A Scale-Free Network of Evoked Words

A. A. A. Ferreira, G. Corso,
Departamento de Biofísica e Farmacologia, UFRN Campus Lagoa Nova, 59078-972, Natal, RN, Brazil

G. Piuvezam, and M. S. C. F. Alves
Departamento de Odontologia, Programa de Pós Graduação em Ciências da Saúde, Av Senador Salgado Filho 1787 - Lagoa Nova 59056-000 - Natal - RN - Brazil

Received on 17 January, 2006

We use a set of evoked words to define the vertices of a network. The connections between vertices are established by individuals in a population that evoke these words. The resulting graph is called an Evoked Words Network, **EWN**. The data of evoked words comes from an epidemiological research in odontological public health. In this research we consider three concept themes or evocative words: mouth, disease, and health. We investigate these words in two populations: an upper middle class and a poor district of the city of Natal. We compare and analyze six **EWN**s. The distribution of connectivities of all of these **EWN**s indicates a scale-free structure, with the data fitting a power law. The analyzed quantities of the **EWN**s depend more on the concept theme than on the income of population. This conclusion is discussed in the context of language-based networks.

Keywords: Networks; Scale-free; Epidemiology; Linguistic

I. INTRODUCTION

Networks have become an important modeling tool in statistical mechanics [1–3]. Scientists are applying network concepts in very diverse areas as linguistics [4, 5], sociology [6, 7], cellular biology [8, 9], computation science [10, 11], and ecology [12]. Networks have been used to find universal laws in nature, which is a specially interesting approach in physics, and also to visualize, describe, and give better insights on many phenomena. In this work we use networks to model a new problem in the interface between linguistics and epidemiology. We develop a network of social representations (evoked words) of mouth, disease, and health, used by two distinct populational groups. Before examining in detail this network, we make some remarks on contemporary network analysis.

A network, also called a graph by the mathematicians, is an object G formed by a pair of sets $G = (V, E)$, where V is a set of vertices (or nodes) and E a set of links (or edges). The total number of vertices is N, and n is the total number of links. A simple measure of a graph is $< k >= 2n/N$, the average number of links of the network or the average linking degree. The quantities N and n give global information about the network. An important local quantity is the connectivity k_i, the number of links of the vertex i. Using k_i we construct $P(k)$, the linking degree distribution, which corresponds to the density of vertices with k links. In the turn of the century, the distribution of links was used by physicists [1, 2] to classify networks into major groups (regular, random, small-word, and power-law networks). Regular graphs are characterized by a constant $P(k)$ as in the usual crystal lattices. In random graphs, $P(k)$ follows a Poisson distribution, which means that links are randomly distributed among all of the vertices of the network. The most recent contribution of physicists to network modeling is the investigation of small-world networks and networks with a power law $P(k)$. In fact, it was found that several phenomena in the world, from Internet linking to cell metabolic processes, can be modeled by a power law $P(k)$ [3].

In this work we construct some networks using data from an epidemiological research in the city of Natal, with a population of about 690,000 inhabitants, in the Northeast of Brazil. This research is about social representations (evoked words) that come from concept themes (evocative words) related to mouth health. Populations from two districts were analyzed, *Lagoa Nova* district, with a high average income, and *Felipe Camarão,* a low-income district. Since the data acquisition was not planned to be used in network analysis, we adapted the epidemiological data for our purposes. We felt quite successful because our modeling has driven the attention of the epidemiologists to new aspects in their statistical analysis.

What is the most relevant factor to define the topology of the underlying network, the evoked word or the social class? In other words, is the economic status of a populational group strong enough to surpass the language influence? We conclude that the social status of a populational group plays a minor role in the topology of the network. The layout of this paper is as follows. In section 2 we show how the epidemiological research was conducted and how we have built a network with the data. In section 3 we give the main results of the network analysis. Finally, in section 4 we give some final remarks and point out the power-law character of language-based networks.

II. THE CONSTRUCTION OF THE NETWORK

In this section we show the construction of the network. In the first subsection we describe the epidemiological research from which we take the raw data. The epidemiological data are related to the social representations of the populations of an upper middle class district and a poor district in Natal. In the second subsection we show the set up of the Evoked Words Network (**EWN**), including the definitions of the vertices and

the connections.

A. The epidemiological research

In contrast to most of the research in epidemiology that is focused on disease statistics in populations, the epidemiological data used in this work corresponds to social representations of health and disease. In fact, there is a new trend in epidemiology that is centered in the analysis of health prevention rather than disease statistics [13]. The understanding of health, by its turn, uses the technique of social representations [14]. The idea behind these social representations is to map and quantify health and disease concepts of a population using standard tools of epidemiology.

The social representation technique is based on the word evocation method. The main idea is that we can map the social representations of a given concept in a population through a set of words evoked by individuals in this population. In this work we use three main concepts, health, disease, and mouth. We ask a question to each individual of our sample: give me three words that come to your mind when you think about health. The same question is also repeated for the concept themes disease and mouth. This questionnaire is randomly applied among individuals belonging to populational groups from an upper middle class and a poor district. Therefore, we have six data sets, H_U, H_P, D_U, D_P, M_U, and M_P. These sets correspond to the three concepts, Health, Disease and Mouth, and to the two populational groups of upper middle and poor districts.

The number of individuals, N_I, is defined in the epidemiological research in order to construct a representative population of the most common dental disease in Brazil, cavity. Using the frequency of dental diseases in Brazil, as well as the populations of the districts where the questionnaires were applied, we found $N_I = 72$ [15]. In this way we argue that, although N_I is not a large number from the point of view of statistical mechanics, it is a representative number in the epidemiological context.

B. The Evoked Words Network

The **EWN** is built in the following way. The vertices of the **EWN** are the evoked words (or word expressions) of the data of the epidemiological research. The connections into the **EWN** are established by the individuals of the population. Each time two individuals share the same concept there is a link between the vertices. Since there are 72 individuals and each one evokes three words, the maximum number of connections is $n = 72 \times 3 = 216$. The number of vertices in the network is the number of distinct words that depend on the specific data set. In Table I we show the actual values of N and n for the six data sets discussed in the previous subsection. In the same Table, we depict the number of isolated vertices, N_{isol}, that is, the vertices that are not connected with the main cluster. Finally, we show in the same table the average connectivity of the network, $< k >= 2n/(N - N_{isol})$. A quick analysis of Table I shows that there is more discrepancy in network parameters among concept themes than social classes. It is interesting to note that the lower-class group has a large value of N and of N_{isol} as compared with the upper middle-class group. As a consequence, the lower-class group is associated with a smaller value of $< k >$. This fact is probably related to the free and creative verbal expression of individuals in the lower-class group. Individuals in the upper class are more inclined to give standard answers to the questionnaire, which in turn reduces the value of N.

TABLE I: The main parameters of the network: the total number of vertices, N, the total number of connections, n, the number of vertices isolated to the main cluster, N_{isol}, and the average connectivity, $< k >$.

	M_U	M_P	D_U	D_P	H_U	H_P
N	77	83	129	151	115	131
n	174	162	209	210	205	208
N_{isol}	8	15	28	43	24	27
$< k >$	5.04	4.76	4.13	3.89	4.51	4.00

Figure 1 depicts a sketch of the **EWN** for the word theme **mouth** in the upper middle class district. We choose this specific network among other data sets because of the relative small number of vertices. A quick view of this **EWN** shows the main poles of the network (teeth, communication, hygiene, and smile). However, most of these words has just two neighbors, since each individual connects each vertex with at least two other ones. There are small clusters that are not connected to the main network. These clusters correspond to individuals using unique expressions that are not shared by people in the main cluster. Although the individuals were asked to evoke three words, some of them evoke two or three times the same word.

III. RESULTS

We start looking the distribution of connectivities $P(k)$ of the **EWN**. The available data for each network is not large, $N \sim 100$, which implies in a poor statistics. In this case the cumulative sum, $\Phi = \int_{k_{max}}^{k} P(k\prime)dk\prime$, is more adequate to work than $P(k)$ because it reduces the fluctuations. If the network follows a power-law, $P(k) \sim k^{\gamma}$, then $\Phi(k) \sim k^{\rho}$, with $\rho = \gamma + 1$. We analyse Φ versus k for the six sets of data. The samples fit very well into a straight line in a log-log plot; the correlation coefficient is above 0.97 for all data. In addition the curves have the parameter ρ in the region: $1 < \rho < 2.$, the full set of values are in Table II.

Table II shows the clustering coefficient, C, the normalized clustering coefficient, $\bar{C} = C/C_{rand}$, the average distance, d, the parameter ρ, and the diameter for the six graphs. The clustering coefficient is normalized by the clustering coefficient of the associated Erdös-Renyi random graph, C_{rand}, which has the same values of N and n but randomly distributed connections. It is not surprising that $\bar{C}$ is at least ten times larger than

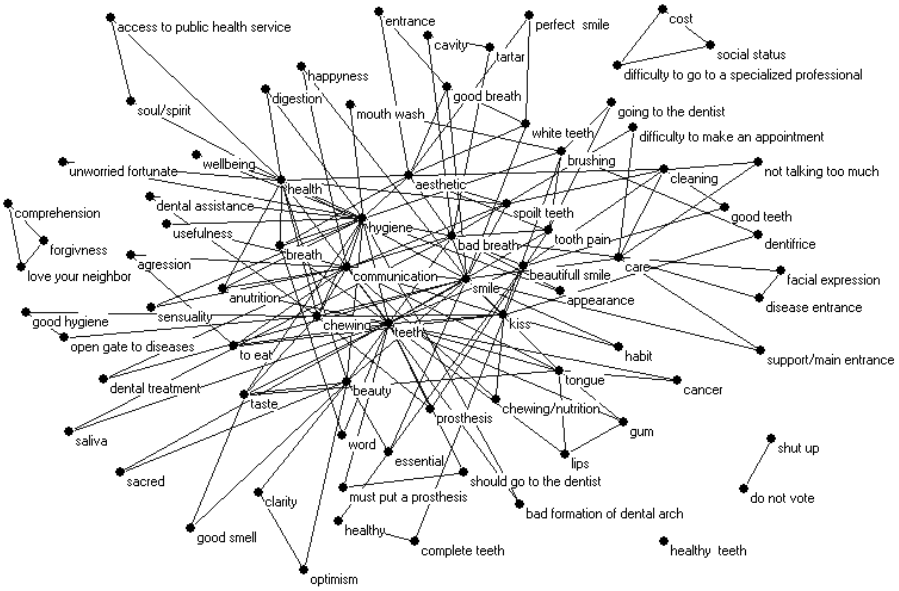

FIG. 1: A sketch of the **EWN** for the word thema **mouth** in a middle class neighborhood. The vertices are the evocated words and connections between words are established by words employed by a same individual.

C, the fact that each individual evokes three words introduces a large number of triangles in the graph. The quantity of triangles is an alternative way to measure the clustering coefficient [16].

TABLE II: Some quantities of the **EWN**: the clustering coefficient, C, the normalized clustering coefficient, $\bar{C}$, the average distance, d, the parameter ρ, and the diameter.

	M_U	M_P	D_U	D_P	H_U	H_P
C	0.77	0.70	0.81	0.87	0.80	0.82
$\bar{C}$	10.5	11.3	32.5	46.1	24.5	32.6
d	2.49	2.71	3.20	2.74	2.89	3.81
ρ	1.22	1.11	1.89	1.81	1.65	2.01
diameter	4	5	8	7	8	10

Finally we compare the properties of **EWN** regarding the differences between concept themes and income. In other words, we ask for the question: what is more important for network properties: the income of the comunity or the evocated word (mouth, disease or health). We use tables I and II to perform the comparision. An analysis of these data shows that the concept theme is more relevant than income in the determination of network properties. In other words, the linguistic phenomenon (evoked word) is more appropriate than income (social group) to describe the properties of network data. The only exception to this trend is the clustering coefficient. This anomalous behavior of C is probably due to the way the network was constructed, with individuals choosing three words, which inflates artificially the number of triangles of the network.

IV. FINAL REMARKS

We have built an Evoked Words Network, **EWN**, whose vertices are formed by evoked words according to three concept themes: health, disease and mouth. Each individual in the research evokes three words whick are linked. We work with six data sets corresponding to the three concept themes applied to individuals of two populations. The populations belong to social groups of an upper middle class and a poor district of a certain town. The distribution of connectivities of all of these **EWN**s follows a power law with an exponent $1.11 < \rho < 2.01$, depending on the concept theme and the social status of the population. The evoked words between the two studied groups are quite different and reflect status, schooling, leisure habits or manner of speech. For the evoked theme mouth, for instance, the five most evoked words in the poor group were chewing, communication, smile, hygiene, and bad breath. In the upper-class group, the analogous words were teeth, communication, hygiene, smile, and health. These words are interesting by themselves and most of the investigations on social representations deal with this kind of problem [17]. Our approach, on the other hand, is focused on network properties.

The question we pose in this paper is: what is the most important factor to determine network properties: the social group or the evoked word? The analysis of the data reveals that, despite economic determinants, the concept theme is

more conclusive than the group income. The quantities N, N_{isol}, $< k >$, γ, d and the diameter support this conclusion. The network structure does not seem to depend on the social group, which suggests that linguistic processes underlying word association are somewhat universal.

It is interesting to compare our results with the work of Motter et al. [18] for the small-world structure of an English thesaurus. These authors construct a network using the entries of a thesaurus as the vertices, and establishing links according to the respective synonyms. This network was shown [18] to present a large clustering coefficient and a power-law distribution of connectivities for large k. In a linguistic perspective, our work is similar, the main difference being the way the organized group of synonyms is constructed. Our work uses an oral synonym model (evoked words) instead of written and rigorous data transcribed in a thesaurus. In addition, we deal with a network of synonyms for three main concepts in specific communities. Motter et al. make a quite similar study on the basis of the full set of concepts of the written English language. The results in both papers, however, indicate a power-law structure for the distribution of connectivities in the network of synonyms. The search for the underlying mechanisms of this power law is a challenging task in cognitive science. It has been claimed [18] that the power-law structure of the network maximizes the efficiency in associative memory processes.

Acknowledgements

The authors gratefully acknowledge the financial support of Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Brazil. We also thank for the utilization of the free program of network computation Pajek, http://vlado.fmf.uni-lj.si/pub/networks/pajek/.

[1] D. J. Watts and S. H. Strogatz, Nature, **393**, 440 (1998).
[2] S. H. Strogatz, Nature, **410**, 268 (2001).
[3] R. Albert and A-László Barabási, Rev. Mod. Physics, **74**, 47 (2002).
[4] R. Ferrer i Cancho and R.V. Solé, Proceedings of the Royal Society of London B **268**, 2261 (2001).
[5] M. Medeiros Soares, G. Corso, and L. S. Lucena, accepted in Physica A (2005).
[6] F. Liljeros, C. R. Edling, L. A. N. Amaral, H. E. Stanley, and Y. Aberg, Nature **411**, 907 (2001).
[7] D. de Lima e Silva, et alli, Physica A **311**, 590 (2004). COMPLETAR
[8] D. Fell and A. Wagner, Nature Biothech. **189**, 1121 (2002).
[9] H. Jeong, B, Tombor, R. Albert, Z. N. Oltvai, and A. -L Barabási, Nature, **407**, 651 (2000).
[10] M. Faloutsos, P. Faloutsos, and C. Faloutsos, Comput. Commun. Rev. **29**, 251 (1999).
[11] R. Albert, H. Jeong, and A.-L. Barabási, Nature **401**, 378 (1999).
[12] J. M. Montoya and R. V. Solé, Journal of Theoretical Biology 214(3), (2002).
[13] C. Howarth, J. Foster, N. Dorrer, J Health Psychol **9(2)**, 229 (2004).
[14] S. Moscovisi and I. Marcova, Culture & Society, **4 3**, 371 (1988).
[15] A. A. A. Ferreira, *A Boca e seus Significados: um estudo de Representações Sociais* PhD Thesis, Universidade Federal do Rio Grande do Norte, (2005).
[16] M. E. J. Newman, S. H. Strogatz, and D. J. Watts, Phys. Rev. E. **64**, 026118 (2001).
[17] R. Goodwin, A. Kozlova, A. Kwiatkowska, L. Anh Nguyen Luu, G. Nizharadze, A. Realo, A. Kulvet, and A. Rammer, Soc Sci Med. **56(7)**, 1374 (2003).
[18] A. E. Motter, A. P. S. de Moura, Ying-Cheng Lai, and P. Dasgupta, Phys. Rev. E. **65**, 065102 (2002).