Network properties of written human language

A. P. Masucci and G. J. Rodgers
Department of Mathematical Sciences, Brunel University,
Uxbridge, Middlesex, UB8 3PH, United Kingdom

(Dated: May 9, 2006)

We investigate the nature of written human language within the framework of complex network theory. In particular, we analyse the topology of Orwell's *1984* focusing on the local properties of the network, such as the properties of the nearest neighbors and the clustering coefficient. We find a composite power law behavior for both the average nearest neighbor's degree and average clustering coefficient as a function of the vertex degree. This implies the existence of different functional classes of vertices. Furthermore we find that the second order vertex correlations are an essential component of the network architecture. To model our empirical results we extend a previously introduced model for language due to Dorogovtsev and Mendes. We propose an accelerated growing network model that contains three growth mechanisms: linear preferential attachment, local preferential attachment and the random growth of a pre-determined small finite subset of initial vertices. We find that with these elementary stochastic rules we are able to produce a network showing syntactic-like structures.

PACS numbers: 89.75.-k, 89.20.Hh, 05.65.+b

arXiv:physics/0605071 v1 8 May 2006

I. INTRODUCTION.

Many systems in nature are composed of a large number of interacting agents that exhibit small world, scale free and hierarchical behavior. Such networks can be found in several disciplines including social human organization, biological and chemical structures, the world wide web, etc...[1].

Words are a good example of simple elements that combine to form complex structures such as novels, poems, dictionaries and manuals that are designed to transport or convey information. The written human language is one of the most important examples of a self-organizing system in nature. It is not only interesting in a linguistic and philosophical sense, but it also has the virtue of being a accessible network system which allows full empirical studies of its structure to be carried out. If we consider a complete single book as our system and treat this book as a finite directed network in which the words are the vertices and two vertices are linked if they are neighbors, then we can analyse this network completely. We are able to know everything about the construction of such a network: we know the sequence in which each vertex was linked to the network and we can follow the growth of each vertex degree without error. Moreover, we already know a great deal about the syntactic and logical structure of grammar which determines how the network organises itself. Finally, plenty of data is available, covering centuries of human literature.

The first quantitative observation that language displays a complex topological structure was due to Zipf in 1949 [20]. Zipf observed that if the words of a prose are ordered in their rank of occurrence, that is the most frequent word in the prose has rank one, the second rank two and so on, and the frequency of the words is plotted against their rank, then, for every text and for every language, we find a skewed distribution that fits a power law (see Fig.2a as an example). This universal empirical law is known as Zipf's Law.

The first attempt to explain this law was due to Simon in 1955 [17]. He proposed a stochastic model for generating a text based on the frequency of occurrence of words, which he was able to solve exactly. In this model a stochastic text was built by adding a previously used word with probability proportional to its frequency and by adding new words with constant rate.

With the introduction of scale free network theory [1, 2], human language structure was examined by a number of authors [3, 7, 8, 13, 19]. There is a straight forward connection between the rank r of a word and the scale free distribution for the vertices degree in a network. If we define the degree k of a word as the number of different words this word is connected to, and $P(k)$ the word degree distribution, we have:

$$r(k) \propto \int_k^\infty P(k')dk'. \tag{1}$$

In language network the degree of a word is equivalent to its frequency, so that Eq. (1) is a direct link to transform the scale free degree distribution into Zipf's Law. Thus, in the context of growing networks, Simon's model is equivalent to the more recent [5], where the network growth is regulated by preferential attachment [2].

In all the models above, growth is based on the global properties of the text. These classical models display a good power law for the degree distribution, but, as it will be stressed later, this power law distribution holds even when

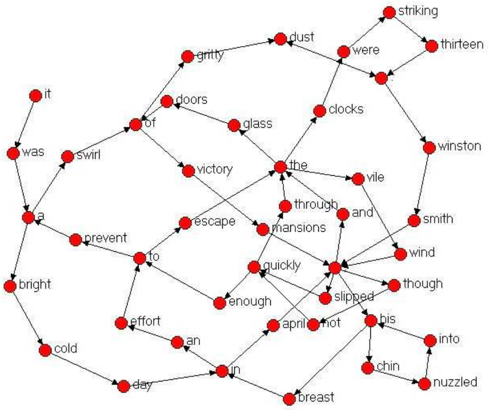

FIG. 1: Illustration of the language network for the first 60 words of Orwell's *1984*.

we randomize the words in the text, that is if we write a meaningless book. Thus, the degree distribution is not the best measure of the self organizing nature of this network.

Nevertheless, everybody who is experienced with the process of writing knows that syntax is the basic rule used to build a sentence. Syntax is nothing more than a set of local rules that give the ensemble of words in a phrase an intelligent and understandable structure.

In this work we analyse in detail the topology of the written language network focusing on the local properties and we find that these local properties are essential elements of the network architecture. We find that to build a stochastic model reproducing the main properties of language we need several growth mechanisms for the network. We obtain the best fits with real data by adding a random attachment mechanism, for a small preselected set of vertices, in an accelerated growing network, that displays both global and local preferential attachment.

The paper is organized as follow: in section II we analyse the growth properties and the nearest neighbors properties of a novel, in section III we generalize the Dorogovtsev-Mendes model for language [3] and extend it to include the local behavior of vertices.

II. TOPOLOGY OF THE NETWORK

We consider the George Orwell's *1984* as our system [16, 21]. This novel can be treated as a finite directed network in which the words are the vertices and two vertices are linked if they are neighbors. Punctuation is also considered as vertices.

For our purposes it is very important to consider written human language as a directed network since the syntactic rules of language are not reflexive. An interesting measure to quantify this behavior of language is the link reciprocity, that quantifies the non-random presence of mutual edges between pairs of vertices [10]. Reciprocity values lie between -1 for perfectly antireciprocal networks, 0 for areciprocal networks and 1 for perfectly reciprocal networks. In our case we find a value of 0.021 for the novel *1984*. It means that written human language is one of the natural networks closest to the absolute areciprocity.

In Fig.1 we show the first 60 vertices of our network, corresponding to the first 60 words of *1984*. A random walk in the whole network defines exactly all the grammatical and syntactical rules that are in the text. In such a directed network the degree and the frequency of a word have the same meaning, and are equal, because every time a new word is added to the text it is the only vertex of the network to acquire an edge.

The network is composed by 8992 vertices representing the different words and 117687 directed edges so that the mean degree is $<k>= 13.1$. The degree distribution follows a power law with slope: $P(k) \propto k^{-1.9}$, and the Zipf's Law, according to eq.(1), has slope -1.1(see Fig.2).

It is important to notice that the feature of being a scale free network doesn't depend on the syntactic structure of the language [12]. In fact, due to the equivalence between frequency and degree of a vertex, if we shuffle the text we obtain the same degree distribution, but loose all the syntactic structure.

BURA

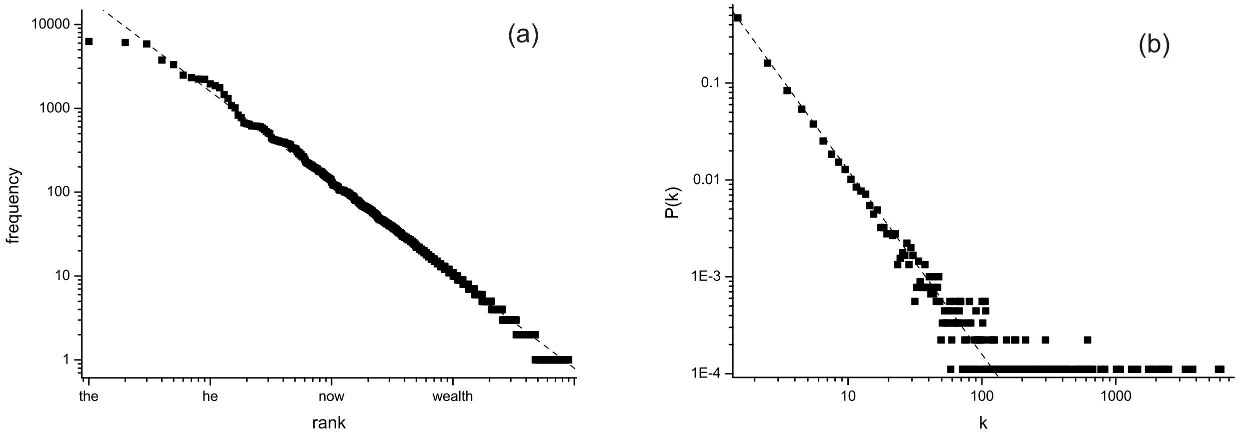

FIG. 2: (a) Measure of the Zipf's Law on *1984*. The dashed line is a power law with slope -1.1. (b) The degree distribution P(k) measured on the same novel. The slope found is -1.9.

A. Growth properties

A basic property of language network is that it is an accelerated growing network in the sense that the number of edges between words grows faster than the number of vertices [4]. If we define the time step t as the time in which a new word is added to the text and $N(t)$ the total number of words in the text at time t, we find empirically that

$$N(t) \sim t^{1.8}. \tag{2}$$

At every time, t will represent the number of different words used to compose the text, or the used vocabulary size, while $N(t)$ will represent the total novel size.

In order to build a stochastic model for language, we will need a random growth mechanism that attaches new words to the text and a preferential attachment (PA) mechanism that attaches previously used words to the text. Moreover we will need a mechanism that could catch the inner structure of the text, the syntax.

B. Nearest neighbor's properties

Syntax is made up of local and selective structures that can be recognizable through the analysis of nearest neighbors. A very important measure to quantify the hierarchical structure of a network is the clustering coefficient [15]. This counts the triangles that form in the unweighted and directed network associated with the network in consideration. We expect language to show a low average clustering coefficient because only a few triangles are present in the related network. The reason for this is the selectiveness of syntactic structures. For example the word "like" is able to link to definite and indefinite articles "a" and "the" but these articles will never be linked to each other.

We define the clustering coefficient c_i for every vertex i of our network as

$$c_i \equiv \frac{e_i}{d_i(d_i - 1)}, \tag{3}$$

where d_i is the number of the different nearest neighbors of vertex i (with $d_i \neq 0, 1$) and e_i is the number of directed edges that connect those nearest neighbors. This formula is a generalization of that for undirected networks [1].

We find that the mean clustering coefficient for our network is $< c >= 0.19$, that is the clustering coefficient is on average very low if compared to that of other real networks [1]. This is due to the syntactic structure of language that tends to create functional structures instead of clustering structures.

In Fig.3 we show the clustering coefficient against the degree of the vertices for our novel. The clustering coefficient values are spread across the graph. If we associate those values to the properties of the subgraph associated to each word, we understand how words can display a very complex organization.

Even if the mean clustering coefficient for this kind of network is very low, modularity is still evident from a global measure of the mean clustering coefficient and the mean nearest neighbors degree as a function of the vertex degrees.

BURA

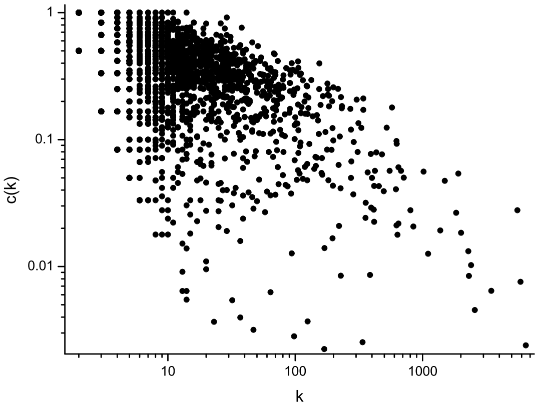

FIG. 3: The clustering coefficient against the degree of the vertices in *1984*.

In particular the mean clustering coefficient as a function of the vertex degree is flat for random graphs [18]. Our measures are shown in Fig.4a. Two different behaviors clearly emerge. For low values of k the data is nearly flat, which means low degree vertices don't display a strong hierarchical behavior. It's not the same for high degree vertices where hierarchical structures are apparent.

The mean nearest neighbor's degree as a function of the vertex degree is a good estimator for the degree-degree correlations [18]. This is also flat in the random graph case. Our measurements in Fig.4b show the existence of global disassortative or negative correlations, where large degree vertices tend to be connected to those with low degree and vice versa. Moreover we can see strong analogies with the measure of the mean clustering coefficient. In fact the mean nearest neighbors degree display two different behaviors, the first for low values of k where the power law fitting curve is nearly flat, revealing very low correlations between the degree of vertices. For large values of k the power law behavior is much stronger, disclosing strong degree-degree correlations for high degree vertices.

The cut off in the power laws for the mean clustering coefficient and the mean nearest neighbor's degree is nearly the same and is around $k \approx 100$. Only around 1% of the total number of vertices in the network posses $k > 100$ and those vertices belong to the 64% of the total number of edges in the network. Those vertices, belonging to the tale of $P(k)$ distribution (see Fig.2b), are essentially articles, punctuation, prepositions and pronouns. They organize the main architecture of the network.

Many attempts to find a statistical measure that can describe the syntactical structure as an emergent property of the language have been proposed [8, 9], but these use a-priory information about the logical role of each word.

The small value of the clustering coefficient implies, as we stressed before, that there are only a few triangles in the network. A particular structure emerging from the network is the directed binary structure, that is when two words are linked together several times in the text.

TABLE I: The most frequent binary structures that are present in *1984* with their relative frequencies.

Rank	Structure	Number of occurrences
1	of the	742
2	, and	717
3	. the	594
4	it was	576
5	in the	570
6	. he	560
7	. it	477
8	’ s	412

We show in Table I the most common binary structures we find in our novel. Some of them can appear trivial, but looking at the number of times they emerge in the text it is understandable why we consider them so important. If we plot a histogram of the occurrence of each binary structure we find a power law distribution. This suggests that the binary structures play an important role in the organization of the complex network.

BURA

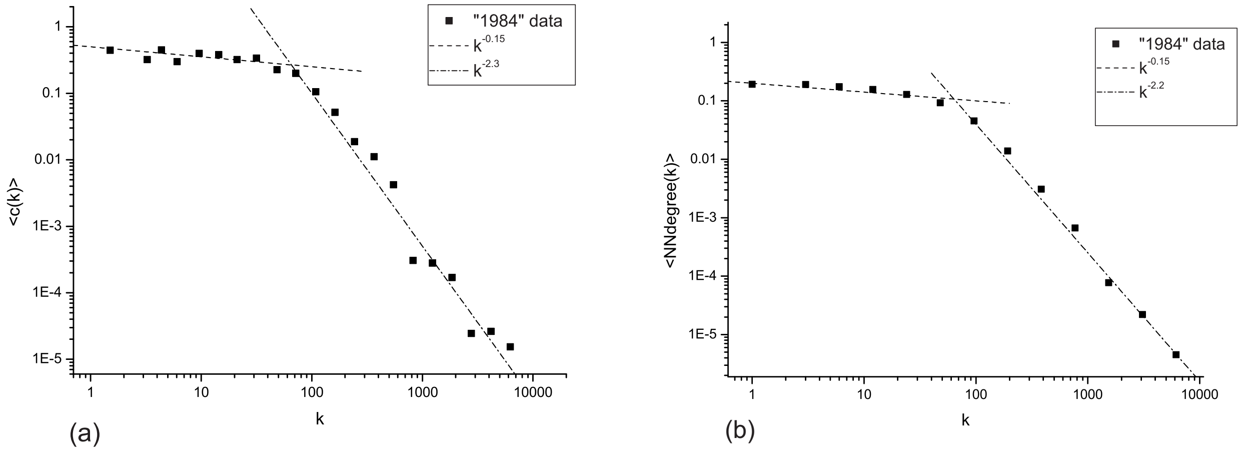

FIG. 4: (a) The average clustering coefficient as a function of the vertex degree. (b) The average nearest neighbors degree as a function of the vertex degree. Noise has been removed using logarithmic binning. In both plots two different regions emerge displaying different power law behavior.

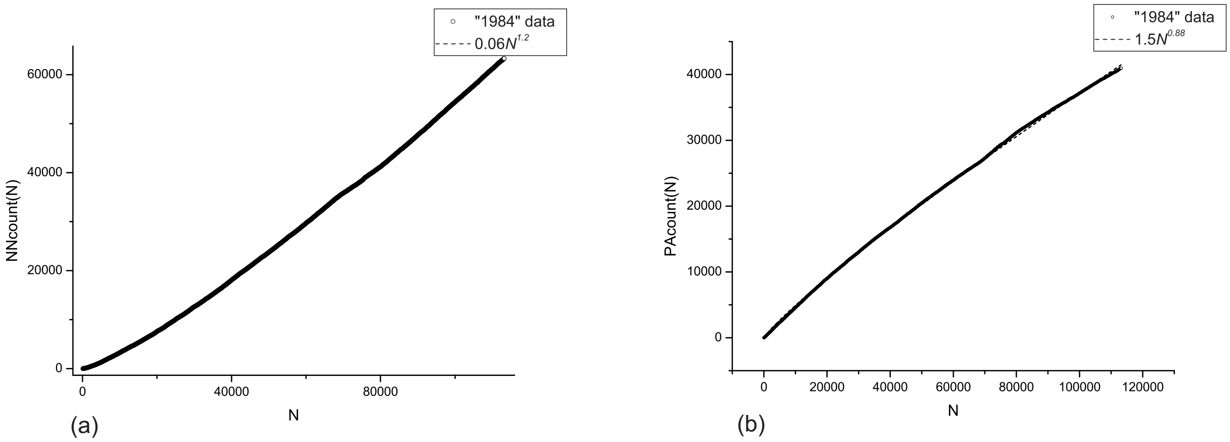

FIG. 5: (a) Number of occurrences of repeated binary structures of words in the text against the total number of words. (b) Number of occurrences of previously existing words in the text, that are not part of previously existing binary structures, against the total number of words.

In Fig.5 we show a measure for the relative occurrence of previously appeared words during the evolution of the text. In Fig.5a we count the occurrence of a repeated word if it belongs to a previously existing binary structure, while in Fig.5b we count the occurrence of a repeated word if it doesn't belong to a previously existing binary structure. We find that more than half the words added to the text form binary structures that are already present, while only 36% of the words entering the text doesn't belong to binary structures.

The standard way to create scale free networks is the PA mechanism [1], that is a new vertex is linked to vertex i with probability proportional to its degree k_i. However, this mechanism doesn't reproduce the massive formation of binary structures observed in the language network.

Creation of binary structures implies that new edges form between previously linked vertices. This behavior can be reproduced through a stochastic process that is not the usual PA, but a *local* PA. By local PA we mean that a vertex will be linked to a node i chosen from its nearest neighborhood with probability proportional to k_i.

It appears quite natural now, for the construction of a model, to split the standard PA [1] in the *local* PA and the *global* PA. For global PA we mean a mechanism for choosing vertices as the standard PA, but excluding the nearest neighborhood, that is a vertex will be linked to a node i that is not part of its nearest neighborhood with probability proportional to k_i.

In the next section we review an important accelerated growing network model that fits very well with global

BURA

properties of language and later we add to it the stochastic behavior we found in the previous analysis to build up a model that captures both the global and local properties of real text.

III. THE MODELS

Dorogovtsev and Mendes introduced a model [3, 4] (hereafter the *D-M model*) for an accelerated growing network as a basic model for language. We generalize this work with the intention to understand the local properties of the language such as the clustering coefficient and the nearest neighbor's properties.

We make a few modifications to the model to suit it to our analysis. We will consider it as a directed network and, based on the equivalence between frequency and degree of a word, we make some simplifications in the implementation of the network.

A. D-M model

The D-M model [3, 4] starts with a chain of 20 connected vertices.

1. At each time step we add a new vertex (word) to the network. We link it to an old vertex of the network through the standard PA, that is it will be linked to any vertex i with probability proportional to k_i.

2. At its birth, $m(t)-1$ new edges emerge between old words, where $m(t)$ is a linear function that can be measured and represents the accelerated growth of the network in consideration. These new edges emerge between old vertices i and j with the probability proportional to the product of their degrees $k_i \cdot k_j$.

We use

$$< m(t) > \approx 0.002t + 2.7, \tag{4}$$

which is a result we measured from *1984*.

This simple model has an analytical solution and can reproduce very well the degree distribution [3, 4]. However it fails to reproduce Zipf's Law (see Fig.7a) and the internal structure of the language. The average clustering coefficient measured in *D-M model* is $< c(k) >= 0.16$, that is smaller then that measured for *1984*, but close to it.

In Fig.8a we show the clustering analysis performed on the model compared to our network. Even though the general behavior of the model follows that of the real data, big differences are quite evident. $< c(k) >$ for the model is much narrower then that from the empirical measurements. This means that it doesn't catch the main complex organization of words. In the text words belong to different subgraphs, reflected in their clustering coefficient, depending on their functional role. In the model all vertices look equivalent in a single global hierarchical organization.

Another measure we are interested in is the counting of the occurrence of binary structures in the model during the evolution of the network. We show this result in Fig.6a comparing it with a line representing the measured value from our network. As we can see the *D-M model* misses the massive formation of binary structures we observe in the real network.

B. Extension of the model

We extend the D-M model to include the local behavior of language. We want to elaborate the D-M approach distinguishing when the PA attachment mechanism is local, that is when a new word is attached to one of its previous neighbors, or global.

We find that the probability that the preferentially chosen edges at each time step are part of a previously existing binary structure follows the power law

$$p(t) \approx 0.1t^{0.16}, \tag{5}$$

quite well. We try to implement this ingredient in the next model.

Model 2: We start with a chain of 20 connected vertices.

1. At each time step we add a new vertex (word) to the network. We link it to an old vertex of the network through the global PA, that is it will be linked to the vertex i, that is not part of its nearest neighborhood, with probability proportional to k_i.

2. For $m(t)-1$ times, where $m(t)$ is the measurable function (4), we perform the following operations: with probability $p(t)$ we link the last linked vertex to an old vertex through local PA, that is we link it to a node i in its nearest neighborhood with probability proportional to k_i; with probability $1-p(t)$ we link the last linked vertex to

BURA

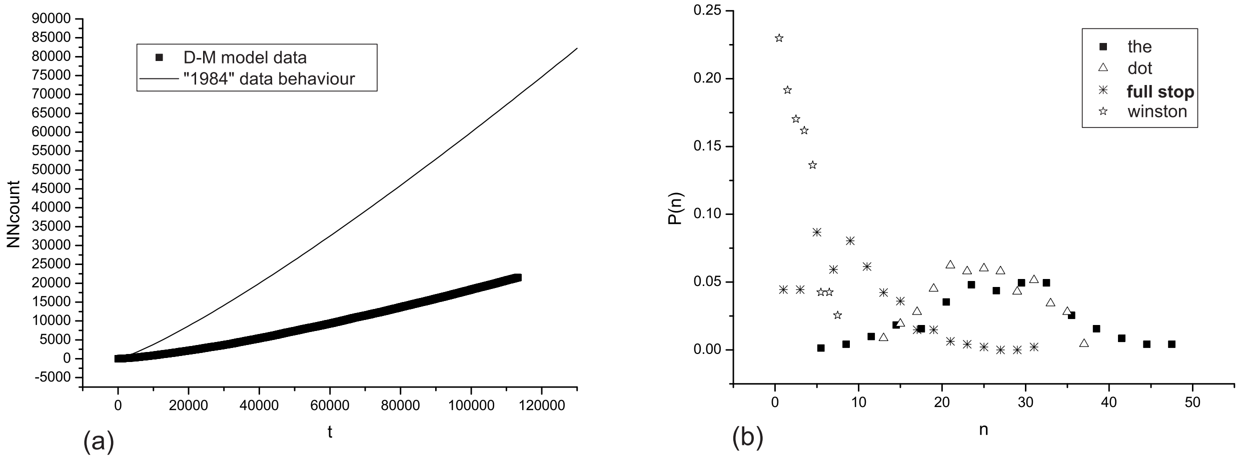

FIG. 6: (a) Count of the occurrence of binary structures during the evolution of the network for the *D-M model* compared to the real network data. (b) Probability distribution for the occurrence of some of the most frequent words in *1984*. To obtain this measure we partitioned the novel, each partition of 500 words, and counted the number of times n that each word appears in each partition. As we can see different words display different distributions. In particular "the" and the "full stop", that are structural words, display a Gaussian distribution while a word as "Winston", who is the first character in the novel, and is a meaningful word, follows an extremely different distribution.

an old vertex through global PA, that is the last linked vertex will be linked to the vertex i that is not part of its nearest neighborhood with probability proportional to k_i .

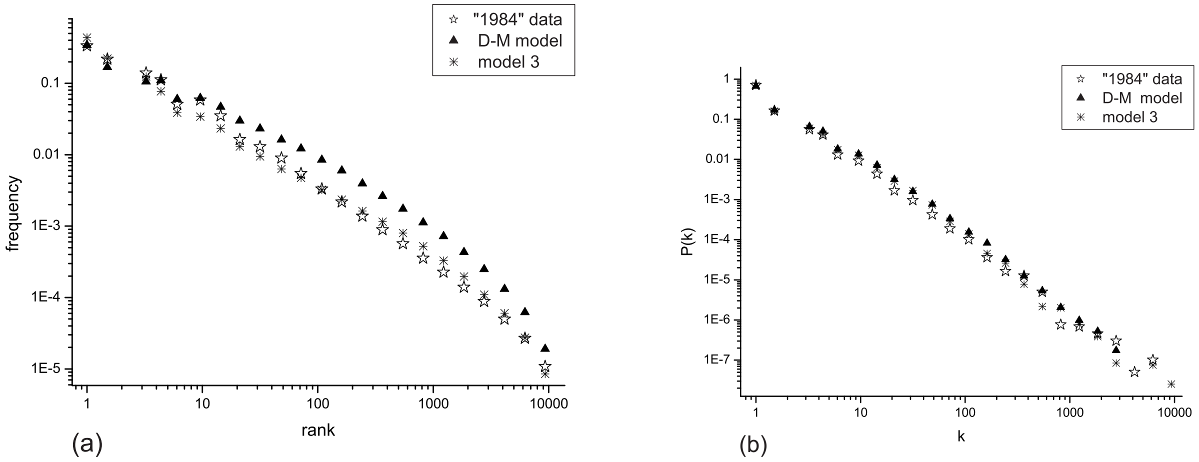

FIG. 7: Comparison between the *D-M model*, *model 3* and real data. (a) Zipf's Law. (b) Degree frequency count. In both plots logarithmic binning is used to reduce noise.

For this model the resulting average clustering coefficient is very low if compared to that of our network. It means, as we could expect, that the introduction of local PA supports selective local rules in the growth of the network that strongly limits the formation of triangles. We find $< c >= 0.08$, while in *1984* $< c >= 0.19$. Thus, by construction, *model 2* matches the analysis performed in Fig.6a for the growth of the network due to repetition of binary structures.

BURA

C. Model 3

The *D-M model* catches the average clustering and the global growth behavior of the network but misses the internal structure while *model 2* catches the global and nearest neighbor growth behavior of the real network but not the characteristic average clustering coefficient. Starting from this last model we would like to find a mechanism to increase the value of the average clustering coefficient.

It is now useful to consider the entropic analysis of language performed by Zanette and Montemurro in [15]. They found that different words in written human language display different statistical distributions, according to their function in the text. Making a partition of the text and counting the occurrence of each word in the partitions, they found that the most frequent words like punctuation and articles follow a Gaussian distribution, that is they are randomly distributed.

We show in Fig.6b a similar measure for the probability distribution for the occurrence of different functional words in the novel *1984*. Our analysis agrees with that in [15]. With this in mind, we import into our next model three apriori selected vertices, representing main punctuation and articles, with different growth properties to the other vertices in the network.

Model 3: We start with a chain of 20 linked vertices.

1. At each time step we add a new vertex (word) to the network. We link it to an old vertex of the network through the global PA, that is it will be linked to the vertex i that is not part of its nearest neighborhood with probability proportional to k_i.

2. For $m(t) - 1$ times, where $m(t)$ is the measurable function (4), we perform the following operations: with fixed probability $q = 0.05$ we link the last linked vertex to one of the three fixed vertices; with probability $p(t)$, given by (5), we link the last linked vertex to an old vertex through local PA, that is we link it to a node i in its nearest neighborhood with probability proportional to k_i; with probability $1 - p(t) - 3q$ we link the last linked vertex to an old vertex through global PA, that is the last linked vertex will be linked to the vertex i that is not part of its nearest neighborhood with probability proportional to k_i.

In the last model, with the introduction of the new random attachment mechanism, the average clustering coefficient increases and becomes $< c >= 0.20$, whilst preserving the global and the nearest neighbors growth counting.

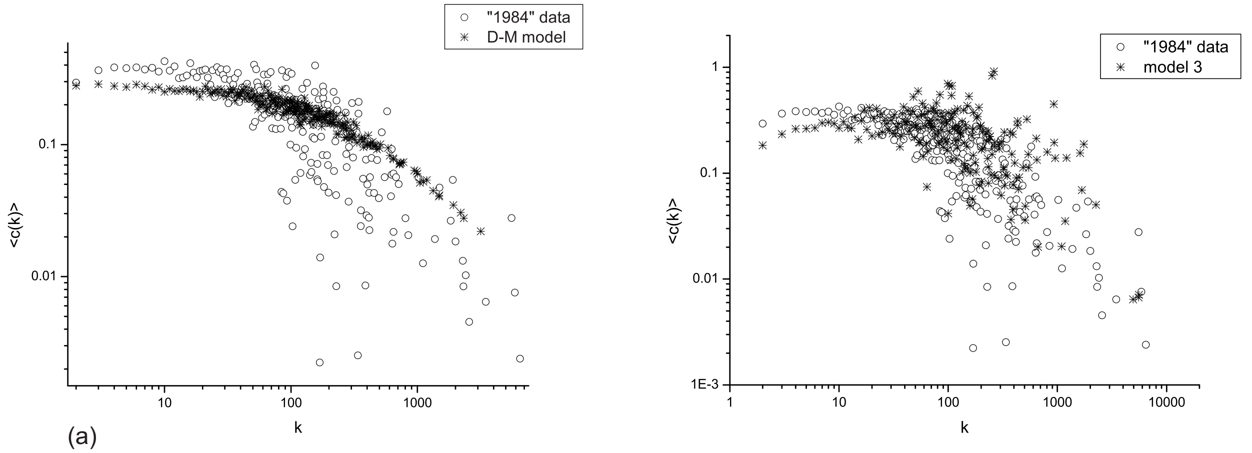

FIG. 8: Average clustering coefficient versus the degree of the vertices. (a) Comparison between *D-M model* and real data. (b) Comparison between *model 3* and real data.

The D-M model, model 3 and *1984* data are compared in Fig.7 and Fig.8. Zipf's analysis, in Fig.7a, reveals that model 3 is better able than the D-M model to describe the hierarchical behavior of the real data. In Fig.8b the measured average clustering coefficient for model 3 results in a spread distribution. This is evidence of a role differentiation in the vertices during the evolution of the network, that is a self-organization of the vertices in the different local structures. Nevertheless the main global statistical properties of the network, as shown in Fig.7b, are preserved.

BURA

IV. CONCLUSIONS

In this work we analysed in detail the topology of human written language, through a network representation of Orwell's *1984*. We performed average clustering coefficient and nearest neighbors analysis, finding that two different vertex behaviors clearly emerge. We performed entropic analysis that allowed us to distinguish different roles of words. We studied the relevance of second order correlations between vertices, finding that those are essential properties of the network architecture.

We proposed a model for matching the identified empirical behavior. This model included new growth mechanisms; a local preferential attachment and the allocation of a set of preselected vertices that have a structural rather than a functional purpose.

The degree of complexity of our model is greater than that of classical models, but we can give a simple explanation for the necessity to deal with such a degree of complexity. First of all, classical scale free network models often don't reflect the emergence of Zipf's Law, that is they don't consider the main hierarchical architecture of the network. The random attachment to a set of preselected vertices, representing the main punctuation, gives a solid structure to the network, between which the scale free system can grow and allows the rhythm of the novel to emerge from their punctuation.

We would like to stress that nearest neighbors analysis for our network display peculiar behavior, much more than those showed in this brief review, and should be considered as the basis for the understanding of network theory with a mixed local-global growth mechanism. These considerations are relevant for all natural systems showing syntactic-like organization rules, that are selective rules creating an intelligent ensemble from simple elements.

Further empirical and theoretical research would be useful. Human language is very important for the general study of network theory because of its great availability, the precision of the data and because we have a detailed knowledge of its local organizational rules.

Acknowledgments

This research is part of the NET-ACE project, supported by the EC. We would like to thank Pierpaolo Vivo for discussions and useful suggestions.

[1] R. Albert, A.L. Barabasi, Rev. Mod. Phys. **74**, 47 (2002).
[2] A.L. Barabasi, R. Albert, H. Jeong, Physica A **272**, 173 (1999).
[3] S.N. Dorogovtsev, J.F.F. Mendes, Proc. Roy. Soc. London B **268**, 2603 (2001).
[4] S.N. Dorogovtsev, J.F.F. Mendes, Contribution to *Handbook of Graphs and Networks: From the Genome to the Internet*, eds. S. Bornholdt and H.G. Schuster, Wiley-VCH, Berlin (2002).
[5] S.N. Dorogovtsev, J.F.F. Mendes, A.N. Samukhin, Phys. Rev. Lett. **85**, 21 (2000).
[6] R. Ferrer, R.V. Solé, Santa Fe Institute working paper 01-03-016 (2001).
[7] R. Ferrer, R.V. Solé, *Two regimes in the frequencies of words and the origin of complex lexicon*, J. Quant. Linguistic **8**, 165 (2001).
[8] R. Ferrer, R.V. Solé, R. Kohler, Phys. Rev. E **69**, 051915 (2004).
[9] R. Ferrer, A. Capocci , G. Caldarelli, cond-mat/0504165 (2005).
[10] D. Garlaschelli, M.I. Loffredo, Phys. Rev. Lett. **93**, 268701 (2004).
[11] P.L. Krapivsky, S. Redner, F. Leyvraz, Phys. Rev. Lett. **85**, 4629 (2000).
[12] W. Li, IEEE Transactions on Information theory **38**, 6 (1992).
[13] M.A. Montemurro, Physica A **300**, 567 (2001).
[14] M.A. Montemurro, D.H. Zanette, Adv. Complex Systems **5** (2002).
[15] E. Ravasz, A.L. Barabasi, Phys. Rev. E **67**, 026112 (2003).
[16] G. Orwell, *Nineteen Eighty-four*, Penguin Books Ltd Paperback, 1990.
[17] H.A. Simon, Biometrika **42**, 425 (1955).
[18] A. Vasquez, Phys. Rev. E **67**, 056104 (2003).
[19] D.H. Zanette, M.A. Montemurro, J. Quant. Linguistics **12**, 29 (2005).
[20] G.K. Zipf, *Human Behaviour and the Principle of Least Effort*, Addison-Wesley Press, 1949.
[21] http://www.online-literature.com/orwell/1984/