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

can always choose a "canonical configuration" where the symbols are ordered according to the increasing probabilities assigned to them (and where each distinct symbol is repeated in a row). A canonical configuration will thus be of the kind $(x_1, \ldots, x_1, \ldots, x_w, \ldots, x_w)$ and it can be divided into sub-groups of symbols where the first sub-group contains all symbols with probability p_1, the second sub-group all symbols with probability p_2 and so forth. In this case edges form only between the vertices x_i and x_{i+1} (for every i) and the vertex degree is mostly accumulated through loops. In this configuration, the length of the degree sequence coincides then with the network's diameter.

The degree distribution invariance has some significance to human language networks. Consider any large prose and the related network, equipped with the adjacency topology. Statistical analysis usually recovers power-law degree distributions (often with with two distinct regimes). It follows from our result that the same distributions would be recovered if the text is randomised. The scale-free properties of human language are thus unrelated to language semantics.

We conclude this section observing that the random sequences we have introduced can be naturally interpreted geometrically. Consider any decomposition of the unit interval $I = [0, 1]$ of the form $I = \bigsqcup_{i=1}^{r} E_i$, where the sets E_i are mutually disjoint Lebesgue measurable subsets of I. Suppose that $\mu(E_i) = p_i$ for every i, where μ is the Lebesgue measure on the real line. Since the measure is additive it must be $\sum_{i=1}^{r} p_i = 1$. On the other hand, given r real numbers $p_1, \ldots, p_r$ such that $\sum_{i=1}^{r} p_i = 1$, it is a simple matter to show that there exist measurable sets $E_1, \ldots, E_r$ such that $I = \bigsqcup_{i=1}^{r} E_i$ and $\mu(E_i) = p_i$. Notice also that, since the measure is in fact countably additive, we could straightforwardly generalise this discussion to countable decompositions of the unit interval.

We can now define an adjacency network through the following random experiment. Suppose to draw at random a number from I at each time-step t, generating a sequence $\{x(0), x(1), \ldots, x(t), \ldots\}$ (where $x(t)$ is the number drawn at time t). Given any Lebesgue decomposition of the unit interval, we can then consider the sequence of sets $\{E(0), E(1), \ldots, E(t), \ldots\}$ defined by the relation $x(t) \in E(t)$ at time t (where $E(t) \in \{E_1, \ldots, E_r\}$ for every t). In this way we have made our probabilistic approach rigorous by considering the *unbiased* random sequence formed by the various measurable sets E_i. Since a Lebesgue decomposition of the unit interval can be associated to each normalised discrete probability, our discussion is completely general.

In some cases the properties of adjacency networks can be expressively illustrated within the geometrical setting we described. For instance a partition of the unit interval of the kind $I = \coprod_{i=1}^{r} E_i$, where $\mu(E_i) = \mu(E_j)$ for every i and j, clearly generates an exponential degree distribution.

Consider instead a partition of the kind $I = \coprod_{i=0}^{r} E_i$ where $E_0 = [0, 1/2)$ and $E_i = (\frac{1}{2} + \frac{i}{2k}, \frac{1}{2} + \frac{i+1}{2k}]$ with $i = 1, \ldots, k-1$ (for large k). Since we can write the generating N-string in any order we please, we can regularly juxtapose E_0 (which appears with frequency 1/2) with sets E_i where $i > 1$. The resulting network will then display a huge hub connected to vertices of low degree, approaching a star network topology when $k \to \infty$.

3. Scale-free properties

In this section we discuss the following problem: given a real number $r \geq 1$ is it possible to determine a probability $p : S \to [0, 1]$ generating a scale-free adjacency network with power-law exponent γ? Intuitively, since we can create hubs just by repeating symbols in a N-string a sufficient number of times, all we need to do is to "match" the hubs' degrees appropriately, in order to satisfy the system constraints. We introduce then the sets

$$S_i = \{x \in S : p(x) = p_i\} \tag{2}$$

and partition S in the form $S = \bigsqcup_{i=1}^{w} S_i$, where w is the number of distinct values taken by p. Notice that S_i is the set of vertices with average degree $2Np_i$.

Normalisation can now be expressed by

$$\sum_{j=1}^{w} p_j \sharp(S_j) = 1 \tag{3}$$

and the partition equation is given by

$$\sum_{j=1}^{w} \sharp(S_j) = r. \tag{4}$$

In general the degree distribution of a finite network is expressed by

$$P(k) = \frac{\sharp(N_k)}{r} \tag{5}$$

where N_k is the set of vertices having degree k. In our case we may clearly limit ourselves to degrees of the form $k = 2Np_i$. Imposing a scale-free constraint

$$P(k) = \frac{\Lambda}{k^{\gamma}} \tag{6}$$

yields then

$$\sharp(S_i) = \frac{r\Lambda}{2^\gamma N^\gamma p_i^\gamma} \tag{7}$$

where for simplicity we have denoted $S_i = S_{2Np_i}$. The constant Λ is uniquely determined by normalisation and takes the form

$$\Lambda = \frac{2^\gamma N^\gamma}{r} \frac{1}{\sum_{i=1}^{w} \frac{1}{p_i^{\gamma-1}}}. \tag{8}$$

The partition equation provides a second expression for Λ

$$\Lambda = \frac{2^\gamma N^\gamma}{\sum_{i=1}^{w} \frac{1}{p_i^\gamma}}. \tag{9}$$

Comparing Eqs. (8) and (9) yields

$$r = \frac{\sum_{i=1}^{w} \frac{1}{p_i^\gamma}}{\sum_{i=1}^{w} \frac{1}{p_i^{\gamma-1}}}. \tag{10}$$

The scale-free constraint can now be expressed in the form

$$\sharp(S_i) = \frac{r}{p_i^\gamma \sum_{j=1}^{w} \frac{1}{p_j^\gamma}}. \tag{11}$$

Notice that normalisation is automatically satisfied whenever Eqs. (10) and (11) hold. We have thus shown that the adjacency network is scale-free and that p is a properly defined probability if and only if Eqs. (10) and (11) hold.

Notice also that Eq. (11) implies $\sharp(S_i) \neq \sharp(S_j)$ for $i \neq j$. Eq. (4) gives then

$$w < \frac{r}{\sharp(S_m)} \tag{12}$$

where

$$\sharp(S_m) = \mathrm{Min}\{\sharp(S_i), i = 1, \ldots, w\}. \tag{13}$$

Notice that the minimum is unique and strictly larger than 1. Indeed, supposing that $\sharp(S_1) = 1$, Eq. (11) gives $r = 1 + \sum_{j=1}^{w} 1/p_j^\gamma$. It follows now from Eq. (10) that $1 + \sum_{j=2}^{w} 1/p_j^\gamma = 1$, which is a contradiction. The same result clearly holds for every other index i and we can thus conclude that it must be $\sharp(S_i) > 1$ for every i.

Eq. (12) is clearly a very rough estimate for the length w of the degree sequence. A slightly better one can be achieved by observing that, since the sets S_i have different sizes, Eq. (4) yields $r > \frac{w(1+w)}{2}$. We have then

$$w < \frac{-1 + \sqrt{1 + 8r}}{2}. \tag{14}$$

It is important to understand that Eqs. (10) and (11) take into account only the *distinct* values taken by p. This suggests we can design a constructive method to generate scale-free adjacency networks. Suppose we wish to produce a scale-free network with a degree sequence of length w. Starting from an injective function (scale-free kernel) $p : S_w \longrightarrow (0, 1)$, defined on the set $S_w = \{x_1, \ldots, x_w\}$, we can then try to extend the kernel on the whole set S by repeating the values already taken by p on S_w in order to satisfy Eqs. (10) and (11). We would then have simultaneously generated a well-defined probability p on S and a scale-free adjacency network. Notice that this method, in the degenerate case where $S_w = \{x_1\}$ and $p(x_1) = 1/r$, yields an exponential network.

We study now two special cases to illustrate how the model can be solved explicitly. Supposing that the coefficients α_i are all distinct and that $\alpha_i < r$ for every i, we first consider kernels of the form

$$p_i = \alpha_i / r \tag{15}$$

with $i = 1, \ldots, w$. Eq. (10) takes now the form

$$\sum_{i=1}^{w} \left(\frac{1}{\alpha_i^\gamma} - \frac{1}{\alpha_i^{\gamma-1}} \right) = 0 \tag{16}$$

which admits solution only if $\gamma = 1$. We now show that it is possible to choose the coefficients α_i in such a way that the resulting adjacency network is scale-free with power law exponent $\gamma = 1$. Eq. (16), when $\gamma = 1$, gives

$$\sum_{i=1}^{w} \frac{1}{\alpha_i} = w \tag{17}$$

which is satisfied by

$$\alpha_i = \frac{1+w}{2i} \tag{18}$$

yielding the degree distribution

$$P(k) = \frac{2N}{wr}\frac{1}{k}. \tag{19}$$

In order to generate scale-free networks with *arbitrary* exponents, we introduce the kernel

$$p_i = \frac{1}{H(n,\alpha)i^{\alpha}} \tag{20}$$

where $i = 1, \ldots, w$. The generalised harmonic numbers $H(n,\alpha) = \sum_{j=1}^{n} \frac{1}{j^{\alpha}}$ depend on the parameter α ranging in the interval $(0, 1]$. Notice that Eq. (10) takes now the form

$$H(n,\alpha) = r\frac{\sum_{i=1}^{w} i^{\alpha(\gamma-1)}}{\sum_{i=1}^{w} i^{\alpha\gamma}}. \tag{21}$$

Since $\sum_{j=1}^{\infty} \frac{1}{j^{\alpha}} = \infty$, the left hand side of Eq. (21) approximates very well its right hand side when an appropriately large n is chosen. Exact equality may then be achieved with a slight change in the definition of p (we just need to change the definition of p on an element of S), whose influence on the network's topology is clearly negligible. We may now proceed to extend the scale-free kernel (Eq. (20)) on S by constructing sets S_i satisfying the scale-free constraint (Eq. (11)). Since Eq. (10) also holds, this extended function p is a well defined probability on S. We have thus generated a scale-free network with degree distribution

$$P(k) = \frac{\Lambda}{k^{\alpha\gamma}} \tag{22}$$

where

$$\Lambda = \frac{(2N)^{\gamma}}{\sum_{j=1}^{w} j^{\alpha\gamma} H(n,\alpha)^{\gamma}}. \tag{23}$$

Notice that the real numbers defined in Eq. (20) resemble the "weights" introduced in the static model [13]. However, while in the static model each vertex i is assigned a different weight w_i, in the adjacency model we explicitly construct sets of vertices (of appropriate size) sharing the same weight. The edge forming mechanisms, in the two models, are then completely different. In the static model edges between vertices i and j form with a probability proportional to $w_i w_j$. Thus it is more likely that pairs of vertices with a high weight connect rather than pairs of vertices with a small weight. This certainly does not happen in the adjacency model where, informally speaking, there are "many" vertices with a small weight and "just a few" vertices with a high weight. It is therefore unlikely (in any system configuration) that edges form between pairs of vertices of high weight. More generally, since the degree distribution is invariant under permutations of the generating N-string, it is not possible to associate a well-defined "probability" to the edge forming process introduced in the adjacency model. Notice also that, in sharp contrast with the static model, when the system is in the canonical configuration the vertex degree is mostly accumulated through loops.

Note that it is possible to generalise the constructive technique described above by associating a scale-free network to any positive injective function p defined on a finite set $S_w \subset S_\infty$ (where S_∞ is a countable set) and satisfying a simple constraint. Suppose for instance is p decreasing. We can always suppose in full generality, up to a proper rescaling of p, that $p(x_i) < 1$ for every x_i in $S_w = \{x_1, \ldots, x_w\}$.

Notice that if

$$p(x_i) < \frac{1}{w} \tag{24}$$

for $i = 1, \ldots, w$ then

$$\sum_{i=1}^{w} \frac{1}{p_i^{\gamma-1}} \left[\frac{1}{p_i} - w \right] > 0. \tag{25}$$

The number r *defined* by the right hand side of Eq. (10) is then larger than w for any choice of γ. We can then choose from $S_\infty r - w$ new symbols $x_{w+1}, \ldots, x_r$ and consider the set $S = \{x_1, \ldots, x_r\}$. It is now a simple matter to show that it is possible to extend the domain of the function p from S_w to S in such a way that Eq. (11) holds. This guarantees at the same time that the resulting adjacency network is scale-free (with power-law exponent γ) and that p is a properly defined probability on S.

We have thus shown that, given any integer w (which will give the desired length of the network's degree sequence) and any real number $\gamma \geq 1$, it is possible to generate a scale-free network with power-law exponent γ and with a degree sequence of length w.

In practice we can start with any decreasing function defined on the real line and consider then the equation $f(x) = w^{-1}$, which admits a unique solution $x = x_0$. Consider then the set $S_w = \{x_0 + 1, \ldots, x_0 + w\}$. Since Eq. (24) is clearly satisfied we can proceed as described above and generate a scale-free adjacency network depending on f. In particular, the dependence of the number r of vertices on the values taken by f is expressed by Eq. (10).

4. Conclusions

We have introduced adjacency networks generalising the edge formation mechanism studied in the context of human language networks. In Section 2 we have shown that their degree distributions are invariant under permutations of the generating N-strings. This property strongly suggests that the adjacency attachment rule cannot be recovered in terms of topologically based attachment rules (as for instance preferential attachment) nor on age-dependent attachment rules. Notice that, whenever an adjacency network displays a scale-free topology, a preferential attachment mechanism may be detected phenomenologically simply because the highest probabilities are then associated to the network's hubs. It is obvious however that this phenomenology has nothing to do with the "real" mechanism at work as clearly shown, for instance, when the canonical configuration is considered.

In Section 3 we concentrated on scale-free properties studying how power-law degree distributions can be explicitly designed. We have shown that a scale-free network can be generated whenever Eqs. (10) and (11) hold. Explicit solutions can be obtained through a constructive method relying on the extension of a scale-free kernel, defined on a set of w symbols.

We have shown two explicit solutions of the model. If p is given as in Eq. (15) it is possible to generate a scale-free network only in the case $\gamma = 1$. If p is given instead as in Eq. (20), it is possible to tune the system's parameters in order to generate scale-free networks with arbitrary power-law exponents. In particular our method shows how scale-free adjacency topologies, when interpreted in the multigraph sense, naturally emerge from Zipf distributed frequencies.

Acknowledgments

We thank the European Union Marie Curie Program (NET-ACE project, contract number MEST-CT-2004-006724) for financial support. We also wish to thank A.P. Masucci for stimulating and helpful discussions.

References

[1] R. Albert, A.L. Barabási, Rev. Modern Phys. 74 (2002) 47.
[2] S.N. Dorogovtsev, J.F.F. Mendes, A.N. Shamukin, Phys. Rev. Lett. 85 (2000) 4633.
[3] S.N. Dorogovtsev, J.F.F. Mendes, Proc. Roy. Soc. London B 268 (2001) 2606.
[4] M. Markosova, Physica A 387 (2008) 661–666.
[5] R.F. Cancho, R. Sole, Proc. Roy. Soc. London B 268 (2001) 2261.
[6] A.P. Masucci, G.J. Rodgers, Phys. Rev. E 74 (2006) 046115.
[7] G. Caldarelli, A. Capocci, P. De Los Rios, M.A. Munoz, Phys. Rev. Lett. 89 (2002) 258702.
[8] V.D.P. Servedio, G. Caldarelli, Phys. Rev. E 70 (2002) 056126.
[9] C. Bedogne', G.J. Rodgers, Phys. Rev. E 74 (2006) 046115.
[10] B. Bollobás, Random Graphs, Academic Press, London, 1985.
[11] G.K. Zipf, Human Behaviour and the Principle of Least Effort, Addison-Wesley, Cambridge, MA, 1949.
[12] M.E. Newman, Phy. Rev. E 70 (2004) 056131.
[13] K.I. Goh, B. Kang, D. Kim, Phys. Rev. Lett. 87 (2001) 278701.