Modeling Discrete Combinatorial Systems as Alphabetic Bipartite Networks (α-BiNs): Theory and Applications

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

Niloy Ganguly
Department of Computer Science and Engineering,
Indian Institute of Technology Kharagpur, Kharagpur 721302, India.

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

Lutz Brusch and Andreas Deutsch
ZIH, TU 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,
Institut des Systémes Complexes de Paris Île-de-France, 57/59, rue Lhomond, F-75005 Paris, France.
(Dated: April 13, 2009)

Life and language 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 DCS can be represented as an Alphabetic Bipartite Network (α-BiN) where there are two kinds of nodes, one type represents the elementary units while the other type represents their combinations. There is an edge between a node corresponding to an elementary unit u and a node corresponding to a particular combination v if u is present in v. Naturally, the partition consisting of the nodes representing elementary units is fixed, while the other partition is allowed to grow unboundedly. Here, we extend recent analytical findings for α-BiNs derived in [Europhys. Lett. **79**, 28001 (2007)] and empirically investigate two real world systems: the codon-gene network and the phoneme-language network. The evolution equations for α-BiNs under different growth rules are derived, and the corresponding degree distributions computed. It is shown that asymptotically the degree distribution of α-BiNs can be described as a family of beta distributions. The one-mode projections of the theoretical as well as the real world α-BiNs are also studied. We propose a comparison of the real world degree distributions and our theoretical predictions as a means for inferring the mechanisms underlying the growth of real world systems.

I. INTRODUCTION

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

The α-BiNs are a subclass of bipartite networks where there are two disjoint partitions of nodes. An edge, in a bipartite network, links nodes that appear in two different partitions, but never those belonging to the same partition. In most of the bipartite networks 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 co-authored 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 bipartite networks, 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 ob-

---

*Corresponding author: peruani@pks.mpg.de ject j and a signal i represents that a possible meaning of the i sign is the object j, the bank-company [16] or donor-acceptor network that accounts for injection and merging of magnetic field lines [17].

Several models have been proposed to synthesize the structure of these bipartite networks, i.e., when both the partitions grow unboundedly over time [2–5, 15]. 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 [18].

Although there have been some works on non-growing bipartite networks [19, 20] those like α-BiNs where one of the partitions remains fixed over time have received comparatively much less attention. Since the set of basic units in DCSs is always finite and constant, α-BiNs have one of its partitions fixed, the one that represents the basic units (e.g. letters, codons). In contrast, the other partition, that represents the unique discrete combinations of basic unit (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 an important and indispensable aspect of the system, which can be modeled within the framework of α-BiNs 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. Fig. 1 illustrates the concepts through the example of genes and codons.

As far as we know, the first empirical evidence of the non-scale free character of the degree distribution of α-BiNs has been reported in [21], while the first systematic and analytical study of such bipartite networks has been presented in [22]. The growth model for α-BiNs, proposed in [22], is based on a 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. For sequential attachment, i.e., when the edges are incorporated one by one, the exact expression for the emergent degree distribution has been derived. Nevertheless, for parallel attachment, i.e., when multiple edges are incorporated in one time step, only an approximate expression has been proposed. It has been shown that for both the cases, the degree distribution approaches a beta-distribution asymptotically with time. Depending on the value of the randomness parameter four distinct types of degree distributions can be observed; these, in increasing order of preferentiality, are: (a) normal distribution, (b) skewed normal distribution with a single mode, (c) exponential distribution, and (d) U-shaped distribution.

In this article, we briefly review these findings and extend the analytical framework. We derive the exact growth model for parallel attachment and study the degree distribution of the one-mode projection of the network onto the alphabet nodes. In order to illustrate how the proposed framework can be used as an analytical tool to study and interpret empirical data, we applied the α-BiN 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 α-BiN as a modeling approach. We start by the the codon-gene network where codons play the role of the basic units while genes are the discrete combinations of them, 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 α-BiN 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 α-BiN growth models. 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 α-BiN model. This indicates that although the α-BiN 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.

Alternatively, DCSs can be studied in the context of Urn models of probability theory. In particular, the α-BiN growth model for sequential attachment presented here is a close relative of the so-called Finite Pólya's process [23, 24]. Moreover, the connection between Urn models and bipartite networks has already been established by Evans and Plato in [20], where the authors developed the theoretical framework for rewiring in non-growing bipartite networks. At the end of the paper, we contextualize the developed α-BiN theory in the framework of Urn models and discuss the similarities and differences with the Finite Pólya's process and the rewiring model suggested by Evans and Plato [20].

The article is organized as follows: Sec. II formally defines α-BiN and introduces two growth models and their 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 α-BiN theory in the context of Urn models of probability theory. The concluding section summarizes the obtained results and discuss the broader consequences of the present work.

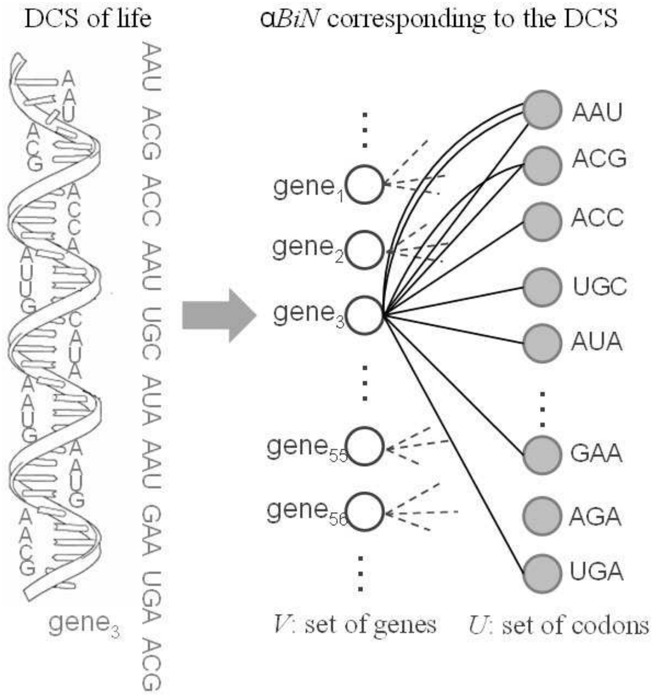

FIG. 1: DNA modeled as a bipartite network α-BiN. The set U consists of 64 codons, whereas the collection V of genes is virtually infinite. Multiple occurrences of a codon in a gene have been represented here by multi-edges. For instance, the codons 'ACG' and 'AAU' have respectively 2 and 3 edges connecting to the node gene$_3$. Alternatively, this could have been represented by single edges with weights 2 and 3, while the weight of the other edges would be equal to 1.

II. THEORETICAL FRAMEWORK FOR α-BINS

A. Formal definition and modeling

A bipartite graph G is a 3-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.

α-BiNs 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 unique 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 $v \in V$ iff u is a part of v. If u occurs w times in v, then there are w edges between u and v, or alternatively, the weight of the edge (u, v) is w. Fig. 1 illustrates these concepts through the example of genes and codons.

Notice that the above model overlooks the order in which the basic units are strung into a particular discrete combination. The order can be taken into account by labeling the basic units in order of appearance in each element of V. However, in this work, we consider only unordered versions of DCSs. As we shall see subsequently, in several real world DCSs, such as the phoneme-language network, the elements of V are just collections of the elements of U, rather than sequences of them.

B. Growth model for sequential attachment

In this subsection, we review the results derived in [22] which apply to sequential as well as parallel attachment. While the results for sequential attachment are exact, for parallel attachment they represent an approximation. In the next subsection the results obtained in [22] are extended and the exact derivation for parallel attachment is presented.

The growth of α-BiNs 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. Let v_i be the node added to V during the ith time step. The theoretical analysis assumes that μ is a *constant* greater than 0. This constraint will be relaxed during the construction of the empirical networks. However, note that if the degrees of the nodes in V are sampled from a Poisson-like distribution with mean μ, the theoretical analysis holds good asymptotically.

Let $\widetilde{A}(k_i^t)$ be the probability of attaching a new edge to a node u_i, where k_i^t refers to the degree of the node u_i at time t. $\widetilde{A}(k_i^t)$ defines the attachment kernel that takes the form:

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

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. Since in a bipartite network the sum of the degrees of the nodes in the two partitions are equal, the denominator in the above expression is equal to $\mu\gamma t + N$. 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* [25].

This means that when a new discrete combination, say a gene, enters the system, it is always assumed to have μ basic units, e.g., a chain of μ codons. The patterns of the codons constituting the newly entered gene depends on the prevalence of the codons in the pre-existing genes as well as a randomness factor $1/\gamma$. At this point it is worthwhile to distinguish between a few basic sub-cases of the growth model. When $\mu = 1$, addition of a node in V is equivalent to addition of one edge in the network and thus the edges attach to the nodes in U in a sequential manner. However, for $\mu > 1$ addition of an edge is no longer a sequential process; rather μ edges are added simultaneously. We refer to the former process as *sequential attachment* and the latter as *parallel attachment*. Depending on the underlying DCS, the parallel attachment process can be further classified into two sub-cases. If it is required that the μ nodes chosen are all distinct, then we call this *parallel attachment without replacement*. On the other hand, if v_i is allowed to attach to the same node more than once, we refer to the process as *parallel attachment with replacement* [45]. Thus, parallel attachment without replacement leads to α-BiNs without multi-edges or weighted edges, while parallel attachment with replacement results in α-BiNs with multi-edges. The two cases collapse for the case of sequential attachment. To motivate the reader further, we provide some examples of natural DCSs from each of the aforementioned classes.

- *Sequential attachment*: Since in the sequential attachment model, every node in V has only one edge, there are no discrete combinations at all. Rather, each incoming v_i is a reinstantiation of some basic unit u_j. However, think of a system where U is the set of languages and V is the collection of speakers, and an edge between $u \in U$ and $v \in V$ implies that u is the mother tongue of v. Although not a DCS, these type of "class and its instance" systems are plentiful in nature and can be aptly modeled using sequential attachment.

- *Parallel attachment with replacement*: Any DCS modeled as a sequence of the basic units can be thought to follow the "with replacement" model. For instance, a gene can have many repetitions of the same codon and similarly, there may be multiple occurrences of the same word in a sentence.

- *Parallel attachment without replacement*: A DCS that is a collection of the basic units can be conceived as an outcome of the "without replacement" model. For instance, the consonants and vowels (partition U) that form the repertoire of basic sounds (phonemes) of a language (partition V), proteins (U) forming protein complexes (V), etc.

In this work, we focus on the topological properties of α-BiNs that are synthesized using the sequential and parallel attachment with replacement. Nevertheless, in section III B we also present some empirical results for the parallel attachment without replacement model in the context of the phoneme-language network.

Any α-BiN 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 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 a degree 0 and there are no nodes in V. Therefore,

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

Here, δ represents the Kronecker symbol. It is interesting to note that unlike the case of standard preferential attachment based growth models for unipartite (e.g., the BA model [18]) and bipartite networks (e.g., [2]), the degree distribution of the partition U in α-BiNs cannot be solved using the stationary assumption that in the limit $t \to \infty$, $p_{k,t+1} = p_{k,t}$. This is because the average degree of the nodes in U, which is $\mu t/N$, diverges with t, and consequently, the system does not have a stationary state.

In [22] it has been shown that $p_{k,t}$ can be approximated for $\mu \ll N$ and small values of γ by integrating:

$$p_{k,t+1} = (1 - A_p(k,t))p_{k,t} + A_p(k-1,t)p_{k-1,t} \tag{3}$$

where $A_p(k,t)$ is defined as

$$A_p(k,t) = \begin{cases} \frac{(\gamma k+1)\mu}{\gamma\mu t+N} & \text{for} \quad 0 \le k \le \mu t \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

for $t > 0$ while for $t = 0$, $A_p(k,t) = (\mu/N)\delta_{k,0}$. The numerator contains a μ because at each time step there are μ edges that are being incorporated into the network rather than a single edge. The solution of Eq. (3) with the attachment kernel given by Eq. (4) reads:

$$p_{k,t} = \begin{pmatrix} t \\ k \end{pmatrix} \frac{\prod_{i=0}^{k-1}(\gamma i+1)\prod_{j=0}^{t-1-k}\left(\frac{N}{\mu} - 1 + \gamma j\right)}{\prod_{m=0}^{t-1}\left(\gamma m + \frac{N}{\mu}\right)} \tag{5}$$

As already mentioned in [22], Eq. (3) cannot describe the stochastic parallel attachment exactly because it explicitly assumes that in one time step a node of degree k can only get converted to a node of degree $k+1$. Clearly, the incorporation of μ edges in parallel allows the possibility for a node of degree k to get converted to a node of degree $k+\mu$. So, for $\mu > 1$, Eq. (5) is just an approximation of the real process for $\mu \ll N$ and small values of γ. However, for $\mu = 1$, i.e. for sequential attachment, Eq. (5) is the exact solution of the process.

Interestingly, for $\gamma > 0$, Eq. (5) approaches, asymptotically with time, a beta-distribution as follows.

$$p_{k,t} \simeq C^{-1}\left(k/t\right)^{\gamma^{-1}-1}\left(1-k/t\right)^{\eta-\gamma^{-1}-1} \tag{6}$$

Here, C is the normalization constant and $\eta = N/(\gamma\mu)$. By making use of the properties of beta distributions, we learn that depending on the value of γ, $p_{k,t}$ can take one of the following distinctive functional forms.

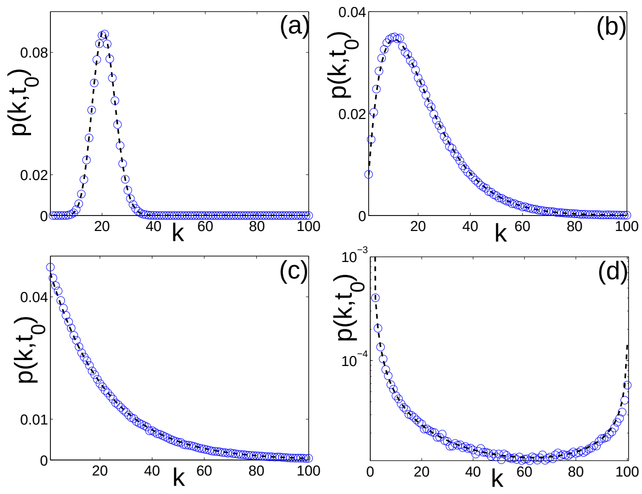

FIG. 2: The four possible degree distributions depending on γ for sequential attachment (and approximated expression for parallel attachment). Circles represent average over 5000, in (a)-(c), and 50000, in (d), stochastic simulations. The dashed curve is the theory given by Eq. (5). From (a) to (c), $t_0 = 1000$, $N = 1000$ and $\mu = 20$. (a) at $\gamma = 0$, $p(k,t)$ becomes a binomial distribution. (b) $\gamma = 0.5$, the distribution exhibits a maximum which shifts with time for $0 \leq \gamma < 1$. (c) $\gamma = 1$, $p(k,t)$ no longer exhibits a shifting maximum and the distribution is a monotonically decreasing function of k for $1 \leq \gamma \leq (N/\mu) - 1$. (d) $\gamma = 2500$, $t_0 = 100$, $N = 1000$ and $\mu = 1$. $p(k,t)$ becomes a u-shaped curve for $\gamma > (N/\mu) - 1$.

a) $\gamma = 0$, a binomial distribution whose mode shifts with time,

b) $0 < \gamma < 1$, a skewed (normal) distribution which exhibits a mode that shifts with time,

c) $1 \leq \gamma \leq (N/\mu) - 1$, a monotonically decreasing (near exponential) distribution with the mode frozen at $k = 0$, and

d) $\gamma > (N/\mu) - 1$, a u-shaped distribution with peaks at $k = 0$ and $k = t$.

Fig. 2 illustrates the possible four regimes of Eq. (5).

In the next subsection we present a generalization of Eq. (3) for $\mu > 1$, i.e. for parallel attachment, and solve it.

C. Growth model for parallel attachment with replacement

Recall that for parallel attachment, t refers to the event of introducing a new node in V with μ edges. Therefore, the correct expression for the evolution of $p_{k,t}$ has the form:

$$p_{k,t+1} = (1 - \sum_{i=1}^{\mu} \widehat{A}(k,i,t))p_{k,t} + \sum_{i=1}^{\mu} \widehat{A}(k-i,i,t)p_{k-i,t} \quad (7)$$

where $\widehat{A}(k,i,t)$ represents the probability at time t of a node of degree k of receiving i new edges in the next time step. The term $\sum_{i=1}^{\mu} \widehat{A}(k,i,t)p_{k,t}$ 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 $\sum_{i=1}^{\mu} \widehat{A}(k-i,i,t)p_{k-i,t}$.

Next we derive an expression for $\widehat{A}(k,i,t)$. 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:

$$\widehat{A}(k,i,t) = \binom{\mu}{i}\left(\frac{1}{N}\right)^i \left(1 - \frac{1}{N}\right)^{\mu-i} \quad (8)$$

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 previous 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. (8) is generalized for $\gamma \geq 0$ as

$$\widehat{A}(k,i,t) = \binom{\mu}{i}\left(\frac{\gamma k + 1}{\mu\gamma t + N}\right)^i \left(1 - \frac{\gamma k + 1}{\mu\gamma t + N}\right)^{\mu-i} \quad (9)$$

Inserting expression (9) into Eq. (7), we obtain:

$$\begin{aligned} &p_{k,t+1} \qquad (10)\\ &= \left[1 - \sum_{i=1}^{\mu}\binom{\mu}{i}\left(\frac{\gamma k + 1}{\mu\gamma t + N}\right)^i \left(1 - \frac{\gamma k + 1}{\mu\gamma t + N}\right)^{\mu-i}\right] p_{k,t} \\ &+ \sum_{i=1}^{\mu}\binom{\mu}{i}\left(\frac{\gamma(k-i) + 1}{\mu\gamma t + N}\right)^i \left(1 - \frac{\gamma(k-i) + 1}{\mu\gamma t + N}\right)^{\mu-i} p_{k-i,t} \end{aligned}$$

The terms between parenthesis in the first line of Eq. (10) can be simplified recalling that

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

Therefore, Eq. (10) can be rewritten by including $i = 0$ in the sum, whereby we obtain

$$\begin{aligned} p_{k,t+1} = \sum_{i=0}^{\mu}\binom{\mu}{i}&\left(\frac{\gamma(k-i) + 1}{\mu\gamma t + N}\right)^i \\ &\left(1 - \frac{\gamma(k-i) + 1}{\mu\gamma t + N}\right)^{\mu-i} p_{k-i,t} \qquad (11) \end{aligned}$$

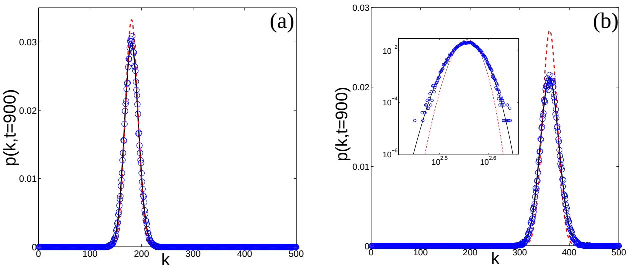

FIG. 3: Comparison between random attachment ($\gamma = 0$) and the approximation given by Eq. (5) (dashed red curve), the exact solution given by the integration of Eq. (11) (solid black curve), and stochastic simulations. Circles correspond to average over 500 simulations. In both the figures $N = 100$. (a) corresponds to $\mu = 20$ while (b) to $\mu = 40$. The inset in (b) shows in log-log scale the deviation of the approximation with respect to the exact solution and simulations.

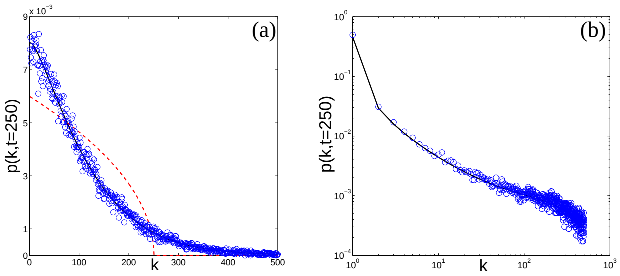

FIG. 4: Comparison between strong preferential attachment ($\gamma \geq 1$) between the approximation given by Eq. (5) (dashed red curve), the exact solution given by the integration of Eq. (11) (solid black curve), and stochastic simulations (circles), averaged over 500 runs, for parallel attachment for $\gamma \geq 1$. In both figures $N = 100$ and $\mu = 40$. (a) corresponds to $\gamma = 1$ while (b) to $\gamma = 16$. Notice that in (b) the approximation falls out of the range of the figure, while the exact solution given by the integration of Eq. (11) describes the simulation data quite well.

Note that Eq. (11) is a generalization of Eq. (2) in [22], which can be obtained from Eq. (11) by assigning $\mu = 1$.

In Fig. 3 we compare for random attachment ($\gamma = 0$) the approximation given by Eq. (5) (dashed red curve), the exact solution is given by the integration of Eq. (11) (solid black curve), and stochastic simulations.

Notice that the approximation, as mentioned above, deviates from the exact solution and simulations as μ increases. Looking at Fig. 3 one may wrongly conclude that the exact solution given by Eq. (11) is a very minor improvement over the approximation given by Eq. (5). However, for large values of γ (see Fig. 4), i.e., for strong preferential attachment, Eq. (5) drastically fails to describe the simulation data, while Eq. (11) accurately explains the data. To summarize, Figs. 3 and 4 validate Eq. (11) and show that, except when $\mu \ll N$ and small γ, the only way to describe the degree distribution for parallel attachment with replacement is through the integration of Eq. (11).

D. One-mode projection

In this section, we analyze the degree distribution of the one-mode projection of α-BiNs onto the set U. Formally, for an α-BiN $\langle U, V, E\rangle$, 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 α-BiN G, then there are w edges linking u_i and u_j in the one-mode projection G_U. Alternatively, one can conceive of a weighted version of G_U, where the weight of the edge (u_i, u_j) is w. In the context of the codon-gene network, the one-mode projection is a codon-codon network, where two codons are connected by as many edges as there are genes in which both of these codons occur. The one-mode projection of an α-BiN provides insight into the relationship between the basic units. For instance in linguistics the one-mode projection of the word-sentence α-BiN 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, [14, 26]).

In [27] a general technique for computing the degree distribution of the one-mode projection of a bipartite network is described. The method has been derived by making use of the concept of generating functions. As we shall see shortly, this technique is only suitable for estimating the weighted degree distribution of the one-mode projection.

Here, we propose a novel technique to derive the *thresholded degree distribution* of the one-mode projection for any arbitrary threshold. We start out by studying first the simple cases of the (non-thresholded) degree distribution of the one-mode projection for sequential and parallel attachment, to finally focus on the new technique to derive the degree distribution of the thresholded one-mode projection for parallel attachment. Notice that in order to distinguish the degree distributions of the one-mode projection from their bipartite counterpart, we shall use the symbol $p_u(k, t)$ to refer to the probability that a randomly chosen node from the one-mode projection of an α-BiN with t nodes in V (i.e., after t time steps) has degree k.

1. *Sequential attachment*

Recall that in the sequential attachment based growth model only one edge is added per time step and consequently, every node in V has degree $\mu = 1$. Therefore, for any two nodes in U, say u_i and u_j, there is no node $v \in V$, which is connected to both u_i and u_j (this is because degree of v is 1). Thus, for α-BiNs that have

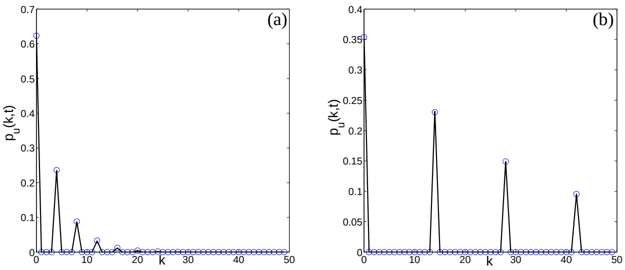

FIG. 5: Comparison between stochastic simulations for the one-mode projection (circles) and Eq. (13) (solid black curve). In both figures $N = 500$ and $\gamma = 1$. Circles correspond to averages over 1000 simulations. In (a) $\mu = 5$ while in (b) $\mu = 15$.

been grown using the sequential attachment model, the one-mode projection is a degenerate graph with N nodes and 0 edges. The degree distribution of this network is

$$p_u(k,t) = \delta_{k,0} \tag{12}$$

2. Parallel attachment

Recall that in the parallel attachment model, at each time step the node which is added to V has μ edges. Consider a node $u \in U$ that has degree k in the α-BiN. 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 $q = k(\mu - 1)$. Consequently, the degree distribution of G_U, $p_u(q,t)$, is related to $p_{k,t}$ in the following way:

$$p_u(q,t) = \left\{ \begin{array}{ll} p_{0,t} & \text{if } q = 0 \\ p_{k=q/(\mu-1),t} & \text{if } \mu - 1 \text{ divides } q \\ 0 & \text{otherwise} \end{array} \right\} \tag{13}$$

Fig. 5 shows a comparison between stochastic simulations (circles) and Eq. (13) (solid black curve). Notice that this mapping simply implies that $p_u(q = 0,t) = p_{0,t}$, $p_u(q = \mu - 1, t) = p_{1,t}$, $p_u(q = 2(\mu - 1), t) = p_{2,t}$, ..., $p_u(q = j(\mu - 1), t) = p_{j,t}$. The same result can be derived by using the generating function based technique described in Eq. 70 of [27]. It is worth noticing that q is the weighted degree of a node (i.e., the sum of the weights of all the edges incident on a node), and therefore, does not give any information about the number of distinct neighbors a node has.

3. Thresholded degree-distribution for parallel attachment

Weighted graphs, such as the one-mode projections of α-BiNs, can be converted to corresponding unweighted version 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, [26, 28]). We shall denote the thresholded degree distribution at threshold 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 number of edges running between u and u' is m, given that the degrees of the nodes are k_u and $k_{u'}$. 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 α-BiNs, 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:

$$p(k_u, k_{u'}, m) = \binom{t}{m} \left(\frac{k_u k_{u'}}{t^2}\right)^m \left(1 - \frac{k_u k_{u'}}{t^2}\right)^{t-m} \tag{14}$$

From Eq. (14), the probability for u and u' of sharing an edge in thresholded G_U is easily computed as:

$$p(k_u, k_{u'}; m > \tau) = \sum_{m=\tau+1}^{t} p(k_u, k_{u'}, m) \tag{15}$$

Consequently, in the thresholded G_U, the expected degree D of a node u whose degree is k in the α-BiN is given by:

$$D(k,\tau) = N \sum_{i=1}^{t} p_{i,t}\, p(k, i; m > \tau) \tag{16}$$

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:

$$p_u(q,t;\tau) = \sum_{q=\lfloor D(k,\tau) \rfloor} p_k \tag{17}$$

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

Fig. 6 shows a comparison between Eq. (17) (solid curves) and stochastic simulations (symbols) for the one-mode projection at different times. The implementation of Eq. (17) was done by summing over the $p_{k,t}$ obtained from the stochastic simulations of the corresponding α-BiN according to $q = \lfloor D(k,\tau) \rfloor$, as indicated by Eq. (16).

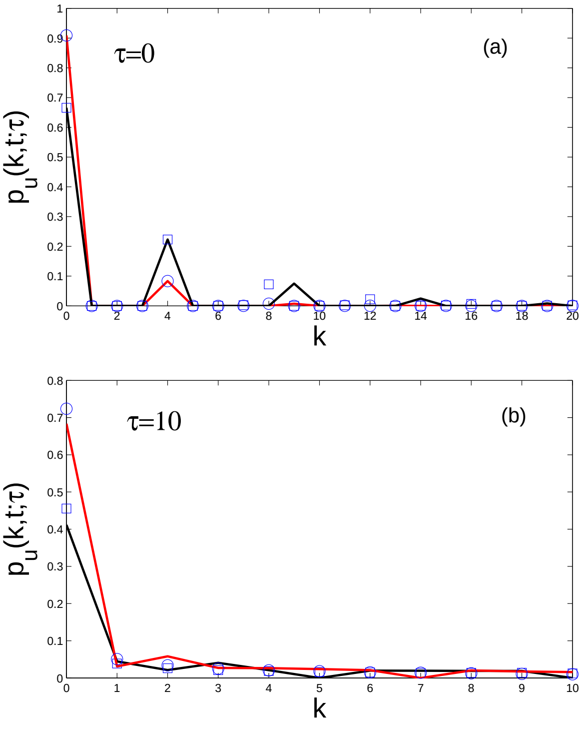

FIG. 6: Comparison between stochastic simulations for the one-mode projection at different times (symbols) and Eq. (17) (solid curve). In (a) $\tau = 0$, $N = 1000$, $\mu = 5$, $\gamma = 1$. The circles and the red curve correspond to $t = 20$, while the squares and the black curve to $t = 100$. The slight deviation of the simulation results from the theoretical predictions is due to rounding of the values. In (b) $\tau = 10$, $N = 100$, $\mu = 20$, $\gamma = 1.5$. The circles and the red curve correspond to $t = 50$, while the squares and the black curve to $t = 100$.

4. *One-mode kernel*

Until now we have been describing growth models for the α-BiNs. The unipartite network G_U is obtained by projecting the α-BiN onto the set of nodes U. We shall now attempt to derive a kernel for the growth of the network G_U, whereby we can construct G_U directly without constructing the underlying α-BiN. Consider a node $v_t \in V$ that has been introduced in the α-BiN in the t^{th} step. There are μ nodes in U to which v_t gets connected. Let us assume that v_t is connected to no node in U more than once. This fact is true in the "parallel attachment without replacement" model that will be described in greater details in Sec. III B. However, as discussed earlier, if $\mu \ll N$ and γ is small, it is quite reasonable to make this assumption even in the case of "parallel attachment with replacement" model.

Introducing v_t in the α-BiN is equivalent to introducing a clique (complete graph) of size μ in G_U. This is because all the nodes that are connected to v_t in the α-BiN are connected to each other in G_U by virtue of sharing a common neighbor v_t. Note that this does not prohibit these μ nodes from having previous connections. The growth process allows formation of multiple edges, that is to say edge weights can be greater than 1.

Let us denote the degree of a node u_i in (the non-thresholded) G_U after t steps as $q_{i,t}$. As discussed in the previous subsection, $q_{i,t} = (\mu - 1)k_{i,t}$, where $k_{i,t}$ is the degree of u_i in the corresponding α-BiN after t steps. Noticing the fact that in the α-BiN the μ nodes are chosen independently of each other solely based on the attachment kernel, we can define a kernel for selecting a collection of μ nodes in G_U as follows.

$$\widetilde{A}(q_{a,t}, q_{b,t}, \dots) = \prod_{j=a,b,\dots} \widetilde{A}(q_{j,t}/(\mu - 1)) \tag{18}$$

where $a, b, \dots$ denotes a randomly chosen collection of μ nodes in G_U. Substituting the expression provided in Eq. (1) for the preferential attachment based kernel we obtain:

$$\widetilde{A}(q_{a,t}, q_{b,t}, \dots) = \prod_{j=a,b,\dots} \frac{\gamma q_{j,t}/(\mu - 1) + 1}{\sum_{i=1}^{N} (\gamma q_{i,t}/(\mu - 1) + 1)} \tag{19}$$

Below we summarize the growth model for the one-mode projection of the α-BiN

- Select a set of μ nodes $a, b, \dots$ with the probability $\widetilde{A}(q_{a,t-1}, q_{b,t-1}, \dots)$ as described by Eq. (19).

- Introduce edges between every pair of the chosen set $a, b, \dots$.

- Advance time by a unit and repeat the process.

We assume an initial condition $q_i = 0$ for all i. Alternatively, but also equivalently, the above growth model can be described as choosing μ nodes independently, each with probability $\widetilde{A}(q_{i,t}/(\mu - 1))$ and then adding edges between them.

Fig. 7 plots the degree distribution obtained from the one-mode kernel and the degree distribution of the one-mode projection of the u_i nodes of the α-BiN built with the same parameters. We can see that the one-mode kernel gives quite similar degree distribution as one-mode projection of the bipartite network. The primary observation from this analysis is that the kernel of the unipartite growth model has the same form as that of the bipartite growth model, with a scaling of the parameter γ by a factor of $1/(\mu - 1)$ in the former. This implies that as μ increases, the extent of degree-based preference decreases in the one-mode projection. The analysis, nevertheless is valid only for the "without replacement" model and holds approximately for the "with replacement" for $\mu \ll N$.

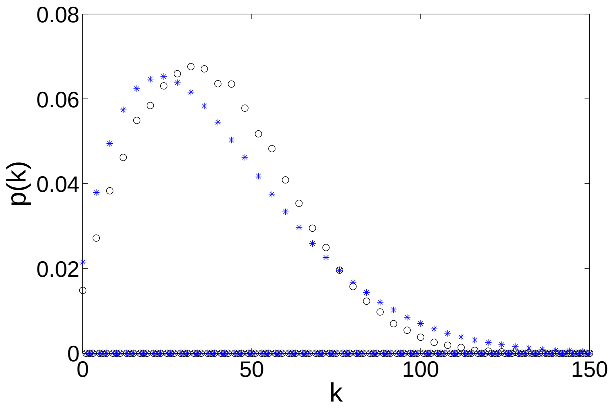

FIG. 7: Comparison between the degree distribution obtained from the stochastic simulation of Eq. (19), averaged over 1000 runs, and the one-mode projection of the α-BiN obtained using Eq. (13), averaged over 100000 runs, with $N = 50$, $\mu = 5$, $\gamma = 0.5$ at $t = 100$. Circles correspond to the one-mode kernel degree distribution, i.e., Eq. (19), while the stars are the one-mode projection of the α-BiN.

III. REAL WORLD α-BIN

A. CoGNet: the 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 some earlier work [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 *random mutation* in a gene sequence. A gene sequence is a string defined over four symbols (A, G, T, and C) that represent the nucleotides. A *codon* is a triplet of adjacent nucleotides (eg. 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 between different phylogenetic groups.

1. *Definition and construction*

We refer to the network of codons and genes as CoGNet and represent it as an α-BiN where V is the collection of *genes*, i.e., 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. Fig. 1 illustrates the structure of CoGNet.

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 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 non-trivial task of estimating the probability distribution p_k, having a very large event space (between 0 and 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 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 number 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 data using both the 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,

$$P_{k,t} = \sum_{i=0}^{k} p_{i,t}\,. \tag{20}$$

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 distribution* as stated in Eq.(20), in 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. (20) will be used. Finally, Fig. 8 shows a comparison between the empirical degree distribution for Xenopus leavis (circles) and the best fit obtained using Eq. (5) (see below for details) for both, $p_{k,t}$ and $P_{k,t}$.

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

TABLE II: The values of γ that yield best fit for the degree distribution under the two different *binning* strategies. The "gene count" serves to provide context but not input data.

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

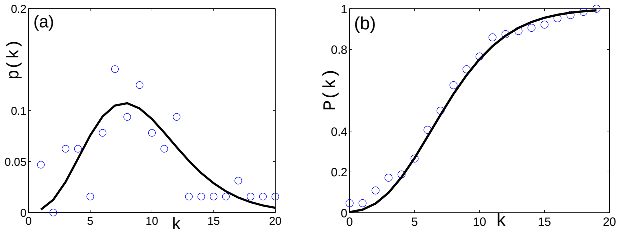

FIG. 8: Degree distribution of the codon nodes for *Xenopus leavis*. In (a) a comparison between the empirical data (circles) and the theoretical $p_{k,t}$ obtained using Eq. (5) (black solid curve) is shown. The cumulative distribution of the real data (circles) and the theory (black solid curve) is shown in (b).

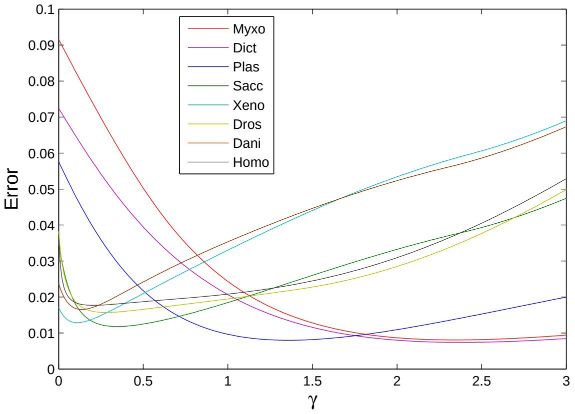

FIG. 9: Error as defined by Eq. (21) as function of γ for the eight organisms using fixed bin size. Steps in γ correspond to 0.01.

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 [35]. Therefore, we choose to apply the "sequential attachment" based growth model for synthesis of CoGNet. This means that we model the CoGNet growth through equations (3) and (5).

For all the CoGNets, the value of N is 64, μ is 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. (5), 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:

$$Error = \sum_{k=0}^{\infty} (p_{k,t}(\gamma) - p^*_{k,t})^2 \,, \tag{21}$$

where $p^*_{k,t}$ represents the empirical distribution, while $p_{k,t}(\gamma)$ is the theoretical distribution given by Eq. (5). 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 it can be seen in Fig. 9, 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. Thus, the minimum was determined with a precision of ± 0.01. Notice that since the fitting of the empirical distribution does not need to rely on the identification of a *heavy tail* dominated by large but rare fluctuations, the description of the empirical CoGNet data (see Fig. 8(a)) is easier and safer than the description of apparent power-law behavior in empirical data [34].

Fig. 10 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 section II, Eq. (5) describes four possible degree distribution depending on γ, two of them corresponding to a range of γ: i) $0 < \gamma < 1$ and ii) $1 \leq \gamma \leq (N/\mu) - 1$ (see Fig. 2 (b) and (c)). 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.

We conclude that at least at the level of codon usage, and according to our α-BiN analysis, *Myxococcus xanthus*, *Dictyostelium discoideum*, and *Plasmodium falciparum* can be put in the same category, while the other five organisms belong to a different class. Moreover, the above 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 to 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 [36]. 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. PlaNet: the phoneme-language network

In this section, we attempt to explain the self-organization of the consonant inventories through α-BiN 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 exceptionally regular patterns across the languages of the world, which is arguably an outcome of the self-organization that goes on in shaping their structures [37]. In order to explain this self-organizing behavior of the sound inventories, various functional principles have been proposed such as *ease of articulation* [38, 39], *maximal perceptual contrast* [38] and *learnability* [39]. The structure of vowel inventories has been successfully explained through the principle of maximal perceptual contrast [38, 39]. Although there have been some linguistically motivated work investigating the structure of the consonant inventories, most of them are limited to certain specific properties rather than providing a holistic explanation of the underlying principle of its organization.

1. *Definition and construction*

A first study of the consonant-language network as an α-BiN can be found in [40]. Here we follow the same definitions given in [40] and refer to the consonant-language α-BiN as PlaNet or Phoneme-Language Network. 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. On the other hand, the one-mode projection of PlaNet onto the consonant nodes is called PhoNet. Fig. 11 illustrates the structures of PlaNet and PhoNet. Note that PlaNet is an unweighted bipartite graph, whereas PhoNet has been represented as a weighted graph.

Many typological studies [38, 41, 42] of segmental inventories have been carried out in the past on the UCLA Phonological Segment Inventory Database (UPSID) [43]. 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 PlaNet. Consequently, there are 317 elements (nodes) in the set V and 541 elements (nodes) in the set U. The cardinality of E, as computed from PlaNet and PhoNet, is 7022 and 30412 edges, respectively. We se-

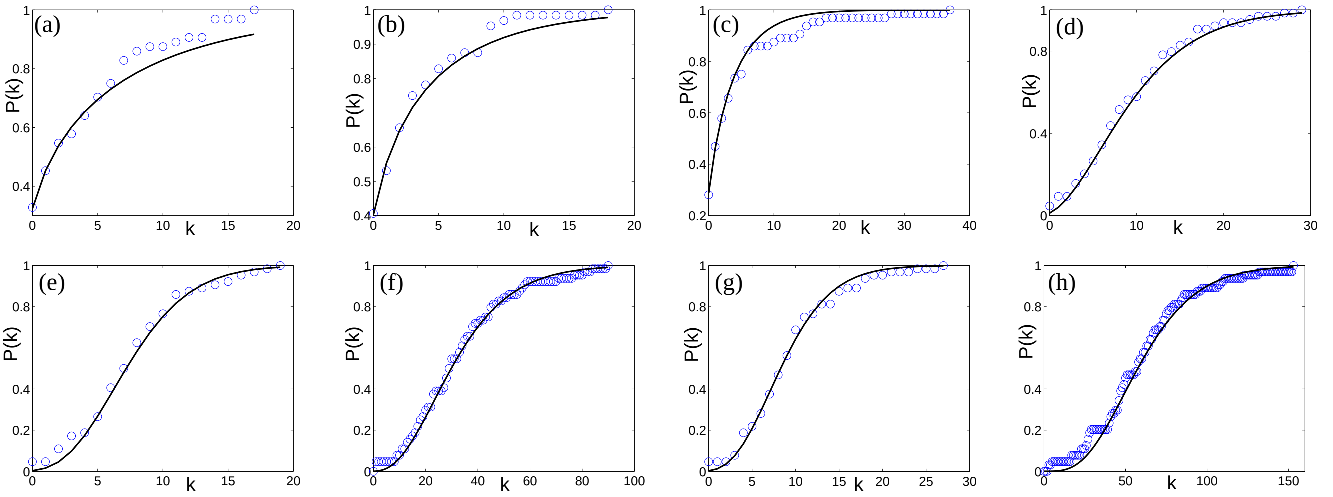

FIG. 10: Cumulative degree distributions for the empirical data (circles) and their corresponding theoretical best γ-fits through Eqs. (3) and (5) (solid curve) for the organisms. (a) *Myxococcus xanthus*, (b) *Dictyostelium discoideum*, (c) *Plasmodium falciparum*, (d) *Saccharomyces cerevisiae*, (e) *Xenopus laevis*, (f) *Drosophila melanogaster*, (g) *Danio rerio*, and (h) *Homo sapiens*.

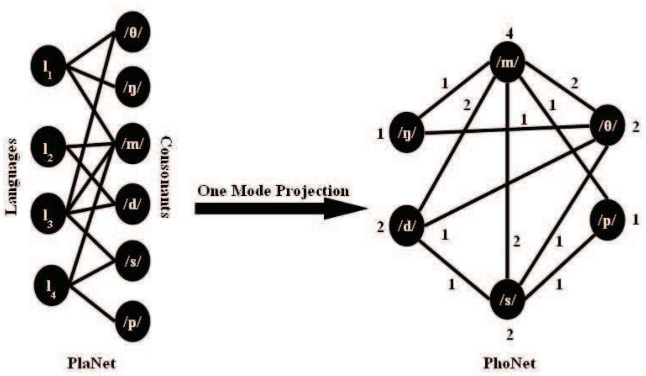

FIG. 11: Illustration of the nodes and edges of PlaNet and PhoNet.

lected 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 that languages share the same origin.

2. *Topological properties*

Fig. 12 illustrates the (cumulative) degree distribution of U. Since the degree of a language node is nothing but the size of the 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. Recall that in the theory for α-BiN the degree of each node in V has been assumed to be a constant μ.

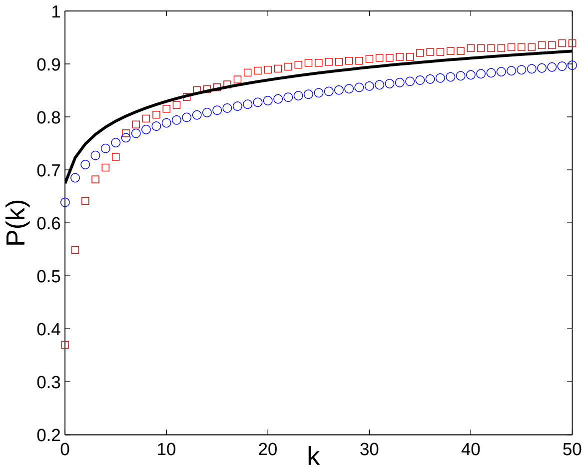

FIG. 12: Cumulative degree distribution of U, i.e., the consonant nodes. Squares correspond to the empirical data, and circles to simulations performed with “parallel attachment without replacement” with $\gamma = 14$ (PlaNet$_{sim}$). The solid line corresponds to the theoretical solution for “parallel attachment with replacement” (PlaNet$_{theo}$) obtained through integration of Eq. (10) with $\gamma = 14$.

3. *Growth models*

In order to obtain a theoretical description of the degree distribution of the consonant nodes in PlaNet (and later on PhoNet), we employ the α-BiN growth model described in Sec. II B. We assume that all the language nodes have a degree $\mu = 22$. Clearly, $N = 541$ is the total number of consonant nodes and $t = 317$ is the total number of languages. Thus, γ is the only free parameter in the model. Notice that, by definition, in PlaNet a consonant can occur only once in a language inventory. Therefore, unlike the case of CoGNet, PlaNet is an α-BiN that has been constructed using a "parallel attachment without replacement" scheme. However, we expect the theory developed in Sec. II B, corresponding to "parallel attachment with replacement", to be a fairly good approximation for the degree distribution of PlaNet. We shall refer this theoretical model of PlaNet as PlaNet$_{theo}$. In order to estimate the free parameter γ, the best fit was obtained with $\gamma = 14$ (see Fig. 12). Since $1 \leq \gamma \leq N/\mu = 24.6$, based on our theoretical analysis we can conclude that the degree distribution is consistent with an attachment kernel where preferetial attachment has a strong weight, i.e., γ takes a large value. Consequently, the degree distribution is consistent with a beta distribution, exhibiting its mode at $k = 1$.

To study the effect of the "parallel attachment without replacement" scheme, we carry out stochastic simulations with such a growth model described below. Suppose that a language node v_i (with degree 22) is added to the system and that $j < 22$ edges of the incoming node have already been attached to u_1, u_2, ..., u_j distinct consonant nodes. Then, the $(j + 1)$th edge is attached to a consonant node based on the same preferential attachment kernel (see Eq. 1), but applied on the reduced set $U - \{u_1, u_2, \ldots, u_j\}$, i.e., the previously selected u_1, u_2, ..., u_j consonant nodes cannot participate in the selection process of the $(j + 1)$th edge of v_i. This ensures that a consonant node is never chosen twice. We shall refer to the degree distributions of the consonant nodes obtained in this way as PlaNet$_{sim}$. The degree distribution of PlaNet$_{sim}$ has the best match with the degree distribution of the real PlaNet when $\gamma = 14$.

We have calculated the error for the aforementioned stochastic simulation model (E_{sim}) as well as the theory of Sec. II B, corresponding to "parallel attachment with replacement" (E_{theo}). The error has been computed using Eq. (21) where $p^*_{k,t}$ stands for the degree distribution of the real PlaNet. It is found that $E_{sim} = 0.0972$ and $E_{theo} = 0.1170$. Since the simulation using the "parallel attachment without replacement" scheme describes the structure of consonant inventories better, the error in this case is smaller than that for "parallel attachment with replacement".

4. *One-mode projection: PhoNet*

Interestingly, when we reconstruct the one-mode projection from either the theory using the "attachment with replacement" scheme (PhoNet$_{theo}$) or stochastic simulation considering the "attachment without replacement" model (PhoNet$_{sim}$), we cannot match the empirical data. Fig. 13 shows the cumulative degree distributions of PhoNet$_{sim}$, PhoNet$_{theo}$ and real PhoNet. We have calculated the error of PhoNet$_{sim}$ and PhoNet$_{theo}$ with respect to the real PhoNet using the Eq. (21) and refer them as (E_{sim}) and (E_{theo}) respectively. Experiments reveal that $E_{sim} = 0.1230$ and $E_{theo} = 0.1438$. The results show a larger quantitative difference between the curves compared to that between their bipartite counterparts. It indicates that the one-mode projection has a more complex structure than that could have emerged from a simple preferential attachment based kernel.

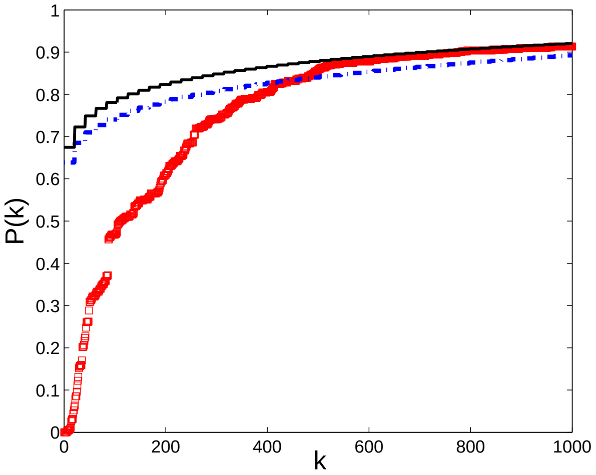

FIG. 13: Cumulative degree distribution of the one-mode projection of PlaNet (PhoNet). Squares correspond to the empirical data (Real PhoNet), dash-dotted line to simulations of one-mode projection model with "attachment without replacement" using kernel Eq. (1) (PhoNet$_{sim}$). The solid curve shows the theoretical degree distribution with the "attachment with replacement" scheme using Eq. (11) (PhoNet$_{theo}$).

Anyway, 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 micro-level dynamics that manifests itself as preferential attachment in PlaNet. However, the fact that the co-occurrence distribution of the consonants, i.e., the degree distribution of PhoNet, is not explained by the growth model implies that there are other organizing principles absent in our current model 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 freamework of Urn models that are popular in probability theory. Particularly relevant to us is the so-called Finite Pólya's process [23, 24]. 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 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 α-BiN growth model for sequential attachment with $\mu = 1$. However, there are 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 $t = 0$. In contrast, in the α-BiN growth model we can assume any initial condition. In particular, we have studied the case where at $t = 0$, all bins are empty, because in the α-BiN growth model the probability of getting a new ball is proportional to $\gamma n_i(t) + 1$. Defining this probability as consisting of two additive terms has another advantage – we can change γ, that is our model parameter, so as to control the weight between both the terms. 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 $\eta = 1$, exactly corresponds to the α-BiN model, under the mapping $n_i^{\alpha BiN} \mapsto n_i^{Polya} - 1/\gamma$. Nevertheless, this generalization, whereby we control the nature of the emergent degree distribution by varying the parameter γ has not been studied previously. Instead, the shape of the urn-size distribution (equivalent to degree distribution in α-BiNs) in a Finite Pólya's process is controlled by varying η. It is important to notice that the use of a linear attachment probability has allowed us to obtain an analytical closed form for the case of sequential attachment, and to derive the exact expression for the time evolution of the degree distribution for parallel attachment, which is not the case for the Finite Pólya's process. The use of a non-linear attachment probability makes the analytical treatment of the latter model much more difficult. However, we stress that it could be interesting to explore the α-BiN growth model that results from combining both degrees of freedom, η and γ, in such a way that the attachment probability becomes proportional to $\gamma n_i^\eta(t) + 1$. Though in this case the analytical treatment of the problem becomes much more involved, it is possible that such a treatment could explain the empirical data more accurately.

Finally, there is another important difference between Pólya's and α-BiN growth models that is worth mentioning. While Pólya's urn model is defined in such a way that at each time step only one ball enters into the system, in the α-BiN growth model μ balls are added into the system per time step. As extensively discussed in section II, the solutions of the problem are remarkably different when the μ balls are added sequentially or in parallel to the system.

Another class of models for non-growing bipartite networks developed by Evans and Plato in [20] (henceforth the EP Model) is based on the concept of rewiring and closely resemble 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 α-BiN 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:

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

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 α-BiN growth model for sequential attachment, 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.

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

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 EP model the distributions reach a steady state, while this does not occur in α-BiNs. In addition, while for α-BiNs, 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. DISCUSSION AND CONCLUSION

In the preceding sections, we have presented growth models for discrete combinatorial systems in the framework of a special class of networks – α-BiNs. To summarize some of our important contributions, we have

- proposed growth models for α-BiNs, which are based on preferential attachment coupled with a tunable randomness component,
- extended the mathematical analysis presented in [22] and derived the exact expression for the degree distribution in case of parallel attachment,
- analytically derived the degree distribution of the one-mode projection,
- and presented case studies for two well-known DCSs from the domain of biology and language, in order to illustrate how to apply our analytical findings to describe the empirical data, and to discuss the limitations of α-BiNs as a modeling tool.

A natural generalization of the α-BiN growth models introduced here would involve the use of non-linear attachment kernels, as discussed above, and the modeling of rewiring during the growth of the α-BiN . In fact, it has been shown through simulations that the degree distribution of the consonant nodes in PlaNet is better explained by having a superlinear kernel as opposed to a linear kernel presented here [44]. An analytical treatment of such a non-linear kernel should be an interesting topic for future research.

There are also some limitations in the study of CoGNet. Selection of correct binning policy to construct the CoGNet is a challenging job. Modeling the CoGNet with parallel attachment where μ is the average number of codons present in the genes is a direct extension of the current work. As a first step, we here classified the eight organisms into two sets and we believe that our new method can further contribute to the reconstruction of phylogenetic relations. Our approach may be especially useful for the analysis of such genome sequences which are so far only available in fragments either due to fragmentary sampling of the biological material or to un-finished sequencing efforts.

Finally, we would like to argue that the condition that one of the partitions in the α-BiN model has to be strictly fixed in size can be relaxed. It is possible to find some real systems where the set of basic units also grow, although at a far slower rate than the collection of their discrete combinations. Under this condition we can expect the reported results to approximately hold. However, it would be very interesting to study to what extent the rate of growth of the two partitions should differ for the current theoretical predictions to remain valid.

Acknowledgments

This work was partially financed by the Indo-German collaboration project DST-BMBT 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 TU-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, 1995).
[2] J.J. Ramasco, S.N. Dorogovstev, and R. Pastor-Satorras, Phys. Rev. E **70**, 036106 (2004).
[3] D.J. Watts and S.H. Strogatz, Nature **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. **1**, 01010 (2006).
[6] L.A.N. Amaral et al., Proc. Natl. Acad. Sci. **97**, 11149 (2000).
[7] M.E.J. Newman, Phys. Rev. E **64**, 016132 (2001).
[8] A.-L. Barabasi 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 **410**, 268 (2001).
[12] Eubank et al., Nature **180**, 429 (2004).
[13] R. Ferrer i Cancho, O. Riordan, O. and B. Bollobás, Proc. R. Soc. B **272**, 561 (2005).
[14] S.M.G. Caldeira, T.G.P. Lobão, R.F.S. Andrade, A. Neme and J.G.V. Miranda, European Physical Journal B **49**, 523-529, (2006).
[15] J.-L. Guillaume and M. Latapy, Information Processing Letters **90**, 215 (2004).
[16] W. Souma, Y. Fujiwara and H. Aoyama, Physica A **324**, 396 (2003).
[17] K. Sneppen, Europhys. Lett. **67**, 349 (2004).
[18] A.-L. Barabasi and R. Albert, Science **286**, 509 (1999).
[19] J. Ohkubo, M. Yasuda and K. Tanaka, Phys. Rev. E **72**, 065104 (2005).
[20] T.S. Evans and A.D.K. Plato, Phys. Rev. E **75**, 056101 (2007).
[21] W. Dahui, Z. Li, and D. Zengru, Physica A **363**, 359 (2006).
[22] F. Peruani, M. Choudhury, A. Mukherjee, and N. Ganguly, Europhys. Lett. **79**, 28001 (2007).
[23] N. Johnson and S. Kotz, *Urn Models and Their Applications: An approach to Modern Discrete Probability Theory*, (Wiley, New York, 1977).
[24] F. Chung, S. Handjani, and D. Jungreis, Annals of Combinatorics **7**, 141 (2003).
[25] S.N. Dorogovtsev and J.F.F. Mendes, *Evolution of Net-*

works: From Biological Nets to the Internet and WWW (Oxford University Press, 2003).

[26] R. Ferrer i Cancho and R.V. Solé and R. Köhler, Phys. Rev. E **69**, 051915 (2004).

[27] M.E.J. Newman, S.H. Strogatz, and D.J. Watts, Phy. Rev. E **64**, 026118 (2001).

[28] M.E.J. Newman, Proc. Natl. Acad. Sci. **101**, 5200 (2004).

[29] P. Sharp et al., Nucl. Acids Res. **16(17)**, 8207 (1988).

[30] S.B. Hedges, Nature Reviews **3**, 838 (2002).

[31] Y. Nakamura, T. Gojobori, and T. Ikemura, Nucl. Acids Res. **28**, 292 (2000).

[32] Codon usage database: http://www.kazusa.or.jp/codon/

[33] M.E.J. Newman, SIAM Review **45**, 167 (2003).

[34] A. Clauset, C. Rohilla Shalizi, M.E.J. Newman, arXiv:0706.1062v2 (2009).

[35] T. Kunkel and K. Bebenek, Annual Review of Biochemistry **69**, 497 (2000).

[36] Fredman et al., Nature Genetics **36**(8), 861 (2004).

[37] P.-Y. Oudeyer, *Self-organization in the Evolution of Speech*, (Oxford University Press, 2006).

[38] B. Lindblom and I. Maddieson, Language, Speech, and Mind, 62 (1988).

[39] B. de Boer, Journal of Phonetics **28**, 441 (2000).

[40] M. Choudhury et al., Proceedings of COLING–ACL **P06**, 128 (2006).

[41] F. Hinskens and J. Weijer, Linguistics **41**, 1041 (2003).

[42] P. Ladefoged and I. Maddieson, *Sounds of the World's Languages*, (Oxford, Blackwell, 1996).

[43] I. Maddieson, *Patterns of Sounds*, (Cambridge University Press, 1984).

[44] A. Mukherjee et al., Journal of Quantitative Linguistics, http://arxiv.org/abs/physics/0610120 (2008).

[45] The names *with* and *without replacement* refer to the fact that in the *without replacement* case, when a basic unit u_k has been selected by one of the μ edges of node v_i, that basic unit is removed from the set of available basic units for the next edges of v_i. In contrast, in the *with replacement* case, if u_k is selected, it is *replaced* back in the set of available basic units for the next edges of v_i. So, the same basic unit can be selected more than once by the same V node.