Modeling discrete combinatorial systems as alphabetic bipartite networks: Theory and applications

Monojit Choudhury
Microsoft Research India, 196/36 2nd Main Sadashivnagar, Bangalore 560080, India

Niloy Ganguly, Abyayananda Maiti, and Animesh Mukherjee
Department of Computer Science and Engineering, Indian Institute of Technology Kharagpur, 721302 Kharagpur, India

Lutz Brusch and Andreas Deutsch
Zentrum für Informationsdienste und Hochleistungsrechnen, Technische Universität Dresden, Zellescher Weg 12, 01069 Dresden, Germany

Fernando Peruani*
CEA-Service de Physique de l'Etat Condensé, Centre d'Etudes de Saclay, 91191 Gif-sur-Yvette, France
(Received 3 November 2008; revised manuscript received 1 December 2009; published 5 March 2010)

Genes and human languages are *discrete combinatorial systems* (DCSs), in which the basic building blocks are finite sets of elementary units: nucleotides or codons in a DNA sequence, and letters or words in a language. Different combinations of these finite units give rise to potentially infinite numbers of genes or sentences. This type of DCSs can be represented as an alphabetic bipartite network (ABN) where there are two kinds of nodes, one type represents the elementary units while the other type represents their combinations. Here, we extend and generalize recent analytical findings for ABNs derived in [Peruani *et al.*, Europhys. Lett. **79**, 28001 (2007)] and empirically investigate two real world systems in terms of ABNs, the codon gene and the phoneme-language network. The one-mode projections onto the elementary basic units are also studied theoretically as well as in real world ABNs. We propose the use of ABNs as a means for inferring the mechanisms underlying the growth of real world DCSs.

DOI: 10.1103/PhysRevE.81.036103 PACS number(s): 89.75.Fb

I. INTRODUCTION

Two of the greatest wonders of evolution on earth, genes and human languages, are *discrete combinatorial systems* (DCSs) [1]. The basic building blocks of DCSs are finite sets of elementary units, such as the letters in a language and nucleotides (or codons) in a DNA sequence. Different combinations of these finite elementary units give rise to a potentially infinite number of words or genes. Here, we analyze a special class of complex networks as a model of DCSs. We shall refer to them as *alphabetic bipartite networks* (ABNs) in order to express the fact that the set of basic units, in both human and genetic languages, can be considered as an *Alphabet*.

The ABNs are a subclass of bipartite networks (BNs). BNs have two disjoint partitions and edges link nodes from one to the other partition, but never nodes belonging to the same partition. In most of the BNs studied in the past both the partitions grow with time. Typical examples of this type of networks include collaboration networks such as the movie actor [2–6], article author [7–9], and board-director [10,11] networks. In the article-author network, for instance, the articles and authors are the elements of the two partitions also known as the *ties* and *actors*, respectively. An edge between an author a and an article m indicates that a has coauthored m. The authors a and a' are *collaborators* if both have coauthored the same article, i.e., if both are connected to the same node m. The concept of *collaboration* can be extended to represent, through BNs, several diverse phenomena such as the city-people network [12], in which an edge between a person and a city indicates that the person has visited that particular city, the signal-object network in linguistics [13], where an edge between an object j and a signal i represents that a possible meaning of the i sign is the object j, the bank company [14] or donor-acceptor network that accounts for injection and merging of magnetic field lines [15].

Several models have been proposed to synthesize the structure of these BNs, i.e., when both the partitions grow unboundedly over time [2–5,16]. It has been found that for such growth models, when each incoming *tie* node *preferentially* attaches itself to the *actor* nodes, the emergent degree distribution of the *actor* nodes follows a power law [2]. This result is reminiscent of unipartite networks where preferential attachment results in power-law degree distributions [17].

Although there have been some works on nongrowing BNs [18,19], those like ABNs, where one of the partitions remains fixed over time, have received comparatively much less attention. In ABNs the partition that represents the basic units in DCSs (e.g., letters, codons) is finite and constant over time. In contrast, the partition that represents the discrete combinations of basic units (e.g., words, genes) can grow unboundedly over time. Notice that the order in which the basic units are strung to form the discrete combination is

*Corresponding author; ferperuani@gmail.com

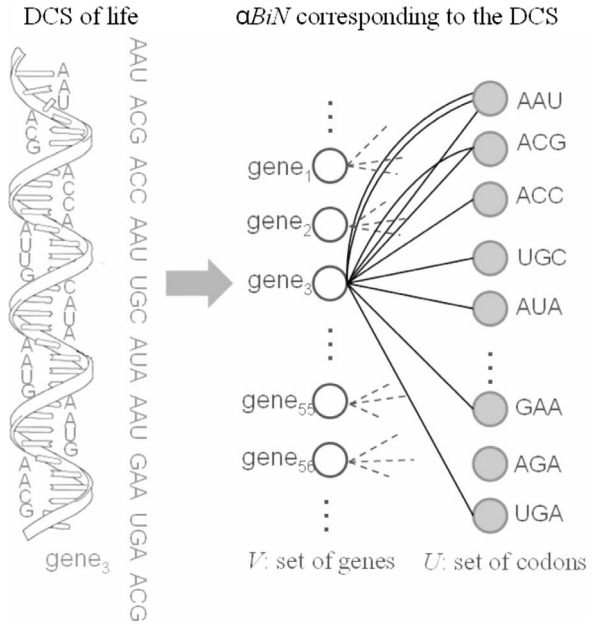

FIG. 1. DNA modeled as a bipartite network (ABN). The set U consists of 64 codons, whereas the collection V of genes is virtually infinite. Multiple occurrences of a codon in a gene have been represented here by multiedges. For instance, the codons "ACG" and "AAU" have, respectively, 2 and 3 edges connecting to the node gene_3. Alternatively, this could have been represented by single edges with weights 2 and 3.

an important and indispensable aspect of the system, which can be modeled within the framework of ABNs by allowing ordering of the edges. Nevertheless, the scope of the present work is limited to the analysis of unordered combinations. Here, we assume a word to be a bag of letters and a gene a collection of codons. Figure 1 illustrates the concepts through the example of genes and codons.

As far as we know, the first empirical evidence of the nonscale free character of the degree distribution of ABNs has been reported in [20], while the first systematic and analytical study of such BNs has been presented in [21]. The growth model for ABNs, proposed in [21], is based on preferential attachment coupled with a tunable randomness component. According to this model, there is a free parameter γ which controls the relative weight of preferential to random attachment, thereby, regulating the randomness present in the connections of the network. The growth model introduced in [21] assumes that the edges are incorporated one by one. Under this assumption, the exact expression for the emergent degree distribution of the basic units has been derived. It has been shown that it approaches a β-distribution asymptotically with time.

In this paper, we generalize the results derived in [21] to include the situation in which multiple edges are incorporated to the system at each time step. This extension accounts for the fact that in DCS the elements representing the discrete combination, as genes or words, are formed by multiple basic units. The use of ABN as a modeling tool of the temporal evolution of DCS, where at each time step, a new discrete combination element is added to the system, requires this generalization. Here, we derive the exact growth model for such processes and study the degree distribution of the one-mode projection of the network onto the basic units. In order to illustrate how the proposed framework can be used as an analytical tool to study and interpret empirical data, we applied the ABN theory to two well-known DCSs from the domain of biology and language. Through the analyses of the empirical data, we show the advantages and limitations of ABNs as a modeling approach. We start with the codon-gene network where codons play the role in the basic units while genes are the discrete combinations of them (see Fig. 1), and observe that the higher the complexity of an organism, the higher the value of the randomness parameter, i.e., γ. The analysis suggests that codon usage can be used to classify organisms. We then apply the ABN theory to the phoneme-language network, where phonemes are the basic units and the sound systems of languages are the discrete combinations, and show that the distribution of consonants over the languages of the world can be satisfactorily described. The study also illustrates certain limitations of the ABN growth model. For instance, we show that the topological characteristics of the network of co-occurrence of phonemes, which is the one-mode projection of the aforementioned network, is different from the theoretical prediction derived from a simple ABN model. This indicates that although the ABN growth model succeeds in explaining the degree distribution of the basic units, the theory fails to describe the one-mode projection. This points to the fact that the real dynamics of the system is much more complex.

Finally, we contextualize the developed ABN theory in the framework of Urn models and discuss the similarities and differences with the Finite Pólya's process [22,23] and the rewiring model suggested by Evans and Plato [19].

The article is organized as follows. Sec. II, formally defines ABN and introduces the growth model and its corresponding theoretical analysis. The two real networks—codon gene and phoneme language—their topology and comparison with the theoretical models are described in Sec. III. In Sec. IV, we place ABN theory in the context of Urn models of probability theory. The concluding section summarizes the obtained results and discusses the broader consequences of the present work.

II. THEORETICAL FRAMEWORK FOR ABN

A. Formal definition and modeling

A bipartite graph G is a three-tuple $\langle U,V,E\rangle$, where U and V are mutually exclusive finite collections of nodes (also known as the two partitions) and $E \subseteq U \times V$ is the collection of edges that run between these partitions. We can also define E as a multiset whose elements are drawn from $U \times V$. Clearly, the last definition of E allows multiple edges between a pair of nodes and the number of times the nodes $u \in U$ and $v \in V$ are connected can be assumed to be the weight of the edge (u,v). Note that although we are defining E to be a collection of ordered tuples, the ordering is an implicit outcome of the fact that edges only run between nodes in U and V. In essence, we do not mean any directedness of the edges.

ABNs are a special type of bipartite networks, where one of the partitions represents a set of basic units while the other partition represents their combinations. The set of basic units is essentially finite and fixed over time. Let us denote the basic units by the nodes in U. Let each unique discrete combination of the basic units be denoted as a node in V. There exists an edge between a basic unit $u \in U$ and a discrete combination $v \in V$ if u is a part of v. If u occurs w times in v, the weight of the edge (u,v) is w. Figure 1 illustrates these concepts through the example of genes and codons.

B. Growth model

The growth of ABNs is described in terms of a simple model based on preferential attachment coupled with a tunable randomness parameter. Suppose that the partition U has N nodes labeled as u_1 to u_N. At each time step, a new node is introduced in the partition V which connects to μ nodes in U based on a predefined attachment rule. Thus, in the model the time refers to the number of nodes in V. Let v_t be the node added to V during the t-th time step, and $\tilde{A}(k_i^t)$ the probability of attaching a new edge to a node u_i, where k_i^t refers to the (weighted) degree of the node u_i at time t. The attachment kernel $\tilde{A}(k_i^t)$ takes the form,

$$\tilde{A}(k_i^t) = \frac{\gamma k_i^t + 1}{\sum_{j=1}^{N} (\gamma k_j^t + 1)}, \tag{1}$$

where the sum in the denominator runs over all the nodes in U, and γ is the tunable parameter which controls the relative weight of preferential to random attachment. Thus, the higher the value of γ, the lower the randomness in the system. Note that the numerator of the attachment kernel could be rewritten as $k_i^t + \alpha$, where $\alpha = 1/\gamma$ is a positive constant usually referred to as the *initial attractiveness* [24].

Any ABN has two characteristic degree distributions corresponding to its two partitions U and V. Here, we assume that each node in V has degree μ and concentrate on the (weighted) degree distribution of the nodes in U. Let $p_{k,t}$ be the probability that a randomly chosen node from the partition U has degree k after t time steps. We assume that initially all the nodes in U have degree 0 and there are no nodes in V. Therefore,

$$p_{k,0} = \delta_{k,0}, \tag{2}$$

where $\delta_{k,0}$ is delta Kronecker. Therefore, the expression for the evolution of $p_{k,t}$ has the form,

$$p_{k,t+1} = \left[1 - \sum_{i=1}^{\mu} \hat{A}(k,i,t) \right] p_{k,t} + \sum_{i=1}^{\mu} \hat{A}(k-i,i,t) p_{k-i,t}, \tag{3}$$

where $\hat{A}(k,i,t)$ represents the probability at time t of a node of degree k of receiving i new edges in the next time step. The term $\Sigma_{i=1}^{\mu} \hat{A}(k,i,t) p_{k,t}$ describes the number of nodes of degree k at time t that change their degree due to the attachment of 1, 2,…, or μ edges. On the other hand, nodes of degree k will be formed at time $t+1$ by the nodes of degree $k-1$ at time t that receive 1 edge, nodes of degree $k-2$ at time t that receive 2 edges, and so on. This process is described by the term $\Sigma_{i=1}^{\mu} \hat{A}(k-i,i,t) p_{k-i,t}$.

Next, we derive an expression for $\hat{A}(k,i,t)$. We start out by a simple case, $\gamma=0$. Since in this case the probability for an edge of attaching to a node is independent of its degree, if we add μ edges, the probability for a node of receiving a single edge is $\mu(1/N)(1-1/N)^{\mu-1}$, the probability of receiving two edges is $\binom{\mu}{2}(1/N)^2(1-1/N)^{\mu-2}$, and for the general case we obtain the expression,

$$\hat{A}(k,i,t) = \binom{\mu}{i} \left(\frac{1}{N} \right)^i \left(1 - \frac{1}{N} \right)^{\mu-i}. \tag{4}$$

Thus, the probability of receiving i new edges is binomially distributed over i irrespective of the degree of the node. To extend this result to $\gamma > 0$, we recall that if we add a single edge, the probability for a node of degree k of receiving that edge is $\phi = (\gamma k + 1)/(\mu\gamma t + N)$, where we have assumed that previously to this edge we had added μt edges to the nodes in U. Clearly, $1-\phi$ is the probability for the edge to attach to some other node. Taking this into account, Eq. (4) is generalized for $\gamma \geq 0$ as

$$\hat{A}(k,i,t) = \binom{\mu}{i} \left(\frac{\gamma k + 1}{\mu\gamma t + N} \right)^i \left(1 - \frac{\gamma k + 1}{\mu\gamma t + N} \right)^{\mu-i}. \tag{5}$$

Inserting expression (5) into Eq. (3), we obtain:

$$p_{k,t+1} = \left[1 - \sum_{i=1}^{\mu} \binom{\mu}{i} \left(\frac{\gamma k + 1}{\mu\gamma t + N} \right)^i \left(1 - \frac{\gamma k + 1}{\mu\gamma t + N} \right)^{\mu-i} \right] p_{k,t} + \sum_{i=1}^{\mu} \binom{\mu}{i} \left(\frac{\gamma(k-i) + 1}{\mu\gamma t + N} \right)^i \left(1 - \frac{\gamma(k-i) + 1}{\mu\gamma t + N} \right)^{\mu-i} p_{k-i,t}. \tag{6}$$

The terms between parentheses in the first line of Eq. (6) can be simplified recalling that

$$\left(1 - \frac{\gamma k + 1}{\mu\gamma t + N} \right)^{\mu} = 1 - \sum_{i=1}^{\mu} \binom{\mu}{i} \left(\frac{\gamma k + 1}{\mu\gamma t + N} \right)^i \left(1 - \frac{\gamma k + 1}{\mu\gamma t + N} \right)^{\mu-i}.$$

Therefore, Eq. (6) can be rewritten by including $i=0$ in the sum, whereby we obtain

$$p_{k,t+1} = \sum_{i=0}^{\mu} \binom{\mu}{i} \left[\frac{\gamma(k-i) + 1}{\mu\gamma t + N} \right]^i \left[1 - \frac{\gamma(k-i) + 1}{\mu\gamma t + N} \right]^{\mu-i} p_{k-i,t}. \tag{7}$$

In [21], a similar growth model was analyzed under the simplistic condition that the degree of an U node cannot receive more than one edge per time step. This assumption is fairly reasonable for $\mu \ll N$ and small values of γ. Then the solution to Eq. (7) can be reasonably approximated by,

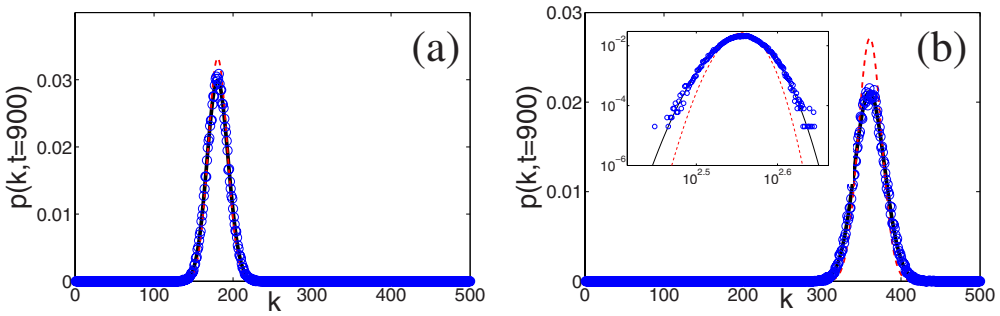

FIG. 2. (Color online) Comparison for random attachment ($\gamma=0$) between the numerical integration of Eq. (7) (solid black curve) and stochastic simulations (blue circles). Circles correspond to average over 500 simulations. In both figures $N=100$. (a) corresponds to $\mu=20$ while (b) to $\mu=40$. The dashed red curve corresponds to the approximation given by Eq. (8). The inset in (b) shows in log-log scale the deviation between Eq. (8), the integration of Eq. (7), and simulations.

$$p_{k,t} = \binom{t}{k} \frac{\prod_{i=0}^{k-1} (\gamma i + 1) \prod_{j=0}^{t-1-k} \left(\frac{N}{\mu} - 1 + \gamma j \right)}{\prod_{m=0}^{t-1} \left(\gamma m + \frac{N}{\mu} \right)}. \tag{8}$$

The expression given by Eq. (8) was derived in [21], and is only an exact solution of Eq. (7) for $\mu=1$. As indicated in [21], for $\gamma>0$, Eq. (8) approaches, asymptotically with time, a β distribution, $p_{k,t} \simeq C^{-1}(k/t)^{\gamma^{-1}-1}(1-k/t)^{\eta-\gamma^{-1}-1}$, where C is the normalization constant and $\eta=N/(\gamma\mu)$. From this expression it is clear that for $\mu=1$ there are four regimes associated to γ, a) A binomial distribution for $\gamma=0$, b) a skewed distribution which exhibits a mode that shifts with time for $0<\gamma<1$, c) a monotonically decreasing distribution with the mode at $k=0$, for $1 \leq \gamma \leq (N/\mu)-1$, and d) a u-shaped distribution, for $\gamma>(N/\mu)-1$.

However, the exact solution of Eq. (3) is not known in general, and Eq. (7) has to be numerically computed. In Fig. 2, we compare for random attachment ($\gamma=0$) the integration of Eq. (7) and stochastic simulations. Equation (8) provides a reasonable approximation of the process as long as γ is very small. Already for $\gamma \geq 1$, Eq. (8) deviates from the exact solution which corresponds to the integration of Eq. (7). Figures 2 and 3 show that Eq. (7) accurately describes stochastic simulations for all parameter values.

C. One-mode projection

In this section, we analyze the degree distribution of the one-mode projection of ABNs onto the set U. Formally, for an ABN $\langle U,V,E\rangle$, the one-mode projection onto U is a graph $G_U:\langle U,E_U\rangle$, where $u_i,u_j \in U$ are connected [i.e., $(u_i,u_j) \in G_U$] if there exists a node $v \in V$ such that $(u_i,v) \in E$ and $(u_j,v) \in E$. If there are w such nodes in V which are connected to both u_i and u_j in the ABN then we say that in the one-mode projection G_U there is an edge linking u_i and u_j with weight w. In the context of the codon-gene network, the one-mode projection is a codon-codon network, where two codons are connected by an edge with a weight that represents the number of genes, in which both of these codons occur. The one-mode projection of an ABN provides insight into the relationship between the basic units. For instance in linguistics the one-mode projection of the word-sentence ABN reveals the co-occurrence of word pairs, which in turn provides crucial information about the syntactic and semantic properties of the words (see, for example [25,26]).

Let us use the symbol $p_u(k,t)$ to refer to the probability that a randomly chosen node from the one-mode projection of an ABN with t nodes in V (i.e., after t time steps) has degree k. Consider a node $u \in U$ that has degree k in the ABN. Therefore, u is connected to k nodes in V, each of which is connected to $\mu-1$ other nodes in U. Defining the degree of a node as the number of edges attached to it, in the one-mode projection, u has a degree of $q=k(\mu-1)$. Conse-

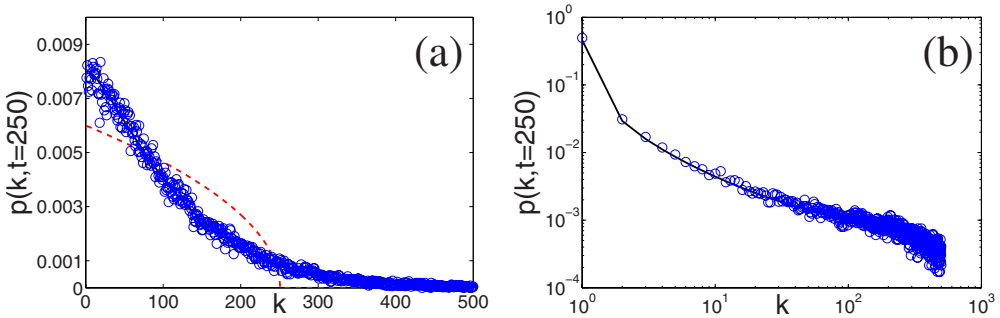

FIG. 3. (Color online) Comparison for strong preferential attachment ($\gamma \geq 1$) between the integration of Eq. (7) (solid black curve) and stochastic simulations (blue circles). Blue circles correspond to an average over 500 runs. In both figures $N=100$ and $\mu=40$. (a) corresponds to $\gamma=1$ while (b) to $\gamma=16$. The dashed red curve in (a) indicates the approximation given by Eq. (8), which falls out of the range of the figure in (b).

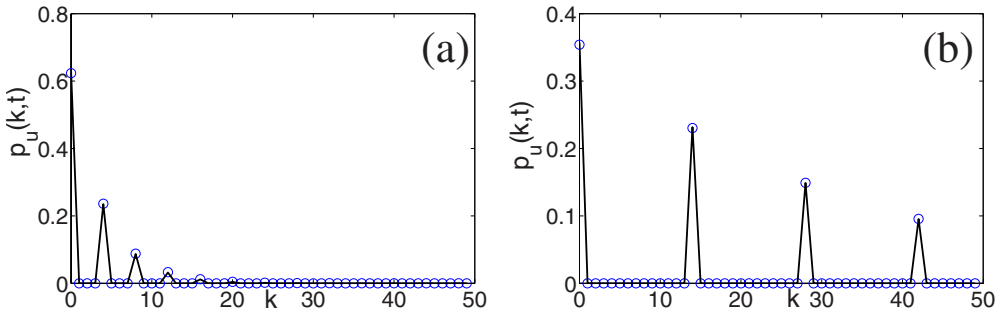

FIG. 4. (Color online) Comparison between stochastic simulations for the one-mode projection (blue circles) and Eq. (9) (solid black curve). In both figures $N=500$ and $\gamma=1$. Blue circles correspond to averages over 1000 simulations. In (a) $\mu=5$ while in (b) $\mu=15$.

quently, the degree distribution of G_U, $p_u(q,t)$, is related to $p_{k,t}$ in the following way:

$$p_u(q,t) = \begin{Bmatrix} p_{0,t} & \text{if } q=0 \\ p_{k=q/(\mu-1),t} & \text{if } \mu-1 \text{ divides } q \\ 0 & \text{otherwise} \end{Bmatrix}. \quad (9)$$

Figure 4 shows a comparison between stochastic simulations and Eq. (9). Notice that this mapping simply implies that $p_u(q=0,t)=p_{0,t}$, $p_u(q=\mu-1,t)=p_{1,t}$, $p_u(q=2(\mu-1),t)=p_{2,t},\ldots,p_u(q=j(\mu-1),t)=p_{j,t}$. The same result can be derived by using the generating function based technique described in Eq. 70 of [27]. It is worth noticing that we have assumed that the weight of edges in G_U is one. Clearly, this is not true in general.

Thresholded one-mode degree distribution

In the general case we have to consider that G_U is a weighted graph. The one-mode projections of ABNs can be converted to an unweighted graph by the process of *thresholding*. A thresholded one-mode projection graph (thresholded G_U) is constructed by replacing every weighted edge in G_U by a single edge iff the weight of that edge exceeds the threshold value τ; otherwise, the edge is deleted. Thresholded degree distributions are more popular in the complex network literature, than their weighted counterparts (see, for example [25,28]). We shall denote the degree distribution, thresholded at τ, as $p_u(q,t;\tau)$.

Let us start by considering two nodes u and u' in U with degrees k_u and $k_{u'}$, respectively. We now try to derive an expression for the probability $p(k_u,k_{u'},m)$ that there are exactly m nodes in V that are linked simultaneously to both u and u'. In other words, $p(k_u,k_{u'},m)$ is the probability that the weight between u and u' is m in G_U, given that the degrees of the nodes are k_u and $k_{u'}$ in U. Let us assume that the μ nodes that each node $v \in V$ is connected to, are all distinct. By the definition of the growth model for ABNs, the event of u being connected to a node v is independent of u' being connected to the same node. Therefore, the probability that a randomly chosen node $v \in V$ is connected to u is k_u/t and the probability that it is connected to u' is $k_{u'}/t$. Recall that t refers to the number of nodes in V. Thus, the probability that v is connected to both u and u' is $k_u k'_u/t^2$. Therefore, the probability that u and u' share m nodes in V takes the form,

$$p(k_u,k_{u'},m) = \binom{t}{m}\left(\frac{k_u k_{u'}}{t^2}\right)^m\left(1-\frac{k_u k_{u'}}{t^2}\right)^{t-m}. \quad (10)$$

From Eq. (10), the probability for u and u' of sharing an edge in the thresholded G_U is easily computed as,

$$p(k_u,k_{u'};m>\tau) = \sum_{m=\tau+1}^{t} p(k_u,k_{u'},m). \quad (11)$$

Consequently, in the thresholded G_U, the expected degree D of a node u whose degree is k in the ABN is given by,

$$D(k,\tau) = N\sum_{i=1}^{t} p_{i,t}p(k,i;m>\tau). \quad (12)$$

Notice that then $p_{k,t}$ can be interpreted as the probability of finding a randomly chosen node with degree $D(k,\tau)$ in the thresholded one-mode projection. Thus, the degree distribution of the thresholded G_U is computed as

$$p_u(q,t;\tau) = \sum_{q=\lfloor D(k,\tau)\rfloor} p_k, \quad (13)$$

where the function $\lfloor a\rfloor$ returns the largest integer smaller than a.

Figure 5 shows a comparison between Eq. (13) and stochastic simulations for the one-mode projection at different times. The implementation of Eq. (13) was done by summing over the $p_{k,t}$ obtained from the stochastic simulations of the corresponding ABN according to $q=\lfloor D(k,\tau)\rfloor$, as indicated by Eq. (12).

III. REAL WORLD ABNS

A. Codon-gene network

As complete genomes of more and more organisms are sequenced, phylogenetic trees reconstructed from genomic data become increasingly detailed. Codon usage patterns in different genomes can provide insight into phylogenetic relations. However, except for a few earlier works [29], studies on the codon usage have not received much attention. One of the main research issues here is to understand the influence of randomness in the growth pattern of genome sequences in connection to biological evolution. A well-known random process in evolutionary biology is a *random mutation* in a gene sequence. A gene sequence is a string defined over four

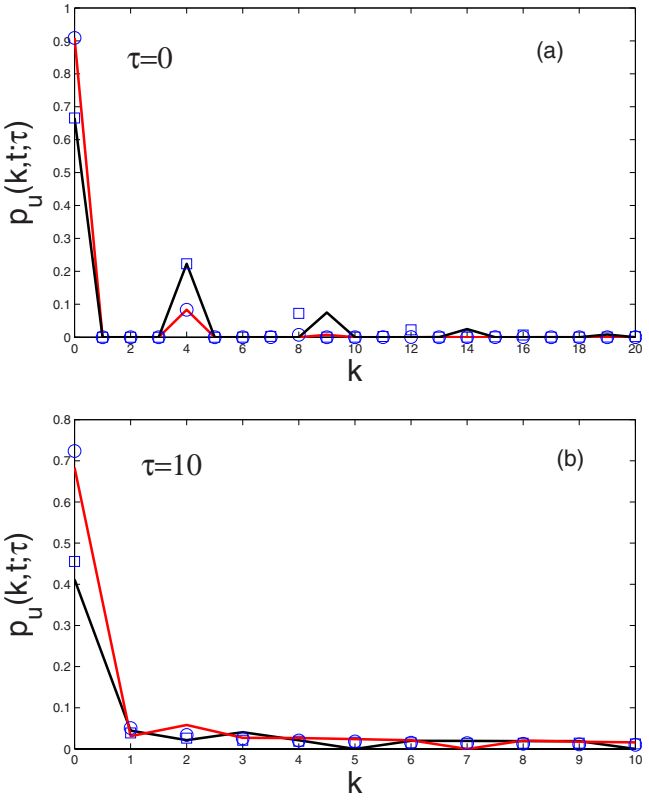

FIG. 5. (Color online) Comparison between stochastic simulations for the one-mode projection at different times (symbols) and Eq. (13) (solid curves). In (a) $\tau=0$, $N=1000$, $\mu=5$, $\gamma=1$. The blue circles and the red (light gray) curve correspond to $t=20$, while the blue squares and the black curve to $t=100$. The slight deviation of the simulation results from the theoretical predictions is due to the rounding of the values. In (b) $\tau=10$, $N=100$, $\mu=20$, $\gamma=1.5$. The blue circles and the red (light gray) curve correspond to $t=50$, while the blue squares and the black curve to $t=100$.

symbols (A, G, T, and C) that represent the nucleotides. A *codon* is a triplet of adjacent nucleotides (e.g., AGT, CTA) and codes for a specific amino acid (plus stop and start sequences). There are only 64 codons. Interestingly, the relation codon-amino acid is not bijective, and several codons can code for the same amino acid. Codon usage in genome sequences varies among different phylogenetic groups.

1. Definition and construction

We represent the codon-gene network as an ABN, where V is the collection of *genes*, i.e., the genome of the organisms, and U is the set of nodes labeled by the codons. There is an edge $(u,v) \in E$ that runs between V and U if and only if the codon u occurs in the gene v. Figure 1 illustrates the structure of the network.

We have analyzed eight organisms belonging to widely different phylogenetic groups. These organisms have been extensively studied in biology and genetics [30] and most importantly, their genomes have been fully sequenced. In Table I, we list these organisms along with a short description and the number of genes and codons (i.e., the cardinality of V). The data have been obtained from the Codon Usage Database [31,32]. The usage of a particular codon in an organism's genome sequence can be as high as one million. In other words, the weighted degree of the nodes in U can be arbitrarily large. This, together with the fact that there are only 64 nodes in U, presents us with the nontrivial task of estimating the probability distribution p_k, having a very large event space (between 0 and a few millions), from very few observations (only 64).

A possible strategy to cope with this situation is through *binning* of the event space. For example, if we use a bin size of 10^4, then degree 1 to degree 10^4 is compressed to a single bin which we label as 1, the next 10^4 degrees are mapped into the bin 2, and so on. Consequently, if for a particular organism the codon count is m, then the maximum degree of a codon node can be at most m. This implies that using a bin size of 10^4, there will be up to $m/10^4$ bins (or possible events), in which the 64 data points will be distributed. If all organisms are analyzed using the same bin size, depending on the length of the organism's genome, i.e., the codon count m, one obtains different numbers of bins. Alternatively, the bin size can be set for each organism in such a way that the resulting number of bins remains the same for all organisms. Thus, if we wish to have b bins for all organisms, the bin size for a particular organism will be m/b. Here, we analyze the

TABLE I. List of organisms along with their probable origin time (in Million Years Ago current time) and codon and gene counts.

Organism's Name	Description	Origin time (MYA)	Gene count	Codon count
Myxococcus xanthus	Gram-negative rod-shaped bacterium	3200	7421	2822743
Dictyostelium discoideum	Soil-living amoeba	2100	3369	1962284
Plasmodium falciparum	Protozoan parasite	542	4098	3032432
Saccharomyces cerevisiae	Single-celled fungi	488	14374	6511964
Xenopus laevis	Amphibian, African clawed frog	416	12199	5313335
Drosophila melanogaster	Two-winged insect, fruit fly	270	40721	21393288
Danio rerio	Tropical fish, zebrafish	145	19062	8042248
Homo sapiens	Bipedal primates, Human	2	89533	38691091

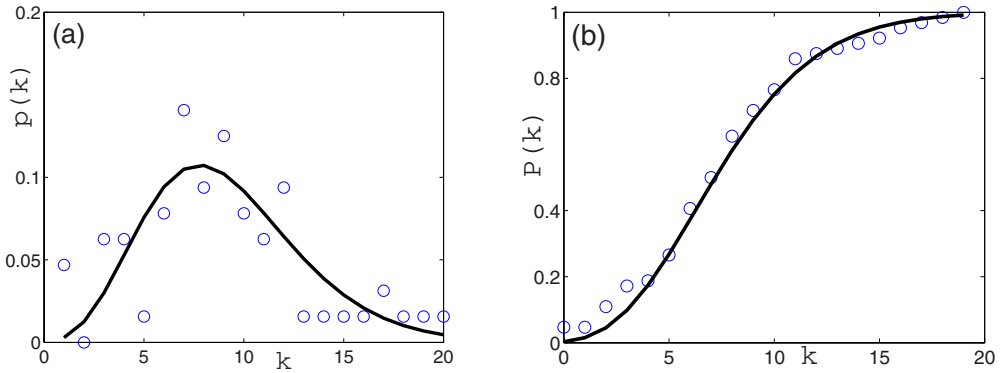

FIG. 6. (Color online) Degree distribution of the codon nodes for *Xenopus leavis*. In (a) a comparison between the empirical data (blue circles) and the theoretical $p_{k,t}$ obtained using Eq. (8) (black solid curve) is shown. The cumulative distribution of the real data (blue circles) and the theory (black solid curve) are shown in (b).

data using both methods: fixed bin size and fixed number of bins.

Apart from *binning*, another way to cope with the problem of data sparseness is to compute the cumulative degree distribution $P_{k,t}$ rather than the standard degree distribution $p_{k,t}$. $P_{k,t}$ is defined as the probability that a randomly chosen node has a degree less than or equal to k. Thus,

$$P_{k,t} = \sum_{i=0}^{k} p_{i,t}. \tag{14}$$

The cumulative distribution is more robust to the noise present in the observed data points, but at the same time it contains all the information present in $p_{k,t}$ [33]. Note that even though it is a standard practice in statistics to define *cumulative distributions* as stated in Eq. (14), in the complex network literature it is often defined as the probability that a randomly chosen node has degree "greater than or equal to" k. In the rest of the paper, the definition given by Eq. (14) will be used. Finally, Fig. 6 shows a comparison between the empirical degree distribution for *Xenopus leavis* and the best fit obtained using Eq. (8) (see below for details) for both, $p_{k,t}$ and $P_{k,t}$.

2. Growth model

A particular gene does not acquire all its constituent codons at a single time instant but evolves from an ancestral gene through the process of mutation, which implies the addition, deletion, or substitution of codons in the ancestral gene [34]. Therefore, we build the networks using the following parameters: $N=64$, $\mu=1$, and t corresponds to the number of codons that appear in the genome of the organism. In our model, we have a single free parameter, γ. In consequence, to describe the degree distribution of the empirical data using Eq. (8), we have to find out the value of γ that best fits the data. The best fitting γ, according to the least-squares method, is the value of γ that minimizes the square error, which is defined as,

$$Error = \sum_{k=0}^{\infty} [p_{k,t}(\gamma) - p^*_{k,t}]^2, \tag{15}$$

where $p^*_{k,t}$ represents the empirical distribution, while $p_{k,t}(\gamma)$ is the theoretical distribution given by Eq. (8). So, to obtain the best fitting γ, the *Error* is computed for all values of γ in the range from 0 to 5, using steps in γ of 0.01. As can be seen in Fig. 7, the *Error* shows a smooth behavior exhibiting a clear global minimum for all eight organisms. For values of γ larger than 5 (data not shown) *Error* gets monotonically larger, confirming the presence of a single (global) minimum. In consequence, the minimum was determined with a precision of ± 0.01 using both methods: fixed bin size and fixed bin count.

Figure 8 shows the cumulative real data and corresponding theoretical distributions of the eight organisms listed in Table I. Table II lists the values of γ for two different methods of binning: fixed bin count (bin count=20) fixed bin size (bin size=10^4). It can be observed that the degree distributions can be classified into two distinct groups corresponding to their γ value. As discussed in Sec. II, Eq. (8) describes

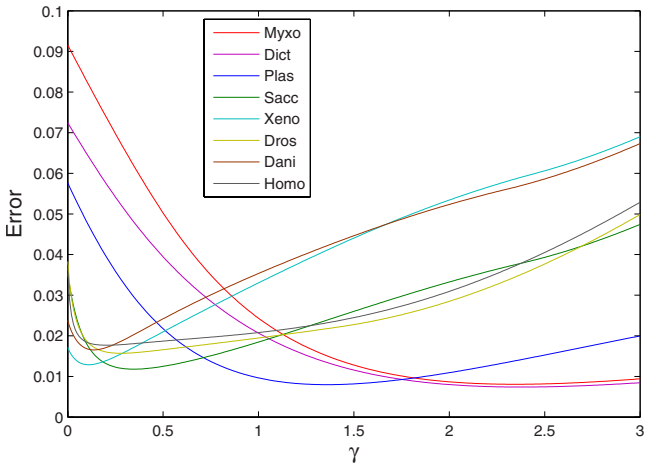

FIG. 7. (Color online) Error as defined by Eq. (15) as function of γ for the eight organisms using *fixed bin size*. Steps in γ correspond to 0.01.

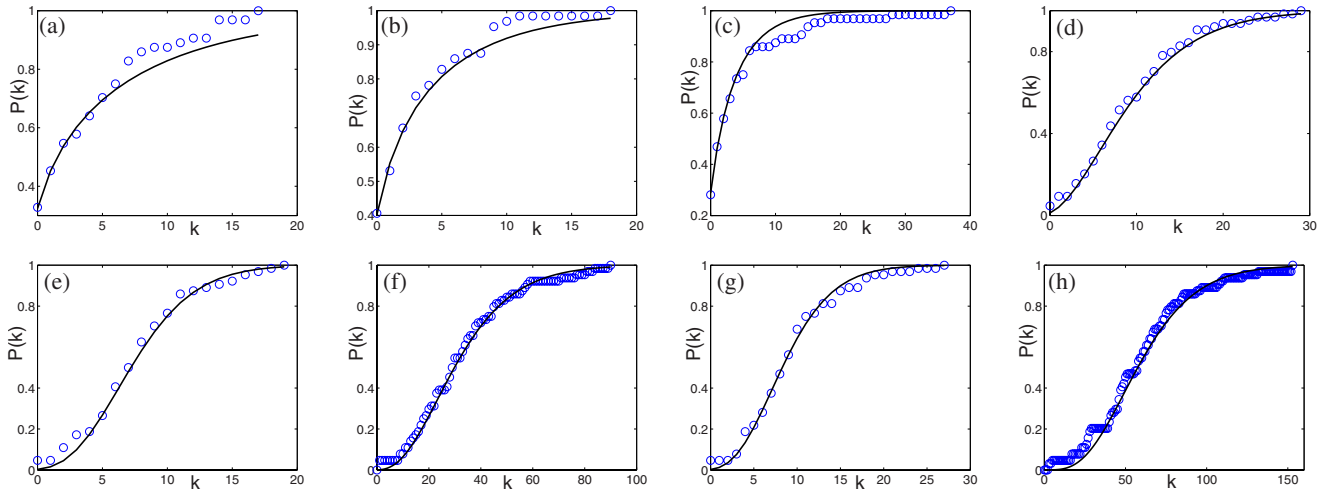

FIG. 8. (Color online) Cumulative degree distributions for the empirical data (blue circles) and their corresponding theoretical best γ fits through Eq. (8) (solid black curve) for the organisms. (a) *Myxococcus xanthus*, (b) *Dictyostelium discoideum*, (c) *Plasmodium falciparum*, (d) *Saccharomyces cerevisiae*, (e) *Xenopus laevis*, (f) *Drosophila melanogaster*, (g) *Danio rerio*, and (h) *Homo sapiens*.

four possible degree distributions depending on γ, two of them corresponding to a range of γ: (i) $0<\gamma<1$ and (ii) $1 \leq \gamma \leq (N/\mu)-1$. Thus, Table II shows that the degree distribution of the eight organisms belongs either to category (i) or (ii). *Myxococcus xanthus*, *Dictyostelium discoideum*, and *Plasmodium falciparum* fall into category (ii), while the rest falls into (i). Note that the classification of the organisms remains the same even after changing the statistics from fixed bin size to fixed bin count. Moreover, the classification is robust against changes in the bin size, as it was observed by repeating the data analysis for various bin sizes.

Interestingly, the three organisms with the larger γ value (between 1.36 and 2.38 for fixed bin size) are the more primitive ones. The rest (with a value of γ between 0.11 and 0.35) came into existence at a later stage of evolution.

This analysis allows us to speculate that in Myxococcus xanthus, Dictyostelium discoideum, and Plasmodium falciparum the degree of randomness during codon selection has been much lower than in *Saccharomyces cerevisiae*, *Xenopus laevis*, *Drosophila melanogaster*, *Danio rerio*, and *Homo sapiens*. These findings are probably correlated with the origin time and the evolutionary processes that shaped the usage of codons as follows. Let us think of evolution as the product of "copy-paste" operations. In this way, new genes emerge as result of imperfect copy-paste operations where the ancestral genes that are being copied are altered by addition, deletion or substitution of codons. Thus, copy-paste operations without defects lead to a high degree of "preferential attachment," while mutations/defects increase the degree of randomness. In consequence, we expect newly born species/organisms to exhibit a higher degree of randomness than their ancestor, given the greater number of mutations experienced by the newly formed organisms [35]. The value of γ in Table II reflects this fact, and suggests that knowledge at the level of codon usage (i.e., γ) can be used as a criterion to classify organisms.

B. Phoneme-language network

In this section we attempt to explain the self-organization of the consonant inventories through ABNs, where the consonants make up the basic units and languages are thought as discrete combinations of them. In fact, the most basic units of human languages are the speech sounds. The repertoire of sounds that make up the sound inventory of a language are not chosen arbitrarily. Indeed, the inventories show excep-

TABLE II. The values of γ that yield best fit for the degree distribution under the two different *binning* strategies.

Organism's Name	Best γ (fixed bin size)	Best γ (fixed bin count)
Myxococcus xanthus	2.35	2.1
Dictyostelium discoideum	2.38	2.57
Plasmodium falciparum	1.36	1.81
Saccharomyces cerevisiae	0.35	0.34
Xenopus laevis	0.11	0.11
Drosophila melanogaster	0.28	0.2
Danio rerio	0.14	0.1
Homo sapiens	0.20	0.09

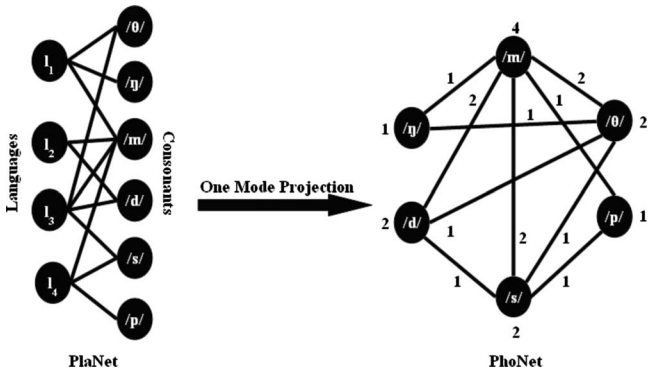

FIG. 9. Illustration of the phoneme-language ABN (referred in the figure as *PlaNet*) and its corresponding one-mode projection (referred as *PhoNet*).

tionally regular patterns across the languages of the world, which is arguably an outcome of the self-organization that goes on in shaping their structures [36]. In order to explain this self-organizing behavior of the sound inventories, various functional principles have been proposed such as *ease of articulation* [37,38], *maximal perceptual contrast* [37] and *learnability* [38]. The structure of vowel inventories has been successfully explained through the principle of maximal perceptual contrast [37,38]. Although there have been some linguistically motivated work investigating the structure of the consonant inventories, most of it is limited to certain specific properties rather than providing a holistic explanation of the underlying principle of organization.

1. Definition and construction

A first study of the consonant-language network as an ABN can be found in [39]. Here we follow the same definitions given in [39] where U is the universal set of consonants and V is the set of languages of the world. There is an edge $(u,v) \in E$ iff the consonant u occurs in the sound inventory of the language v. Figure 9 illustrates the structure of this ABN together with its associated one-mode projection onto the consonant nodes.

Many typological studies [37,40,41] of segmental inventories have been carried out in the past on the UCLA phonological segment inventory database (UPSID) [42]. UPSID records the sound inventories of 317 languages covering all the major language families of the world. In this work, we have used UPSID consisting of these 317 languages and 541 consonants found across them, for constructing ABN. Consequently, there are 317 elements (nodes) in the set V and 541 elements (nodes) in the set U. We selected UPSID mainly due to two reasons—(a) it is the largest database of this type that is currently available and, (b) it has been constructed by selecting one language each from moderately distant language families, ensuring a good balance between representatives from different origins.

2. Topological properties

Figure 10 illustrates the (cumulative) degree distribution of U. Since the degree of a language node is actually representing the size of a consonant inventory, we take as μ, i.e., the degree of each V node, the average number of consonants in human languages which is 22. Actually, the inventory size distribution of languages follow a skewed β distribution with mean=22 (see Fig. 2 of [39]). About 90% of the inventories in UPSID have sizes between 18 and 30. Therefore, the assumption that the each node in V has a constant degree does not render a big difference in the results. We have even simulated the model using the exact distribution of the consonant inventory sizes, and the results were slightly, but not significantly better (see Fig. 7 of [39]). This averaging was necessary because, recall that in the theoretical analysis of the ABN growth model the degree of each node in V has been assumed to be a constant (i.e., μ).

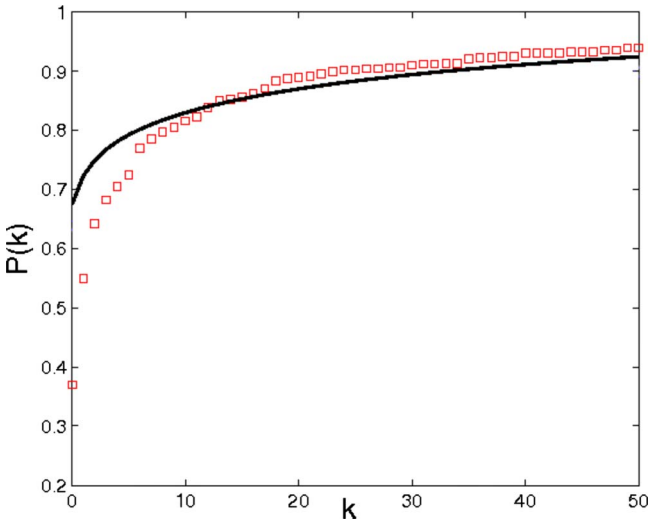

FIG. 10. (Color online) Cumulative degree distribution of U, i.e., the consonant nodes. Red squares correspond to the empirical data and the solid black line corresponds to the theoretical solution obtained through integration of Eq. (6) with $\gamma=14$.

3. Growth model

In order to obtain a theoretical description of the degree distribution of the consonant nodes in the ABN we employ the growth model described in Sec. II B with parameters $\mu=22$ and $N=541$. This means that we are assuming that each language has 22 consonant. The total number of consonants in the database is 541. Since there are 317 languages, we take $t=317$. Thus, γ is again the only free parameter in the model. The best fit of the data was obtained with $\gamma=14$ (Fig. 10). As before, the fitting was performed by minimizing error [see Eq. (15)] between the theoretical prediction and the emperical data.

4. One-mode projection

Interestingly, when we reconstruct the one-mode projection from the theory we find that we cannot match the empirical data. Figure 11 shows the cumulative degree distributions of the one-mode projection of the real data and theoretical one. The results show a larger quantitative difference between the curves compared to that between their bi-

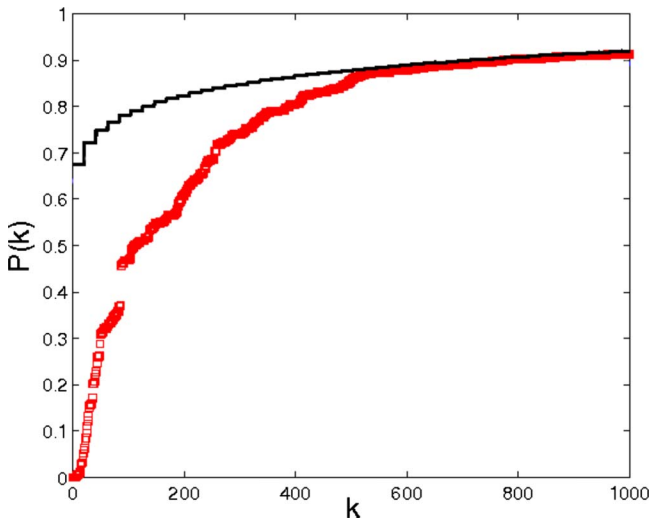

FIG. 11. (Color online) Cumulative degree distribution of the one-mode projection. Red squares correspond to the empirical data, while the solid black curve refers to the theoretical approximation.

partite counterparts. It indicates that the one-mode projection has a more complex structure than that, which could have emerged from a simple preferential attachment based kernel.

Nevertheless, we observe that preferential attachment can explain the occurrence distribution of the consonants over languages to a good extent. One possible way to explain this observation would be that a consonant, which is prevalent among the speakers of a given linguistic generation, tends to be more prevalent in the subsequent generations with a very little randomness involved in this whole process. It is this microlevel dynamics that manifests itself as preferential attachment in this ABN. However, the fact that the co-occurrence distribution of the consonants, i.e., the degree distribution of the one-mode projection, is not fully explained by the growth model implies that there are other organizing principles that are involved in shaping the structure of the consonant inventories.

IV. RELATED MODELS

As mentioned in the introduction, DCSs can be alternatively studied in the framework of Urn models that are popular in probability theory. Particularly relevant for us is the so-called Finite Pólya's process [22,23]. The process is defined as follows:

(1) Imagine a system consisting of N bins, each of them containing one ball at time $t=0$,

(2) at the time step $t+1$, place a new ball in the i-th bin with a probability proportional to $n_i^{\eta}(t)$, where $n_i(t)$ is the number of balls in the i-th bin at time t, and the exponent η is a model parameter.

The Finite Pólya's process is closely related to the developed ABN growth model for $\mu=1$. In order to make a comparison between both models, we consider the urns to be equivalent to the nodes in the fixed partition, i.e., the elementary units, and the balls to the degree of these nodes. Though, there are several similarities between both models, there are also several differences. For instance, notice that by definition, Pólya's process requires step (1), which means every bin is assumed to have one ball at time $t=0$, and $\mu=1$. In contrast, in the ABN growth model we can assume any initial condition, and $\mu \geq 1$. It is easy to see that if in the step (1) of the Finite Pólya's process, the urns are assumed to contain $1/\gamma$ balls instead of one ball, then this modified process, and for $\eta=1$, corresponds to the ABN model, under the mapping $n_i^{ABN} \mapsto n_i^{Polya} - 1/\gamma$, where n_i^{ABN} refers to the degree of the i-th node in the ABN model, n_i^{Polya} to the number of balls in the Finite Pólya's process, and $1/\gamma$ to the initial number of balls. Nevertheless, by definition the Pólya's process cannot start with a fractional number of balls in the urns. Therefore, the attachment process in ABN is a stronger generalization of the finite Pólya's process with $\eta=1$, with an extra parameter, γ. In summary, the shape of the urn-size distribution (equivalent to the degree distribution in ABNs) in a Finite Pólya's process is controlled by varying η, while in ABNs, the emergent degree distribution is determined by γ. It is also important to notice that the use of a linear attachment probability has allowed us to obtain an analytical closed form for the case of $\mu=1$, and to derive the exact expression for the time evolution of the degree distribution for $\mu>1$. However, we stress that it could be interesting to explore the ABN growth model that results from combining both degrees of freedom, η and γ, in such a way that the attachment probability becomes proportional to $\gamma n_i^{\eta}(t)+1$.

Another class of models for nongrowing bipartite networks has been developed by Evans and Plato in [19] (henceforth, the EP model). The EP model is based on the concept of rewiring and closely resembles the Urn model. In this study, one of the partitions, which the authors refer to as the set of *artifacts*, is fixed. The nodes in the other partition are referred to as individuals, all of which have degree one. The names artifacts and individuals reflect the fact that the model was initially conceived to describe cultural transmission. Note that artifacts and individuals are comparable in the context of an ABN to the basic units and their discrete combinations, respectively. In the EP model, there are fixed number of edges. At every time step, an artifact node is selected following a distribution Π_R and an edge that is connected to the chosen artifact is picked up at random. This edge is then rewired to another artifact node which is chosen according to a distribution Π_A. During the rewiring process the other end of the edge is always attached to the same *individual* node. The authors derive the exact analytical expressions for the degree distribution of the artifact nodes at all times and for all values of the parameters for the following definitions of the removal and attachment probabilities:

$$\Pi_R = \frac{k}{E}, \quad \Pi_A = p_r \frac{1}{N} + p_p \frac{k}{E},$$

where E, N, and k stands for the number of edges, the number of artifacts, and the degree of an artifact node, respectively. Furthermore, p_r and p_p, which add up to one, are positive constants (model parameters) that control the balance between random and preferential attachment.

The EP model is comparable to the ABN growth model for $\mu > 1$, except for the fact that the total number of edges in the latter case *diverges with time*, which changes the scenario completely. If we rewrite the attachment probability for the sequential growth model in a form similar to that of Π_A, we obtain the following expressions for the parameters p_r and p_p.

$$p_r = \frac{1}{1 + \gamma t/N}, \quad p_p = \frac{\gamma t/N}{1 + \gamma t/N}.$$

Clearly, as $t \to \infty$, $p_r \to 0$, and $p_p \to 1$, whereas in the EP model these parameters are fixed. Thus, apart from the two extreme cases of $p_r = 0$ and $p_r = 1$, the two models are fundamentally different, a fact which is also manifested in their emergent degree distributions. For instance, in the EP model the distributions reach a steady state, while this does not occur in ABNs. In addition, while for ABNs, we observe four distinct types of degree distributions, the equilibrium degree distribution of the EP model shows only two patterns: inverse power law with exponential cut off (comparable to the case when $\gamma < 1$), and a u-shaped distribution (comparable to the case of very large γ).

V. CONCLUSION AND DISCUSSION

In the preceding sections we have shown that DCSs can be described in terms of a special class of networks, the ABNs. In particular, we have generalized and extended previous results for ABNs [21] to include growth models where at each time step a discrete combination of $\mu > 1$ basic units is added to the system [43]. In addition, we have studied the properties of the one-mode projection network onto the DCS elementary units. Finally, and very importantly, we have shown how ABNs can be applied to analyze real-world DCSs. We have used the ABN analytical framework to characterize the codon-gene and phoneme-language network. The ABN approach has proven to be a very powerful tool to understand the selection process of elementary units during the generation of DCS discrete combination in real systems. We have shown that the model parameter γ gives a good measure of the relative weight between random and preferential attachment during the selection process.

From the codon-gene network analysis, we have learned that codon usage can help to classify organisms. Our study has revealed that the eight analyzed organisms can be classified into two sets according to its γ value. The results suggest that an ABN approach can further contribute to the reconstruction of phylogenetic relations. The use of ABN may be especially useful for the analysis of genome sequences which are so far only available in fragments either due to fragmentary sampling of the biological material or to unfinished sequencing efforts.

On the other hand, from the phoneme-language network analysis we have learned that the occurrence distribution of the consonants over languages can be explained in terms of an ABN with a strong degree of preferential attachment. However, the simple ABN growth models analyzed here have failed to explain co-occurrence distribution of the consonants, i.e., the degree distribution of the one-mode projection. This suggests more complex organizing principles absent in the current ABN growth models. There are some natural generalizations of the current ABN growth models that can certainly help to improve the description of the real phoneme-language network in terms of ABNs. For instance, rewiring during the growth of the ABN can be added to the theoretical description. The use of nonlinear attachment kernels can also help to improve the match between theory and real data. In fact, it has been recently shown through simulations that the degree distribution of the consonant nodes in the phoneme-language ABN can be better explained by using a superlinear kernel [44]. We expect the study of nonlinear kernels and rewiring in ABNs to be the focus of future research.

ACKNOWLEDGMENTS

This work was partially financed by the Indo-German collaboration project DST-BMB through grant "Developing robust and efficient services for open source Internet telephony over peer to peer network." N.G., A.N.M., and A.M. acknowledge the hospitality of Technische Universit at Dresden. A.M. would also like to thank Microsoft Research India for financial assistance. F.P. acknowledges the hospitality of IIT-Kharagpur and funding through grant ANR BioSys ("Morphoscale").

[1] S. Pinker, *The Language Instinct: How Mind Creates Language* (Perennial, New York, 1995).
[2] J. J. Ramasco, S. N. Dorogovtsev, and R. Pastor-Satorras, Phys. Rev. E **70**, 036106 (2004).
[3] D. J. Watts and S. H. Strogatz, Nature (London) **393**, 440 (1998).
[4] R. Albert and A.-L. Barabasi, Phys. Rev. Lett. **85**, 5234 (2000).
[5] M. Peltomäki and M. Alava, J. Stat. Mech.: Theory Exp. (2006), P01010.
[6] L. A. N. Amaral *et al.*, Proc. Natl. Acad. Sci. U.S.A. **97**, 11149 (2000).
[7] M. E. J. Newman, Phys. Rev. E **64**, 016132 (2001).
[8] A.-L. Barabási *et al.*, Physica A **311**, 590 (2002).
[9] R. Lambiotte and M. Ausloos, Phys. Rev. E **72**, 066117 (2005).
[10] G. Caldarelli and M. Catanzaro, Physica A **338**, 98 (2004).
[11] S. H. Strogatz, Nature (London) **410**, 268 (2001).
[12] S. Eubank *et al.*, Nature (London) **429**, 180 (2004).
[13] R. Ferrer i Cancho, O. Riordan, and B. Bollobás, Proc. R. Soc. London, Ser. B **272**, 561 (2005).
[14] W. Souma, Y. Fujiwara, and H. Aoyama, Physica A **324**, 396 (2003).
[15] K. Sneppen, Europhys. Lett. **67**, 349 (2004).

[16] J.-L. Guillaume and M. Latapy, Inf. Process. Lett. **90**, 215 (2004).
[17] A.-L. Barabási and R. Albert, Science **286**, 509 (1999).
[18] J. Ohkubo, M. Yasuda, and K. Tanaka, Phys. Rev. E **72**, 065104(R) (2005).
[19] T. S. Evans and A. D. K. Plato, Phys. Rev. E **75**, 056101 (2007).
[20] W. Dahui, Z. Li, and D. Zengru, Physica A **363**, 359 (2006).
[21] F. Peruani, M. Choudhury, A. Mukherjee, and N. Ganguly, Europhys. Lett. **79**, 28001 (2007).
[22] N. Johnson and S. Kotz, *Urn Models and Their Applications: An Approach to Modern Discrete Probability Theory* (Wiley, New York, 1977).
[23] F. Chung, S. Handjani, and D. Jungreis, Ann. Comb. **7**, 141 (2003).
[24] S. N. Dorogovtsev and J. F. F. Mendes, *Evolution of Networks: From Biological Nets to the Internet and WWW* (Oxford University Press, New York, 2003).
[25] R. Ferrer i Cancho, R. V. Solé, and R. Köhler, Phys. Rev. E **69**, 051915 (2004).
[26] S. M. G. Caldeira, T. G. P. Lobão, R. F. S. Andrade, A. Neme, and J. G. V. Miranda, Eur. Phys. J. B **49**, 523 (2006).
[27] M. E. J. Newman, S. H. Strogatz, and D. J. Watts, Phys. Rev. E **64**, 026118 (2001).
[28] M. E. J. Newman, Proc. Natl. Acad. Sci. U.S.A. **101**, 5200 (2004).
[29] P. Sharp *et al.*, Nucleic Acids Res. **16**, 8207 (1988).
[30] S. B. Hedges, Nat. Rev. Genet. **3**, 838 (2002).
[31] Y. Nakamura, T. Gojobori, and T. Ikemura, Nucleic Acids Res. **28**, 292 (2000).
[32] Codon usage database: http://www.kazusa.or.jp/codon/
[33] M. E. J. Newman, SIAM Rev. **45**, 167 (2003).
[34] T. Kunkel and K. Bebenek, Annu. Rev. Biochem. **69**, 497 (2000).
[35] D. Fredman *et al.*, Nat. Genet. **36**, 861 (2004).
[36] P.-Y. Oudeyer, *Self-organization in the Evolution of Speech* (Oxford University Press, New York, 2006).
[37] B. Lindblom and I. Maddieson, in *Language, Speech, and Mind*, edited by L. M. Hyman and C. N. Li (Routledge, London, 1988), pp. 62–78.
[38] B. de Boer, J. Phonetics **28**, 441 (2000).
[39] M. Choudhury *et al.*, Proceedings of COLING–ACL P06, 128 (2006).
[40] F. Hinskens and J. Weijer, Linguistics **41**, 1041 (2003).
[41] P. Ladefoged and I. Maddieson, *Sounds of the World's Languages* (Blackwell, Oxford, 1996).
[42] I. Maddieson, *Patterns of Sounds* (Cambridge University Press, Cambridge, England, 1984).
[43] In [21] the rigorous analysis of ABNs was restricted to the case $\mu=1$.
[44] A. Mukherjee *et al.*, J. Quant. Linguist. **16**, 157 (2009).