Physica A 387 (2008) 6863–6868

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Adjacency networks

C. Bedogne'*, G.J. Rodgers

Department of Mathematical Sciences, Brunel University, Uxbridge, Middlesex UB8 3PH, UK

ARTICLE INFO

Article history:
Received 19 February 2008
Received in revised form 21 July 2008
Available online 18 September 2008

PACS:
89.75.Da
89.75.Efa

Keywords:
Complex networks
Language networks
Scale-free networks

ABSTRACT

We consider a finite set $S = \{x_1, \ldots, x_r\}$ and associate to each element x_i a probability p_i. We then form sequences (N-strings) by drawing at random N elements from S with respect to the probabilities assigned to them. Each N-string generates a network where the elements of S are represented as vertices and edges are drawn between adjacent vertices. These structures are multigraphs having multiple edges and loops. We show that the degree distributions of these networks are invariant under permutations of the generating N-strings. We describe then a constructive method to generate scale-free networks and we show how scale-free topologies naturally emerge when the probabilities are Zipf distributed.

1. Introduction

In the last few years much attention has been drawn to the application of network theory to the study of human language [1–6]. Various network structures can be associated to a piece of prose, based on the different relations that can be established between words. For instance one could consider a network where words (represented as vertices) are connected in terms of a semantic relationship (such as synonymity) or in terms of a syntactical relationship (such as position or co-occurrence). In the following we will discuss only this second class of models, referring only to the *collocation* of words within a piece of prose. A network structure is then defined by representing all symbols (words and punctuation) appearing in a piece of prose as vertices and by drawing edges between pairs of adjacent symbols. Empirical work on the positional word web usually recovers both small-world properties (such as small characteristic path length and high clustering coefficients) and power-law degree distributions. It is very remarkable that these results do not depend on the particular language nor on the piece of prose considered. It also turns out, perhaps surprisingly, that some Asian languages (such as Mandarin) seem to behave much in the same way as Indo–European languages. It is therefore tempting to conjecture that a universal property of human language is being uncovered.

There have been attempts to explain the empirical data by introducing a modified preferential attachment mechanism interpolating between "pure" preferential attachment and an age-dependent edge formation process [2]. Other edge forming mechanisms have also been introduced, as for instance a combination of global and local preferential attachment [3], and a combination of preferential and random attachment [5].

Several conceptual objections, however, can be made to the universality of the preferential attachment mechanism [7–9]. For instance, in many situations it is not realistic to assume that a vertex has the complete information about the degree distribution it would need in order to know where to attach preferentially [7,8].

In the following, we will then introduce a new model attempting to capture some of the properties of language networks without relying on any preferential attachment mechanism. Although our model can only be considered as a toy model with

* Corresponding author.
E-mail address: cesare.bedogne@gmail.com (C. Bedogne').

0378-4371/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.physa.2008.09.001

respect to the complexities of human language, we believe that it may however help to capture some universal features of language networks (referring not only to human languages but to any "language" where the adjacency relationship is relevant, for instance in formal languages and DNA codes).

It is important to understand, however, that the mathematical structure underlying our model is a multigraph [10] rather than a simple graph. Vertices are thus permitted to have multiple edges and loops. The degree of a vertex is still defined as the number of distinct edges incident to the vertex but, in contrast with simple graphs, it is generally larger than the number of neighbours of the vertex. Since in the positional word web multiple edges are prevalent, a multigraph approach appeared to us more natural than the simple graph approach often encountered in literature.

In Section 2 we formally introduce adjacency multigraphs and study the invariance properties of their degree distributions. We also give a geometrical interpretation of the model in terms of Lebesgue measurable subsets of the unit interval.

The statistical properties of human language were first studied in Ref. [11], which can be considered the foundation of quantitative linguistics. It was shown empirically that, if the words of a piece of prose are ordered in their rank of occurrence (by assigning rank 1 to the most frequent word, rank 2 to second most frequent one, and so on) the frequency of words scales as a power-law of the rank (Zipf's Law)

$$f(r) = A\frac{1}{r^{\gamma}} \tag{1}$$

where r represents the rank and A is a constant. Zipf's law then allows one to derive the actual frequencies of words from just their *ordering* through ranking. Exponents $1 \leq \gamma \leq 2$ are typically encountered in the analysis of different languages. The significance of Zipf's law, however, is not specific to language as Zipf's distributed frequencies are also ubiquitous in demography, economics and geography [12].

In Section 3 we study the necessary and sufficient conditions for generating scale-free adjacency multigraphs. In particular, we show that scale-free topologies naturally follow from Zipf's law.

2. The adjacency model

We consider a set of r distinct elements (or "symbols") $S = \{x_1, \ldots, x_r\}$ and define a discrete probability $p : S \to [0, 1]$ by assigning to each element x_i a number p_i such that $\sum_{i=1}^{r} p_i = 1$. We then form ensembles of sets (N-strings) by drawing at random (with respect to the probabilities introduced above) an element from S at each time-step. Supposing that each draw is independent from the previous one, we assume that (when N/r is large enough) the frequency of a symbol x_i in a N-string is proportional to p_i. Note that we consider here a large but *finite* N.

Informally speaking the symbols $x_1, \ldots, x_r$ will represent the elemental units (atoms, words, numbers) of a "physical" system and the N-strings possible configurations (or states) of the system itself. We could formalise in this way a DNA sequence, an encoded message or a piece of language prose. In the last case the set S will consist of all symbols (words and punctuation) appearing in the text. In the following we will denote the generic N-string by the symbol $S^i_{N,r}$. Thus $S^i_{N,r}$ is a set of the form $S^i_{N,r} = (x_{i1}, \ldots, x_{iN})$, where x_{ij} belongs to S for every j.

Generalising the edge forming process usually introduced in language networks, we define an N-string dependent network by representing the elements of S as vertices and by drawing edges between adjacent vertices. The topology thus defined is related only to the local property of adjacency but not to global properties of the string such as its total linear order. In the following, we will then call the structures thus defined "adjacency networks". We will also say that N^i_S is the network induced (or generated) by the N-string $S^i_{N,r}$.

A remarkable property of adjacency networks is that, under a very mild statistical hypothesis, all random permutations of a given N-string give rise to the same multigraph degree distribution. Although it would be easy to give a formal proof of our claim, it is perhaps more convincing to illustrate our statement through an elementary example. Consider the symbols A, B, C, D, E, F and fix for instance $N = 11$. Consider then the two configurations (which will be related to probabilities which is not important to specify now) $S^1_{11,6} = (A, B, C, A, A, B, D, E, D, F, A)$, $S^2_{11,6} = (A, A, B, B, D, C, D, A, E, F, A)$ and notice that they differ only by a permutation of their elements. The networks induced by the two 11-strings (which for brevity will be denoted by N_1 and N_2) are clearly not isomorphic under *any* bijection $\sigma : N_1 \to N_2$. The two N-strings induce thus very different topologies but nevertheless the degree distribution in both cases is the same. This property clearly follows from the degree of x_i being proportional to p_i and from the fact that the extremal symbols happen to be the same in both strings. Suppose now that the first and the last symbols differ. It is a simple matter to show then that there exist then at most 4 vertices whose degree may change by ± 2. These variations are clearly irrelevant when N is large.

Notice that this small difficulty could also be more elegantly ruled out by representing the symbols along an oriented circle instead of in a string. In this case our result would be exact for all N-strings. Equivalently we could also generalise the adjacency relation by *defining* the first and last element of each string as being adjacent to each other. In this case, since each vertex's degree is expressed by an even number and since the multigraph we introduced is connected, it is possible (by the Euler–Hierholzer theorem [10]) to join all vertices with a circuit traversing each edge only once. We have thus implicitly defined an Eulerian multigraph [10]. Notice that, although the degree distribution is invariant under permutations, the distances between vertices are clearly not invariant. Notice also that, without altering the degree distribution, we