The social network induced by the common knowledge of proverbs

Matthias Funk; Armando B. Mendes
e-mail: mfunk@uac.pt; amendes@uac.pt
Department of Mathematics
University of the Azores
Ponta Delgada, Portugal

***Abstract*— In a series of interviews, we 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. One of the interesting questions was if we could find a minimum base of proverbs as an indicator to decide from which place a person came due to their specific profile of proverbial knowledge. Before trying this challenge, we will analyse, in this article, the probability of achieving such an idea by trying to find out if a homomorphism between the proverbial knowledge and the geographical location of a person could exist.**

To solve this question, 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 found, in the present pilot-project based on small data sets, 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 population.

Keywords: Proverbs; Social Network; Clique Analysis

I. Geography and universe of the interviews

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. Due to big waves of emigration into the USA, which have taken 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 biggest part of the emigrants is located in the area of Toronto (Canada), in Bermuda, in California and in New England.

The survey data was organized in a relational data base, which centralizes all information about the recognition of proverbs inside the Azorean cultural space. 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 proverb inside the cultural (sub-)space of the community as a whole or of the single location.

The data was collected between 1997 and 2000 and was analysed [in 1 and 2] in order to discover the local proverbial treasure of the main Island and also of the two hotspots of Azorean emigration in the U.S.A. (California/New England). Later [3], the properties of the central geographic group were analysed. The 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].

The sample of 221 people (83 males/138 females) used in this article show the following distribution taking in account, in the first place, mainly the 158 fixed residents (which lived only in one location) and, in the second place, the so called mobile residents (which lived in, at least, two places over a period of more than 5 years):

California (4 as fixed residents/12 as mobile residents), Corvo (36/2), Faial (14/13), Flores (15/6), Graciosa (20/2), New England (1/36), Pico (18/18), St. Jorge (24/2), St. Mary (7/1), St. Michael (2/32) and Terceira (18/11).

The average knowledge of the 40 selected proverbs is 64% over the whole universe and 63% over the universe of all fixed residents. Therefore, the data shows a globally insignificant tendency to an augmentation of competence by migration.

II. The measure of cultural homogeneity and the Clique Analysis

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. 1 shows the proverbial DNA for 2 interviewees and the cultural proximity defined by the Hamming distance of their DNA. The DNA of two persons is matched row by row in order to find their proverbial divergence. While we have 10 cases of mismatches (marked in Fig. 2 by a grey background), we find the Hamming Distance of 10 between the respective people.

Person 573 →
0011010010110010000000100000011 01010 01010
Person 575 →
0001110010110010000100101100011 01001 10011

Figure 1: Comparison of two proverbial DNA to get the Hamming distance of 10.

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). We can see that this matrix can be interpreted as a description of a totally connected, undirected, but weighted graph with n nodes. The excess of information given by this graph should be reduced by marking 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. One of them is subgraph 20 in Fig. 2.

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 on 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 cases, we have to mark the direction of interest with an arrow. The weight of an edge is given by the lowest score of both nodes involved.

When we follow the (omitted) arrows in the diagram, we find local attraction maximums that are stuck on a level without finding a way to a higher ranked person, as in pair 802/795 or 583/587. In such cases, the set of all the nodes interconnected on the same level is called (directional) Clique.

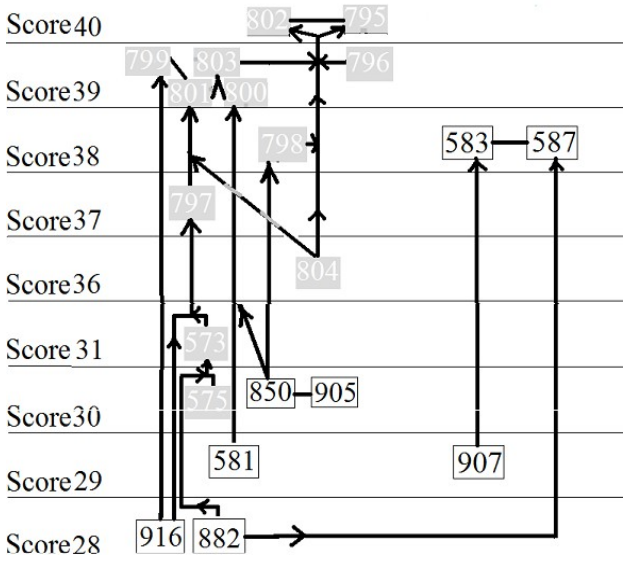

Figure 2: Subgraph surrounding Clique #2 (802/795) and #6 (583/597)

In a Clique, every relation is symmetric and every one of its nodes finds its best peer inside it. Besides, 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. It is easy to show that each subgraph has at least one Clique.

Per definition, the union of all nodes of a Clique (seen as a super node) behaves like a sink. Therefore, the Cliques inside the subgraph 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, while some nodes can achieve more than one Clique.

We note that, in Fig. 2, 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 hand, the Clique formed by the persons with numbers 583 and 587 is entirely composed of people from St. Mary Island. Such a homomorphism between the cultural proximity (expressed in the graph) and the geographical distance indicates that there are strong correlations between the cultural and the geographic background of these 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.

III. Homomorphism in the Clique

The division of the graph in its subgraphs is the most structural aspect of the analysis; therefore, it is interesting that one subgraph containing two thirds of all nodes exists. We will call this the central subgraph and we will call the Cliques in it the central Cliques (#04, #05, #08, #11, #12, #13, #15 and #17). All the other subgraphs 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. 3. 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 common sense, we note down an "x".

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 a high common knowledge in 5 of the 8 central Cliques; in #15, there is a moderate tax of common knowledge.

In the Cliques of outsiders, we have a high tax of common ignorance of 55% for #01, 75% for #02, 65% for #03, 88% for #06, 93% for #07 and a moderate ignorance of 40% for #09, 33% for #10, 43% for #14 and 36% for #16.

Clique 04	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Clique 05	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Clique 08	1	1	x	x	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	0	1	1	0	1	1	1	1	1	1	1
Clique 11	x	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	x	1	1	0	x	1	1	1	1	1	1	0	1	1	0	0	1	1	1	1	1	1
Clique 10	0	0	1	1	0	1	0	0	1	0	1	0	1	1	1	0	1	1	1	1	x	0	1	0	x	1	1	0	1	1	1	1	1	1	1	0	0	x	1	1
Clique 09	1	x	0	0	0	1	1	1	1	1	1	0	0	1	0	0	0	0	0	x	0	0	0	1	1	0	1	1	1	1	0	1	1	1	1	0	1	1	1	1
Clique 01	0	0	1	0	0	1	0	0	1	0	0	1	1	1	1	0	0	0	0	0	1	0	1	0	0	1	1	0	1	0	1	1	1	0	0	1	0	0	0	1
Clique 03	1	0	0	0	0	0	0	1	0	1	0	1	1	1	0	1	0	0	0	0	0	1	0	0	0	1	1	1	0	0	0	1	0	1	0	0	0	1	0	0
Clique 02	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	1	1	1	0	0	0	0	0	1	1	0	1
Clique 06	0	x	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	x	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Clique 07	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	x	0	x
Clique 12	0	x	0	1	0	0	x	0	0	0	1	0	0	0	0	x	0	0	0	0	1	1	1	0	0	0	0	0	1	1	1	1	x	1	1	1	0	x	1	1
Clique 13	1	x	0	x	0	1	1	1	0	1	x	1	0	1	1	1	1	1	x	1	1	0	1	1	0	1	0	1	1	1	1	x	0	1	1	1	0	0	1	1
Clique 14	0	1	x	0	1	1	x	x	1	0	0	1	x	0	x	0	0	0	1	1	1	0	1	0	0	0	1	1	0	x	1	0	1	1	1	0	0	0	1	x
Clique 15	x	1	0	1	x	1	1	x	1	0	1	x	1	0	0	0	x	x	1	0	1	x	1	x	x	x	x	1	1	1	1	1	1	0	1	0	0	1	1	x
Clique 16	1	1	0	1	0	x	1	x	0	x	x	x	1	0	x	0	0	0	1	x	0	0	x	0	0	1	1	1	1	1	1	1	0	1	1	1	0	0	x	1
Clique 17	x	x	x	1	0	1	1	1	0	1	x	0	x	x	x	0	0	x	0	0	0	0	0	x	0	0	1	x	1	0	1	0	0	0	0	0	0	0	0	1

Figure 3: The DNA of the Clique

Generally, we can confirm that the Cliques of outsiders are connected according to ignorance and the central Cliques according to high knowledge of proverbs, since the outsiders don't follow the rule in cases #09 and #10 and the central Cliques also don't do it in cases #12 and #17. In the following moderated cases we can verify the expected dominance of knowledge over ignorance for the central Clique #15 (ratio 19 to 9) and ignorance over knowledge for the outsider Clique #14 (ratio 17 to 16). So in 12 of 17 cases, the rule applies.

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 the cultural and the geographical structure. The outsider Cliques #01, #02, #10 and #14 are each formed by people from Corvo Island. Three more Cliques of outsiders have this degree of homomorphism: #03 is formed by 5 mobile habitants from St. Michael and New England, #06 includes only habitants from St. Mary, as we had already seen, and #09 has only habitants from Pico Island. While 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 subgraphs 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. However, the central Cliques #12 and #17 are formed by persons who once lived at least 5 years in Pico Island and we can 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 subgraph, half of the Cliques show a geographical separability. The remaining four Cliques are a mixture of people living in the central Islands of the Azores, which seems to be the best characteristic of the central subgraph. 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. Almost half of the nodes are involved in this structure; therefore, we will concentrate especially on its flux. We know that this Clique is characterized by the highest possible score of 40 and therefore by the highest common knowledge tax of respectively 100%.

As we have also seen in Fig. 2, nodes that are near the Clique could also have an interesting structure for a 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.

Fig. 4 relates the distribution inside Clique #04, in its Inner and Extended Circle. Outside the brackets, we count the presence of fixed residences and, inside the brackets, the number of mobile ones. 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 so called Laplace estimator [5, p. 91] and add value 1 to both parts of the fraction.

Local	Clique	Inner Circle	Extended Circle	Real/Estimated Occurrence ratio
California	0	2 (+3)	3 (+6)	131,6% (124,1%)
Corvo	0	5 (+1)	9 (+1)	51,8% (143,9%)
Faial	0	4 (+1)	8 (+6)	110,8% (116,1%)
Flores	1 (+1)	6 (+1)	13 (+2)	142,0% (90,4%)
Graciosa	0	3 (+0)	10 (+0)	98,4% (56,5%)
New England	0 (+1)	0 (+5)	0 (+13)	66,2% (93,8%)

Pico	1	7 (+3)	10 (+7)	108,3% (100,4%)
Ter-ceira	2 (+2)	4 (+3)	10 (+4)	108,3% (95,1%)
St. Jorge	0	1 (+0)	15 (+0)	121,0% (56,5%)
St. Mary	0	0 (+0)	0 (+1)	21,9% (143,9%)
St Mich-ael	0	0 (+4)	2 (+10)	148,5% (82,2%)

Figure 4: Characteristics of the affluence to Clique #4

The biggest peak in the Extended Circle for this extract of residents can be found in St. Michael and Flores, but there is also, in California and in St. Jorge, a significant high performance. Besides the fact that we should ignore the result of St. Michael and California, due to the insignificance of their local universe, we were surprised because, 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 reduced occurrence: St. Mary, Corvo and New England. This seems to be a natural consequence of the fact that in the group of higher performers the location with less knowledge rate has lower representation.

The most interesting detail is the distribution inside the Extended Circle. While members from St. 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 by another 1/3 of mobile inhabitants from Terceira. The existence of such slides in local clusters forces the idea that the dispersion of those proverbs is dominated by physical and not by virtual transmission channels.

V. Conclusion

By selecting the 40 best known proverbs among a universe of thousands of proverbs, we can define a measure due to the proximity of common knowledge between every pair of two persons among the 221 inquired people.

By interpreting the set of this data as an incidence matrix of a graph, we can draw a diagram about 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 subgraphs, which distinguishes the society in a kind of different families of proverbial users. 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 interviewees.

In summary, we find, in almost all groups surrounding the 17 Cliques, local patterns which justify the idea that it could be realistic to choose a small base of proverbs to achieve a geographic indicator for the residency of a person, deduced only by its proverbial knowledge. An interesting point of start could be a reduction of the Clique DNA, for example, by methods like LAD to a small nucleus of relevant proverbs.

References

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

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

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

[4] Armando Mendes, Gebriela Funk and Matthias Funk, "Extrair Conhecimento de Provérbios". Awaiting publication for the Temas em Métodos Quantitativos series. Sílabo, 2008, Lisbon.

[5] I,H. Witten and Frank Eibe, Data Mining – Practical Machine Learning Tolls and Techniques, 2nd Edition, 2005, Morgan Kaufman Publ., San Fransisco.