Chapter 7
A Local Structure-Based Method for Nodes Clustering: Application to a Large Mobile Phone Social Network

Alina Stoica, Zbigniew Smoreda, and Christophe Prieur

Abstract In this paper we present a method for describing how a node of a given graph is connected to the network. We also propose a method for grouping nodes into clusters based on the structure of the network in which they are embedded, so on the description provided by the first method. We apply these methods to a mobile phone communications network. When confronting the obtained clusters of individuals to their age and to their intensity of communication, the results are quite promising: the two measures are correlated to the social network cluster. We finish by providing a typology of the mobile phone users based on social network cluster, communication intensity and age.

7.1 Introduction

In this paper, we want to describe how each individual of a given social network is connected to the network and to cluster nodes that are connected in a similar way to the network. One can see this distribution of nodes into clusters as an identification of network "roles". Without pretending to have solved the problem of identification of roles, we present a method to distribute nodes into clusters based on the local structure of the network.

A. Stoica (✉)
EDF R&D, 1 av. du Gen. de Gaulle Clamart, Clamart, France
e-mail: alina.stoica@edf.fr

Z. Smoreda
Orange Labs, 38-40 rue du Gen. Leclerc, Issy les Moulineaux, France
e-mail: zbigniew.smoreda@orange-ftgroup.com

C. Prieur
LIAFA, Paris-Diderot, 75 rue du Chevaleret, 75013 Paris, France
e-mail: prieur@liafa.jussieu.fr

T. Özyer et al. (eds.), *The Influence of Technology on Social Network Analysis and Mining*, Lecture Notes in Social Networks 6, DOI 10.1007/978-3-7091-1346-2_7,

We apply the method to a large mobile phone network. The obtained results are quite promising, in particular when the clusters are confronted to other characteristics of the individuals. Indeed the probability that an individual belongs to a certain cluster depends on his or her age; even more, using these probabilities we are able to group together different ages, thus discovering four groups containing consecutive ages, corresponding to four life stages. The probability that a person belongs to a certain cluster also depends on his or her mobile phone communication intensity; moreover the intensity of communication allows us to predict with rather high accuracy the cluster membership of a person.

We begin by recalling some work related to the subject. Next we present the method for describing how a node of a given graph is embedded in the network. We then propose a method for clustering nodes based on the structure of the network surrounding them. Next we present the results of the application of the methods to the mobile phone communications network: the obtained clusters of individuals, the correlations with the age and the intensity of communication and a typology of mobile phone users based on social network cluster, communication intensity and age.

7.2 Related Work

Social roles. The notion of role refers to the position of an actor in society and it is based on the relationships that the actor in question has with other actors. Actors playing a particular social role are connected in the same way to the network. Generally, the nodes in a network can be grouped into equivalence classes based on the roles they play, so nodes having the same role have to be equivalent or similar to each other by some metric. Probably the best known equivalence relations for this purpose are the structural, the automorphic and the regular equivalence.

Structural equivalence [11]. Two nodes are considered equivalent if and only if they have exactly the same neighbors in the graph, so they are linked to exactly the same set of nodes with (in the case of directed graphs) the arrows pointing in the same directions. Thus, two structurally equivalent actors can exchange their positions without changing the network.

However, it is not frequent to find two persons with identical relations. There are examples of actors who play the same role without being connected to exactly the same people, but rather have similar relations with people who have themselves a same role. The two following relations express this idea.

Automorphic equivalence. Two nodes are considered equivalent if one is the automorphic image of the other one. Formally, two vertices u and v of a graph G are *automorphically equivalent* if there is an automorphism φ of G such that $\varphi(u) = v$.[1]

[1] The notion of automorphism is presented in Sect. 7.3.

Regular equivalence [3, 18]. Two nodes are considered equivalent if they are connected to equivalent nodes. Imagine that nodes having the same role are given the same color (and nodes with distinct roles are given distinct colors). If two nodes are equivalent, the colors found in the neighborhood of one node are also found (possibly in different numbers) in the neighborhood of the other node. Also, the definition can be understood in the following way: every equivalence class is represented by a single node in an "image graph" (also called "blockmodel" or "role model"). The nodes of the image graph are connected (disconnected) if the nodes in the corresponding classes are connected (disconnected) in the original graph.

A lot of research has been devoted to blockmodeling. Some authors focused on efficient algorithms for blockmodeling [2], others on its mathematical foundations [1, 5, 18], others proposed problem relaxations [15] or generalizations for different types of relations [5]. The method we propose here is different from the research on blockmodels. Although the goal is the same, to cluster nodes that share some network characteristics, our method can be applied to nodes that belong to the same graph as well as to nodes belonging to different graphs (as for example nodes in personal networks obtained by interviews). The blockmodeling, on the other hand, searches for roles in a same graph. Also, our method can be easily applied to (very) large networks, taking a few dozens of minutes to compute clusters of nodes in a graph containing millions of nodes.

The method we propose here is somehow related to the equivalence of roles introduced by Burt [4] who published in English the work of Hummell and Sodeur [8]. In this paper the authors characterized each node of a given network by the number of occurrences of the node in triads. One looks for the presence/absence of links between the given node and every other two nodes of the network. As the graph where the triads are computed is directed, one counts the presence of the node in 36 types of triads. Then the Euclidian distance is used in order to find similar nodes. In a way, the method we propose here follows this idea. However, we look for patterns of a higher order than the triads. Also, one major difference is that, when characterizing a node, we look only at its neighbors and the connections between them, while Hummell and Sodeur look at its relation with every pair of nodes in the network. Their characterization is therefore more detailed, but way too complex for large networks. Looking at $\binom{n-1}{2}$ nodes in order to characterize one node of a network with n nodes is impossible when n is high, so this method cannot be applied to large social networks. Another difference is that our method is designed for undirected networks, but it can be easily modified to take into consideration directed graphs.

Mobile phone social networks. We apply the method proposed here to a large network built from mobile phone communications. Different properties have been already identified in such networks [12, 13]. Onnela et al. [13] show with no surprise that the distributions of degree and of the duration of calls are power-laws. They also give a definition for the strength of ties depending on the duration of calls and they analyze the connection between the strength and the connectivity or the community structure.

Using mobile phone communication data, Lambiotte et al. [10] were able to test the sociological hypothesis that the existence of a call between two persons depends on the geographical distance between them. They thus show that the probability of a mobile phone call is inversely proportional to the square of the geographical distance between the two persons. Several researchers analyzed the temporal dynamics of mobile phone networks e.g. the temporal stability of links [7, 14]. In [7], Hidalgo and Rodriguez-Sickert define the persistence of a link over a set of time periods as the number of periods where the link is activated (i.e. there are reciprocal calls between the two persons during that period). They find that persistent links are more common with people with low degree and high clustering.

However, the properties of mobile phone social networks computed in these studies are global, characterizing the network as a whole. Here, our aim is to characterize the local structure of the graph. We thus propose methods to describe the way each node is embedded into the network and to find similarly connected nodes. We then apply the methods to a large mobile phone network.

7.3 Preliminaries

7.3.1 Basic Graph Notions

Let $G = (V, E)$ be a graph; V is the set of its vertices, $E \subseteq V \times V$ is the set of its edges. The graph G is *undirected* if for all $(u, v) \in E$ also $(v, u) \in E$ i.e. edges are unordered pairs of nodes. G is *connected* if there exists a finite path between every two vertices and it is *simple* if it has no multiple edges (i.e. for all $u, v \in V$ there is at most one edge connecting u to v) and no self-loops ($(v, v) \notin E$, for all $v \in V$). All the graphs we consider here are simple and undirected.

Given a vertex $v \in V$, a vertex $u \in V$ is a *neighbor* of v if and only if $(u, v) \in E$. The set of neighbors of v represents its *neighborhood* denoted by $N(v) = \{u \in V, (u, v) \in E\}$ and the cardinal of this set represents its *degree*. Two graphs $G = (V, E)$ and $H = (V', E')$ are *isomorphic* if and only if there exists a bijective function $\varphi : V \to V'$ such that, for any two vertices u and v in V, $(u, v) \in E$ if and only if $(\varphi(u), \varphi(v)) \in E'$. If G and H represent the same graph, the function φ is called *automorphism* of the graph G. The subgraph *induced* by a set of vertices $V' \subseteq V$ in G is the graph $H = (V', E')$ with $E' = \{((u, v) \in E \mid u, v \in V')\}$.

7.3.2 Data Mining Notions

ANOVA test. This test measures the correlation between a continuous variable and a categorization. It tells if the mean of the continuous variable is the same for the different categories. If this is true then the two variables are independent.

For instance, one can use the ANOVA test in order to see if the salary (the continuous variable) is independent from the gender (the categories, male and female). However, this test says only if the means are different or not, but it does not say for which categories the means are significantly different and for which they are not. A test that can provide such information is called a multiple comparison test. Such tests are the Bonferroni [6] and the Scheffé tests [16].

We also present briefly two widely-used methods for clustering objects.

K-means clustering. Given a set of objects, the *k-means* algorithm groups them into a given number k of clusters using the distance between the objects. If each object is characterized by a feature vector with n elements, one usually uses the Euclidian distance between the feature vectors as distance between objects. The Euclidian distance is defined as

$$d(u,v) = \sqrt{(u_1 - v_1)^2 + (u_2 - v_2)^2 + \cdots + (u_n - v_n)^2} = \sqrt{\sum_{i=1}^{n}(u_i - v_i)^2}$$

where u and v are two objects characterized by two feature vectors with n elements $(u_1, \ldots u_n)$ and $(v_1, \ldots v_n)$ respectively.

Given a cluster K containing n_K objects characterized by feature vectors of n elements, the center (or centroid) C_K of the cluster is a vector representing the average of all the objects in the cluster i.e. for each variable i from 1 to n, the i-th value of the vector is the arithmetic mean of the i-th values of the feature vectors of the objects in the cluster: $C_K(i) = \frac{1}{n_K} \sum_{v \in K} v_i$ where v is an object in the cluster and v_i is the i-th value of its feature vector.

Kohonen self organizing maps. Given a set of p individuals (or objects) characterized by feature vectors with n variables, the aim of the *Kohonen self-organizing map*[9] is to cluster the individuals and also to build a bi-dimensional map with n layers (a layer for each variable describing the individuals) where the individuals are placed depending on their topological proximity. The map's smallest entity is a cell, and each individual is placed in only one cell (the individual has the same position and therefore cell on all the layers); there are $\sqrt{|p|}$ cells. The method has three steps. The first one is the learning. The feature vectors of the cells are randomly initialized. Then a subset of the population to model is randomly selected; for each individual in this selection the SOM finds the ("winner") cell whose feature vector is the most similar (i.e. is the closest by a given distance). The feature vector of the winner cell is updated to take into account the feature values of the individual. The feature vector of the neighbor cells are then modified to reduce the vectors gradient with the new values of the cells' feature vector. The second step of the algorithm is the processing of the global population to model: each individual is placed in the cell with the closest feature vector. Finally the last step is the clustering of the cells with, for instance, a k-means algorithm, based on the similarity of their feature vectors.

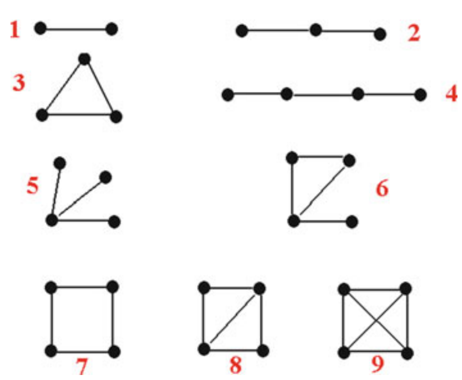

Fig. 7.1 The nine patterns with at most four vertices and at least one edge

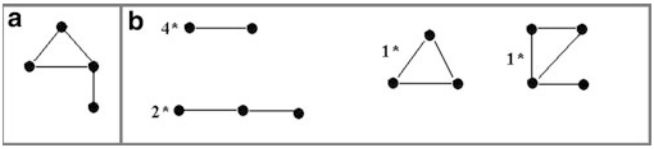

Fig. 7.2 A graph (*a*) and its patterns (*b*)

7.4 Local Structure-Based Node Characterization

Suppose that we are given a network that represents a set of individuals and some connections between them. We want to study how each one of the individuals is connected to the network. For that, we analyze the connections between each node and its neighbors and between these neighbors. More precisely, we characterize the egocentred network of each one of the nodes in the network. By egocentred network of a given node we mean the network formed by its neighbors and the links between them.

Formally, let $G = (V, E)$ be a simple undirected graph such that V corresponds to the set of individuals and E to the set of connections between them: two vertices u and v are connected by an edge (u, v) if there is a connection between the two individuals u and v. We call *egocentred network* of the node $v \in V$ the graph $Eg(v)$ induced by the neighbors of v in G i.e. the graph whose vertices are the neighbors of v and whose edges are the edges between these neighbors in G.

We call *patterns* the nine non-isomorphic undirected connected graphs with at most four vertices and at least one edge (Fig. 7.1). We say that a pattern P appears in a graph $G = (V, E)$ if there exists a set of vertices $V_P \subseteq V$ such that the subgraph induced by V_P in G is isomorphic to P. Listing all the occurrences of the pattern P in the graph G means finding all the sets of vertices V_P according to the previous definition. As an example, Fig. 7.2 represents a graph (a) and the patterns it contains and their number of occurrences (b).

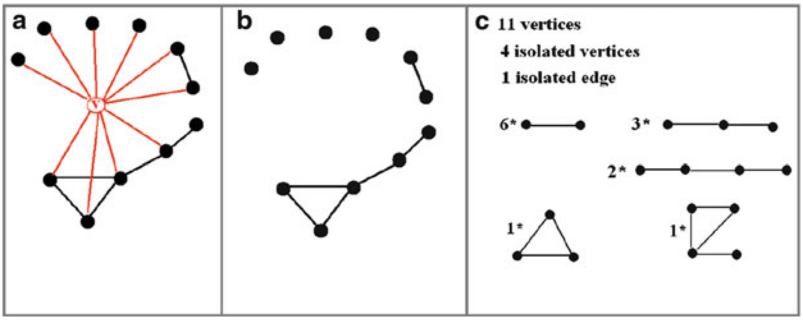

Fig. 7.3 A vertex v and its neighbors (a), the egocentred network $Eg(v)$ of v (b) and the patterns of $Eg(v)$ (c)

Now, to characterize a node v of a graph G we proceed as it follows (method ***characterize(v)***; this is part of the method *local_structure* that we introduced in [17]):

Step 1. Extract the egocentred network $Eg(v)$ of v i.e. the subgraph induced by the neighbors of v in G;

Step 2. List the patterns of $Eg(v)$;

Let us explain the two steps of the method with an example. In Fig. 7.3a, the black circles correspond to the neighbors of v. The egocentred network $Eg(v)$ of v is represented in Fig. 7.3b and the patterns of $Eg(v)$ in Fig. 7.3c.[2] We chose not to include v in its egocentred network because we know that it is connected to all the vertices in this graph, its presence doesn't bring any information. After performing the two steps of the method one has a rich description of the way v is connected to the graph G. For a more detailed description of the local structure of G around v one can list the patterns of a higher order (with five vertices or more); the patterns with four vertices are however a good compromise between the variety of forms and their number, providing, in many cases, a detailed enough picture.

We call *pattern-frequency vector* of v the vector containing the number of occurrences of the different patterns in its egocentred network (along with the number of isolated vertices and edges in its network):

Definition 1. Given a vertex v of a graph $G = (V, E)$, we call pattern-frequency vector of v the vector

$$f(v) = (f_{iv}(v), f_{ie}(v), f_{\sqcup}(v), f_{\llcorner}(v), f_{\triangle}(v), f_{\sqsubset}(v), f_{\angle}(v), f_{\boxslash}(v), f_{\square}(v), f_{\boxslash}(v), f_{\boxtimes}(v))$$

[2]We have also counted the number of isolated vertices and edges in $Eg(v)$.

where:

- $f_{iv}(v)$ is the number of isolated vertices in the egocentred network $Eg(v)$,
- $f_{ie}(v)$ is the number of isolated edges

and the subsequent components are the numbers of occurrences of the patterns as induced subgraphs in the egocentred network $Eg(v)$ of v:

- $f(v)$, pattern 1, edges,
- $f(v)$, pattern 2, paths with two vertices,
- $f(v)$, pattern 3, triangles,
- $f(v)$, pattern 4, paths with three vertices,
- $f(v)$, pattern 5, stars,
- $f(v)$, pattern 6,
- $f(v)$, pattern 7, chordless squares,
- $f(v)$, pattern 8, squares with one chord,
- $f(v)$, pattern 9, four-cliques.

For instance, for the vertex in Fig. 7.3a, its pattern-frequency vector is

$$f(v) = (4, 1, 6, 3, 1, 2, 0, 1, 0, 0, 0).$$

Note that this method provides a description of how a given vertex is connected to the network; it can be applied to the set of all the vertices of the graph or only to some of them. As it is local, one doesn't need to have all the vertices and edges in the graph, but only the neighbors of each studied vertex and the edges between them.

7.5 Pattern Frequency Clustering of Nodes

In this section we want to group together nodes that are connected in a similar way to the network. We use the previously defined pattern-frequency vectors in order to describe how the different nodes are connected to the network. Now we have to define what "connected in similar ways" represents.

One possibility is to define an equivalence relation on nodes using the pattern-frequency vectors. For instance:

Definition 2. Two vertices of the graph G are said to be equivalent if and only if they have identical pattern-frequency vectors. We call this pattern equivalence.

Note that this equivalence is less strict than the structural and the automorphic equivalences. Indeed, vertices that have exactly the same neighbors in the network (so are structurally equivalent) have identical egocentred network, so identical feature vectors, and therefore are pattern equivalent. Also, vertices that are automorphically equivalent have isomorphic egocentred networks, so identical feature vectors and are thus pattern equivalent. For the two definitions, the opposite is not always true, so one can say that the pattern equivalence is included in the structural and automorphic equivalences. This means that the pattern equivalence is less strict than these two relations. However, it is still not enough flexible for real-world networks where nodes having exactly the same patterns in exactly the same amounts are rare (at least for values of the degree higher than, let's say, 10). Thus, the equivalence classes obtained when applying the definition to large graphs are much too numerous. Here we want to group the nodes of a given large network into a small number of classes (i.e. smaller than a given constant, for instance 20). Each class should contain similar nodes in terms of network structure. It is the local structure of the network surrounding the node that should matter when attributing a node to a class, and not its degree or the fact of being connected to other nodes in the class. The interest of computing such classes is that they are very easy to use. Thus, one can measure correlations with other properties of the nodes or make predictions (e.g. predict a property when knowing the class and vice-versa).

Instead of grouping together nodes that have identical pattern-frequency vectors (as in the pattern equivalence), we cluster nodes that have similar vectors. For that, we use a classical clustering method, the k-means algorithm. The advantage of performing a clustering to define vertex equivalence is its flexibility: one can distribute the vertices into a small number of clusters (if this is his or her goal) or a large number of clusters (where vertices in the same cluster are very similar to each other).

Before performing the clustering, we filter out vertices that have identical pattern-frequency vectors. These vertices are not distinguishable by using only the patterns; their egocentred networks contain exactly the same patterns in exactly the same number. By default, they belong to the same cluster. The elimination of multiple copies of the same pattern-frequency vector insures a smaller complexity of computation and also allows us to perform a finer clustering. Of course, after having clustered the remaining vertices (we call them the reduced population), we put the filtered out vertices into the clusters where the vertices with identical vectors have been already placed.

Definition 3. Given a graph G, we call reduced population of G a maximal set of vertices of G that have distinct pattern-frequency vectors. Given a positive integer d, we denote by $Pop_d(G)$ the set of vertices in the reduced population of G that have degree d (in G).

7.5.1 The Issue of the Degree

There is an important factor that must be taken into consideration prior to the clustering: the degree of vertices. It is difficult to compare the number of occurrences of patterns in egocentred networks of vertices with different degrees because these values are biased by the degree. For vertices with high degrees, the number of

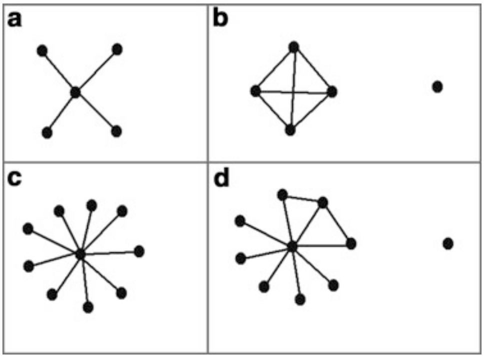

Fig. 7.4 An example of four egocentred networks with five vertices (*a* and *b*) and ten vertices (*c* and *d*) respectively (ego has been removed)

occurrences can have high values, too. Actually, for a vertex (ego) with degree d, a pattern with k vertices can occur at most $\binom{d}{k}$ times in its egocentred network. So, while the minimal value of the number of occurrences of a pattern is always 0, the maximal value depends on the degree of ego. Therefore, the exact values of the number of occurrences of patterns can be misleading. Look, for instance, at the four egocentred networks in Fig. 7.4 (ego has been removed). Their pattern-frequency vectors are presented in Table 7.1 where one can see that the values of many variables are higher for C and D than for A and B. Even more, the networks C and D look more similar to each other than A and B, so the vectors of C and D should be closer to each other than those of A and B. However, the Euclidian distance between the pattern-frequency vectors is 74 for A and B and 1,726 for C and D.

In order to avoid the problem of the degree, we choose to perform a clustering for each degree. Thus, the distance between the vertices C and D in the previous example will be compared to the distances between other pairs of vertices of degree 10 and not to all the input vertices. If we manage to group together the vertices of each degree in a same number of clusters and to match together the clusters obtained for the different degrees, then we have that each cluster contains vertices of all the degrees. This is exactly our goal here: we want a vertex to belong to a given cluster because it has a certain type of connection to the network and not because it has a certain degree. Thus, if a vertex gets another degree during time, we can see if the type of structure in which it is connected also changes by checking if its cluster changes. It is not the difference of degree that we want to capture but the difference of structure. If we don't have exactly the same clusters for all the degrees, we cannot do this. And this is exactly what might happen if we perform a single clustering for all the degrees (and not for each degree separately): there might be clusters with no vertices of some degrees (because, for instance, there are fewer vertices of that degree).

Table 7.1 The pattern-frequency vectors of the egocentred networks in Fig. 7.4

net.	f_{deg}	f_{iv}	f_{ev}	$f_{\multimap}$	$f_{\llcorner}$	$f_{\triangle}$	$f_{\sqsubset}$	$f_{\angle}$	$f_{\boxslash}$	$f_{\square}$	$f_{\boxbslash}$	$f_{\boxtimes}$
A	5	0	0	4	6	0	0	4	0	0	0	0
B	5	1	0	6	0	4	0	0	0	0	0	1
C	10	0	0	9	36	0	0	84	0	0	0	0
D	10	1	0	10	26	2	0	45	10	0	1	0

7.5.2 Pattern-Frequency Clustering of Nodes

We proceed as it follows:

1. For each degree, we perform several k-means clusterings on the vertices with that degree in the reduced population, using different numbers of clusters; we compute the best number of clusters;
2. We keep as final number of clusters the number indicated as best for most degrees; let this number be n_c;
3. For each degree, we divide the vertices with that degree into n_c clusters;
4. We finally match the clusters found for the different degrees.

Let us explain the different steps.

STEP 1. Given that we base our clustering on occurrences of patterns with four vertices and less, we cluster only vertices with degree at least 4. For each degree, we use the k-means algorithm on modified versions of the pattern-frequency vectors of the nodes. As k-means starts by randomly picking the first centers, we perform 50 clusterings for each degree and each number of clusters and choose the clustering with the lowest intra-cluster variance. The best number of clusters is computed by comparing the average silhouette values obtained for the different numbers of clusters.

Let us explain why and how we modify the pattern-frequency vectors. The k-means algorithm uses a given distance between elements in order to compute the clusters; this distance is usually the Euclidian distance between the feature vectors of the elements. We need to modify the pattern-frequency vectors before computing the Euclidian distance on them. There are several reasons for that.

(a) Modifying the ranges of values. Even if we focus on each degree at a time, the numbers of occurrences of the different patterns are not placed in the same ranges of values. For instance, the maximal number of occurrences of the $\angle$–pattern is generally a lot higher than the maximal number of the $\boxtimes$–pattern. We need to place the ranges of values of all the variables participating to the Euclidian distance between the same extreme values. This can be done for instance by centering and scaling the variables or by giving them new values, obtained from a computation of slices. It is the second solution that we adopt here.

Generally, given a group of n elements that have values $a_1, a_2 \dots a_n$ for a given attribute (or variable) a, one can compute k bins (or slices) such that there is a fairly equivalent number of elements whose values are placed in each bin. For that, one needs to compute $k + 1$ ascendant values (called limits) such that the first limit is the minimal value of a_i for $i \in \{1, 2, \dots n\}$, the last limit is the maximal value of a_i and there is a fairy equivalent number of elements (i.e. $\frac{n}{k}$) whose values are placed between two consecutive limits. Now, one can use instead of the values $a_1, a_2 \dots a_n$ the corresponding slices: instead of the value a_i one uses the value x if a_i belongs to the $x-$th bin. Note that the computation of only two bins ($k = 2$) is equivalent to the computation of the median value of the attribute a. In this case, one can use, instead of the real value a_i of the attribute, a value that is either 1 or 2 depending on a_i : if a_i is inferior to the median value, then one uses 1, otherwise 2.

This is the technique that we apply here. Instead of using the real values of the pattern-frequency vectors, we compute and use slices of values. There are several advantages in doing this. First, we eliminate the problem of comparing very different values for different patterns: now we have, for all the patterns, the same possible values. Second, the new values are established using the ranges of values, as found in the network. Thus, the number of occurrences of a given pattern in a given egocentred network can be very small comparing to the maximal possible value and, in the same time, very high comparing to its value in the other egocentred networks. We want to emphasize the fact that this value is high in *our* network, which the slices do. Third, the extreme values (often difficult to handle) are simply put in the marginal slices and are no longer seen as extreme.

For each degree d and each one of the 11 components of the pattern-frequency vector, we choose five bins such that an equivalent number of nodes in Pop_d (the reduced population with degree d) have values in each one of the bins.

(b) Using the absent patterns. By using the pattern-frequency vectors we take into consideration the presence of different structures in the egocentred networks. Besides this, it can be useful to take into consideration also the absence of different structures. Thus, two nodes are similar if they have many common patterns in their egocentred networks, but also if patterns that are not present in one are not present in the other one either. To take this information into consideration, we add to the pattern-frequency vector of each node the pattern-frequency vector of the complement graph of its egocentred network. The complement graph of a graph $G = (V, E)$ is a graph $G' = (V', E')$ where the vertices are the same as in G (i.e. $V' = V$) and the edges are all the possible edges between vertices in V that are not present in E (i.e. $E' = \{(u, v), u, v \in V \text{ and } (u, v) \notin E\}$).We thus have, for each vertex v, a vector containing the number of occurrences of patterns in the egocentred network $Eg(v)$, followed by the number of occurrences of patterns in the complement graph $Eg'(v)$ of the egocentred network. Next we replace the real values in this new vector by the corresponding slices as previously explained; we thus obtain the *extended pattern-frequency vector*.

Definition 4. Given a vertex v of a graph G, we call extended pattern-frequency vector of v the vector with 22 components containing first the slice values of the pattern-frequency vector of v and then the slice values of the pattern-frequency vector of the complement graph $Eg'(v)$ of the egocentred network $Eg(v)$ of v.

It is on the extended pattern-frequency vectors that we compute the Euclidian distance and we perform the k-means clustering.

STEP 3. Suppose n_c was found as best number of clusters for most degrees, so we need to divide the nodes with each degree in the reduced population into n_c clusters. We perform again 50 k-means clusterings with $k = n_c$ for each degree and we keep the clustering with the lowest intra-cluster variance.

STEP 4. We have now n_c clusters for each degree greater than 3. We need to match the clusters obtained for the different degrees so that, every node, no matter its degree, belongs to one of the n_c clusters. In order to do the matching, we compute the center (or centroid) of each cluster for each degree. Recall that the center of a cluster is the average of all the points in the cluster i.e. a vector where each component is the arithmetic mean of the values of that component for all the elements in the cluster.

We match clusters for consecutive degrees by using the centers: for each degree $d > 4$, we compute the centers of the clusters obtained for d (let C_i be the center of the ith cluster, with i from 1 to n_c) and for $d - 1$ (let C_i' be the center of the ith cluster, with i from 1 to n_c) and the Euclidean distances between these centers. For each one of the clusters obtained for degree d we have to choose exactly one cluster from those obtained for degree $d - 1$, and each one of these clusters must be chosen exactly once. This corresponds to a permutation of n_c elements: each cluster with index 1 to n_c obtained for degree d is given a new index, also from 1 to n_c, corresponding to the cluster for degree $d - 1$ with which it is matched. We choose the permutation σ that minimizes the sum of distances between centers of matched clusters: $\sum_{i=1,\ldots,n_c} dist(C_i, C'_{\sigma(i)})$. For that, let us observe that if there is a valid permutation σ such that, for all i from 1 to n_c, $dist(C_i, C'_{\sigma(i)})$ is the minimum distance between C_i and any C_j', with j from 1 to n_c, then σ is the permutation that minimizes the sum of distances. This case may occur for many pairs of consecutive degrees, so in this case no other computation is needed. After having computed the permutation σ that minimizes the sum of distances, one has a bijective matching of clusters for the given pair of consecutive degrees. By doing this for each pair, we obtain a matching of all the clusters.

Each vertex in the reduced population thus belongs to one of the n_c clusters. We now distribute into clusters the vertices that we have previously filtered out by putting them in the clusters of the vertices with the same pattern-frequency vector.

7.6 Mobile Phone Communications Network

We apply the previously presented methods to a large mobile phone communication network. We first present the dataset and some statistics on duration and frequency of communication depending on the age of the person. Then we describe how each node is connected to the network using the method that we have introduced in Sect. 7.4. Next, we group the individuals in the social network into clusters using the method presented in Sect. 7.5. Finally, we compare the obtained clusters to the age and communication intensity of the people in the social network.

7.6.1 Data Description

The analyzed dataset contains the CDRs (call detail records) of the mobile phone communications of the customers of a mobile phone operator (we call it O) in a European country during the month of October 2006. O has approximately 30 % share market in the country. The dataset contains several details of each mobile phone communication in the O network: the identifiers of the two persons in communication, their mobile phone operators (for the communication to be recorded, at least one of the two persons must be a O customer), the type of communication (this can be call or short message SMS), the time when the communication began and its duration (in the case of a phone call). The phone numbers have been hashed and each person has been given a unique identifier that does not allow finding the identity of the person. The dataset contains over one billion records involving ten millions users. As we do not have the mobile phone communications between the persons not belonging to O, we keep in our analysis only the communications where the two persons are both O customers. We thus analyze 3.3 million users that have exchanged more than 170 million phone calls and SMS during the followed month. For these customers the database contains also their age. We compared the age distribution of the mobile phone customers in our dataset to the census distribution in the population of the analyzed country. The differences between the two are very small, so there is no systematic bias in our data as regarding this characteristic (except for people over 55 who are underrepresented among mobile phone users). The following analysis has been done on customers aged 18–55 thus the older age groups under-representation in the mobile phone dataset does not affect our results.

First, we computed some statistics of mobile phone usage by age. The age seems an interesting variable here as different generations of people began to use the mobile phone at different ages. As the mobile phone diffusion started in the mid-1990s, there is only the nowadays youngest part of population who entered in their "communication age" directly with a cell phone at hand. We thus expected a different usage of the mobile phone between age groups. Figure 7.5 shows the average number of out-going calls and SMS by age during the studied month. We observe no important difference in the number of calls by age. The main distinction concerns SMS usage: younger users send more SMS than older ones. In the age group 18–25 this tendency is really impressive: the SMS is used four times more frequently than a conversational exchange. This means that to recognize a younger customer in the dataset, one can look at the number of SMS sent. Figure 7.6 shows the average of the total duration of calls by age. We observe that people from 22 to 34 have in average the greatest total durations (these are out-going calls, so the age is that of the caller); for older people, the total duration of calls decreases with the age.

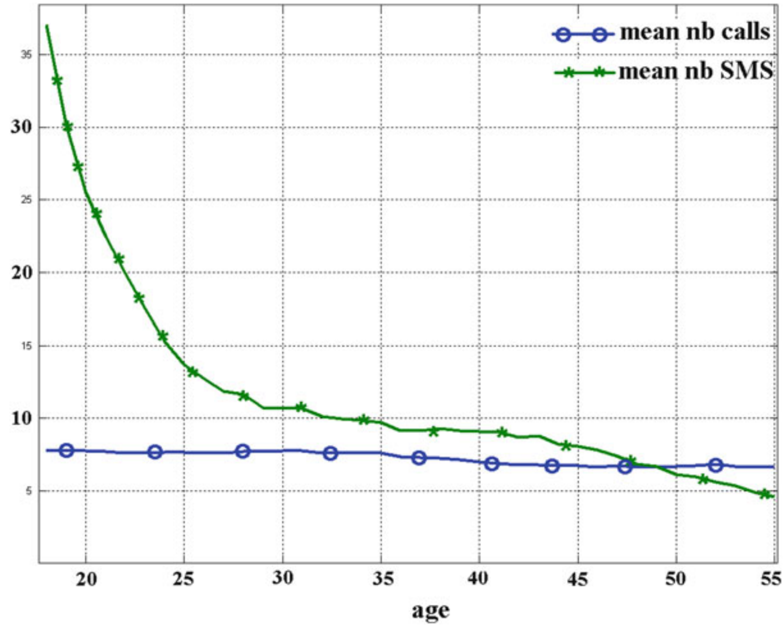

Fig. 7.5 Average number of calls and SMS as a function of phone user's age

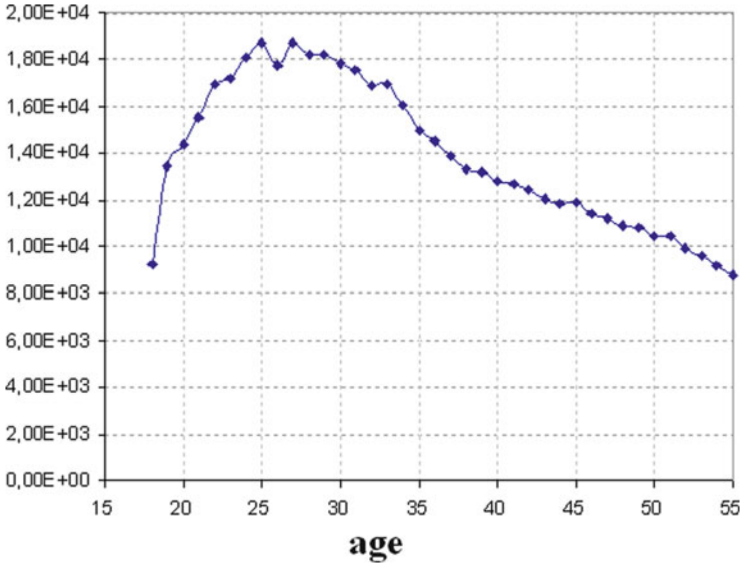

Fig. 7.6 Average total duration of calls (in seconds) as a function of phone user's age

While these measures represent a first analysis of the mobile phone communication data, our purpose is to study the social network modeling this data. The remaining part of this paper deals with the clustering of individuals using the social network structures and with the correlation between clusters and intensity of communication or age.

7.7 Personal Network Clusters

We model the mobile phone communications set by a simple undirected graph G. In this graph the vertices are the customers; we connect such two vertices by an undirected link if there had been at least one communication in each direction between the two persons during the followed period. This way we do not take into consideration the one-way contacts (calls or messages), single events in most of the cases suggesting that the two individuals do not know each other personally. We keep only the vertices with degree greater than 0, thus obtaining a graph G with 2.7×10^6 vertices and 6.4×10^6 edges.

In this graph, we apply the method *characterize* introduced in Sect. 7.4 in order to analyze how each one of the $2.7M$ vertices is connected to the network. Next, we apply the method introduced in Sect. 7.5 in order to group the nodes into clusters based on their network insertion.

The best number of clusters is found to be 6. Figure 7.7 represents the distribution into clusters of the egocentred networks of vertices with degrees 4 (up) and 5 (bottom). In our graph, all the possible egocentred networks for these degrees are present; these are all the possible undirected graphs with four and five vertices respectively. For each network, we have marked the cluster to which it belongs.

We observe that cluster 1 contains dense networks, while cluster 6 contains very sparse networks. Networks in cluster 2 seem to have a high number of stars, while those in cluster 5 have both isolated vertices and a rather dense group. For clusters 3 and 4 we can say that networks in cluster 3 are denser than those in cluster 4. These observations have been made by simply analyzing the clusters obtained for degrees 4 and 5. When looking at the centers of the clusters obtained for the different degrees, we observe that, for all degree:

- The center of cluster 1 has the maximal value for the number of edges and for the number of triangles i.e. vertices in cluster 1 have the highest average of f_{-} and of $f_{\triangle}$;
- The opposite situation happens for cluster 6: the center of this cluster has the minimal value for the number of edges and for the number of triangles i.e. vertices in cluster 6 have the lowest average of f_{-} and of $f_{\triangle}$;
- From the remaining clusters, the center of cluster 5 has the maximal value for the number of isolated vertices multiplied by the number of edges i.e. vertices in cluster 5 have the highest average of $f_{iv} \times f_{-}$;

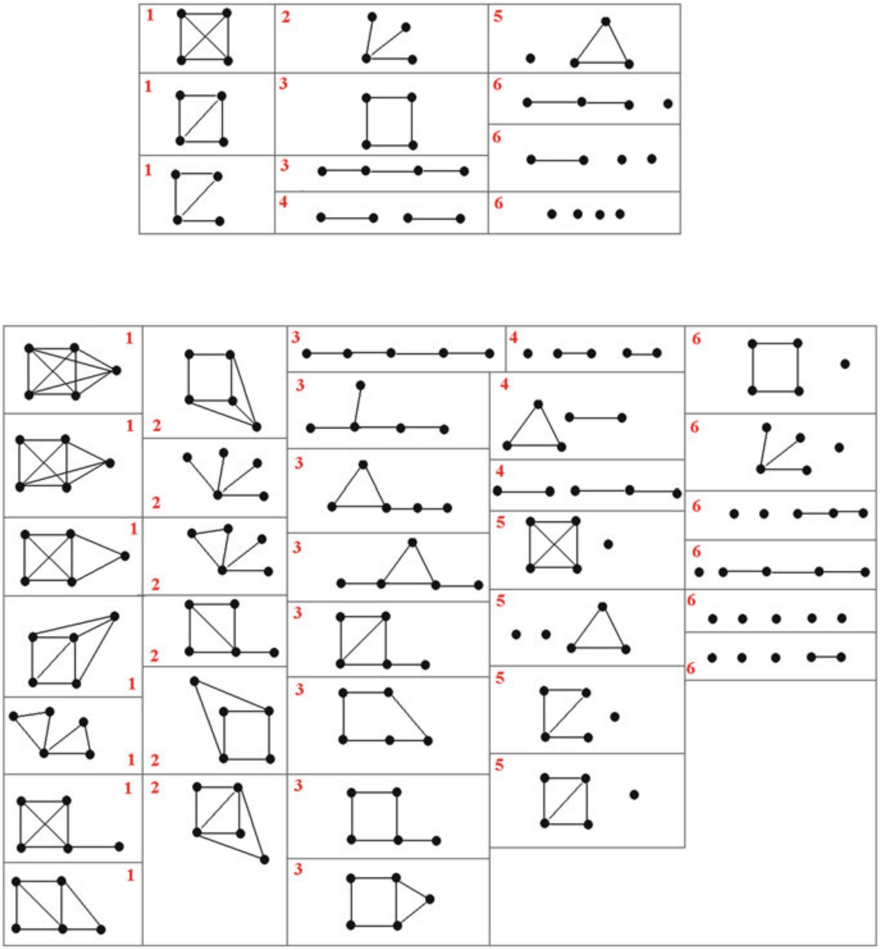

Fig. 7.7 All the possible egocentred networks of vertices with degrees 4 (*up*) and 5 (*bottom*) and their clusters

- The center of cluster 2 has the maximal value for the number of stars i.e. vertices in cluster 2 have the highest average of $f_{\angle}$;
- From the remaining two clusters, the center of cluster 3 has a higher value for the number of edges than the center of cluster 4 i.e. vertices in cluster 3 have a higher average of f_{-} than vertices in cluster 4.

This sustains our previously made observations for degrees 4 and 5: cluster 1 contains the densest networks, while cluster 6 contains the sparsest ones. Networks in cluster 2 have many stars, while those in cluster 5 have both isolated vertices and a dense group. Finally, networks in cluster 3 are denser than those in cluster 4.

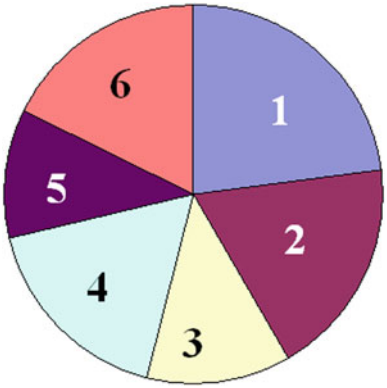

Fig. 7.8 The distribution of the reduced population into the six clusters

Table 7.2 The distribution of the reduced and total population into the six clusters

Cluster	% of the reduced population	% of the total population
1	23.16	4.15
2	18.6	2.91
3	12.24	2.54
4	17.05	26.93
5	11.12	5.04
6	17.83	58.43

Remember that before computing the clusters we have eliminated the multiple copies of pattern-frequency vectors. It is in this reduced population that we have computed the six clusters. The different resulting clusters contain fairly similar percentages of the reduced population (see Fig. 7.8 and Table 7.2).

However, when reintroducing the filtered out vertices, the population is not equally divided into clusters any more. This is caused by the low local density of the graph: most vertices have very sparse egocentred networks, so the different patterns occur in their networks in a small number. Thus the majority of the eliminated vertices belongs to cluster 6. After the introduction of the previously filtered out vertices, the new repartition into clusters becomes very unbalanced (Table 7.2).

In the following sections we confront the identified clusters to other characteristics of the mobile phone customers.

7.7.1 Clusters Versus Age

For the mobile phone customers who have provided their birth year when subscribing to the studied operator, we want to see if there is a connection between the age of a person and his or her cluster. Remember that in Sect. 7.6 we have presented some statistics on mobile phone use. There are some differences in call frequency and duration between ages, but the main distinction concerns SMS usage,

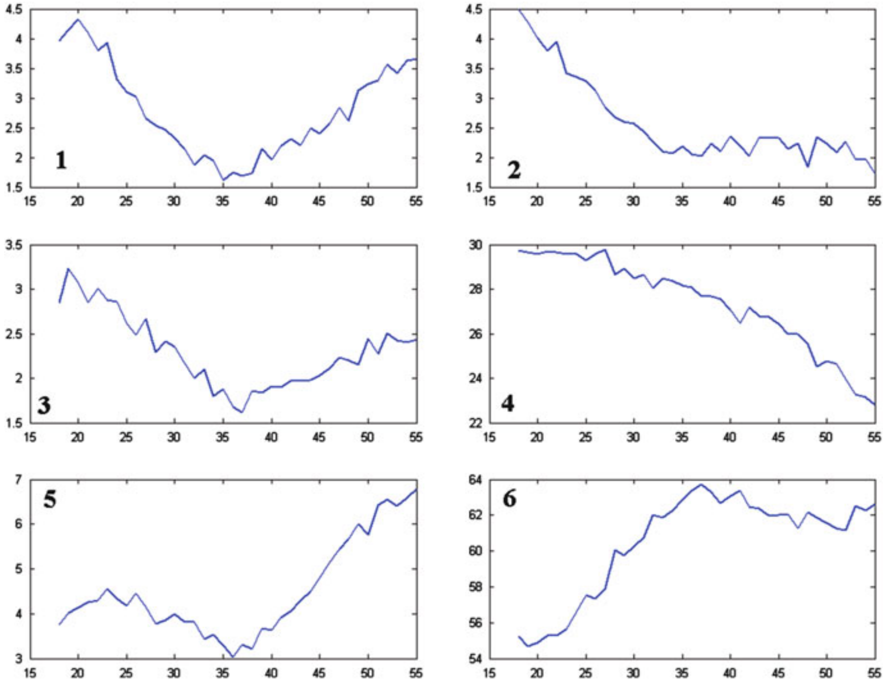

Fig. 7.9 For each cluster (each image), the probability of belonging to that cluster by age (on x-axis)

the younger users sending a lot more SMS than the older ones. Here we want to see if these differences in mobile phone uses are visible in the structure of the network surrounding each person.

We compute, for each cluster k from 1 to 6 and for each age a from 18 to 55,[3] the probability that a person of age a who has at least four contacts belongs to cluster k :

$$P(a,k) = \frac{\text{nb. persons of age } a \text{ and cluster } k}{\text{nb. persons of age } a \text{ and degree} > 3}$$

The plot of these probabilities is presented in Fig. 7.9. We observe that middle age people (30–45) have the lowest probability of belonging to cluster 1, so generally they are not involved in dense structures. This can be seen also in the plot for cluster 6 (the cluster containing the sparsest networks), where there is a peak for 35–40. Younger people belong generally to clusters 2, 3 and 4 and rarely to cluster 6 (in any case, a lot less frequently than older people). The oldest people are generally

[3] 18 is the minimal age to have a mobile phone subscription, while for persons of more than 55 years old, 70 % of them have degree smaller than 4.

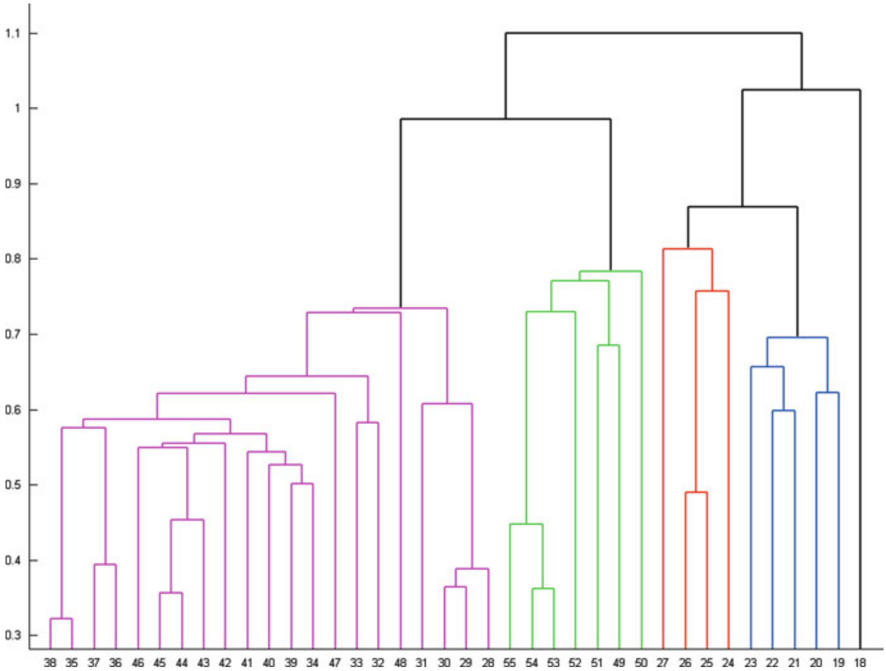

Fig. 7.10 Hierarchical clustering of ages on probabilities of belonging to the six clusters

placed in cluster 5: there is an increasing probability of having a densely connected group and some isolated contacts when going from 40 years old to 55.

Let us now group together the ages that have similar probabilities for the six clusters. We perform a hierarchical clustering on the ages using the cluster probabilities previously computed, after having centered and scaled the probabilities so that they have the same mean and standard deviation for each profile. The result of this analysis is shown in Fig. 7.10. We observe that there are four principal, homogeneous age groups similar to life stages categories: 19–23 (who can be associated with “students”), 24–27 (young people starting their active life), 28–48 (the age of living in couple, often with children), and 49–55 (people at an advanced stage of the professional life, whose children are adult or living apart). Note that this classification is based exclusively on structural characteristics of the local communication network where the degree was neutralized.

To sum up, there are some differences in the mobile phone usage and in the network structure depending on the age that allow recovering homogeneous age groups from mobile phone data. The personal network structure changes with age and the communication practices follow this transformation. Our analysis shows that the particular ways of interconnection to a personal network are in fact markers of age-related sociability of the mobile phone user.

7.7.2 Clusters Versus Intensity of Communication

7.7.2.1 Basic Statistics

We compute for each person (ego) the total number of calls he or she had during the followed period (both in-coming and out-going calls), the total duration of the calls and the total number of SMS (similarly, in-coming and out-going SMS). Also, we compute the average number of calls, total duration and number of SMS he or she had with each one of the contacts. We limit the contacts to the persons who initiated at least one communication (call or SMS) with ego and who also received at least one call or SMS from ego; these persons correspond to ego's neighbors in our graph. Besides the average values, we also compute the standard deviation for the number of calls, the duration and the number of SMS per contact. We thus have for each ego a vector with nine variables characterizing ego's communications. We use these vectors to measure the relation between communication intensity and the previously obtained clusters.

We begin by testing, for each one of the nine variables, the independence of the variable and the clusters by performing an ANOVA test: we test the hypothesis that the mean value of the variable is the same for the different clusters. As the distributions for the nine components are heavily right-skewed, we use the log values instead of the real ones. The ANOVA test rejects the hypothesis of equal means for each one of the components with $p = 0$. However, the ANOVA test specifies only that the means are different (i.e. they are not all equal) but does not say for which pairs of clusters these means are significantly different and for which they are not. In order to find this information, we perform a Bonferroni multi-comparison test for each one of the nine variables. We thus have:

- For the total number of calls, all the means are significantly different, except for the clusters 1 and 2; the order of the mean values of the total number of calls for the six clusters is, from low to high: 6, 4, 5, 3, 2, 1;
- Similarly, for the total duration of calls and the total number of SMS, all the means are significantly different, except for the clusters 1 and 2; in this case the order is 6, 5, 4, 3, 1, 2;
- Very similar results are obtained for the other variables; the ascending order of the values is always 6, 4, 5, 3, 2, 1, maybe with an interchange of 4 and 5 and of 1 and 2; the average duration of calls per contact is the only variable for which there isn't a significant difference between the mean values for the six clusters.

So, for each one of the nine components, cluster 6 has the lowest mean, followed by clusters 5 and 4 (or 4 and 5), cluster 3 and finally 2 and 1 (or 1 and 2). However, using the mean values is not satisfying as the different variables have a right-skewed distribution. Therefore, for each variable, we compute ten slices as we did in Sect. 7.5: we divide its spectrum of values into ten slices or bins such that a fairly equal number of values belong to each one of the bins. Then, we compute the probability that an individual belonging to a given cluster has values in a certain bin:

$$P(\text{variable, cluster, bin}) = \frac{\#\text{individuals} \in \text{cluster s.t. value(variable)} \in \text{bin}}{\#\text{individuals} \in \text{cluster.}}$$

We plot these probabilities for the first three variables in Fig. 7.11: the number of calls in (a), the total duration of calls in (b) and the number of SMS in (c). Each bar corresponds to a bin, going from the bin with the lowest values (left side) to the bin with the highest ones (right side). For each cluster, the height of each bin represents the previously computed probability i.e. the probability that an individual in that cluster has values in that bin; the sum of heights of bins of one cluster is thus equal to 1. For the three variables, individuals in clusters 1, 2 and 3 have a greater probability to have values in the highest bins than in the lowest ones, while for cluster 6 the opposite situation happens. Cluster 4 has values especially in the intermediate bins, while cluster 5 has values both in high and low bins, but fewer in the intermediate ones.

7.7.2.2 Predicting the Cluster from the Communications

Given these differences in quantity of communications for the different clusters, we want to see if we can guess in which cluster an individual is placed given his or her communications. For that, we use a decision tree to unveil the relation between communication intensity and cluster and thus to predict the cluster of each individual. The explanatory variables are the nine characterizing the communications of an individual: the number of calls, the total duration of calls, the number of SMS, the average number of calls, duration and number of SMS per contact, and the standard deviation of the number of calls, duration and number of SMS per contact. Based on the learning population, the tree learns the associations between intensity of communication and cluster; then it predicts the cluster of the individuals in the test population. If the predicted cluster is the same with the real cluster of the person, then the prediction is correct; otherwise the prediction is false. To measure the accuracy of the tree, one counts the correct predictions as compared to the size of the test population: the higher this number, the better the prediction. This number is then compared to the random prediction, where one attributes individuals into clusters randomly, with an equal probability.

Remember that the number of individuals in the six clusters is very uneven, with cluster 6 over-represented. If the decision tree learns and tests its rules of association on populations with such uneven distribution of clusters, it will associate everybody with cluster 6: no matter the communication characteristics of the different persons, if everybody is put in cluster 6, the tree gives the correct class to all the individuals in cluster 6 and the wrong cluster to all the others. As the individuals in cluster 6 are much more numerous then the others, the tree has a high rate of success. We want to avoid this situation and impose to the tree to search for associations between communications and clusters. Therefore, we give it a learning population where there is an equal number of individuals belonging to each cluster; the individuals are randomly chosen from the individuals in each cluster. We do the same thing

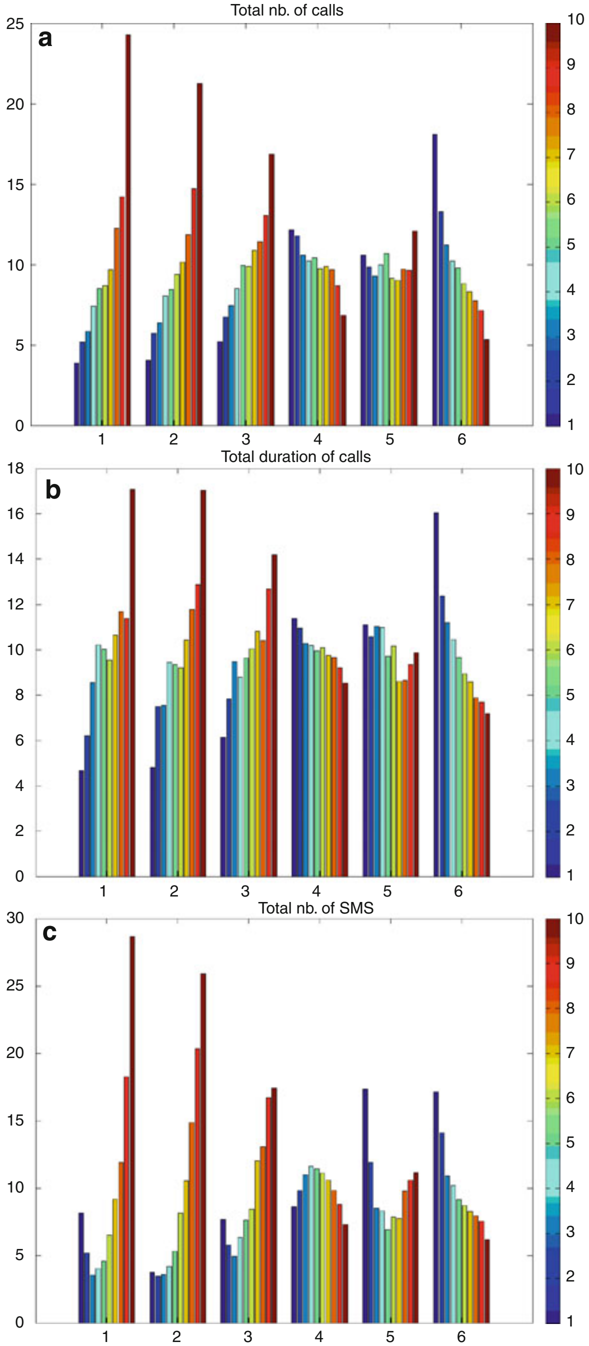

Fig. 7.11 For each cluster (Ox-axis), the probability that the communications of an individual in that cluster are in a given slice of values of the number of calls (*a*), total duration of calls (*b*) and number of SMS (*c*). In each image, the ten slices for each cluster are grouped together, with the lowest values in the *left side*

Table 7.3 The proportion of correct predictions in the six clusters

Cluster	Rate of success (%)
1	31.2
2	22.6
3	24.3
4	40.4
5	51.8
6	37.1

for the test population. As we want to predict six clusters, the rate of success of the random prediction is $\frac{100}{6} = 16.66\,\%$. Our decision tree has a rate of success of 34.6 %, so more than twice the random one. The rate of correct predictions in the different clusters is presented in Table 7.3.

This result shows that there is a correlation between the intensity of communication and the cluster to which an individual belongs. Even more, we are able to predict the cluster with a rather high accuracy (as compared to the random prediction) given a set of variables characterizing the communications of each person.

7.7.3 A Typology of Customers

In the previous two sections we compared the social network clusters first to customers' age and then to their communication intensity, observing that the probability that an individual belongs to a given cluster is not independent from these measures.

Here we want to take into consideration, in the same time, all the three dimensions characterizing the individuals: the age, the communication intensity and the social network cluster. We want to see how these characteristics are distributed in the population and also to create a typology of customers based on these three dimensions. We would thus obtain groups of individuals such that the persons in a same group have similar communication practices and about the same age and cluster.

We use the Kohonen self organizing map. Remember that this clustering method produces a map with several layers, one for each variable characterizing the individuals. This shows how the different variables are distributed in the population. Also, the algorithm produces cells grouping individuals with close characteristics. In a second step, the algorithm computes a clustering of the individuals. The obtained clustering will represent our typology.

We choose the following parameters to characterize the individuals:

- Thier age (socio-demographic variable);
- Cluster membership (social network variable, from 1 to 6, as obtained in the previous sections); as it takes only six values, this variable can be seen as a class or a label of each individual;
- Communication intensity: number of calls, total duration of calls and number of SMS; (three communication variables).

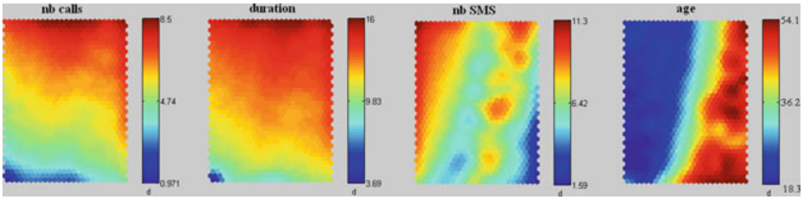

Fig. 7.12 SOM results: the individuals are grouped into cells depending on their communication intensity and age

Each individual is thus characterized by a vector with five elements. For the communication variables, we use a log transformation instead of the values themselves as these variables are heavily right-skewed. Also, recall that the distribution of individuals into clusters is very uneven, with cluster 6 being overrepresented. As we want to measure the influence of the variable “cluster”, too, we randomly choose a same number of individuals in each cluster.

The set of individuals is then processed by the Kohonen self organizing map. This algorithm does not take labels into consideration when building the map, so it builds the map using only the other variables.

The processing of the set of individuals by the Self Organizing Map (SOM) provides Fig. 7.12. We observe that, unsurprisingly, the number of calls and the total duration are highly correlated, with increasing values on the south-north axis: the individuals with the lowest number of calls and total duration are placed in the south part of the map, while those with the highest values are placed in the north part. The number of SMS, however, is not correlated to the two previous ones, its values increasing from east to west. This variable seems to be correlated to the age: the highest values of the number of SMS are in the west part, where the youngest people are placed, while the lowest values are placed in the east part, where the oldest persons are placed. All these observations sustain our previous ones, presented in Sect. 7.6: there is no influence of the age on the call frequency and duration, but there is a high influence on the number of SMS.

Let us now analyze the distribution of the variable “cluster” in the different cells. Figure 7.13 shows this distribution, cluster by cluster. Each image in the figure corresponds to a cluster: the dark cells contain mostly individuals of the given cluster, while the white cells contain mostly individuals of other clusters. Recall that the different clusters are not taken into consideration when building the map; the cells are colored depending on the clusters of the people present in the cell, after all the computations. We observe that clusters 1, 2 and 3 are present especially in the north-west side of the map, while clusters 4, 5 and 6 are placed especially in the south-east side. Most of the cells labeled cluster 1 contain individuals with very high number of SMS or very high number of calls and total duration. Cluster 2 is generally associated with cells containing individuals with a high number of SMS or a high number of calls and total duration. Clusters 3 and 4 are generally present in cells where the individuals have a medium number of calls, total duration and

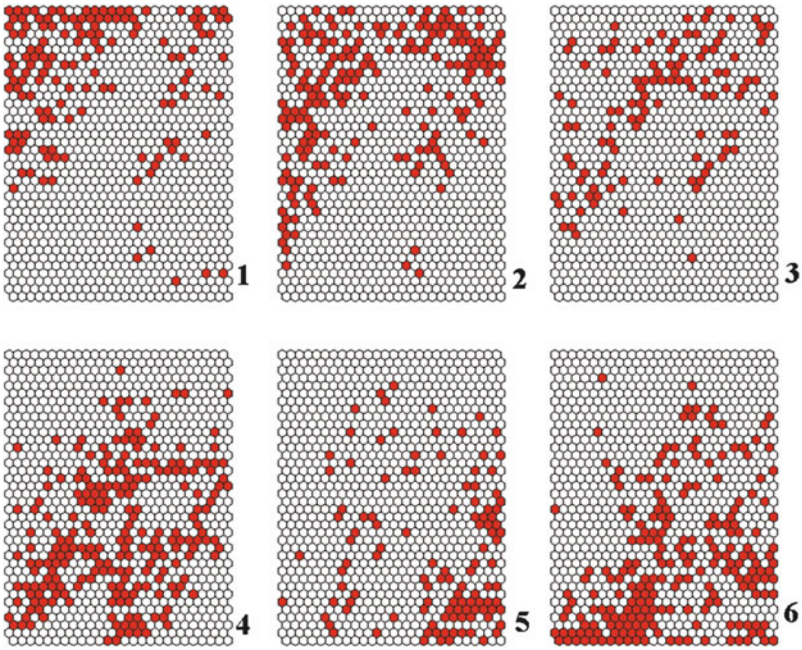

Fig. 7.13 For each cluster (each image), the cells where the cluster is in the majority (the *dark cells*)

number of SMS. Cluster 6 is especially placed in the south-east part of the area, where there are individuals with low number of calls, total duration and number of SMS. There seems to be no clear relation between the label of the cell and the average age of the persons in the cell, except for cluster 5 which is present especially in the cells containing the oldest people.

We now cluster the cells using the k-means algorithm. We thus obtain nine profiles, as showed in Fig. 7.14. We present the different characteristics of the people with each profile in Table 7.4. This result represents a typology of individuals based on their age, communication intensity and social network cluster.

7.8 Conclusions

In this paper we presented a method for clustering nodes, thus relating to the problem of identification of roles in a network. In this problem often encountered in social network analysis, one wants to group together the nodes of the network that are connected in similar ways to the network. There are however several questions that make this problem difficult to solve: What is a good characterization of the way a node is connected to the network? What does "similar connections" mean?

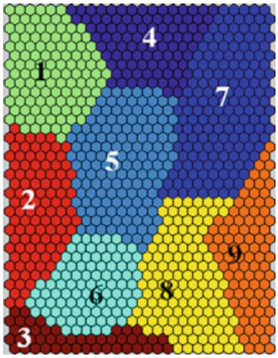

Fig. 7.14 The nine profiles produced by the Kohonen SOM

Table 7.4 The different characteristics of the individuals in the nine profiles produced by the SOM

Profile	Age	Nb. calls and duration	nb. SMS	Most represented cluster(s)
1	Youngest	High	Very high	1(45 %), 2(41 %)
2	Youngest	Medium	High	2(38 %), 3(20 %)
3	Young-middle	Very low	Low	6(70 %)
4	Young-middle	Very high	Medium	1(31 %), 2(31 %)
5	Young-middle	Medium-high	Low	4(39 %), 6(24 %)
6	Young-middle	Low	Low	4(45 %), 6(43 %)
7	Oldest	High	High	2(29 %), 1(19 %)
8	Oldest	Low	Low	4(34 %), 6(29 %)
9	Oldest	Low	Very low	5(42 %), 6(35 %)

Can the solution be applied to large graphs? How can one check the relevance of the different groups of nodes? In which conditions can one say that there is no better way of grouping the nodes?

We have made several choices in order to answer the different questions. First, we have characterized the way a node is connected to the network by counting the patterns present in its egocentred network; we have stored the number of occurrences of the different patterns in a pattern-frequency vector characterizing the node. Second, we have considered that nodes connected in a similar way to the network have close pattern-frequency vectors; here “close” is defined with respect to a set of transformations made on the pattern-frequency vectors. We have thus proposed a method for nodes clustering that groups together vertices that are embedded in similar egocentred networks. The clustering is done efficiently, so the method can be applied to large graphs. As said before, we have made several choices in order to answer the different questions. The proposed method gives promising results when applied to our real-world graph. As always, in this kind of methods, the solution validation is a delicate problem, but the results we have obtained for our large social network sustain the relevance of our method.

We have applied the proposed method to a mobile phone graph. This graph models 1-month mobile phone communications between the three million individuals.

The clusters produced by the method can be seen as a segmentation of the set of customers based on their social network insertions. We have compared the different clusters to the other information we had on the individuals (age and communication intensity), showing that the different parameters characterizing the individuals are not independent. Thus, the probability that a node belongs to a given cluster is not independent from the age and the mobile phone use of the person represented by the node. These results confirm the soundness of our method, even though, as always, many concurrent clusterings for various purposes may as well be relevant.

References

1. Batagelj, V.: Notes on blockmodeling. Soc. Netw. **19**, 143–155 (1997)
2. Batagelj, V., Ferligoj, A., Doreian, P.: Direct and indirect methods for structural equivalence. Soc. Netw. **14**, 63–90 (1992)
3. Borgatti, S.P., Everett, M.G.: The class of all regular equivalences: algebraic structure and computation. Soc. Netw. **11**(1), 65–88 (1989)
4. Burt, R.S.: Detecting role equivalence. Soc. Netw. **12**, 83–97 (1990)
5. Doreian, P., Batagelj, V., Ferligoj, A.: Generalized Blockmodeling. Cambridge University Press, Cambridge, England (2005)
6. Dunnett, C.W.: A multiple comparisons procedure for comparing several treatments with a control. J. Am. Stat. Assoc. **50**, 1096–1121 (1955)
7. Hidalgo, C.A., Rodriguez-Sickert, C.: The dynamics of a mobile phone network. Phys. A Stat. Mech. Appl. **387**(12), 3017–3024 (2008)
8. Hummell, H., Sodeur, W.: Strukturbeschreibung von positionen in sozialen beziehungsnetzen. In: Pappi, F.U. (ed.) Methoden der Netzwerkanalayse. Oldenbourg, Munich (1987)
9. Kohonen, T.: The self-organizing map. Proc. IEEE **78**(9), 1464–1480 (1990)
10. Lambiotte, R., Blondel, V.D., de Kerchove, C., Huens, E., Prieur, C., Smoreda, Z., Van Dooren, P.: Geographical dispersal of mobile communication networks. Phys. A **387**(21), 5317–5325 (2008)
11. Lorrain, F., White, H.: Structural equivalence of individuals in social networks. J. Math. Sociol. **1**, 49–80 (1971)
12. Onnela, J.P., Saramäki, J., Hyvönen, J., Szabó, G., de Menezes, A.M., Kaski, K., Barabási, A.L., Kertész, J.: Analysis of a large-scale weighted network of one-to-one human communication. New J. Phys. **9**(6), 179+ (2007)
13. Onnela, J.P., Saramäki, J., Hyvönen, J., Szabó, G., Lazer, D., Kaski, K., Kertész, J., Barabási, A.L.: Structure and tie strengths in mobile communication networks. Proc. Natl. Acad. Sci. **104**(18), 7332–7336 (2007)
14. Palla, G., Barabasi, A.-L., Vicsek, T.: Quantifying social group evolution. Nature **446**(7136), 664–667 (2007)
15. Reichardt, J., White, D.R.: Role models for complex networks. EPJ Manuscr. **60**, 217–224 (2007)
16. Scheffé, H.: The Analysis of Variance. Wiley, New York (1959)
17. Stoica, A., Couronné, T., Beuscart, J.S.: To be a star is not only metaphoric: from popularity to social linkage. In: Proceedings of the 4th International AAAI Conference on Weblogs and Social Media (ICWSM). AAAI, Washington, DC (2010)
18. White, D., Reitz, K.: Graph and semigroup homomorphisms on netwoks and relations. Soc. Netw. **5**, 193–234 (1983)