \in C$ and $g \in G$.
- $S_g = \{< i, j > \mid i \text{ selects } j \text{ in } g \text{ group}\}$: the set of actual selection relationships within students of the specific g group.
- $R_g = \{< i, j > \mid i \text{ rejects } j \text{ in } g \text{ group}\}$: the set of actual rejection relationships within students of the specific g group.

Firstly, this module considers the ratios calculated as the number of selection relationships from students of a particular type to students of another type divided by the number of possible relationships in the specific pair of types, in a particular group. These ratios are denoted as Q_{xyg} for each pair of types $< t_x, t_y >$ in a g group, and are calculated with equation 3.

$$Q_{xyg} = \frac{|\{< i, j > \in S_g | i \in T_{xg} \wedge j \in T_{yg}\}|}{|T_{xg}| \cdot |T_{yg}|} \tag{3}$$

The aforementioned ratios are calculated for each group of students. However the training data is composed of several groups. In particular, these ratios need to be combined to obtain global ratios for each pair of types. These ratios are denoted as Q_{xy}. This work proposes two options, in which the first option is referred as Q'_{xy} to distinguish it. The first option is to calculate Q'_{xy} as the average of Q_{xyg} for all the groups $g \in G$. This option is calculated with equation 4.

$$Q'_{xy} = \frac{\sum\limits_{g \in G} Q_{xyg}}{|G|} \tag{4}$$

The second option is to calculate Q_{xy} as the sum of all selection relationships for a pair of types within all the groups divided by the sum of possible relations in all groups. This can be calculated with equation 5.

$$Q_{xy} = \frac{\sum\limits_{g \in G} |\{< i, j > \in S_g | i \in T_{xg} \wedge j \in T_{yg}\}|}{\sum\limits_{g \in G} |T_{xg}| \cdot |T_{yg}|} \tag{5}$$

The second option Q_{xy} has been selected for CLUS-SOCI as it considers all the students of each type with a similar weight regardless their groups.

By contrast the other option Q'_{xy} had the problem that it did not satisfy this condition. The average of all groups was considered equally for all the groups. Students of small groups influenced more in the corresponding Q_{xyg} of its group than students of large groups. Hence, this difference of influence was also present in the global average of Q'_{xy}.

Therefore, CLUS-SOCI uses the ratios Q_{xy} for all the possible pairs of types. These ratios are mathematically represented with a matrix of two dimensions called $\mathbf{Q}$. In this matrix, each position (x, y) contains Q_{xy}. This matrix has as many rows and columns as the number of student types, and is shown below:

$$\mathbf{Q} = \begin{pmatrix} Q_{00} & Q_{01} & ... & Q_{0N} \\ Q_{10} & Q_{11} & ... & Q_{1N} \\ ... & ... & ... & ... \\ Q_{N0} & Q_{N1} & ... & Q_{NN} \end{pmatrix}$$

In a similar way, this module also obtains the ratios of rejections in each possible pair of types. These ratios were calculated for each group, and these are denoted as W_{xyg} with a similar meaning to Q_{xyg} but considering rejections (i.e. R_g) instead of selections (i.e. S_g). This work considered two similar alternatives for combining W_{xyg} as in Q_{xyg}. CLUS-SOCI also selected the alternative that equally takes the rejections of all students into account regardless their groups. These ratios are denoted as W_{xy}, and are defined with equation 6. All these values are stored in a matrix called $\mathbf{W}$, which contains the rejection ratios W_{xy} for all the possible pairs of types t_x and t_y.

$$W_{xy} = \frac{\sum_{g \in G} |\{< i, j > \in R_g | i \in T_{xg} \wedge j \in T_{yg}\}|}{\sum_{g \in G} |T_{xg}| \cdot |T_{yg}|} \tag{6}$$

Finally, the present approach also takes into account how probable is that a relationship changes between two students for each possible pair of types. For this purpose, the training data must contain sociograms of the same groups in two different times. For instance, each group of students can be surveyed at the beginning of the academic year (e.g. just after its first month) for conforming an initial sociogram and then be surveyed several months after to conform the final sociogram (e.g. in the sixth month of the academic year). In particular, this approach considers the ratios of relationships that changed for each pair of possible types.

In order to express these ratios, the groups of both selections and rejections now also need to explicit mention the time (i.e. zero for initial sociograms and one for final sociograms). It is worth mentioning that the previous definitions considered the selections and rejections of final sociograms. The next equations use the following definitions of selections and rejections considering the time:

- $S_{g0} = \{< i, j > \mid i$ selects j in g group at time zero $\}$: the set of actual selection relationships within students of the specific g group in the initial time.
- $S_{g1} = \{< i, j > \mid i$ selects j in g group at time one $\}$: the set of actual selection relationships within students of the specific g group in the final time.
- $R_{g0} = \{< i, j > \mid i$ rejects j in g group at time zero $\}$: the set of actual rejection relationships within students of the specific g group in the initial time.
- $R_{g1} = \{< i, j > \mid i$ rejects j in g group at time one $\}$: the set of actual rejection relationships within students of the specific g group in the final time.

In order to count the relationships that changed in a sociogram of a group, this work uses the union of (1) the set with differences in selection and (2) the set with the differences in rejections. The sets with differences consider both directions: the relations of a particular kind (i.e. either selections or rejections) that appeared and the ones that disappeared. In particular, F_g denotes the set of relationships that have changed in the g group from the initial time to the final time. This set is calculated with equation 7.

$$F_g = (S_{g1} - S_{g0}) \cup (S_{g0} - S_{g1}) \cup (R_{g1} - R_{g0}) \cup (R_{g0} - R_{g1}) \tag{7}$$

The ratios are calculated as the cardinality of the resulting changing relations set for a particular pair of types (i.e. subset of F_g) divided by the number of possible relations for this pair of types. Specifically, the set of Z_{xyg} denotes this ratio of relation changes for the t_x and t_y types for a particular g group, and is calculated with equation 8.

$$Z_{xyg} = \frac{|\{< i, j > \in F_g | i \in T_{xg} \wedge j \in T_{yg}\}|}{|T_{xg}| \cdot |T_{yg}|} \tag{8}$$

This ratio of changes needs to be calculated for all the groups. This is proposed to be performed either by calculating the average for all the groups (see the definition of Z'_{xy} in equation 9) or by summing all the relations changed and dividing it by the number of possible relations for a specific pairs of types (see the definition of Z_{xy} in equation 10).

$$Z'_{xy} = \frac{\sum_{g \in G} Z_{xyg}}{|G|} \tag{9}$$

$$Z_{xy} = \frac{\sum_{g \in G} |\{< i, j > \in F_g | i \in T_{xg} \wedge j \in T_{yg}\}|}{\sum_{g \in G} |T_{xg}| \cdot |T_{yg}|} \tag{10}$$

CLUS-SOCI selected the second option Z_{xy} as it equally considers the changes of all students regardless their group, which is not satisfied by Z'_{xy} for similar reasons as the ones mentioned for Q'_{xy}. CLUS-SOCI learns the ratios Z_{xy} for all possible pairs of types, and stores these in a matrix called **Z**. In this matrix, the position (x, y) contains the value for the pairs of types $< t_x, t_y >$. This **Z** matrix has the same number of rows and columns as the number of types, and is shown below:

$$\mathbf{Z} = \begin{pmatrix} Z_{00} & Z_{01} & ... & Z_{0N} \\ Z_{10} & Z_{11} & ... & Z_{1N} \\ ... & ... & ... & ... \\ Z_{N0} & Z_{N1} & ... & Z_{NN} \end{pmatrix}$$

4.3. ABS for simulating sociograms

The simulator module provides an ABS for simulating sociograms. This system was developed following PEABS (A Process for developing Efficient Agent-Based Simulators) [19].

The agents of this ABS can play three different roles that are respectively denoted as simulator, student and sociogram manager. In particular, the simulator agent guides the simulation according to certain parameters introduced by the user. The student agents simulate the social behavior of students considering their types based on their psychological features. The interactions between student agents simulate their collaborative participation in class. The sociogram manager agent updates a sociogram taking the social interactions of student agents into account.

The current ABS considers the following common facts about sociograms proven in the literature:

- The student profiles influence the establishment of relations among students [13].
- The evolution of the grades of relationships usually follows some trends, although these trends vary sometimes [30].
- The trend changes are less frequent in long-term relationships than in short-term relationships [23].
- The group size commonly influences the relationships between the different profiles [41].
- In sociograms, there is usually a high proportion of relationships that are reciprocal in comparison to all the relationships [17]. In other words, a person X is more likely to choose a person Y when Y chooses X [12].

The sociograms are represented as weighted graphs. The nodes represent the agents, and the edges represent their relationships with real weighs in the [-1, 1] interval. The weighs near zero according to certain threshold represent neutral relationships. Otherwise, positive weights represent selection relationships, while negative weights represent rejection relationships.

In a conversation between two student agents, their messages are influenced by their psychological profiles. These messages can be either friendly, neutral or hostile, and consequently these can affect their relationship in different ways. The probability of sending friendly messages depends on the affinities of student types (i.e. the **Q** matrix obtained from the clustering as described in section 4.2), while the probability of sending hostile messages depends on the dislikes between student types (i.e. the **W** matrix from the clustering).

In order to consider the relationship trends, there is another graph that represents these trends. The relationships usually evolve following their trends. In particular, the relationships either increases or decreases a percentage of the distance to the approaching limit for respectively positive and negative trends. In this way, the relationship can continuously evolve without reaching the limits. In case of a neutral trend, the corresponding relationship does not vary.

In some cases, these trends can vary considering different aspects such as the aforementioned types of messages in student conversations. More specifically, the following steps determine whether each relationship trend changes:

- *Changeability between student types*: It considers the **Z** matrix from the clustering for determining the changeability between the corresponding student types. The corresponding matrix value is multiplied with a constant regarding its size level. In this step and some of the following ones, the corresponding probability is simulated by generating a random number in the [0, 1] interval, and the test passes only if it is lower than the probability value. If the test fails in this step, the current process stops without changing the relationship trend.

- *Test of long-term relations.* Since the long-term relations are considered to change less frequently, the probability of $CST\_TREND/iteration$ is simulated if this value is lower than one. The iteration is represented with its chronological position, and consequently this probability decreases when the iteration increases. If this test fails, the process ends without altering the relationship trend.

- *Consideration of the type of messages*: It changes the trend of the relationship into either positive, neutral or negative regarding the type of messages of the conversation.

- *Change of the reciprocal relationship trend*: It simulates the probability of the CST_COHERENCE constant value according to the size level. If the test passes, the trend of the reciprocal relationship is set to the same value as the trend mentioned in the previous step.

In order to consider the size of the group, CLUS-SOCI applies certain lists of constant values that multiply **Q**, **W** and **Z** matrices for certain size levels. These lists of constants are respectively called Q_FACTOR, W_FACTOR and Z_FACTOR. The constants CST_TREND and CST_COHERENCE also have different values regarding the size levels. A data structure determines the size levels by indicating their limits. All these constants are adjusted following the White Box Calibration process proposed by Fehler et al. [18]. This process considers the different influences of these constants in the different features of sociograms such as cohesion, dissociation and coherence.

Another alternative would have been to learn **Q**, **W** and **Z** matrices separately for sociograms of different size levels. However, the values of these matrices are usually more accurate for larger numbers of participants (i.e. members of all groups). In particular, CLUS-SOCI obtained appropriate results when learning **Q**, **W** and **Z** from all the participants and then adjusting these values with the aforementioned multiplier constants concerning size levels.

In this way, the current approach can be efficiently applied for a group of any size even when there are only one or very few training examples of the same size level. In case of applying the current approach to huge-sized networks outside the size levels of the training examples, the current approach proposes to apply the following formula for obtaining the corresponding factor for the **Q**, **W** and **Z** matrices:

$$X\_FACTOR = K/n \tag{11}$$

In this equation, X_FACTOR can be either Q_FACTOR, W_FACTOR or Z_FACTOR. The n value represents the size of the group (i.e. the number of students). K represents the K_Q, K_W and K_Z constants for the different matrices. In this manner, the probabilities of these factors decrease for larger size levels as it is common in sociograms. K is obtained from the largest size level of the training sample with the following equation:

$$K = X\_FACTOR' \cdot n' \tag{12}$$

where X_FACTOR' is the corresponding trained factor from the largest size level of the training data, and n' is the average size of this level.

In the case of CST_TREND and CST_COHERENCE, the approach uses the same values as in the largest trained size level, since these constants have neither a clear increasing nor decreasing trend depending on the size level as it was observed in the experimentation.

Therefore, in this way, the current approach can be applied to any size of groups, even for complex huge-sized networks with larger sizes than the training data. However, the results will probably be more accurate when the training sample contains some networks with similar sizes.

5. Experimentation

In order to assess the reliability of the current approach, CLUS-SOCI has been experienced and compared with real data from student sociograms.

In particular, the real data were obtained by surveying 714 students (350 males and 364 females) with a mean age of 14.0 years (SD = 1.44). The students were attending four different schools in the Autonomous Community of Aragón (Spain) and 38 classrooms. Students were distributed in first, second, third and fourth year of secondary school classes according to the following percentages: 23.4%, 17.8%, 31.8% and 27.0%, respectively. These students were surveyed two times with about five months of difference. In this manner, the evolution of relations was also registered. This experimentation extracted 38 real sociograms in a particular time (referred as time one), and the same number of real sociograms were captured for the same student classes in a previous time (denoted as time zero).

The survey was conducted in compliance with the ethical standards of the American Psychological Association (APA). Firstly, the approval from the Provincial Board of Education and Science was obtained to perform the study. Secondly, we contacted the principal of each school to explain the aim of the research and requested their permission to conduct the study at their school. Next, passive consent was obtained from parents or guardians; they received written notice from the school that their children would be participating and were invited to contact the school if they did not want their child to participate. On the day of the survey, students were invited to participate and assured that the survey was confidential and voluntary. With the purpose of trying to reduce the possible effect of social desirability, they were informed that the researchers were interested in knowing what they thought and felt about themselves and that there were not right or wrong answers. Students filled out the questionnaires in classrooms two times at different occasions. The first time took place in November 2010, while the second time took place in April 2011.

At least one qualified researcher (i.e. with Ph.D.) was present during the administration of the instruments to provide students with the necessary support to complete the questionnaires. To nominate a classmate on the sociometric test, students had to select the code number associated to the classmate. To make this possible, a roster of the grade, with the names and their associated code numbers, was written on the blackboards of the classrooms. The psychological features of the participants were measured with the instruments previously introduced in section 3.

In order to validate the current approach, the real data were divided into a training set and a validation set. In particular, 26 out of 38 student groups were randomly selected as training data, and the remaining 12 student groups

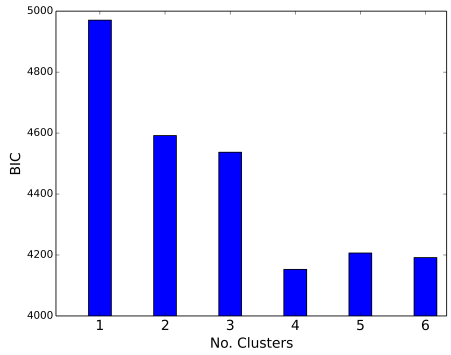


Figure 2: BIC as a function of the number of clusters.

were used later in the validation phase.

The clustering process was performed from the students and their psychological features by means of the module of CLUS-SOCI introduced in section 4.2. This clustering process was only applied in the training set of data. Figure 2 shows the results of BIC considering the different numbers of clusters. According to the BIC criterion, the right number of clusters is four. The clusters are referred as student types from this point forward. These types are labeled as numbers from zero to three.

The relations of students were automatically examined by the clustering module of CLUS-SOCI, and then these were learned by the management module. As result, CLUS-SOCI incorporated the **Q**, **W** and **Z** matrices that were calculated as described in section 4.2. Table 1 presents the values of these matrices learned from the set of training data. As mentioned before, these matrices determine respectively the affinities, dislikes and changeability between each possible pair of student types.

Then, CLUS-SOCI was trained to learn some constants for certain size levels, as previously introduced in section 4.3. Table 2 indicates the size intervals and the constants that were learned. In order to support scalability, some of the values of the last size level are expressed in terms of the size (referred as n). In this manner, the ABS of CLUS-SOCI is completely configured to perform simulations.

After the training, the ABS was executed for the same student groups of the training data as an additional part of validation (the results will be discussed later in this section). Two groups out of 26 were discarded be-

Q matrix (affinities)				
	Type 0	*Type 1*	*Type 2*	*Type 3*
Type 0	0.45439	0.37011	0.32348	0.43659
Type 1	0.36322	0.34009	0.28936	0.33157
Type 2	0.31199	0.27872	0.28533	0.29573
Type 3	0.46014	0.33157	0.36128	0.40331
W matrix (dislikes)				
	Type 0	*Type 1*	*Type 2*	*Type 3*
Type 0	0.16047	0.17011	0.15764	0.07609
Type 1	0.16092	0.18919	0.17447	0.09700
Type 2	0.18227	0.18511	0.15067	0.08841
Type 3	0.11051	0.09877	0.16159	0.08011
Z matrix (changeability)				
	Type 0	*Type 1*	*Type 2*	*Type 3*
Type 0	0.37692	0.40193	0.41020	0.35952
Type 1	0.38264	0.37086	0.39628	0.32360
Type 2	0.37694	0.32508	0.31076	0.28542
Type 3	0.41190	0.34063	0.40041	0.35273

Table 1: Learned values of **Q**, **W** and **Z** matrices from the training data.

Size Intervals	**[0, 10]**	**[11, 17]**	**[18, 20]**	**[21, 25]**	**[26, 30]**	**[31, ∞]**
Q_FACTOR	5.00000	1.00000	0.80000	0.60000	0.50000	$14.000/n$
W_FACTOR	1.70000	1.13000	0.85000	0.75000	0.50000	$14.000/n$
Z_FACTOR	2.00000	1.58000	1.00000	0.90000	0.30000	$8.400/n$
CST_TREND	10	10	10	10	10	10
CST_COHERENCE	0.40000	0.40000	0.40000	0.25000	0.40000	0.40000

Table 2: Learned size intervals and constants for these intervals, where n denotes the group size.

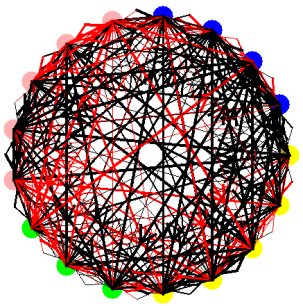
cause of their sizes, since sociograms of too small sizes are not considered representative in the training process.

After the whole training process, the ABS of CLUS-SOCI was executed for the set of validation data. These data are constituted with different students and different groups from the ones used in the training process. The classification of each student was performed with the clusters (i.e. student types) already learned in the previous training phase.

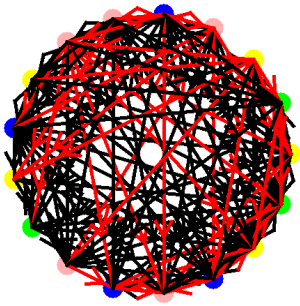
In the validation set, Figure 3 shows some simulated sociograms and the corresponding real ones. In particular, this figure shows the sociograms for the groups identified as 013A, 032P and 032A as examples. These sociograms were either simulated or loaded with the CLUS-SOCI. All the sociograms of these examples use the same colors for the same student types. The black arrows represent selection relationships, while red arrows represent rejection relationships. It is worth mentioning that the ABS of CLUS-SOCI arranges all the students of each type together as it only receives the number of students of each type. However, the real sociograms place the students in a different order, since it uses the order in which the students were loaded from the corresponding file. One can observe that each example of simulated sociogram is quite similar to the corresponding real one. However, proving their similarity requires further analysis and taking all the validation cases into account.

In order to assess whether the ABS of CLUS-SOCI actually provides sociograms similar to the real ones, a statistical test has been performed for comparing their sociometric indices in all the cases, presenting the results separately for the training set and the validation set. This test was carried out with a nominal alpha level of .05. Since in this work all sociometric indices were proportions, except for IIg, the work required a statistical test that assesses whether the difference between two proportions (a real and an estimated one) is significant. The binomial test is an adequate test for this. To apply the same statistical test to all indices, IIg was converted into a proportion (named T-IIg) by dividing it by the total number of the members of the group (the maximum value that this index can take).

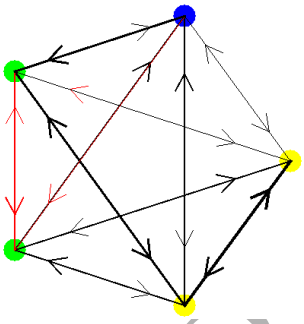
Table 3 presents the results of comparing the real values and the values estimated with CLUS-SOCI, in the training set of data. Table 4 presents the comparison of the true values and the estimated values for the validation set. Both comparisons applied the binomial tests for each sociometric index. Reported alpha levels were adjusted using the Bonferroni procedure to control for family-wise error rate. Specifically, all the tests performed within the



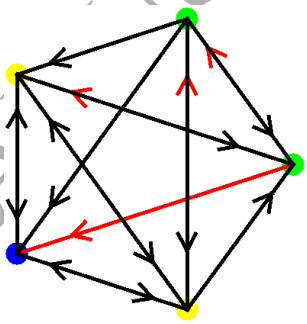
(a) Simulated sociogram for 013A group.



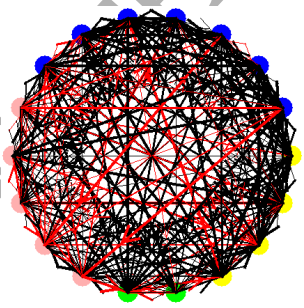
(b) Real sociogram of 013A group.



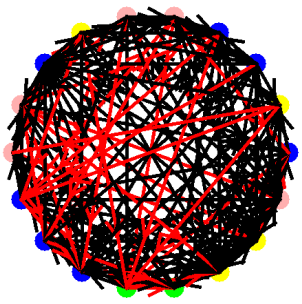
(c) Simulated sociogram for 032P group.



(d) Real sociogram of 032P group.



(e) Simulated sociogram for 032A group.



(f) Real sociogram of 032A group.

Figure 3: Graphical comparisons of simulated and real sociograms.

Group ID	Group Size	Real Value					Estimated Value					Difference Risk			
		IAg	ICg	IDg	IIg	T-IIg	IAg	ICg	IDg	IIg	T-IIg	IAg	ICg	IDg	T-IIg
042A	28	.13	.52	.02	6.78	.24	.15	.52	.01	5.67	.20	-.02	.00	.01	.04
05E3C	28	.12	.49	.02	5.26	.19	.14	.48	.02	5.47	.20	-.02	.01	.00	-.01
042B	27	.18	.53	.01	7.38	.27	.15	.51	.02	5.52	.20	.03	.02	-.01	.07
043C	25	.12	.41	.01	4.42	.18	.17	.49	.03	7.66	.31	-.05	-.08	-.02	-.13
05E3A	24	.07	.29	.04	4.91	.20	.19	.51	.03	7.65	.32	-.12	-.22	.01	-.12
024B	23	.18	.46	.02	5.91	.26	.18	.50	.04	7.03	.31	.00	-.04	-.02	-.05
023B	22	.27	.58	.08	9.57	.44	.15	.47	.05	6.73	.31	.12	.11	.03	.13
043B	22	.15	.61	.08	7.57	.34	.20	.51	.03	6.80	.31	-.05	.10	.05	.03
05E4C	22	.17	.55	.00	5.90	.27	.14	.45	.04*	6.22	.28	.03	.10	-.04	-.01
043A	20	.11	.51	.01	3.89	.19	.14	.47	.03	4.86	.24	-.03	.04	-.02	-.05
05E4A	20	.16	.45	.02	5.63	.28	.20	.55	.02	5.39	.27	-.04	-.10	.00	.01
032B	19	.18	.42	.01	6.11	.32	.18	.54	.03	5.32	.28	.00	-.12	-.02	.04
021A	18	.15	.55	.04	5.65	.31	.17	.52	.04	4.99	.28	-.02	.03	.00	.03
021B	18	.15	.56	.01	4.88	.27	.20	.55	.02	5.18	.29	-.05	.01	-.01	-.02
031A	18	.24	.57	.06	8.00	.44	.19	.54	.03	4.82	.27	.05	.03	.03	.17
034B	18	.13	.45	.05	5.12	.28	.16	.49	.04	4.86	.27	-.03	-.04	.01	.01
011A	17	.32	.60	.03	7.69	.45	.32	.66	.03	7.12	.42	.00	-.06	.00	.03
031B	17	.27	.60	.05	7.31	.43	.29	.63	.04	6.85	.40	-.02	-.03	.01	.03
033B	17	.32	.66	.09	10.00	.59	.22	.59	.05	5.98	.35	.10	.07	.04	.24
012A	14	.35	.71	.08	7.00	.50	.26	.60	.04	5.39	.39	.09	.11	.04	.11
034A	13	.33	.68	.00	5.42	.42	.29	.64	.03*	5.30	.41	.04	.04	-.03	.01
032C	12	.30	.66	.02	5.45	.45	.26	.61	.03	4.49	.37	.04	.05	-.01	.08
014A	11	.37	.61	.00	4.60	.42	.25	.58	.04*	4.03	.37	.12	.03	-.04	.05
024DV	9	.38	.61	.00	4.38	.49	.43	.73	.00	3.90	.43	-.05	-.12	.00	.06
Mean		.21	.55	.03	6.20	.34	.21	.55	.03	5.72	.31	.01	.00	.00	.03
SD		.09	.10	.03	1.60	.11	.07	.07	.01	1.07	.07	.11	.07	.07	.01

Table 3: Real values, estimated values in the training phase and difference risks for sociometric indices.
*$p < .05$, based on the binomial test with the correction of Bonferroni.

same index were considered as a family of tests. Both tables report the significant differences from the real values, with an asterisk in the corresponding estimated value. With the purpose of measuring the similarity between the estimated value and the true value, this work calculated the difference risks between the predicted risk and the true risk. It also calculated the averages and the standard deviations of real, estimated and difference risk values.

In the training data, the binomial tests showed that none of the estimated values were statistically different from the true values for IAg, ICg and T-IIg indices. On the contrary, the IDg index had statistically significant differences in three out of twenty four estimations (11.5%).

Results obtained on validation data were parallel to those results observed in training data. None of the estimated values were statistically different from the true values for any sociometric indices, except for IDg index. Binomial tests performed on this index revealed that only one out of twelve estimations

Group ID	Group Size	Real Value					Estimated Value					Difference Risk			
		IAg	ICg	IDg	IIg	T-IIg	IAg	ICg	IDg	IIg	T-IIg	IAg	ICg	IDg	T-IIg
05E3B	29	.12	.43	.01	6.11	.21	.16	.52	.01	6.05	.21	-.04	-.09	.00	.00
041A	28	.18	.43	.04	9.11	.33	.18	.53	.01	6.29	.22	.00	-.10	.03	.11
041B	26	.15	.51	.06	9.20	.35	.14	.49	.02	5.22	.20	.01	.02	.04	.15
041C	24	.22	.59	.03	8.04	.34	.18	.50	.04	7.60	.32	.04	.09	-.01	.02
044A	22	.11	.45	.01	4.00	.18	.21	.52	.03	7.14	.32	-.10	-.07	-.02	-.14
05E4B	22	.08	.27	.02	4.57	.21	.14	.44	.04	6.23	.28	-.06	-.17	-.02	-.07
044C	19	.20	.52	.04	5.67	.30	.31	.64	.03	7.57	.40	-.11	-.12	.01	-.10
032A	18	.27	.56	.02	7.00	.39	.26	.62	.04	6.81	.38	.01	-.06	-.02	.01
013A	17	.29	.62	.04	7.69	.45	.26	.63	.04	6.61	.39	.03	-.01	.00	.06
033A	17	.27	.49	.00	6.00	.35	.23	.59	.06*	6.41	.38	.04	-.10	-.06	-.03
044B	16	.15	.58	.02	4.33	.27	.32	.66	.03	6.69	.42	-.17	-.08	-.02	-.15
032P	5	.45	.60	.05	3.50	.70	.30	.62	.00	1.60	.32	.15	-.02	.05	.38
Mean		.21	.50	.03	6.27	.34	.23	.56	.03	6.19	.32	-.02	-.06	.00	.02
SD		.10	.10	.02	1.96	.14	.07	.07	.02	1.59	.08	.09	.07	.03	.15

Table 4: Real values, estimated values in the validation phase and difference risks for sociometric indices.
*$p < .05$, based on the binomial test with the correction of Bonferroni.

(8.3%) reached a significant level of difference.

This parallelism between results obtained on training phase and on validation phase support the idea that CLUS-SOCI performs equally well on original sample as on new sample, discarding good performance as a consequence of over-fitting on original data. Results on training and on validation phases indicated that CLUS-SOCI estimates correctly IAg, ICg and T-IIg indices. However, for the IDg index, around 10% of the estimations were wrong. This advocates that for CLUS-SOCI group disassociation is harder to estimate than group cohesion, group coherence and group intensity.

6. Conclusions and future work

In conclusion, the CLUS-SOCI tool implements a hybrid approach that combines agent-based simulation and clustering for simulating sociograms. The clustering allows it to obtain a classification of individuals in the training phase. CLUS-SOCI also learns the behaviors of the different types by learning certain matrices and constants. These matrices and constants determine the way that agents of each type will socially interact within the ABS. Practitioners can classify each new individual according to their psychological features by means of CLUS-SOCI. They can also execute the ABS for a particular group with certain members classified with this approach. In this manner, users can simulate sociograms given a group of individuals with

certain psychological features.

CLUS-SOCI has useful practical applications on secondary education. Concretely, CLUS-SOCI has been designed to predict social interactions of classes based on student aggression and victimization levels. The final target of this tool is to predict social dynamics of classes, with the purpose of facilitating interventions to produce positive change on these. By knowing in advance some features of social interactions of classes, teachers can distribute students in classes, according to their psychological features, in a way that conflictual relationships would be diminished and friendly relations would be facilitated. In this way, not only well-being of students could be improved, but also academic performance.

The current approach has been experienced with 38 real sociograms that were composed of 714 students in total from four different secondary schools of the Aragón region of Spain. Two thirds of these data were used in the training phase, while the remaining third was used for validation. The validation experimentation concluded that the current approach provided simulated sociograms that did not have statistically significant differences from the real sociograms in terms of cohesion, coherence of reciprocal relationships and intensity. These results were corroborated with the binomial test applying the correction of Bonferroni.

In the future, the current approach is planned to be experienced in other levels of education such as higher education in universities. The experimentation of the current work is also planned to be enhanced by simulating sociograms in companies after training the system with employees and their corresponding psychological features. In this way, CLUS-SOCI will be evaluated if it properly works for simulating sociograms in different education levels and other contexts such as companies. CLUS-SOCI will be trained separately for each context or education level. After this experimentation, CLUS-SOCI may (1) improve the clustering process, (2) enhance the training phase, and (3) incorporate new rules for simulating sociograms.

Finally, the current approach will be improved by simulating more kinds of relations among people in sociograms. In particular, it will consider the perceptions of peer acceptance and peer rejection, which are obtained respectively with questions (a) "Who do you think have chosen you as team mate in the class?", and (b) "Who do you think have rejected you as a team mate in the class?" CLUS-SOCI will be trained taking into consideration some sociometric indices that use these new kinds of relations, and will simulate sociograms with these relations.

Acknowledgments

We acknowledge the "Fondo Social Europeo" and the "Departamento de Tecnología y Universidad del Gobierno de Aragón" for their support.

References

[1] Ahmed, N. M., Chen, L., 2016. An efficient algorithm for link prediction in temporal uncertain social networks. Information Sciences 331, 120–136.

[2] Andrei, A. L., Comer, K., Koehler, M., 2014. An agent-based model of network effects on tax compliance and evasion. Journal of Economic Psychology 40, 119–133.

[3] Ayala-Cabrera, D., Herrera, M., Izquierdo, J., Pérez-García, R., 2014. GPR data analysis using multi-agent and clustering approaches: A tool for technical management of water supply systems. Digital Signal Processing 27, 140–149.

[4] Barrasa, A., Gil, F., 2004. A software application for the calculus and representation of sociometric indexes and values (In Spanish Un programa informático para el cálculo y la representación de índices y valores sociométricos). Psicothema 16 (2), 329–335.

[5] Becu, N., Perez, P., Walker, A., Barreteau, O., Le Page, C., 2003. Agent based simulation of a small catchment water management in northern thailand: description of the catchscape model. Ecological Modelling 170 (2), 319–331.

[6] Bierman, K. L., 2004. Peer rejection: Developmental processes and intervention strategies. New York: Guilford Press.

[7] Bierman, K. L., Smoot, D. L., Aumiller, K., 1993. Characteristics of aggressive-rejected, aggressive (nonrejected), and rejected (nonaggressive) boys. Child development 64 (1), 139–151.

[8] Bishop, C. M., 2006. Pattern Recognition and Machine Learning. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

[9] Bliss, C. A., Frank, M. R., Danforth, C. M., Dodds, P. S., 2014. An evolutionary algorithm approach to link prediction in dynamic social networks. Journal of Computational Science 5 (5), 750–764.

[10] Borch, C., Hyde, A., Cillessen, A. H., 2011. The role of attractiveness and aggression in high school popularity. Social Psychology of Education 14 (1), 23–39.

[11] Crick, N. R., Grotpeter, J. K., 1996. Children's treatment by peers: Victims of relational and overt aggression. Development and Psychopathology 8 (02), 367–380.

[12] Davis, J. A., 1977. Sociometric triads as multi-variate systems. Journal of Mathematical Sociology 5 (1), 41–59.

[13] DeGenaro, J. J., 1988. Condensing the student profile. Academic Therapy 23 (3), 293–96.

[14] Di Nardo, A., Di Natale, M., Giudicianni, C., Musmarra, D., Santonastaso, G. F., Simone, A., 2015. Water distribution system clustering and partitioning based on social network algorithms. Procedia Engineering 119, 196–205.

[15] Dos Santos, D. S., Bazzan, A. L., 2012. Distributed clustering for group formation and task allocation in multiagent systems: A swarm intelligence approach. Applied Soft Computing 12 (8), 2123–2131.

[16] Elhabyan, R. S., Yagoub, M. C., 2015. Two-tier particle swarm optimization protocol for clustering and routing in wireless sensor network. Journal of Network and Computer Applications 52, 116–128.

[17] Engel, A., Coll, C., Bustos, A., 2013. Distributed teaching presence and communicative patterns in asynchronous learning: Name versus reply networks. Computers & Education 60 (1), 184–196.

[18] Fehler, M., Klügl, F., Puppe, F., 2005. Techniques for analysis and calibration of multi-agent simulations. In: Engineering Societies in the Agents World V. Vol. 3451 of Lecture Notes in Computer Science. Springer, pp. 305–321.

[19] García-Magariño, I., Gómez-Rodríguez, A., González-Moreno, J. C., Palacios-Navarro, G., 2015. PEABS: A Process for developing Efficient Agent-Based Simulators. Engineering Applications of Artificial Intelligence 46, 104–112.

[20] García-Magariño, I., Plaza, I., 2015. FTS-SOCI: An agent-based framework for simulating teaching strategies with evolutions of sociograms. Simulation Modelling Practice and Theory 57, 161–178.

[21] Garruzzo, S., Rosaci, D., 2008. Agent clustering based on semantic negotiation. ACM Transactions on Autonomous and Adaptive Systems (TAAS) 3 (2), 7.

[22] Gifford-Smith, M. E., Brownell, C. A., 2003. Childhood peer relationships: Social acceptance, friendships, and peer networks. Journal of School Psychology 41 (4), 235–284.

[23] Goldstone, S., Boardman, W. K., Lhamon, W. T., Fason, F. L., Jernigan, C., 1963. Sociometric status and apparent duration. The Journal of social psychology 61 (2), 303–310.

[24] Grund, T. U., 2012. Network structure and team performance: The case of english premier league soccer teams. Social Networks 34 (4), 682–690.

[25] Hoff, K. E., Reese-Weber, M., Schneider, W. J., Stagg, J. W., 2009. The association between high status positions and aggressive behavior in early adolescence. Journal of School Psychology 47 (6), 395–426.

[26] Howe, C., 2010. Peer groups and children's development. Oxford: Wiley-Blackwell.

[27] Jain, A. K., 2010. Data clustering: 50 years beyond K-means. Pattern recognition letters 31 (8), 651–666.

[28] Johnson, M., Paulusma, D., van Leeuwen, E. J., 2015. Algorithms for diversity and clustering in social networks through dot product graphs. Social Networks 41, 48–55.

[29] Kumpulainen, K., Räsänen, E., Puura, K., 2001. Psychiatric disorders and the use of mental health services among children involved in bullying. Aggressive Behavior 27 (2), 102–110.

[30] Leinhardt, S., 1972. Developmental change in the sentiment structure of children's groups. American Sociological Review, 202–212.

[31] Little, T., Henrich, C., Jones, S., Hawley, P., 2003. Disentangling the "whys" from the "whats" of aggressive behaviour. International Journal of Behavioral Development 27 (2), 122–133.

[32] López-Ortega, O., Rosales, M.-A., 2011. An agent-oriented decision support system combining fuzzy clustering and the AHP. Expert Systems with Applications 38 (7), 8275–8284.

[33] Macy, M. W., Willer, R., 2002. From factors to actors: Computational sociology and agent-based modeling. Annual review of sociology, 143–166.

[34] Maslow, A. H., 1943. A theory of human motivation. Psychological review 50 (4), 370.

[35] Moreno, J. L., 1951. Sociometry, experimental method and the science of society. An Approach to a New Political Orientation. New York: Beacon House, Beacon.

[36] Mynard, H., Joseph, S., 2000. Development of the multidimensional peer-victimization scale. Aggressive Behavior 26 (2), 169–178.

[37] Newcomb, A. F., Bukowski, W. M., Pattee, L., 1993. Children's peer relations: a meta-analytic review of popular, rejected, neglected, controversial, and average sociometric status. Psychological bulletin 113 (1), 99.

[38] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12, 2825–2830.

[39] Prenkert, F., Følgesvold, A., 2014. Relationship strength and network form: An agent-based simulation of interaction in a business network. Australasian Marketing Journal (AMJ) 22 (1), 15–27.

[40] Raberto, M., Cincotti, S., Focardi, S. M., Marchesi, M., 2001. Agent-based simulation of a financial market. Physica A: Statistical Mechanics and its Applications 299 (1), 319–327.

[41] Roberts, S., 2008. Using practitioner research to investigate the role of the teacher in encouraging student interaction within group work. Nurse education today 28 (1), 85–92.

[42] Roeder, K., Wasserman, L., 1997. Practical Bayesian density estimation using mixtures of normals. Journal of the American Statistical Association 92 (439), 894–902.

[43] Ryan, R. M., Deci, E. L., 2000. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. American psychologist 55 (1), 68.

[44] Serban, A., Yammarino, F. J., Dionne, S. D., Kahai, S. S., Hao, C., McHugh, K. A., Sotak, K. L., Mushore, A. B., Friedrich, T. L., Peterson, D. R., 2015. Leadership emergence in face-to-face and virtual teams: A multi-level model with agent-based simulations, quasi-experimental and experimental tests. The Leadership Quarterly.

[45] Serrano, E., Moncada, P., Garijo, M., Iglesias, C. A., 2014. Evaluating social choice techniques into intelligent environments by agent based social simulation. Information Sciences 286, 102–124.

[46] Trimm, D., Rheingans, P., desJardins, M., 2012. Visualizing student histories using clustering and composition. Visualization and Computer Graphics, IEEE Transactions on 18 (12), 2809–2818.

[47] Xu, K., Zou, K., Huang, Y., Yu, X., Zhang, X., 2015. Mining community and inferring friendship in mobile social networks. Neurocomputing 174 (B), 605–616.

[48] Xu, R., Wunsch, D., 2005. Survey of clustering algorithms. Neural Networks, IEEE Transactions on 16 (3), 645–678.

[49] Yu, Z., Dongsheng, C., Wen, W., 2014. The heterogeneous effects of ability grouping on national college entrance exam performance – evidence from a large city in china. International Journal of Educational Development 39, 80–91.

[50] Zimmer-Gembeck, M. J., Pronk, R. E., 2012. Relation of depression and anxiety to self-and peer-reported relational aggression. Aggressive behavior 38 (1), 16–30.

[51] Zimmer-Gembeck, M. J., Pronk, R. E., Goodwin, B., Mastro, S., Crick, N. R., 2013. Connected and isolated victims of relational aggression: Associations with peer group status and differences between girls and boys. Sex roles 68 (5-6), 363–377.