Knowledge Discovery in the Virtual Social Network Due to Common Knowledge of Proverbs

Armando B Mendes, Matthias Funk, *Department of Mathematics, University of the Azores, 9501-801 Ponta Delgada (telephone: 351296650073, fax: 351296650072, {amendes, mfunk}@uac.pt)*
Luís Cavique, *Department of Informatics, Open University, Portugal, (lcavique@univ-ab.pt)*

***Abstract*— In a series of interviews, it was collected a heterogeneous set of several million relations of positive and negative knowledge that a group of thousands of people has about a set of circa twenty-two thousand Portuguese Proverbs. This is a unique source for socio-cultural analysis of the mechanisms of transmission of oral culture in geographic discontinuous spaces. We present in this article some results on the problem of finding a homomorphism between proverbial knowledge and geographical locations.**

To find this relation, we chose an approach based on the Analysis of social networks where the broadcast of oral culture, at least historically, could be interpreted as a trace of direct social contact between some of their users. We can simply give the Hamming Distance between two people by comparing their proverbial knowledge and, then, choose for every person only those relations to a peer where this distance is minimal. The resulting graph is analysed by a new Clique Analysis procedure, proposed in this work, design to work on very dense networks.

The procedure was tested on a subset of data and we found that there are clusters where the neighbourhood relation inducted by the minimum Hamming Distance could be a reflex of the geographical distribution and of some migration flux of the Azorean population. When we compare the cliques with high geographic proximity, we found some proverbs which are good discriminators between the different clusters.

Keywords: Proverbs; Social Network; Clique Analysis. Positional Analysis.

I. INTRODUCTION

This case study is based on data collected in 11 locally disconnected areas inside the cultural space of the Azorean community. This community is centred on the Portuguese archipelago located on the middle Atlantic rift. The specific situation of the Azores is very interesting to analyse the balance between local and global knowledge inside a common linguistic and cultural space. On one side, there is the geographical dispersion and the isolation imposed by the natural barrier of the sea in an archipelago formed by 9 populated islands over its 2,330 km^2 area wide, spreading over a rectangle of 630 Km in the West-East and 130 km in the North-West extension. This most important neighbourhood relationship is present in the aggregation of the islands in three geographical groups, as can be seen in Fig. 1.

On the other side, we find a geographical continuity in the central group that is composed by 5 islands (Faial [Acronym: Fai/Habitants: 15,063], Pico [Pic/14,806], S. Jorge [SJo/9,674], Terceira [Ter/55,823] and Graciosa [Gra/4,780]) which are close to each other. The same happens with the two western islands, Corvo [Cor/425] and Flores [Flo/3,995]. But, only in exceptional climatic conditions the view penetrates the 80 km between Sa. Maria [SMa/5,578] and S. Miguel [SMi/131,609].

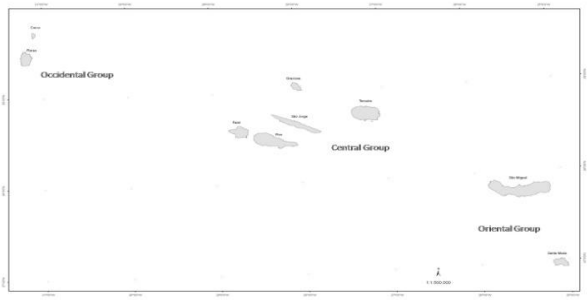


Fig. 1. The Azores Archipelago
showing the space dispersion of the nine islands

Due to big waves of emigration into the USA, which took place since the end of the 19th century until the end of the 20th century, the group of emigrated people is two times bigger than the resident population on the archipelago with circa 250,000 habitants. The population flux includes also the inner-Azorean migration which is mainly characterized by the attraction to urban centres which are the former administrative capitals: Ponta Delgada, in S. Miguel, Angra, in Terceira and Horta, in Faial.

The survey data was organized in a relational data base, which centralizes all information about the recognition of proverbs. This exhaustive survey was taken to clarify the relevance of over 22,000 Portuguese proverbs in the Azorean society and to calculate the recognition percentage of this proverbs inside the cultural (sub-)space of the community as a whole or of the single location. The data collected between 1997 and 2000 was analysed in [1]-[2]. The same authors in [3], analysed the properties of the central geographic group. This data base was also restructured, cleaned and statistically analysed, using simple description statistics, hypothesis testing and cluster analysis, as described in a previous publication [4].

For practical reasons, it was impossible to submit the 22,000 proverbs to a single person. Therefore, we divided the corpus into 14 packages with circa 1500 proverbial phrases each. These corpora were submitted, in each location, to, at least, 10 and, at most, 40 interviewees with preferentially more than 40 years old. Some interviewees were later excluded because there was inconsistency in their responses.

We utilize one of these packages where we can see that in small islands like Corvo, Sa. Maria, Graciosa, Flores and S. Jorge the population shows low mobility as the larger majority lived their whole life in only one island. The islands with local capitals are characterised by a high mobility, probably because of their high attraction for the population of the surrounding islands. A very strong relation can be verified between Faial and Pico where 61% and 72% respectively of the mobile interviewees were also related with the other island. We call mobile persons the ones which have lived at least in 2 different islands or locations outside Azorean archipelago for at least 5 years in each other.

The mobility of 22 of the 32 mobile people associated to S. Miguel is caused by the (time wise) emigration to New England. It is clear that the mobility in emigration zones, namely New England and California, is connected primarily to islands of origin which are, in the first case, mainly S. Miguel and, in the second case, mainly the islands of the central group.

The objective of the analysis proposed is to delineate a procedure capable to confirm or reject the working hypothesis that is possible to identify the geographical origin of people if they know (or not) a group of core proverbs. In the following paragraphs the procedure is described using only a small data set. The same procedure is briefly presented in the following illustration.

Given a set of individuals I
Given a set of proverbs P with cardinality *n*
Given a binary function f(I, P) $\in$ {0, 1}

Build the relational table T (I, P) $\in$ {0, 1} where each line is the individual DNA

Build the symmetrical matrix of weights M(I,I) = n-h_{ij} by calculating the Hamming Distance h_{ij} of each pair of lines
Interpreting M as an incidence matrix we get a complete network where the links have weighs n- h_{ij}

For each line *i* identify the maximum weight
Calculate M' by deleting all non maximum entries of matrix M
Interpreting M' as an incidence matrix we get an directed network
Ranking the nodes by the weights of the links they connect we get a set of directed networks.
All paths in these networks end in a node with the higher rank. We call this nodes sinks.
Some sinks of different networks are connected by undirected links
Identify as a clique any cluster of sinks and confirm the hypothesis for the individuals in this hierarchical cliques.

Fig. 2. Outline of the procedure described in the next paragraphs.

In section II a measure for hierarchical networks is presented. In section III and IV real data is analysed and in section V we draw the conclusions.

II. From a Measure of Cultural Homogeneity to Hierarchical Networks

First, we need a measure for the proximity between all pairs of two persons due to their knowledge of the 40 selected proverbs. This is done by ordering the proverbs in a line of attributes for every person, which results in a 40 row vector with an entrance of 1 or 0 when the person knows or ignores the respective proverb. We call this vector the proverbial DNA of an interviewee. The following Fig. 3 shows the proverbial DNA for a well chosen sample of 20 interviewees.

The cultural proximity was defined by the Hamming distance of their DNA, as is exemplified in Figure 3 where the DNA of two persons is matched row by row in order to find their proverbial divergence. Whereas we have 10 cases of mismatches (marked in Fig. 4 by a grey background), we found the Hamming Distance of 10 between the respective people.

573 (alias 1) → 00110 10010 11001 00000 00100 00011 01010 01010
575 (alias 2) → 00011 10010 11001 00001 00101 10011 01001 10011
581 (alias 3) → 00000 00010 01010 10010 01100 01000 11000 11100
583 (alias 4) → 01000 00000 01010 00000 00000 00000 00000 00001
587 (alias 5) → 00000 00000 01010 00000 00000 10000 00000 00001
795 (alias 6) → 00010 00000 10000 00000 01100 00011 10000 01101
796 (alias 7) → 00010 00000 10000 00000 01100 00011 10000 01001
797 (alias 8) → 00000 00000 10001 00000 00100 00011 00010 01101
798 (alias 9) → 00010 00000 10000 00000 01101 00011 11000 01101
799 (alias 10) → 00000 00000 10000 00000 00100 00011 10100 01101
800 (alias 11) → 00000 00000 10000 00000 01100 00010 10000 01101
801 (alias 12) → 00000 00000 10000 00000 00100 00011 10000 01101
802 (alias 13) → 00010 00000 10000 00000 01100 00011 10000 01101
803 (alias 14) → 00000 00000 10000 00000 01100 00011 10000 01101
804 (alias 15) → 00010 00000 00000 00000 00100 00001 10000 01111
850 (alias 16) → 00000 00100 10100 01000 01001 00010 01010 10101
882 (alias 17) → 00100 10101 01000 00001 00100 10011 01110 00001
905 (alias 18) → 00000 00101 10101 01111 00010 00010 00011 10101
907 (alias 19) → 01001 00101 11010 10010 00111 01010 00000 00001
916 (alias 20) → 01010 01001 11010 10100 00100 00111 01100 00101

Fig. 3. A small sample of f(I, P) or Proverbial DNA for 20 people.

person 1 → 00110 10010 11001 00000 00100 00011 01010 01010
person 2 → 00011 10010 11001 00001 00101 10011 01001 10011

Fig. 4. Comparison of two proverbial DN to get the Hamming distance of 10 (score=30).

For *n* people, we can find a symmetric *n* times *n* matrix with the entry of the respective score (that is the difference between the dimension of the DNA-Vector and the Hamming Distance), as is exemplified in Fig. 5 for the sample of Fig. 3. We see that this matrix can be read as an incidence matrix of a totally connected, undirected, and weighted graph with *n* nodes. This representation would be of low usefulness due to excess of information, so some simplification is in order. In the Fig. 5, we mark with a grey background the line-maximums of the non-reflexive relations as the best cultural peer for the line element.

If we erase all non-marked entries from this matrix, the correspondent graph would be a directed and weighted graph. By doing the described procedure for the whole sample of the 221 interviewees, we get a graph with 8 isolated sub-graphs, presented in the appendix.

Person	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	x	30	23	24	24	28	29	31	28	27	26	28	28	27	28	22	28	18	19	23
2	30	x	19	22	24	24	25	25	27	23	22	24	24	23	24	22	20	20	22	21
3	23	19	x	27	27	27	26	24	28	26	29	27	27	28	27	23	19	18	24	22
4	24	22	27	x	38	28	29	29	26	29	30	30	28	29	31	26	26	22	30	27
5	24	24	27	38	x	28	29	30	26	28	30	29	28	29	30	28	28	22	27	25
6	28	24	27	28	28	x	39	35	38	37	38	37	40	39	36	28	24	22	23	27
7	29	25	26	29	29	39	x	35	37	36	37	37	39	38	35	29	25	21	24	26
8	31	25	24	29	30	35	35	x	33	36	35	37	35	36	33	29	27	27	24	26
9	28	27	28	26	26	38	37	33	x	35	36	37	38	37	35	30	24	21	23	27
10	27	23	26	29	28	37	36	36	35	x	37	39	37	38	35	27	27	24	25	28
11	26	22	29	30	30	38	37	35	36	37	x	38	38	39	35	30	24	24	27	25
12	28	24	27	30	29	37	37	37	37	39	38	x	37	39	36	28	26	24	26	27
13	28	24	27	28	28	40	39	35	38	37	38	37	x	39	36	28	24	22	23	27
14	27	23	28	29	29	39	38	36	37	38	39	39	39	x	35	29	25	23	24	26
15	28	24	27	31	30	36	35	33	35	35	35	36	36	35	x	25	24	20	23	25
16	22	22	23	26	28	28	29	29	30	27	30	28	28	29	25	x	24	30	23	21
17	28	20	19	26	28	24	25	27	24	27	24	26	24	25	24	24	x	25	24	25
18	18	20	18	22	22	22	21	27	21	24	24	24	22	23	20	30	25	x	23	19
19	19	22	24	30	27	23	24	24	23	25	27	26	23	24	23	23	24	23	x	26
20	23	21	22	27	25	27	26	26	27	28	25	27	27	26	25	21	25	19	26	x

Fig. 5. Score for common knowledge of proverbs for DNAs in Fig. 3.

One of them is that sub-graph which contains exactly the 20 people mentioned in Fig. 3 and Fig. 5. The respective diagram is shown in Fig. 6.

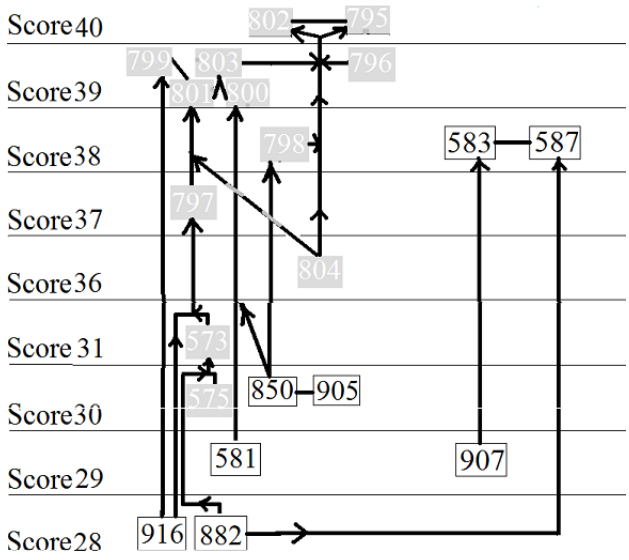


Fig. 6. Sub-graph surrounding Clique #2 (→ 802/795) and #6 (→ 583/587).

In this diagram, we use the hierarchy of proximity induced by the score of each node for a vertical order which means that all nodes with the same score are in the same level. An existing link between nodes of the same level is necessarily bidirectional and will be represented by a non-directional edge. Links between distinct levels indicate an asymmetric attraction between the involved nodes (because the higher graded one finds an even closer partner). In such a case, we have to mark the direction of interest with an arrow. The weight of an edge is given by the lowest score between the nodes involved.

When we follow the (omitted) arrows in the diagram, we find local attraction maximums that are stuck in a level without finding a way to a higher ranked person, as in pair 802/795 or 583/587. In such a case, the set of all the nodes interconnected on the same level is called (directional) Clique. Formally, the Clique, in the theory of social network, has a slightly different definition based on a extreme dense cluster (see for instance [5] and [6]), but both concepts aren't so different when we use it in a sequence of maximal Clique with a score-level in downfall, so we maintain the term clique in this text.

In such a finite graph there exists necessarily at least on the highest layer an undirected sub-graph between at least two vertices. We call a Clique any such a symmetric sub-graph if none of its vertices has an outgoing edge to a higher rank. At the highest rank there will be a Clique. But in any layer a Clique imply a transitive set of best neighbours with a local attraction maximum as in pair 802/795 or 583/587.

Inside a Clique, every node finds its best peer. And, by having a path via a sequence of best peer relations, between every two members of this set, we found a cohesive set of people with the biggest cultural proximity.

Per definition, the union of all nodes of a Clique (seen as a super node) behaves like a sink. Therefore, the Cliques inside the sub-graph induce a system of connected flows. So, the graph could be compared to a river system which runs lastly to different seas. This system is efficient to divide the universe in distinct fluent systems which are not strictly separated, whereas some nodes can achieve more than one Clique. If a local homogeneity exists, then, these subsystems should show some conceptual coherence.

Formally, the Clique, in the theory of social networks, has a slightly different definition based on an extreme dense cluster (see for instance [5]). Our concept isn't so different when we use a sequence of maximal Clique with a score-level in downfall. Because of that we maintain the term clique in this text.

We note that, in Fig. 6, all people with a white identity number come from Corvo Island and have their flux directed to the Clique formed by the two members of this island whose personal identity is 802 and 795. On the other side, the Clique formed by the persons with numbers 583 and 587 is entirely composed of people from Sa. Maria Island. Such a homomorphism between the cultural proximity (expressed in the graph) and the geographical region indicates that there are strong correlations between the cultural and geographic background of this people. In this study, we will analyse this kind of correlations and see what could be used to retrieve information about the habitat of a person by its cultural knowledge.

We use the word homomorphism in this text as a structure preserving the mapping between location where respondents lived for more than 5 years and the identified Cliques. Is also a way of concept generalization and the main objective for the development of the procedure described in this text. These homeomorphisms can be measured as any other supervised classification as using the rate of correct classifications or error rate.

III. CLIQUE - ISLAND HOMOMORPHISM

The division of the graph in its sub-graphs is the most structural aspect of the analysis; therefore, it is interesting that one sub-graph containing two thirds of all nodes exist. This, we will call the central sub-graph and, the Cliques in it, we will call the central Cliques (#04, #05, #08, #11, #12, #13, #15 and #17). All the other sub-graphs and their Cliques will be referred as the outsiders (#01, #02, #03, #06, #07, #09, #10, #14 and #16). To see why such a rough distinction makes sense, we should compare the structure of both sorts of Cliques by their Clique-DNA in Fig. 7. Value "1" indicates a common knowledge and value "0" indicates a common ignorance of the specific proverb due to all Clique members. If there is no agreement, we note down an "x".

It seems that the distinction between central Cliques and the Outsiders could be generally expressed by the rule: central Cliques are more related with common knowledge while the Outsiders are more related with common ignorance.

We see that the tax for common knowledge in the central Cliques is 100% for #04, 98% for #05, 88% for #08, 83% for #11, 35% for #12, 63% for #13, 48% for #15 and 24% for #17. So, we have in 5 of the 8 central Cliques a high common knowledge; in #15 there is a moderate tax of common knowledge, but whereas the tax of 53% of common ignorance in each of the remaining two central Cliques, #12 and #17, is very high, we have here two exceptions to the rule. So, we have in 62,5% of the central Cliques a high common knowledge (more or equal 63%); in 12,5% a moderate tax of common knowledge, but even in the 25% with a low tax of common knowledge the tax of common ignorance is not high.

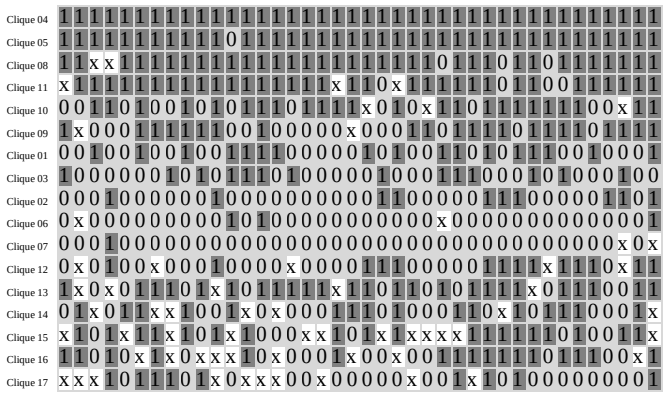


Fig. 7. Example of Clique DNAs.

In the Cliques of outsiders, we have the following taxes of common ignorance: 55% for #01, 75% for #02, 65% for #03, 88% for #06, 93% for #07, 40% for #09, 33% for #10, 43% for #14 and 36% for #17. So we have in 55% of the cases a higher common ignorance as in the worst case of central Cliques. And in the rest 45% cases none overtake the barrier of 60% for common knowledge which is surpass by 62% of the central Cliques.

In Clique #03, we have five, in #4, six and, in #15, tree nodes. All the remaining Cliques are constituted by two nodes.

With 8 of 17 Cliques, there is, in almost half part, a total homomorphism between cultural and geographical structure. The outsider Cliques #01, #02, #10 and #14 are each formed by people from Corvo Island. This is congruent with the general observation that the population of Corvo has really the status of an outsider due to their low level of proverbial knowledge. And, whereas #10 is one of the exceptions to the previous mentioned rule, we have a simple explanation since even the high competent habitants seem to have a so characteristic profile that they did not connect to the central subgroup.

Three more Cliques of outsiders have this degree of homomorphism: #03 is formed by 5 mobile habitants from S. Miguel and New England, #06 includes only habitants from Sa. Maria, as we had already seen, and #09 has only habitants from Pico Island. Whereas we found in 7 out of 9 Cliques of outsiders a homomorphism attached so directly to a quarter of all nodes of outsiders, we can conclude that these seven sub-graphs are highly separable due to the origin of the involved people.

In the central Cliques, only #13 is homomorphous, in this case, due to the geographical origin of Graciosa Island. We can find two more central Cliques (#12 and #17) where every node represents a person which once lived in Pico Island, but the most part of the involved people are mobile habitants. And we see that four out of six people represented in central Clique #04 lived in Terceira Island for at least 5 years. So, in the central sub-graph, only half of the Cliques show a geographical separ1313ability. The remaining four Cliques are a mixture of persons living in the central Islands of Azores. And this seems to be the best characteristic of the central sub-graph. It represents a big coherency between the people living in the five central Islands.

IV. HOMOMORPHISM IN THE FLUX TO THE CLIQUE

The biggest single affluence in the whole graph happens in Cliques #4 and then #5. More than half of the nodes are involved in this structure; therefore, we will concentrate especially on these two fluxes. We know that both Cliques are characterized by the highest possible score of 40 and by the highest common knowledge tax of respectively 100% and 98%, which means that there is a distinction by only one proverb: "Ninguém diga: desta água não beberei e deste pão não comerei."[=Nobody should say: I will not drink this water or eat this bread.].

As we have also seen in Fig. 6, nodes that are near the Clique could also have an interesting structure in terms of geographical separation. So, we will define two circular areas. First, the Inner Circle, in which all elements are situated at most at 3 score-points above the Clique (including itself), and, second, the Extended Circle, with all elements situated at most at 10 score-points above the Clique (including the Inner Circle).

It is clear that Clique elements are closer to the Inner

Circle than to the Extended Circle. In this scheme, we did not take into account all members of the Clique group outside the distance of 10 score-points because these are equivalent to, at least, a difference of 25% of the common knowledge.

TABLE I relates to Clique #5 including 19 people with a fixed residence and 6 people which are more mobile over their lifetime. It shows the distribution inside this Clique, in its Inner and Extended Circle. Outside the brackets, we count the presence of fixed residences and, inside the brackets, the number of mobile persons. The last row calculates, for the Extended Circle, the ratio between the expected and the real occurrence of people among the different localities. To get the expected value, we must multiply the number of people from the location times the ratio of the population attached to the respective Extended Circle and the total population. But, to avoid some numerical deformation, what could happen for values between 0 and 1, we use the Laplace estimator [7] and add value 1 to both parts of the fraction.

TABLE I
CHARACTERISTICS OF CLIQUE #5

Local	Clique	Inner Circle	Extended Circle	Real/Estimated Occurrence ratio
California	0	0	0(+2)	1/1.48= 67,6% (2/2,09=95,7%)
Corvo	0	0	0(+1)	1/5.32= 18,8% (2/1,18=169,5%)
Faial	0	2(+1)	5(+1)	6/2,68=223,9% (2/2.18= 91,7%)
Flores	0	0(+1)	1(+4)	2/2.8=71,4% (5/1,54=324,7%)
Graciosa	1	3	6(+1)	7/3.4=205,9% (2/1.18=169,5%)
New England	0	0	0(+1)	1/1.12= 89,3% (2/4.26=46,9%)
Pico	0	0	1(+1)	2/3,16= 63,3% (2/2.63= 76.0%)
Terceira	0	0	3(+2)	4/3.16=126,6% (3/2=150%)
S. Jorge	1	1(+1)	3(+1)	4/3.88=103,1%(2/1 .18 =169,5%)
Sa. Maria	0	0	0(+0)	1/1.84= 54,3% (1/1.09 = 91,7%)
S. Miguel	0	0(+1)	0(+2)	1/1.24= 80,6% (3/3.9= 76,9%)

We note that neither the geographic western group, nor the eastern group, or the areas of emigration in the US have a significant representation in the group of fixed persons. The most relevant areas due to fixed persons are Faial and Graciosa. But, due to mobile habitants, Flores gets an extremely high relevance. This means that we can characterize this flux that is near the sink by its geographical inclination to the Islands of Faial and Graciosa and the emigrants from Flores. This shows an interesting proximity between Flores and Faial, probably induced by inhabitants of Flores that are attracted to the administrative centre - Faial. In such a situation, the long geographical distance will be overcame by a more specific and periodical contact. It is very interesting that the geographical Islands, Pico and Faial, have no homologous tendency.

In TABLE II, we present the respective results for the affluence to Clique #4. The most important indicator is the local distribution among the fixed residents. The biggest peak in the Extended Circle for this extract of residents can be found in S. Miguel and Flores, but there is also in California and in S. Jorge a significant high performance. Besides the fact that we should ignore the result of S. Miguel and California due to the insignificance of their local universe, we were surprised for the fact that, in the first case, the totality and, in the second case, 75% of the fixed residents are involved. There are only three localities with a concentrated occurrence: S. Maria, Corvo and New England. This seems to be the natural consequence for the group of higher performers, the location with less knowledge rate has lower representation.

TABLE II
CHARACTERISTICS OF CLIQUE #4

Local	Clique	Inner Circle	Extended Circle	Real/Estimated Occurrence ratio
California	0	2(+3)	3(+6)	4/3,04=131,6% (7/5,64=124,1%)
Corvo	0	5(+1)	9(+1)	10/19,32=51,8% (2/1,39=143,9%)
Faial	0	4(+1)	8(+6)	9/8,12=110,8% (7/6,03=116,1%)
Flores	1(+1)	6(+1)	13(+2)	14/8,64=142,0% (3/3,32=90,4%)
Graciosa	0	3(+0)	10(+0)	11/11,18=98,4% (1/1,77=56,5%)
New England	0(+1)	0(+5)	0(+13)	1/1,51=66,2% (14/14,93=93,8%)
Pico	1	7(+3)	10(+7)	11/10,16=108,3% (8/7,97=100,4%)
Terceira	2/(+2)	4(+3)	10(+4)	11/10,16=108,3% (5/5,257=95,1%)
S. Jorge	0	1(+0)	15(+0)	16/13,22=121,0% (1/1,77=56,5%)
Sa. Maria	0	0(+0)	0(+1)	1/4,56=21,9% (2/1,39=143,9%)
S Miguel	0	0(+4)	2(+10)	3/2,02=148,5% (11/13,38=82,2%)

But the most interesting detail is the distribution inside the Extended Circle. Whereas members from S. Jorge are almost exclusively outside the Inner Circle, members of Pico are mainly inside. On the other hand the Clique itself is dominated by 1/3 of fixed inhabitants and another 1/3 of mobile inhabitants of Terceira. The existence of such local clusters forces the idea that the dispersion of those proverbs is dominated by physical and not by virtual transmission channels.

Clique #8 has a common knowledge of 90% and a score of 38 between the two persons, one is from Terceira and

the other is a mobile inhabitant from Faial and Pico. The Inner Circle contains no other person and the Extended Circle counts with one more mobile inhabitant from Faial and Pico.

The last Clique of interest is #11 with a score of 37 and a common knowledge of 83% between the fixed habitant from Graciosa and the mobile one from Terceira/Pico. The Inner Circle counts with two more inhabitants from Faial (score 36 and 34) and one from Flores (score 34). On the other hand the Extended Circle is completed by three more inhabitants from S. Jorge (score 33 and 32), two from Corvo (score 31 and 29) and one from Flores (score 28). This seems highly inhomogeneous but the different locations are distinct through the score-layers.

V. Conclusion

Due to the cultural space related to the Archipelago of the Azores formed by 9 physical islands and some virtual islands of immigration located mainly on the American continent, we analysed a huge amount of data about the recognition of proverbs.

For that data we develop a new procedure designed to analyse very dense social networks. The algorithm needs extensible testing but the preliminary results reported here are encouraging.

The procedure starts by selecting the 40 best known proverbs among a universe of more than 1500. Using a measure due to the proximity of common knowledge between every pair of two persons among the 221 inquired people we can define an incidence matrix of a graph. This social network shows relationships that translates the proverbial proximity inside this community.

This image will be clearer when we maintain only the most proximate relations. Such a reduction divides the graph in 8 oriented and isolated sub-graphs which distinguishes the society in a kind of different families of proverbial users. There are seven families with 20 members, in maximum, and a very big one which involves 2/3 of the universe. While we selected only high frequent proverbs, such types of continuity were expected. But by applying a hierarchical Clique analysis we can structure this apparently continuous space in sharply distinct clusters with a high inner homogeneity due to the location of the involved interviewee.

In general we can say that groups around the Cliques #04, #05 and #08 are nearly identical centres of accumulation for people that know most proverbs, while groups around Clique #06 and #07 lead to people which ignore almost all proverbs. Especially in the group of outsiders the local aggregation is more evident.

In summary we find in almost all groups surrounding the 17 Cliques local patterns which justify the idea that is realistic to choose a small base of proverbs to achieve a geographic indicator for the residency of person deduced only by its proverbial knowledge. From these partial results obtained only from a small set of data, we can conclude that there is evidence that our working hypothesis can be true and knowing or not knowing a small set of proverbs can be utilised to discriminate between different regions.

We propose to use the algorithm in bigger data sets and explore other methods for data reduction from the Clique DNA. One example we have beginning to work with is LAD – Logical Analysis of Data method [8] or even rough sets to identify a small nucleus of relevant proverbs. Other way we intend to explore is the relation of this work with what Faust S. K. Wasserman call positional analysis [9]. Essentially the concept is the same as our objective but he uses different techniques.

Appendix

In the next page we present the full network with all the nodes referred to in the text.

Acknowledgment

The authors wish to thank all the collaborators responsible for survey design and data collection with a very special mention to Professor Gabriela Funk, specialist in proverbs.

References

[1] G. Funk and M. Funk, *Pérolas da sabedoria popular: Os provérbios Açoreanos nos EUA*. Salamandra: Lisbon, 2001.

[2] G. Funk and M. Funk, *Pérolas da sabedoria popular: Os provérbios de S. Miguel*. Salamandra: Lisbon, 2001.

[3] G. Funk and M. Funk, *Pérolas da sabedoria popular: Provérbios da Ilhas do Grupo Central dos Açores (Faial, Graciosa, Pico São Jorge e Terceira)*, Salamandra: Lisbon, 2003.

[4] A. Mendes, G. Funk and M. Funk "Extrair Conhecimento de Provérbios". In *Temas em Métodos Quantitativos* M. F. Salgueiro (Eds.). Sílabo: Lisboa, 2009, pp. 89-107.

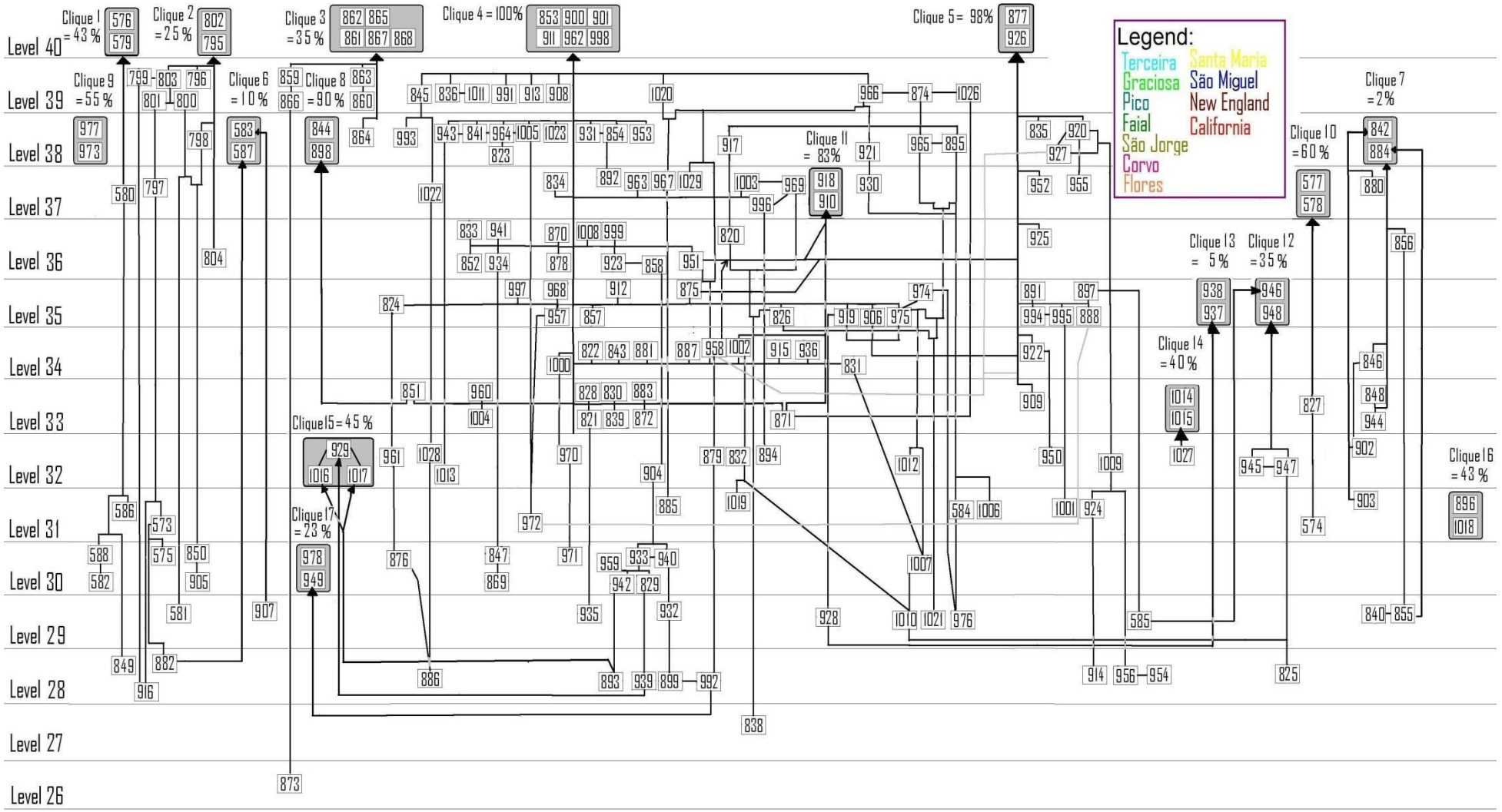
[5] P. Mika, *Social Networks and the Semantic Web*. Semantic Web and Beyond, New York, USA: Springer-Verlag, 2007, pp. 37-38.

[6] L. Cavique, C. Rego and I. Themido, *A Scatter Search Algorithm for the Maximum Clique Problem*, in Essays and Surveys in Metaheuristics, C. Ribeiro and P. Hansen (Eds), Kluwer Academic Publishers, 2005, pp.227-244.

[7] I. H. Witten and E. Frank *Data Mining: Practical machine learning tools and techniques.* The Morgan Kaufmann Series in Data Management Systems, San Francisco, USA: Morgan Kaufmann Pubs, 2005, pp. 91.

[8] E. Boros, P.L. Hammer, T. Ibaraki, A. Kogan, E. Mayoraz and I. Muchnik, An implementation of logical analysis of data, *IEEE Transactions on Knowledge and Data Engineering*, 12, 2000, pp. 292-306.

[9] Faust S. K. Wasserman, *Social Network Analysis: Methods and Applications*, Cambridge University Press, 1994.


Clique 1 = 43 %
576 579
Clique 2 = 25 %
802 795
Clique 3 = 35 %
862 865 861 867 868
Clique 4 = 100%
853 900 901 911 962 998
Clique 5 = 98%
877 926
Clique 6 = 10 %
583 587
Clique 7 = 2%
842 884
Clique 8 = 90 %
863 860
Clique 9 = 55 %
977 973
Clique 10 = 60 %
577 578
Clique 11 = 83%
918 910
Clique 12 = 35 %
946 948
Clique 13 = 5 %
938 937
Clique 14 = 40 %
1014 1015
Clique 15 = 45 %
929 1016 1017
Clique 16 = 43 %
896 1018
Clique 17 = 23 %
978 949
Legend:
Terceira
Graciosa
Pico
Faial
São Jorge
Corvo
Flores
Santa Maria
São Miguel
New England
California
Level 40
Level 39
Level 38
Level 37
Level 36
Level 35
Level 34
Level 33
Level 32
Level 31
Level 30
Level 29
Level 28
Level 27
Level 26