A clustering procedure for mixed-type data to explore ego network typologies: an application to elderly people living alone in Italy

Elvira Pelle[1] · Roberta Pappadà[2]

Accepted: 30 August 2021

Abstract

The analysis of ego networks has attracted a great attention recently and found application in many areas of the social sciences. In particular, the identification of network typologies has become a crucial task and a powerful tool to capture aspects of the social space or personal community in which people are embedded. In this work, we propose a distance-based clustering procedure to identify homogeneous groups of ego networks that are only described by a small number of compositional variables. The proposed approach is motivated by the empirical study of ego networks of contacts extracted from the "Family and Social Subjects" (FSS) Survey conducted by the Italian National Statistical Institute in 2016, which is not specifically oriented to network analysis. We focus on elderly respondents living alone, which can be regarded as a vulnerable category, with the aim to describe their network of contacts. First, mining relational information in FSS data, we derive the ego networks of respondents. Then, we develop a methodology for coping with the presence of heterogeneous data and small amount of information from a network perspective. To this aim, we introduce a dissimilarity measure for mixed-type data, and exploit hierarchical clustering for grouping ego networks according to their composition. In doing so, we intend to make our approach applicable to various surveys.

Keywords Ego network · Hierarchical clustering · Prototype · Mixed-type data

✉ Elvira Pelle
elvira.pelle@unimore.it

Roberta Pappadà
rpappada@units.it

[1] Department of Communication and Economics, University of Modena and Reggio Emilia, Reggio Emilia, Italy

[2] Department of Economics, Business, Mathematics and Statistics "B. de Finetti", University of Trieste, Trieste, Italy

1 Introduction

Recent years have witnessed an increased interest in the analysis of personal networks and their application in many areas of the social sciences. Research in this field has shown the importance of network characteristics, such as size and composition, to assess the impact of social relationships on many aspects of everyday life, e.g., social support, well-being, health, and mobility (see, among others, Vacca 2020; Dykstra et al 2016; Amati et al 2015; Gallagher and Vella-Brodrick 2008; Taylor 2007). In the social network literature (Breiger 2004), an ego-centered (or egocentric) network (hereafter, ego network) forms around a particular social actor (the ego), and involves all other actors (the alters) with whom an ego enjoys a specific type of relation (e.g. emotional, support, etc.). Typical ego network data provide information on the ties between the ego and alters, as well as specific information on the alters, including how they are connected (see McCarty et al 2019; Perry et al 2018).

Ego networks have been studied extensively following different lines of research. One of these is focused on studying ego network typologies with the aim to capture aspects of the social "space" or personal community in which people are embedded (see, e.g., Crossley et al 2015). Traditionally, methods to extract ego network typologies are based on the composition of the network (i.e., the characteristics of the alters and the relation of ego to its alters). From another perspective, the relations of specific individuals are analyzed, in order to examine the same structural properties usually described in complete networks (Molina et al 2014; Domínguez and Maya-Jariego 2008; McCarty 2002). In this context, clustering methods have been exploited to identify and characterize existing types of personal networks in a sample. Methods to analyze and compare a set of networks have been discussed, for instance, in Brandes et al (2011) who developed a clustering procedure for a network ensemble (i.e. a collection of attributed graphs with some substantive commonality), and discussed its application to personal networks of migrants in the context of acculturation strategies. Among recent works that provided a strong contribution in this field, the authors in Bidart et al (2018) presented a typology of personal networks only based on indicators related to the structure of relations between alters, and analyzed data from a longitudinal study on young French people. Vacca (2020) reviewed existing methods to identify types of ego network structure, and proposed a novel approach to detect typologies based on three measures summarizing the overall structural configuration of a personal network; finally, he compared the results with those from the method introduced by Bidart et al (2018). When typologies of support are of interest (e.g., with the elderly and with the immigrant population), the most common approach is to use indicators of size and composition of the social support providers system. Maya-Jariego (2021) reviewed the most relevant classifications with regard to social support and personal networks, then used structural indicators and cluster analysis to capture the diversity of personal networks in a representative sample of individuals from a medium-sized city. A different approach can be found in Giannella and Fischer (2016), where random forests are applied to survey data of social relations, to reduce the large number of variables to a suitable combination into fewer dimensions used for typology detection.

Although the ego network design can be easily embedded as part of a representative survey of a large population (Marsden 2011), in some situations the paucity of information on alter features and alter-alter ties leads to new methodological challenges, especially when the survey goals are not specifically oriented to network analysis. In particular, focusing on the Italian context, the "Family and Social Subjects" (FSS) Survey[1] conducted by the Italian National Institute of Statistics (ISTAT) constitutes the principal statistical source of information on socio-demographic characteristics of Italian households, along with their over time dynamics, since it is based on a wide probability sample, including specific groups (by age, by living arrangements, etc.) of population. The FSS survey is not conceived to collect data according to the conventional approach used for gathering ego networks, however mining the FSS relational questions allows investigation of the personal networks of respondents at different stages of their life course or family formation (e.g., young, adults, singles or partner in a couple). This is done in Amati et al (2015), where the 2003 FSS data are used in a network perspective to construct potential and effective ego networks of Italian young adults. A similar strategy to derive ego networks was adopted in Amati et al (2017), who studied support networks of individuals living in Italy in the first stages of their family life, by using the FSS Survey carried out in Italy in 2009. The authors compared two different clustering methods for defining network typologies described by five categorical variables on the relationships role of alters with the ego.

In this article, we present a clustering procedure to identify groups of similar ego networks based on egocentric data. We use data from the FSS 2016 edition[2], the most recent among the FSS Italian surveys. As noted by Amati et al (2017), several studies have shown that the analysis of support networks cannot disregard the age of individuals and the transitions of the life course; among these, Sherman et al (2013) and Kalmijn and Vermunt (2007) attested the changes in size and composition of social networks at each transition of the life course. Although our research is not specifically oriented to support network, we assume that a similar rationale can apply to networks of contacts. In particular, we focus on elderly respondents living alone, i.e., unmarried, separated/divorced or widowed individuals (hereafter, we will sometimes refer to elderly respondents living alone as elderly singles). Here, the choice of elderly people as target group is motivated by the increasing interest in the study of social interaction of older adults (see, among others, Dykstra et al 2016; Ayalon and Levkovich 2019; Pelle et al 2021). In addition, elderly individuals living as singles are generally more vulnerable than other population groups (Djundeva et al 2019). While related works on previous editions of the FSS data have explored the support networks for Italian young adults or young couples (see

[1] Since 1998, The FSS is a thematic survey of the Multipurpose Survey Program delivered every 5 years. The questionnaire covers several topics on living arrangement and socio-demographic behaviors (life cycle, relations inside family, transition to adulthood, social mobility, fertility intentions, work histories, economic and social support, etc.) of the Italian population.

[2] https://www.istat.it/it/archivio/185678.

Amati et al 2015, 2017), to the best of our knowledge, there are no contributions that exploit the latest edition of the FSS survey to analyze the elderly population from a network of contacts perspective. Our research is motivated by the features we observe from the latest FSS data, whose major limitation is the lack of information on alter-alter ties. Moreover, the FSS survey adopts conventional formats to collect relational data only for a few items.

Starting from the construction of the ego network of contacts of elderly individuals living alone on the basis of the available compositional information, the scope of the paper is twofold: first, we develop a clustering procedure in the hierarchical framework in order to identify a partition of ego networks; second, we describe and characterize the resulting network types, and highlight gender differences within the target group. The proposed approach has the main advantage to be particularly suitable when the involved variables are heterogeneous, both in range and type, which can easily happen if ego networks are derived from secondary data, rather than using ad-hoc designs. Moreover, the choice of an appropriate clustering method and the selection of the number of clusters are addressed by considering aspects that are highly relevant in the specific context, namely, (1) a large within-cluster homogeneity, and (2) a representation of clusters through prototypical units.

The remainder of the article is organised as follows. Section 2 presents FSS data and the definition of ego networks of contacts for elderly respondents living alone. In Sect. 3 we introduce a dissimilarity measure for mixed-type data, and in Sect. 4 we describe the clustering procedure used to identify groups of similar ego networks with respect to their compositional features. Clustering validation and interpretation of main results are also discussed. Section 5 ends the paper with some concluding remarks.

2 Data and network construction

The FSS Survey conducted in 2016 provides data from 24753 individuals living in 852 Italian municipalities of different demographic size. The target group of our analysis is composed of respondents living alone and aged 65 years and over (i.e., elders for which the household structure is composed only of the individual). This data set originally consisted of 1851 individuals, along with their relational information and socio-demographic characteristics (missing values occur in about 1.7% of the subjects). After removing the records containing missing values, the resulting data set is formed of 1820 subjects (522 males and 1298 females): 70% is widowed, 13% is separated or divorced, and the remaining 17% is unmarried. Other socio-demographic characteristics of the elderly living as single are reported in Table 1, where we consider the distribution for males and females to highlight gender differences.

Among the elderly single respondents, women appear to be older than men, with 57.7% of men older than 75 years compared to the 70% of women aged 75 years and over. Looking at the distribution of the perceived health at the time the survey was conducted, some differences can be traced among women and men: about 78%

Table 1 Socio-demographic characteristics of the 1820 elderly respondents living alone, from FSS 2016, ISTAT

	Elders (%)	
	Men ($n = 522$)	Women ($n = 1298$)
Age		
65-74	42.3	29.8
75+	57.7	70.2
Health		
Bad	22.4	28.8
Fair	38.9	44.0
Good	38.7	27.2
Education		
High	9.4	4.7
Medium	21.3	15.8
Low	69.3	79.5
Place of residence		
Metropolitan area	17.2	16.8
Municipality (<10000)	41.6	37.1
Municipality (> 10000)	41.2	46.1

of men declare that they perceive their own health as good or fair, while more than 28% of women perceive a bad health status for themselves. With respect to educational level, more than 30% of men present a medium (high school diploma) or high (university degree/PhD) level of education, while the same level can be found in 20.5% of female respondents, with the remaining 79.5% of women declaring to have a low level of education (compulsory or none). No striking differences can be found looking at the distribution of the place of residence[3].

2.1 Ego network of contacts

As mentioned in the introduction, FSS Survey data are not collected using a network perspective, which thus provides limited information concerning relational data. Despite such a limitation, recent works (Amati et al 2015, 2017) have provided a general approach to construct ego networks using the presence, residential proximity, and frequency of face-to-face contacts with non-cohabitant persons collected for each respondent.

In particular, to derive the ego network of contacts of our target group (older people living as singles) from the FSS 2016 edition, we combine the information on non-cohabitant kin (siblings, children, grandchildren[4]), other relatives (if any), and

[3] In FSS 2016, place of residence is coded in three categories: metropolitan areas, municipalities up to 10000 inhabitants, and municipalities larger than 10000 inhabitants. The category “metropolitan area” includes the main big cities (e.g., Turin, Milan, Venice, Genoa, Bologna, Florence, Rome, Naples, Bari, Palermo, Catania, Cagliari).

[4] For siblings, children, and grandchildren, the FSS questionnaire allows for listing a maximum of three members for each category.

non-kin (asked only as the number of friends and presence of neighbors), to whom the respondent "is close" or on whom the ego "can count". Specifically, following Amati et al (2015), we assume that frequent contacts (at least once in a week), and close residential proximity (even in a different municipality but not farther than 16 km), allow definition of a group of alters composed of siblings, children, and grandchildren with whom the respondent can entertain social contacts. In addition, we consider the number of (non-cohabitant) relatives and friends, as well as information on the presence of neighborhood relationships (non-kin), on which the respondent "can count" if necessary. Doing so, we aggregate alters by their role relation with each respondent (the ego), and derive an ego network of contacts composed of a maximum of 6 different alter roles: *Siblings*, *Children*, *Grandchildren*, *Relatives*, *Friends*, and *Neighbors* (see Fig. 1). Note that, in the case of neighbors, only the presence (yes/no) is available. Consequently, we treat the information on the availability of neighbors as a dichotomous variable, taking the value 1 if a neighbor is present, and 0 otherwise. It is worth noting that the networks are built in such a way that they are not typical ego networks, such as the ones derived using ad-hoc designs. Indeed, the absence of alter characteristics and alter-alter ties represents a major limitation of networks extracted from FSS Survey data, which are not designed to incorporate structural information.

Table 2 reports the distribution of the number of different alter roles in the ego network of contacts of elderly singles by gender. It can be noted that the majority of the elderly (either men or women) are embedded in an ego network of contacts composed of one to three alter roles, and only a small percentage of individuals presents four or more types of alters. Nevertheless, Table 2 reveals some differences regarding the distribution by gender. In particular, a higher proportion of men presents only one role compared to women (24.5% vs 17%); women rely on 4 roles or more in about 24% of cases, compared to 14% for male respondents, indicating that, in general, networks of female respondents show a higher number of alter roles than men in the target group (gender differences in the distribution of personal networks were discussed, for instance, in Agneessens et al (2006), Moore (1990)).

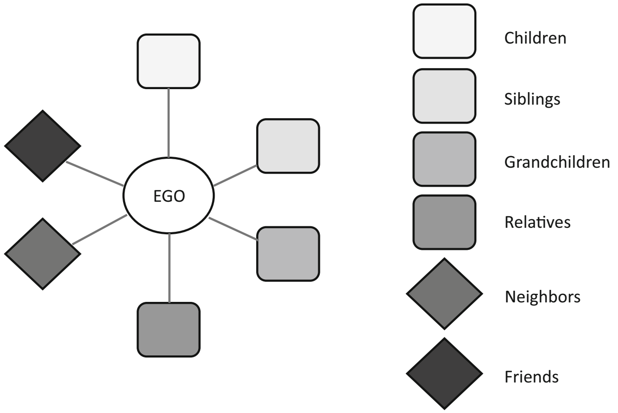

Fig. 1 Ego and types of alters in ego networks of Italian elderly singles

Table 2 Distribution of the number of different alter roles in the ego network of contacts of elderly respondents living alone, from FSS 2016, ISTAT

n. alter roles	Elders (%)	
	Men $n = 522$	Women $n = 1298$
0	15.3	9.0
1	24.5	17.1
2	25.5	28.4
3	20.3	21.6
4	8.8	14.6
5	4.1	7.1
6	1.5	2.2

Furthermore, the proportion of *empty* networks (i.e., networks with no alters) is 11% of the total networks of older people living alone, with a higher percentage for men (15.3%) than for women (9.0%). Since our research interest is the identification of network types based on the composition of the ego network of contacts, we do not consider egos that are not related to any type of alter and, therefore, we finally select $n = 1623$ individuals, for whom at least one alter is present. Moreover, on the light of previous considerations, it is reasonable to carry out separate analyses for male and female respondents living alone and having a non–empty network, which represent 27.2% ($n_m = 442$) and 72.8% ($n_f = 1181$), respectively. As a result, the two data sets we analyzed consisted of ego networks with six alter categories for which the number of people in each role—specifically, siblings, children, grandchildren, relatives, and friends—and the absence/presence information for neighbors is available. Table 3 reports some descriptive statistics on the ego networks of males and females living as single from the FSS Survey (2016), defined via the illustrated approach. We note that most of the variables present a distribution that is right skewed, with a large number of zeros, which makes the data set sparse (mainly due to the presence of few ties in the network of aged respondents).

In what follows, each ego network is described in terms of five numeric variables and one dichotomous variable (coded as 0-1) describing the different alter roles. Note that the different range of alter categories is due to specific limitations imposed by the FSS questionnaire, an issue that will be addressed in the next section when a suitable definition of similarity between two ego networks is introduced, by taking into account the heterogeneity in the data.

3 Definition of dissimilarity

To apply a dissimilarity-based clustering method, a formal definition of ‘similarity’ between ego networks is needed. Typically, a dissimilarity measure for mixed-type data can be defined by combining distinct components, each related to a different type of attribute (e.g., numeric or categorical) to be taken into account. A crucial task is then how to aggregate the different contributions, eventually performing normalization (standardization) of data for comparison’s sake across variables, or

Table 3 Characteristics of alter roles in the ego networks (excluding the empty ones) extracted for the elderly (FSS 2016) living as single, according to gender (1623 subjects)

Men	Alter role	Mean	Median	SD	Range
n_m = 442	Siblings	0.58	0.00	0.89	0–3
	Children	0.88	1.00	0.96	0–3
	Grandchildren	0.77	0.00	1.14	0-3
	Relatives	0.97	0.00	2.62	0–23
	Friends	1.10	0.00	2.14	0–10
		Levels	Frequency		
	Neighbors	Absent (0)	205		
		Present (1)	237		
Women	Variable	Mean	Median	SD	Range
n_f = 1181	Siblings	0.51	0.00	0.88	0–3
	Children	1.19	1.00	1.00	0–3
	Grandchildren	1.09	1.00	1.20	0–3
	Relatives	1.04	0.00	2.72	0–27
	Friends	1.06	0.00	1.99	0–10
		Levels	Frequency		
	Neighbors	Absent (0)	521		
		Present (1)	660		

assigning suitable weights to the features, according to a prior knowledge of the relative importance of the different variables. Common approaches for coping with variables of mixed type include Gower's distance (Gower 1971) and its extensions (see, e.g., Podani 1999). More recent works that mainly focused on partitional clustering based on mixed data include De Amorim and Mirkin (2012), Hennig and Liao (2013), D'urso and Massari (2019). For a complete review of distance-based methods with mixed data, see van de Velden et al (2019) and the references therein.

Here, we introduce a dissimilarity measure between two ego networks in terms of two metrics, one for numeric variables and the other for the categorical part, to be used in a hierarchical clustering (HC) framework. Without ambiguity, we assume that the ensemble of ego networks constructed in Sect. 2 can be represented as a matrix of size $N \times p$, where N is the number of respondents, and p is the total number of features. The i-th row of such a matrix represents the i-th unit of the data set $(1 \leq i \leq N)$, whose values $x_{i,1}, \ldots, x_{i,m}$ are numerical, whereas the values $x_{i,m+1}, \ldots, x_{i,p}$ are categorical. We define the dissimilarity between units i and i' as follows

$$d_{ii'} := d(\mathbf{x}_i, \mathbf{x}_{i'}) = \sum_{s=1}^{m} \alpha_s |x_{i,s} - x_{i',s}| + \sum_{t=m+1}^{p} \alpha_t \delta(x_{i,t}, x_{i',t}), \tag{1}$$

where $\alpha_j > 0, j = 1, \ldots, p$, is the weight associated with the j-th attribute type, and δ

is a suitable dissimilarity measure for categorical objects. In Eq. (1), the contribution of numeric variables is measured via the L_1 norm (also known as the 'city block' distance). This is a well-established option, although other dissimilarity measures could also be used, such as the Euclidean distance. Compared to the latter, the L_1 norm avoids a large dissimilarity on one variable from having a strong influence on the total dissimilarity. Concerning the contribution of the categorical variables to the mixed metric, we chose the weighted simple matching distance (see, e.g., Huang 1998), where the dissimilarity between two units (egos) based on the t-th categorical attribute, $x_{1,t}$ and $x_{2,t}$, is

$$\delta(x_{1,t}, x_{2,t}) = \begin{cases} 0, & \text{if } x_{1,t} = x_{2,t} \\ 1, & \text{if } x_{1,t} \neq x_{2,t}. \end{cases}$$

Evidently, for two vectors of categorical features, the smaller the number of mismatches, the more similar the two objects. Note that, for binary variables represented by numerical codes 0-1, the dissimilarity adopted assumes that both 0-0 matches and 1-1 matches carry equivalent information, that is, $\delta(x_{1,t}, x_{2,t}) = 0$ if $x_{1,t} = x_{2,t} = 0$ or $x_{1,t} = x_{2,t} = 1$.

Two main issues when applying the mixed metric in Eq. (1) are how to make the variables with different ranges comparable for aggregation, and how to set the weights in order to reflect the relative importance of the variables for the specific clustering problem. Typically, classic normalization (standardization) procedures can be adopted for numeric features (e.g., min–max normalization that transforms the data to the range [0, 1]), in order to prevent the largest-scaled features from dominating the others. However, as Suarez-Alvarez et al (2012) noted, classic normalization procedures do not, in general, guarantee equal contributions of all features to the results. Alternative approaches have been considered in the framework of clustering of mixed data, for instance by introducing a user-defined weight of the entire group of categorical variables, to avoid favoring either type of attribute, as proposed by Huang (1998).

Here, without transforming the original data or relying on heuristic methods for the specification of the weight for the categorical part, we tackle the issue of balancing the contributions of the different features by adopting the statistical approach to normalization of feature vectors proposed in Suarez-Alvarez et al (2012). Specifically, the authors suggest how to set appropriate weights in such a way that the average contributions of the variables to the measure, regardless of their type, will be statistically the same. The normalization of the metric in Eq. (1) is fulfilled by defining the weights

$$\alpha_s = 1/E|X_{i,s} - X_{i',s}|, \text{ for } s = 1, \ldots, m \tag{2}$$

and

$$\alpha_t = 1/E\delta(X_{i,t}, X_{i',t}), \text{ for } t = m+1, \ldots, p \tag{3}$$

where E denotes expectation, and $X_{i,j}$, $X_{i',j}$ are independent random variables whose values are distributed in accordance with the distribution of the j-th attribute,

$j = 1, \ldots, p$. To compute the weights α_j one can use sample estimators (see Prostov et al 2015). We prefer this normalization method to standard range normalization (which is also implicit in Gower's method), with more evidence in its favor presented below, where the effect of variables is evaluated.

4 Clustering of FSS-based ego networks

This section focuses on clustering ego networks extracted from the latest edition of the FSS Survey, based on the dissimilarity measure defined in the previous section, with the final aim of characterizing non-overlapping groups of homogeneous networks. In particular, when hard clustering is adopted (Kaufman and Rousseeuw 2009), data are divided into distinct clusters, where each data element belongs to exactly one cluster. As noted before, it makes sense to apply the same clustering procedure on the data sets derived for males and females, separately, which consist of $n_m = 442$ and $n_f = 1181$ individuals in the target group, respectively. Here, according to the ego network construction described in Sect. 2, the total number of variables is $p = 6$, $m = 5$ numerical features (*Siblings*, *Children*, *Grandchildren*, *Relatives*, *Friends*), and one binary variable (*Neighbors*).

Starting from the two data sets of $n_m \times p$ and $n_f \times p$ observations, we compute pairwise dissimilarities by using Eq. (1) and the weights defined in Eqs. (2)–(3). Note that the normalization procedure described in Sect. 3 is regarded as a kind of weighting, used to balance the contribution of numerical and categorical variables, as opposed to weighting used to express domain-specific knowledge on variables' importance. Regarding the latter, we assume that all variables are relevant to our problem, i.e., we do not consider upweighting or downweighting certain variables in our application.

Once the $n_m \times n_m$ and $n_f \times n_f$ dissimilarity matrices are obtained, standard distance-based methods can be adopted. In the context of mixed-type data, partitional dissimilarity-based methods, such as K-medoids or its variants (Everitt et al 2011), have been largely exploited, due to their efficiency for clustering large data sets. Since the dimensionality of the problem is moderate, we consider the HC framework (Kaufman and Rousseeuw 2009), which allows visualization of the produced nested partitions, and the relationships among the clusters as well. In such a framework, the core idea is to construct the hierarchical aggregation among objects to be grouped, starting from k clusters, each containing a singleton, and ending when all objects form a cluster, where a linkage rule is used to compare clusters. Standard agglomerative hierarchical linkages that only require a dissimilarity matrix are the single, the complete, and the average one (see, e.g., Everitt et al 2011), each giving rise to a different inter-group distance measure. Recently, the so-called *minimax linkage* has been investigated by Bien and Tibshirani (2011), which shares many of the desirable theoretical properties of the standard linkages, while adding interpretative value by means of the prototypes—i.e. units chosen from the original data set—associated with each cluster in the final solution.

Formally, the distance between two clusters based on the minimax linkage is defined to be the radius of the smallest enclosing ball, centered at a point chosen from the two clusters, C_1 and C_2:

$$d(C_1, C_2) = \min_{\mathbf{x}_i \in C_1 \cup C_2} \left\{ \max_{\mathbf{x}_{i'} \in C_1 \cup C_2} d_{ii'} \right\}, \tag{4}$$

where $d_{ii'}$ is the dissimilarity between units i and i'. The prototype of the newly formed cluster, $C_1 \cup C_2$, is the unit $\mathbf{x}^\star = \arg\min_{\mathbf{x}_i \in C_1 \cup C_2} (\max_{\mathbf{x}_{i'} \in C_1 \cup C_2} d_{ii'})$. As discussed in Bien and Tibshirani (2011), the minimax linkage (1) can be applied directly to a dissimilarity matrix, (2) does not produce inversions and satisfies the reducibility property, and (iii) yields a clustering for which no point will be farther than h from its prototype, h being the height at which the minimax linkage tree is cut.

The dendrograms produced by the above-described procedure applied to the dissimilarity matrices of male and female respondents in the target group are shown

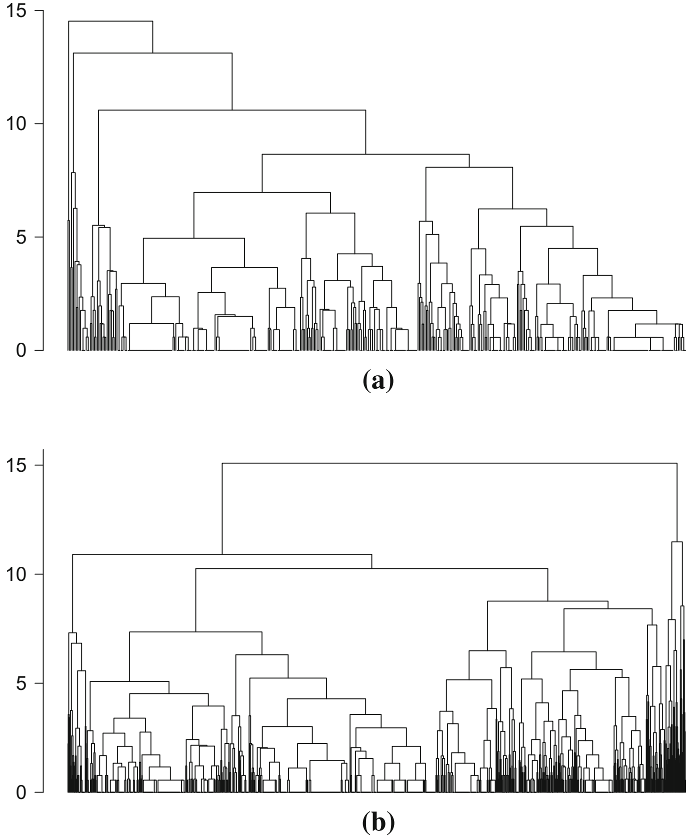

Fig. 2 Dendrogram resulting from agglomerative HC (minimax linkage) of the ego networks of (**a**) male respondents (**b**) and female respondents

in Fig. 2a, b, respectively. The minimax method has been implemented by using the R package protoclust (Bien and Tibshirani 2019). Once a satisfactory solution in k clusters has been found, the prototypes $\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_k$, associated with the groups $C_1, \ldots, C_k$, can be used to summarize the clusters via the corresponding feature vectors. Each cluster prototype can be interpreted as the "central unit" of the cluster, i.e., a unit of the data set that is found to be the most highly representative element of that cluster (prototypes from minimax linkage are indeed similar in concept to medoids from the K-medoids algorithm). Because prototypes are units from within the original data set, they clearly offer a meaningful way to summarize the clusters, which can be more appropriate than using centroids.

To see that the minimax prototype is the central point associated with the cluster, we illustrate an example that considers a branch of the tree in Fig. 2a, made up of 14 ego networks, with associated prototype given by the unit with row index "177" in the considered data set. The left panel of Fig. 3 shows the subtree obtained by cutting the whole tree at height $h = 7.8$, where each leaf corresponds to one of the original data points (labels are row indices), and every node has an associated prototype. The prototype with label "177" is such that the dissimilarity with all the points in the cluster is no larger than the height of the cut, a property that is visually demonstrated in the right panel of Fig. 3. Specifically, we applied multidimensional scaling to map the dissimilarities computed for the 14 data points in the subtree based on $p = 6$ variables into $\mathbb{R}^2$, thus allowing the visualization of the prototype as the center of the ball covering all of the data points in the cluster.

4.1 Cluster validation

To decide how many clusters to select by cutting the dendrogram at a certain height, several clustering validity measurements (Halkidi et al 2001) in conjunction with visual inspection of the tree structure can be adopted. As discussed, for instance, in

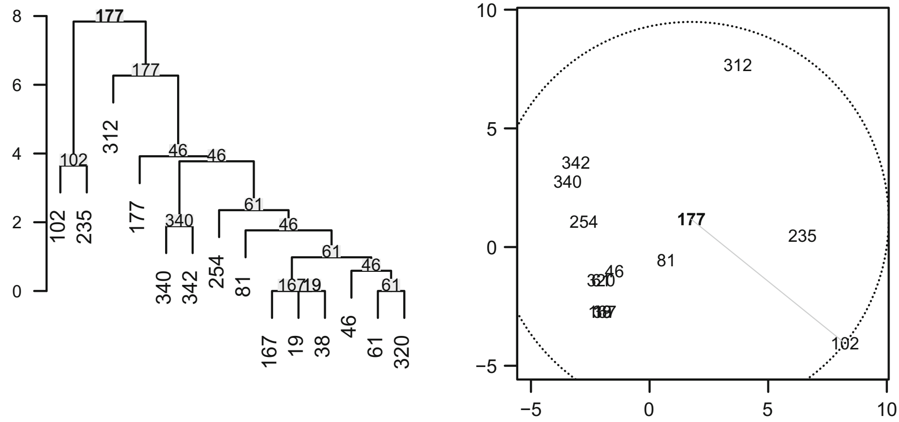

Fig. 3 View of a subtree (left) from Fig. 2a with the branch's prototype in bold and prototypes associated with each leaf (labels are row indices of data points in the original data set) and representation of dissimilarities (right) via multidimensional scaling (minimax linkage is the radius drawn on the plot)

Akhanli and Hennig (2020), the exploration of the quality of the achieved clusterings can be pursued by considering indices that measure a single isolated aspect of clustering that can be of particular interest to the researcher. Here, networks belonging to the same group need to have a good internal homogeneity, which can be measured by the average within-cluster dissimilarity, $I_{ave.wit}$, computed as in Akhanli and Hennig (2020) in such a way that every observation has the same overall weight. We use a normalized version of such an index, and subtract it from 1, so that large values (corresponding to smaller within-cluster dissimilarities) are better:

$$I_w^* = 1 - \frac{I_{ave.wit}}{\max_{\mathbf{x},\mathbf{y}\in\mathbf{X}} d(\mathbf{x},\mathbf{y})},$$

where $\mathbf{X}$ is the set of data objects and d is the mixed dissimilarity, as defined above. The index I_w^* cannot be optimised over k, because increasing k will normally increase the index. However, examining the index values for different numbers of clusters, ranging from 2 to 10, we observe that, in the case of male respondents, I_w^* reaches high values for $k \geq 5$ ($I_w^* = 0.838$ for $k = 5$, $I_w^* = 0.854$ for $k = 6$) and does not improve much for $k > 6$ ($I_w^* = 0.855$ for $k = 7$), implying that the solution with $k = 6$ can be selected, since it corresponds to the value after which the curve flattens out, while still increasing. Similarly, for female respondents, going from $k = 6$ to $k = 7$ does not substantially improve clustering quality in terms of within-cluster homogeneity, and the index is much lower for $k<6$ ($I_w^* = 0.829$ for $k = 5$, $I_w^* = 0.847$ for $k = 6$, $I_w^* = 0.848$ for $k = 7$). In addition, we compute the dissimilarity-based Calinski-Harabasz (CH) index (Hennig and Liao 2013) (a generalization of the index originally proposed in Caliński and Harabasz (1974)), which compares squared within-cluster dissimilarities with all between-cluster dissimilarities (measuring separation). A better clustering is indicated by a large value of the index, so that the best number of clusters can be chosen by maximizing CH over k. We found that, for both data sets, $k = 6$ is a global optimum of the CH index (for males, CH is less than 127.7 for $k \leq 5$, is 143 for $k = 6$, and decreases for $k \geq 7$; for females, CH goes from 173.2 when $k = 5$ to 227.8 when $k = 6$, and falls below 194 for $k \geq 7$; see Fig. 6 in Appendix A). Thus, we finally select a partition of ego networks into 6 groups, for both males and females, by cutting the dendrograms at height 7.8 and 8.5 of the trees, as displayed in Fig. 2a, b, respectively.

4.2 Characterization of clusters

For interpreting the results of the solution with $k = 6$ from HC, we focus on the characterization of clusters by studying the ego networks composition and size (i.e., the total number of alters regardless of their role), as presented in Fig. 4, and the frequency of specific alter roles, as reported in Table 4. In particular, a high group frequency of the presence of a tie with a specific alter category implies that the group is mostly formed of individuals who exchange relationships with that alter type, thus allowing the description of the clusters according to the most relevant role relations. Next, examining the distribution of the number of alter categories (ranging

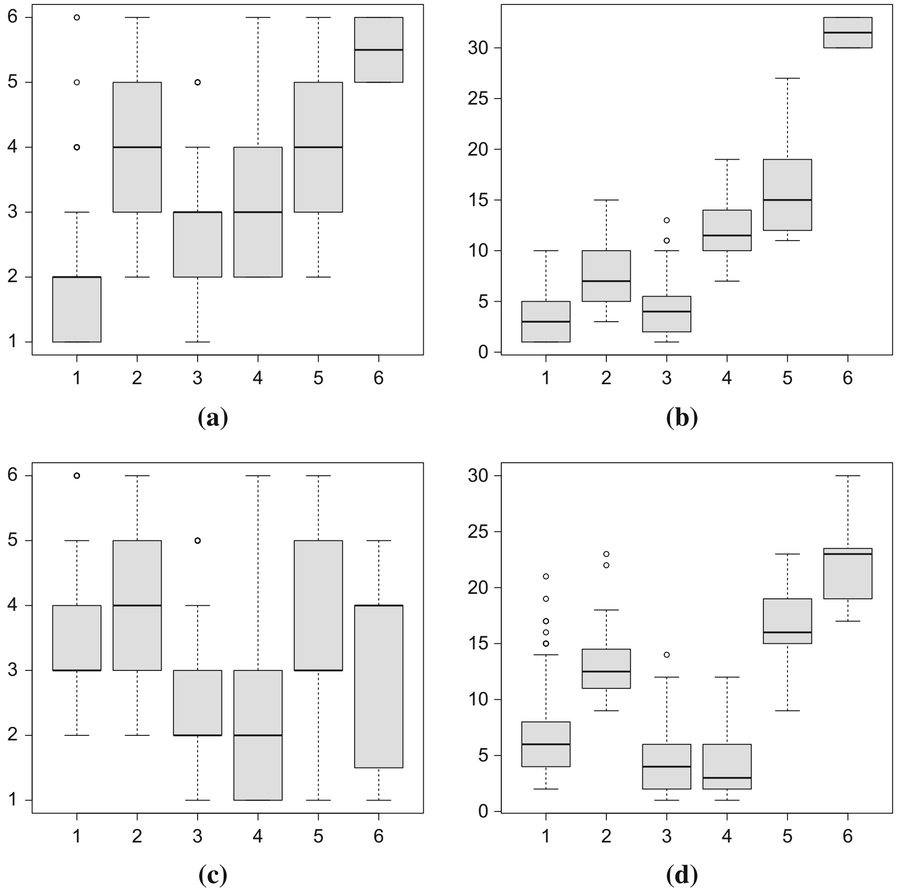

Fig. 4 Number of alter roles and network size (*y*-axis) across the six clusters (*x*-axis) of ego networks of elderly males (panels (**a**) and (**b**), respectively) and female respondents (panels (**c**) and (**d**), respectively) obtained by applying the HC with minimax linkage and mixed dissimilarity measure

from the situation where only one type of alter is present, to the case in which the ego is connected to all the six alter types), as shown in Fig. 4a–c, and ego network size (see Fig. 4b–d) enables a fairly easy interpretation with enough differentiation between the clusters. When interpreting the network size, it is important to consider that, for alter categories of siblings, children and grandchildren, the number of people an ego can declare is limited to a maximum of three, whereas for friends and other relatives, such a limitation does not exist. Moreover, when neighbors are present, the value 1 is added to the ego network size. Finally, Table 5 describes the clusters in terms of (1) average values on each attribute (i.e., cluster centroids), and (2) prototypes resulting from HC and minimax linkage. Additional Tables and Figures in Appendix A provide further details on the composition of the ego networks of contacts in the resulting clusters (Tables 7 and 8), and the distribution of age and health condition by cluster (Figs. 7 and 8).

Table 4 Proportion (%) of presence of each alter for each cluster and in the entire sample, for men (*M*) and women (*W*)

Cluster ($\lvert C_i \rvert$)	Siblings	Children	Grandchil.	Relatives	Friends	Neighbors
M						
1 (214)	45.79	64.49	46.26	14.95	1.87	6.54
2 (37)	21.62	97.30	97.30	27.03	48.65	100.00
3 (155)	25.16	32.26	9.68	20.00	63.87	100.00
4 (20)	70.00	20.00	10.00	100.00	60.00	75.00
5 (14)	7.14	71.43	35.71	57.14	100.00	100.00
6 (2)	100.00	100.00	50.00	100.00	100.00	100.00
Sample	36.65	54.30	35.75	23.30	33.71	53.62
W						
1 (280)	38.57	52.14	32.50	45.00	82.50	100.00
2 (36)	47.22	63.89	55.56	19.44	100.00	100.00
3 (667)	11.99	82.01	66.57	17.84	16.04	40.93
4 (159)	100.00	50.94	25.79	14.47	10.06	27.04
5 (28)	14.29	57.14	46.43	100.00	67.86	82.14
6 (11)	36.36	63.64	45.45	100.00	9.09	45.45
Sample	31.50	69.43	51.99	26.59	34.72	55.88

The number of individuals in each cluster is reported in parentheses

By combining the aforementioned information, we can summarize the results of the clustering procedure conducted on ego networks of FSS Survey respondents. We begin by lining up the characteristics of the ego networks in each cluster for the elderly males considered in the analysis. *Cluster 1* ($n = 214$, about 60% aged 75 years and over) is formed of individuals whose network of contacts is mainly kinship, with a high presence of children and almost always characterized by the absence of friends and neighbors, with a network mean size around 3 (see Table 5, where the prototype $\mathbf{b}_1$ has value 1 on close kin, e.g., siblings, children, and grandchildren, and 0 on the other roles); such a cluster is quite homogeneous in terms of number of alter roles and network size, with low variability of these two characteristics (see the upper row in Fig. 4). *Cluster 2* ($n = 37$) is much smaller than *Cluster 1*, and it is more balanced on alter categories (the network mean size is around 8, with the median number of alter roles equal to 4); the prototype is the vector $\mathbf{b}_2 = (0, 2, 3, 1, 1, 1)$ with all alter types in the network, except siblings, and with a higher number of children and grandchildren compared to other roles. The network size of this cluster is moderate and weakly affected by the relationships with other relatives or friends, which are only marginally available (see Table 4). Almost all of the 80% of respondents in this cluster are aged 75 years and over, and perceived their own health status as "bad" or "fair." *Cluster 3* ($n = 155$) is the second largest cluster; it is healthier (more that 80% of the respondents declared a "good" or "fair" health status) and characterized by a high presence of friends and

Table 5 Centroids and prototypes ($\mathbf{b}_k$) for the 6 clusters based on the hierarchical method with minimax linkage applied to ego networks of elderly men (M) and women (W)

		Siblings	Children	Grandchil.	Relatives	Friends	Neighb.
M							
C_1	Centroid	0.69	1.04	0.98	0.37	0.02	–
(3.2)	$\mathbf{b}_1$	1	1	1	0	0	Absent
C_2	Centroid	0.38	2.11	2.57	0.73	1.27	–
(8.1)	$\mathbf{b}_2$	0	2	3	1	1	Present
C_3	Centroid	0.39	0.41	0.13	0.45	1.66	–
(4.1)	$\mathbf{b}_3$	1	0	0	0	3	Present
C_4	Centroid	1.45	0.25	0.10	7.95	1.45	–
(11.9)	$\mathbf{b}_4$	2	0	0	8	2	Present
C_5	Centroid	0.14	1.14	0.71	3.57	10.00	–
(16.6)	$\mathbf{b}_5$	0	2	0	6	10	Present
C_6	Centroid	2.50	1.50	1.00	22.00	3.50	–
(31.5)	$\mathbf{b}_6$	3	1	2	23	3	Present
W							
C_1	Centroid	0.49	0.71	0.55	1.49	2.29	–
(6.5)	$\mathbf{b}_1$	2	1	1	3	2	Present
C_2	Centroid	0.72	1.28	1.08	0.69	8.50	–
(13.3)	$\mathbf{b}_2$	1	2	3	0	10	Present
C_3	Centroid	0.17	1.48	1.46	0.35	0.28	–
(4.1)	$\mathbf{b}_3$	0	2	2	1	0	Absent
C_4	Centroid	2.03	0.81	0.48	0.34	0.24	–
(4.2)	$\mathbf{b}_4$	3	1	0	0	0	Present
C_5	Centroid	0.21	1.00	1.11	10.96	2.79	–
(16.9)	$\mathbf{b}_5$	0	1	0	11	2	Present
C_6	Centroid	0.55	1.64	1.18	17.91	0.18	–
(21.9)	$\mathbf{b}_6$	1	3	3	19	0	Absent

The network mean size for clusters $C_1, \ldots, C_6$ is reported in parentheses (if the alter category "neighbors" is present, then a value of one is added in the resulting network size of the respondent)

neighbors, although the limited number of friends (the mean is 1.66, the prototype value is 3) determines networks of overall small size. *Cluster 4* ($n = 20$, 70% aged 75 and older) and *Cluster 5* ($n = 14$, about 60% aged between 65 and 74) are similar in the number of elements, and are formed of individuals whose contacts include non-cohabitant kin as well as non-kin (e.g., friends and neighbors); among kinship roles, *Cluster 4* is dominated by siblings and other relatives, whereas in *Cluster 5* sibling are almost absent and the dominant ties are with children and other relatives. One major distinction between these two clusters is the number of friends, which is much larger for *Cluster 5* than for *Cluster 4* (see the prototypes $\mathbf{b}_4$ and $\mathbf{b}_5$ in Table 5). Finally, an extremely small group stands out for having an especially high number of alter roles and size (*Cluster 6*, $n = 2$), which can be classified as a cluster of atypical points or outliers.

For women, the clusters obtained can be characterized as follows. *Cluster 1* ($n = 280$) is composed of networks with a moderate number of alter roles and a large variability on the network size (see the bottom row in Fig. 4). The totality of the respondents in this cluster declared the availability of neighbors, and 80% the availability of friends, although in about 90% of cases the number of friends is less than four; the distributions of age and health condition are similar to that of the whole sample of elderly females. *Cluster 2* ($n = 36$) is healthier (only 11% perceives its health as "bad") and younger than *Cluster 1*, characterized by networks with a larger size, due to the presence of children and grandchildren, and to the large number of friends (see the prototype $\mathbf{b}_2$ for women in Table 5). *Cluster 3* ($n = 667$) is the largest group, composed of almost 70% of individuals older than 75 years, declaring bad health conditions in 30% of cases. The cluster is dominated by alter categories "children" and "grandchildren", as indicated by the prototype's values on these attributes. The networks in *Cluster 4* ($n = 159$) have a size similar to that of *Cluster 3*, but show a larger variability on the distribution of the number of alter roles (bottom-left panel of Fig. 4). This indicates that, in terms of role relations, the networks belonging to this group are less narrow than the ones from *Cluster 3*. However, the prototype $\mathbf{b}_4 = (3, 1, 0, 0, 0, 1)$ indicates that *Cluster 4* is more oriented to close kin and neighbors. *Cluster 5* and *Cluster 6* are made up of a small number of networks (2.4% and 0.9%, respectively), with median network size given by 16 and 23, respectively, due to the large influence of non-cohabitant relatives. Regardless of the size, the individuals from *Cluster 5* and *Cluster 6* present a large variability of the number of alter types. In particular, *Cluster 5* shows a high prevalence of non-kin (connections with friends and neighbors are present in 68% and 82% of cases, respectively); by contrast, the egos in *Cluster 6* have contacts with non-kin members in a small proportion of cases (for instance, friends are present in only 9.09% of networks). This distinction is also clear from the resulting prototypes for these clusters (see $\mathbf{b}_5$ and $\mathbf{b}_6$ in Table 5).

4.3 Contribution of variables to clustering

To gain insight into the importance of each variable for the overall clustering, we compare the clustering recovered by all variables and the clusterings obtained by considering one variable at a time that is left out. To do this, we use the Adjusted Rand Index (Hubert and Arabie 1985) (ARI) to measure the agreement between the partition into $k = 6$ clusters, as identified in Sect. 4.1, and the clusterings obtained by eliminating each variable in turn. In addition, to evaluate the sensitivity of the variables' impact against the choice of the dissimilarity measure, we compared the results from HC applied to the dissimilarity described in Sect. 3, with the commonly used approach wherein numerical attributes are preliminarily range-normalized to lie in the interval [0, 1] (i.e., min–max normalization) and then Eq. 1 is used (in such a case, we set the weights to be $\alpha_j = 1$, for all $j = 1, \ldots, p$). For both methods (which are referred to as `Range-norm` and `Stat-norm`), we consider the results from HC with minimax linkage. According to the validity criteria illustrated in Sect. 4.1, when applying the range-based normalization approach to the data concerning male respondents, the solution with $k = 2$ optimizes the CH index, whereas a local optimum is reached with $k = 4$. In terms of cluster homogeneity, the solution with $k = 4$ works significantly better than the one for $k = 2$, supporting the solution with $k = 4$. For females, CH is optimal for $k = 3$, and therefore we consider the partition into three groups. The ARIs for each omitted variable are reported in Table 6 (the maximum ARI is 1 for perfect agreement between two clusterings; a value of 0 is the expected value for comparing two unrelated random clusterings; negative values can occur as well). Here, high values of ARI indicate that omitting a variable does not substantially change the clustering, implying that the variable has a low impact on the clustering.

Table 6 shows that the biggest difference between the two methods in this respect is the impact of *Relatives* for males, which turns out to be an influential variable for `Stat-norm` but with almost no impact with `Range-norm`. For females, the relevance of *Siblings* with method `Stat-norm` is high (the ARI is 0.077) but not particularly strongly involved for the `Range-norm` approach. In general, `Range-norm` seems to be strongly dominated by the categorical variable (*Neighbors*), since omitting such variable produces partitions having low values of ARI with the clusterings on the full data set (ARI is 0.095 and 0.206 for males and females, respectively). By contrast, with `Stat-norm` none of the values are particularly high, and thus the clustering recovered with all variables seems to derive from a combined effect rather than from being dominated by one single variable.

4.4 Clustering results by aggregated alter types

With the aim to better characterize clustering results on the basis of the role of the alters in the ego network, we aggregate the six alter categories into three main categories: (1) “Immediate family,” composed of children, (2) “Extended family,” composed of alters (at least one) among siblings, grandchildren, and relatives, (iii) “No family,” with only neighbors and/or friends (a similar aggregation is also considered in Amati et al (2015), Pelle et al (2021)). The pairwise combination of these three types produces seven distinct types of network of contacts: (T1) *Immediate family*, (T2) *Extended family*, (T3) *No family* (Non-kin), (T4) *Immediate and Extended family* (Kin), (T5) *Immediate and No family*, (T6) *Extended and No*

Table 6 ARI between clustering on all variables and clustering with a variable omitted, according to the range-based normalization (`Range-norm`) and the normalization method described in Sect. 3 (`Stat-norm`) for male and female egos

Method	Variable					
	Siblings	Children	Grandch.	Relatives	Friends	Neighb.
Men						
Range-norm, $k = 4$	0.817	0.875	0.884	0.970	0.688	0.095
Stat-norm, $k = 6$	0.525	0.564	0.551	0.489	0.540	0.127
Women						
Range-norm, $k = 3$	0.661	0.697	0.590	0.668	0.797	0.206
Stat-norm, $k = 6$	0.077	0.204	0.137	0.124	0.220	0.058

family, (T7) *Comprehensive* (Kin and Non-kin). Note that this classification takes into account only the presence of an alter role, regardless of the number of alters embedded in the ego network.

Figure 5 presents relative frequency distributions of estimated clusters described in Sect. 4.2, according to the seven types from T1 to T7. The fill patterns reflect the number of networks in the clusters, i.e., the largest group is plotted with the densest horizontal lines; for groups with less units lines are sparser and the angle is raised counter-clockwise; the smallest cluster (i.e., *Cluster 6* for both men and women) is plotted with the sparsest vertical lines. Looking at the left panel of Fig. 5, it can be clearly seen that the kin network type for male respondents (considering T1, T2, T4) is dominated by the largest cluster (*Cluster 1*), whereas non-kin ties and networks that include the *Immediate family* (and less frequently the *Extended family*) are well represented by the second largest group (*Cluster 3*). We note that the network type with highest frequency is that of individuals having kin and non-kin interactions simultaneously (i.e., the comprehensive type, T7), and is represented by all the six clusters in the HC solution, albeit they differ in the network size and other compositional characteristics, as discussed in Sect. 4.2. The top three clusters are *Cluster 4*, *5*, and *6* that only contain 8% of units (see Table 4). Among female respondents (right panel of Fig. 5), the comprehensive ego networks (T7) are mainly from Cluster 1 and Cluster 3, which correspond to the two largest clusters. Almost 28% of female respondents have a kin-oriented type of network (in the form of both the *Immediate* and the *Extended family* type), which is dominated by Cluster

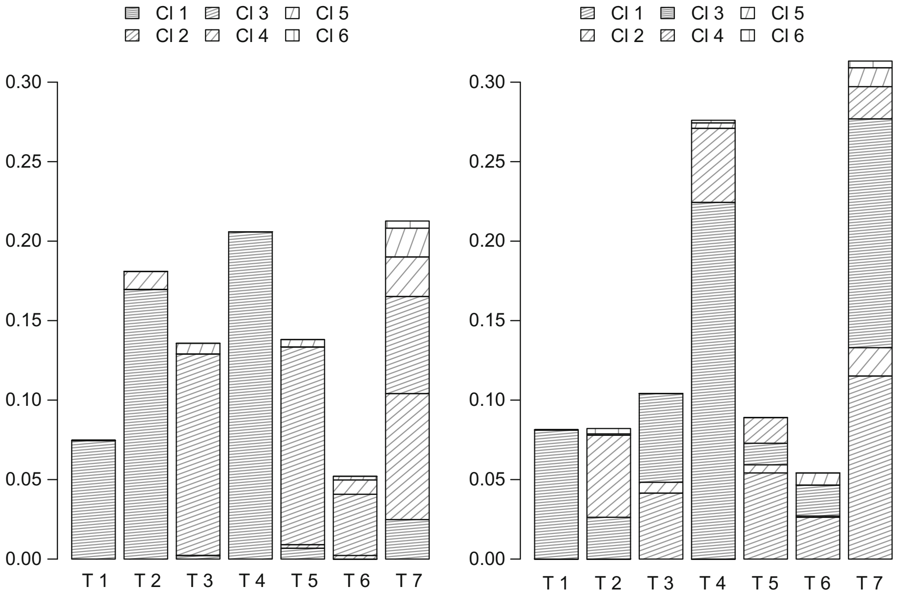

Fig. 5 Relative frequency distributions of estimated clusters (from *Cluster 1*,"Cl1", to *Cluster 6*, "Cl 6") for male (left) and female (right) respondents, according to network types (T1) Immediate, (T2) Extended, (T3) No family, (T4) Immediate+Extended, (T5) Immediate+No family, (T6) Extended+No family, (T7) Comprehensive

3 and 4. The other types of networks obtained by aggregating alter roles are less frequent than networks within the family circle or networks where all alter types are present.

We observe that, in general, networks where the ego is connected to the *Extended family* but not to children are more widespread among male respondents than females; in the same direction, having only friends and/or neighbors (no family network) is more common among males than females. Interestingly, with regard to *Comprehensive* ties for the elderly males, the clustering procedure allows the distinction between networks wherein the presence of children and grandchildren is predominant (i.e., the networks in *Cluster 2*) and egos interact more with relatives and/or friends (*Cluster 4* and *5*), thus yielding an overall large network size. Finally, concerning elderly females, we observe that the *No family* type is mainly represented by egos from *Cluster 1* and *Cluster 3*, where the main difference is the role of friends that are almost absent for the egos of this type belonging to *Cluster 3*, coherently with previous considerations on the separation between clusters in terms of the number of alter roles.

5 Conclusions

A clustering procedure to be used in the framework of ego network analysis is presented. The proposed approach is motivated by the need of non-standard methods to deal with atypical network data, as in the context of the Family and Social Subject (FSS) Survey carried out by ISTAT. We use the 2016 FSS questionnaire, and focus on the elderly living alone at the time the survey was conducted. Despite the lack of information provided by the FSS data, at both the individual and the relational level, we first build the ego networks of respondents by using the information on not-cohabiting people (along with their role relations) with whom the egos are connected. Then, a suitable dissimilarity measure for ego networks is introduced. The dissimilarity measure adopted allows for consideration of different types of variables, measured on different scales. In addition to the choice of an appropriate metric, the issue of feature weighting is addressed in terms of the variables' effects on the final results, by means of the comparison between two alternative normalization approaches. To identify clusters of ego networks, we explore a hierarchical approach yielding a "prototype" associated with each cluster in the final solution. Such prototypes are representative units of the group to which they belong, thus enhancing cluster interpretation.

In analyzing the ego networks of contacts of elderly singles in the FSS 2016 data, we found six well interpretable clusters that bring enough separation, for both male and female respondents. In particular, the identified clusters well characterize the elderly living alone in terms of the size and composition of their networks. Notably, the proposed procedure allows to isolate outlying observations in small clusters, which enables the detection of homogeneous clusters and causes the prototypes to be less affected by outliers within clusters.

Even though the proposed method is designed from very peculiar data concerning the Italian context, it can be applied to standard ego network data, provided that they are summarized in terms of the availability and/or the number of alters with specific roles.

Appendix A

Figure 6 reports the validity measures used to choose the number of clusters, i.e., the $I_w^\star$ index, which emphasizes small within-cluster dissimilarities, and the CH index, which attempts to balance internal homogeneity and cluster separation. For these indices, a better clustering is indicated by a larger value.

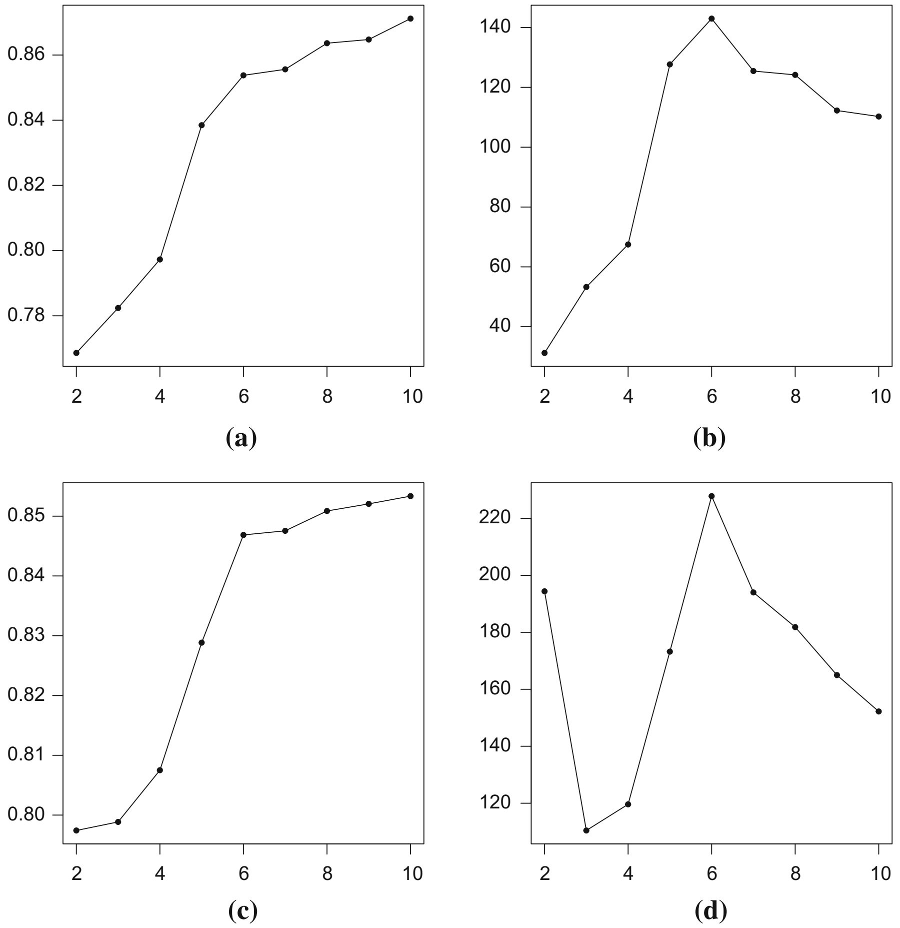

Fig. 6 I_w^* index (panels (**a**) and (**c**) for male and female respondents, respectively) and CH index (panels (**b**) and (**d**) for male and female respondents, respectively) used for cluster validation versus the number of groups ($k \in \{2, 3, \ldots, 10\}$ on the x-axis) in agglomerative HC (with minimax linkage)

Tables 7 and 8 report the distribution of the attributes (alters on six roles in the ego network of contacts of respondents) for male and female respondents in the target group, respectively. For both groups, the values observed for each variable are in the first column (e.g., *Siblings* can take on values 0, 1, 2, or 3), where for *Relatives* and *Friends* the distribution into classes is reported; for the binary attribute *Neighbors*, the value 0 refers to an absence of neighbors on whom the ego

Table 7 Proportion (%) and frequency count (in parentheses) of the variables (alter roles) by cluster for elderly males living alone (percentages may not add to 100 due to rounding)

Variable	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
Siblings						
0	54.2 (116)	78.4 (29)	74.8 (116)	30.0 (6)	92.9 (13)	0.0 (0)
1	29.4 (63)	10.8 (4)	15.5 (24)	5.0 (1)	0.0 (0)	0.0 (0)
2	9.8 (21)	5.4 (2)	5.2 (8)	55.0 (11)	7.1 (1)	50.0 (1)
3	6.5 (14)	5.4 (2)	4.5 (7)	10.0 (2)	0.0 (0)	50.0 (1)
Children						
0	35.5 (76)	2.7 (1)	67.7 (105)	80.0 (16)	28.6 (4)	0.0 (0)
1	34.1 (73)	10.8 (4)	23.2 (36)	15.0 (3)	35.7 (5)	50.0 (1)
2	21.0 (45)	59.5 (22)	9.0 (14)	5.0 (1)	28.6 (4)	50.0 (1)
3	9.3 (20)	27.0 (10)	0.0 (0)	0.0 (0)	7.1 (1)	0.0 (0)
Grandchildren						
0	53.7 (115)	2.7 (1)	90.3 (140)	90.0 (18)	64.3 (9)	50.0 (1)
1	12.2 (26)	13.5 (5)	6.4 (10)	10.0 (2)	14.3 (2)	0.0 (0)
2	16.4 (35)	8.1 (3)	3.2 (5)	0.0 (0)	7.1 (1)	50.0 (1)
3	17.8 (38)	75.7 (28)	0.0 (0)	0.0 (0)	14.3 (2)	0.0 (0)
Relatives						
[0, 1]	90.2 (193)	81.1 (30)	88.4 (137)	0.0 (0)	64.3 (9)	0.0 (0)
(1, 5]	8.9 (19)	16.2 (6)	11.6(18)	10.0 (2)	7.1 (1)	0.0 (0)
(5, 10]	0.90 (2)	2.7 (1)	0.0 (0)	70.0 (14)	14.3 (2)	0.0 (0)
> 10	0.0 (0)	0.0 (0)	0.0 (0)	20.0 (4)	14.3 (2)	100.0 (2)
Friends						
[0, 1]	100.0 (214)	67.6 (25)	51.0 (79)	45.0 (9)	0.0 (0)	0.0 (0)
(1, 4]	0.0 (0)	27.0 (10)	43.2 (67)	50.0 (10)	0.0 (0)	100.0 (2)
> 4	0.0 (0)	5.4 (2)	5.8 (9)	5.0 (1)	100.0 (14)	0.0 (0)
Neighbors						
0	93.5 (200)	0.0 (0)	0.0 (0)	25.0 (5)	0.0 (0)	0.0 (0)
1	6.5 (14)	100.0 (37)	100.0 (155)	75.0 (15)	100.0 (14)	100.0 (2)

Table 8 Proportion (%) and frequency count (in parentheses) of the variables (alter roles) by cluster for elderly females living alone (percentages may not add to 100 due to rounding)

Variable	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
Siblings						
0	61.4 (172)	52.8 (19)	88.0 (587)	0.0 (0)	85.7 (24)	63.6 (7)
1	28.9 (81)	30.6 (11)	8.2 (55)	35.2 (56)	7.1 (2)	18.2 (2)
2	8.9 (25)	8.3 (3)	2.9 (19)	27.0 (43)	7.1 (2)	18.2 (2)
3	0.7 (2)	8.3 (3)	0.9 (6)	37.7 (60)	0.0 (0)	0.0 (0)
Children						
0	47.9 (134)	36.1 (13)	18.0 (120)	49.1 (78)	42.9 (12)	36.4 (4)
1	34.6 (97)	19.4 (7)	32.8 (219)	28.9 (46)	21.4 (6)	0.0 (0)
2	15.7 (44)	25.0 (9)	32.8 (219)	14.5 (23)	28.6 (8)	27.3 (3)
3	1.8 (5)	19.4 (7)	16.3 (109)	7.6 (12)	7.1 (2)	36.4 (4)
Grandchildren						
0	67.5 (189)	44.4 (16)	33.4 (223)	74.2 (118)	53.6 (15)	54.5 (6)
1	14.3 (40)	22.2 (8)	16.5 (110)	10.7 (17)	7.1 (2)	9.1 (1)
2	13.9 (39)	13.9 (5)	20.2 (135)	8.2 (13)	14.3 (4)	0.0 (0)
3	4.3 (12)	19.4 (7)	29.8 (199)	6.9 (11)	25.0 (7)	36.4 (4)
Relatives						
[0, 1]	63.6 (178)	86.1 (31)	91.0 (607)	89.3 (142)	0.0 (0)	0.0 (0)
(1, 5]	30.7 (86)	8.3 (3)	8.4 (56)	10.7 (17)	0.0 (0)	0.0 (0)
(5, 10]	5.4 (15)	5.6 (2)	0.6 (4)	0.0 (0)	39.3 (11)	0.0 (0)
> 10	0.4 (1)	0.0 (0)	0.0 (0)	0.0 (0)	60.7 (17)	100.0 (11)
Friends						
[0, 1]	27.5 (77)	0.0 (0)	92.7 (618)	93.7 (149)	35.7 (10)	90.9 (10)
(1, 4]	61.4 (172)	2.8 (1)	7.2 (48)	5.0 (8)	42.9 (12)	9.1 (1)
> 4	11.1 (31)	97.2 (35)	0.1 (1)	1.3 (2)	21.4 (6)	0.0 (0)
Neighbors						
0	0.0 (0)	0 (0)	59.1 (394)	73.0 (116)	17.9 (5)	54.5 (6)
1	100.0 (280)	100.0 (36)	40.9 (273)	27.0 (43)	82.1 (23)	45.5 (5)

"can count if necessary," and is 1 otherwise. Columns labelled from *Cluster 1* to *Cluster 6* refer to the 6 clusters obtained by applying the HC with minimax linkage.

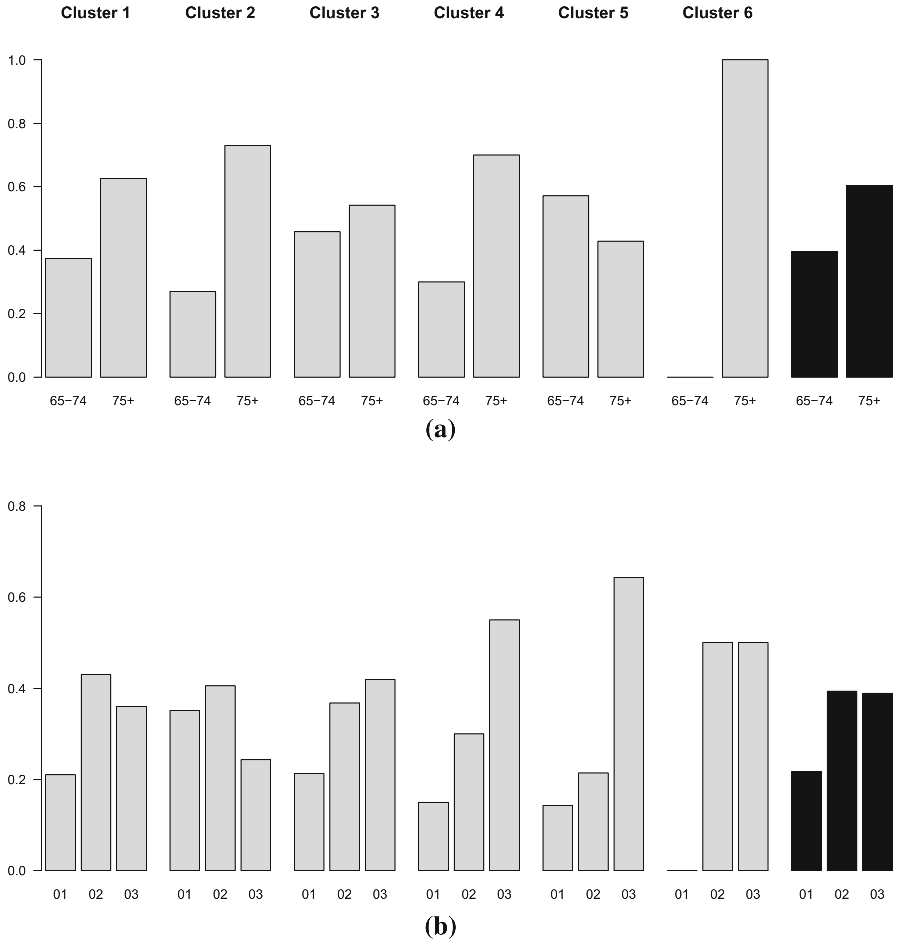

Fig. 7 Male respondents. Distribution (relative frequency) of (**a**) age, and (**b**) health condition (01 = bad, 02 = fair, 03 = good) by cluster. The distribution for the whole sample ($n_m = 442$) is given in black

Figures 7 and 8 display the distribution of age and perceived health conditions by cluster (from *Cluster 1* to *Cluster 6*) for the male and female respondents in the target group, respectively.

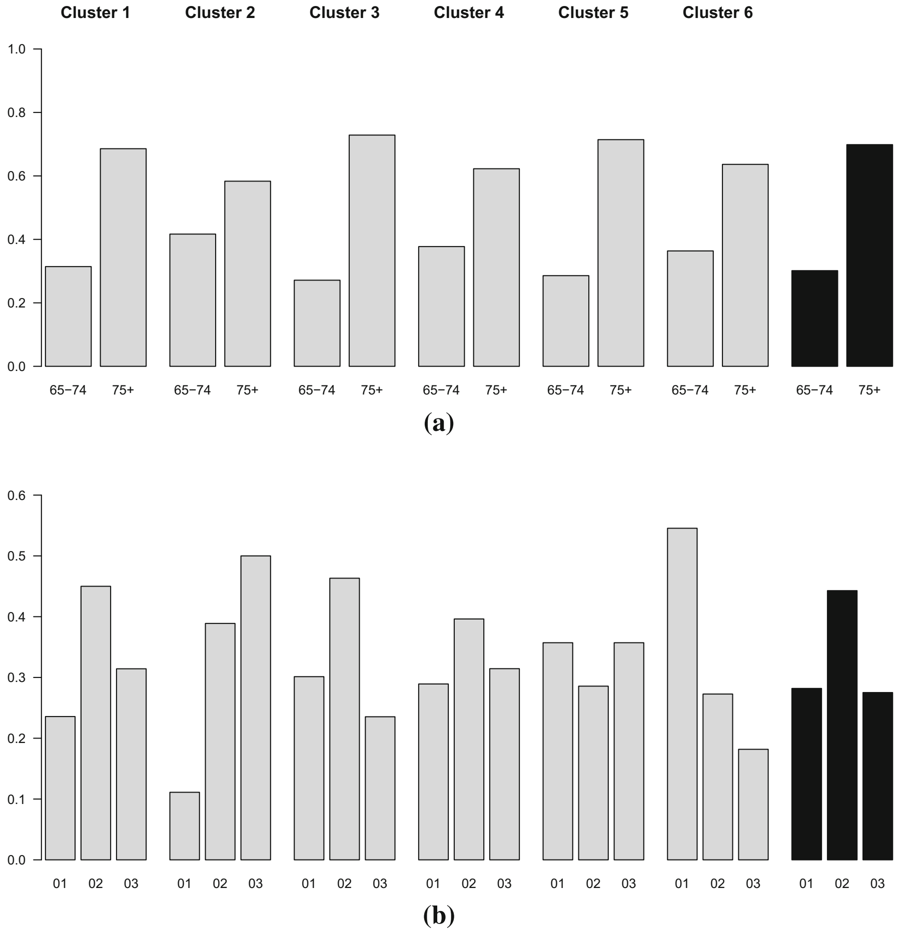

Fig. 8 Female respondents. Distribution (relative frequency) of (**a**) age, and (**b**) health condition (01 = bad, 02 = fair, 03 = good) by cluster. The distribution for the whole sample ($n_f = 1181$) is given in black

Acknowledgements The authors wish to thank the associate editor and the anonymous referees for their valuable comments and suggestions, which helped to improve the quality of this manuscript.

References

Agneessens F, Waege H, Lievens J (2006) Diversity in social support by role relations: a typology. Social Netw 28(4):427–441

Akhanli SE, Hennig C (2020) Comparing clusterings and numbers of clusters by aggregation of calibrated clustering validity indexes. Stat Comput 30(5):1523–1544

Amati V, Rivellini G, Zaccarin S (2015) Potential and effective support networks of young Italian adults. Social Indic Res 122(3):807–831

Amati V, Meggiolaro S, Rivellini G, Zaccarin S (2017) Relational resources of individuals living in couple: evidence from an Italian survey. Social Indic Res 134(2):547–590

Ayalon L, Levkovich I (2019) A systematic review of research on social networks of older adults. Gerontologist 59(3):e164–e176

Bidart C, Degenne A, Grossetti M (2018) Personal networks typologies: a structural approach. Social Netw 54:1–11

Bien J, Tibshirani R (2011) Hierarchical clustering with prototypes via Minimax Linkage. J Am Stat Assoc 106(495):1075–1084

Bien J, Tibshirani R (2019) Protoclust: hierarchical clustering with prototypes. r package version 1.6.3. https://CRAN.R-project.org/package=protoclust

Brandes U, Lerner J, Nagel U (2011) Network ensemble clustering using latent roles. Adv Data Anal Classif 5(2):81–94

Breiger RL (2004) The analysis of social networks. In: Hardy M, Bryman A (eds) Handbook of data analysis. Sage, London, pp 505–526

Caliński T, Harabasz J (1974) A dendrite method for cluster analysis. Commun Stat Theory Methods 3(1):1–27

Crossley N, Bellotti E, Edwards G, Everett MG, Koskinen J, Tranmer M (2015) Social network analysis for ego-nets: social network analysis for actor-centred networks. Sage, London

De Amorim RC, Mirkin B (2012) Minkowski metric, feature weighting and anomalous cluster initializing in K-Means clustering. Pattern Recogn 45(3):1061–1075

Djundeva M, Dykstra PA, Fokkema T (2019) Is living alone “aging alone”? Solitary living, network types, and well-being. J Gerontol B Psychol 74(8):1406–1415

Domínguez S, Maya-Jariego I (2008) Acculturation of host individuals: immigrants and personal networks. Am J Comm Psychol 42(3–4):309

Dykstra PA, Bühler C, Fokkema T, Petrič G, Platinovšek R, Kogovšek T, Hlebec V (2016) Social network indices in the generations and gender survey: an appraisal. Demogr Res 34:995–1036

Durso P, Massari R (2019) Fuzzy clustering of mixed data. Inform Sci 505:513–534

Everitt B, Landau S, Leese M, Stahl D (2011) Cluster analysis, 5th edn. Wiley, New York

Gallagher EN, Vella-Brodrick DA (2008) Social support and emotional intelligence as predictors of subjective well-being. Pers Indiv Differ 44(7):1551–1561

Giannella E, Fischer CS (2016) An inductive typology of egocentric networks. Social Netw 47:15–23

Gower JC (1971) A general coefficient of similarity and some of its properties. Biometrics 27:857–871

Halkidi M, Batistakis Y, Vazirgiannis M (2001) On clustering validation techniques. J Intell Inf Syst 17:107–145

Hennig C, Liao TF (2013) How to find an appropriate clustering for mixed-type variables with application to socio-economic stratification. J Roy Stat Soc C Appl 62(3):309–369

Huang Z (1998) Extensions to the k-means algorithm for clustering large data sets with categorical values. Data Min Knowl Discov 2(3):283–304

Hubert L, Arabie P (1985) Comparing partitions. J Classif 2(1):193–218

Kalmijn M, Vermunt JK (2007) Homogeneity of social networks by age and marital status: a multilevel analysis of ego-centered networks. Social Netw 29(1):25–43

Kaufman L, Rousseeuw PJ (2009) Finding groups in data: an introduction to cluster analysis, vol 344. Wiley, New York

Marsden PV (2011) Survey methods for network data. SAGE Handb social Netw Anal 25:370–388

Maya-Jariego I (2021) Building a structural typology of personal networks: individual differences in the cohesion of interpersonal environment. Social Netw 64:173–180

McCarty C (2002) Structure in personal networks. J Soc Struct 3(1):20

McCarty C, Lubbers MJ, Vacca R, Molina JL (2019) Conducting personal network research: a practical guide. Guilford Publications, New York

Molina JL, Maya-Jariego I, McCarty C (2014) Giving meaning to social networks: methodology for conducting and analyzing interviews based on personal network visualizations. In: Mixed methods social networks research Design and applications, pp 305–335

Moore G (1990) Structural determinants of men’s and women’s personal networks. Am Sociol Rev 726–735

Pelle E, Zaccarin S, Furfaro E, Rivellini G (2021) Support provided by elderly in Italy: a hierarchical analysis of ego networks controlling for alter-overlapping. Stat Method Appl 1–26

Perry BL, Pescosolido BA, Borgatti SP (2018) Egocentric network analysis: foundations, methods, and models, vol 44. Cambridge University Press, Cambridge

Podani J (1999) Extending Gower’s general coefficient of similarity to ordinal characters. Taxon 48(2):331–340

Prostov MY, Suarez-Alvarez MM, Prostov YI (2015) Properties of the sample estimators used for statistical normalization of feature vectors. Data Min Knowl Discov 29(6):1815–1837

Sherman CW, Webster NJ, Antonucci TC (2013) Dementia caregiving in the context of late-life remarriage: support networks, relationship quality, and well-being. J Marriage Fam 75(5):1149–1163

Suarez-Alvarez MM, Pham DT, Prostov MY, Prostov YI (2012) Statistical approach to normalization of feature vectors and clustering of mixed datasets. P Roy Soc A Math Phys 468(2145):2630–2651

Taylor SE (2007) Social support. In: Friedman HS, Silver RC (eds) Foundations of health psychology. Oxford University Press, Oxford, pp 145–171

Vacca R (2020) Structure in personal networks: constructing and comparing typologies. Netw Sci 8(2):142–167

van de Velden M, Iodice D'Enza A, Markos A (2019) Distance-based clustering of mixed data. WIREs Comput Stat 11(3):e1456