Statistics of Shared Components in Complex Component Systems

Andrea Mazzolini,[1] Marco Gherardi,[2,3] Michele Caselle,[1] Marco Cosentino Lagomarsino,[2,3,4] and Matteo Osella[1,*]

[1]*Physics Department and INFN, University of Turin, via P. Giuria 1, 10125 Turin, Italy*
[2]*Sorbonne Universités, UPMC Univ Paris 06, UMR 7238,*
Computational and Quantitative Biology, 4 Place Jussieu, Paris, France
[3]*CNRS, UMR 7238, Paris, France*
[4]*FIRC Institute of Molecular Oncology (IFOM), 20139 Milan, Italy*

(Received 27 July 2017; revised manuscript received 29 January 2018; published 20 April 2018)

Many complex systems are modular. Such systems can be represented as "component systems," i.e., sets of elementary components, such as LEGO bricks in LEGO sets. The bricks found in a LEGO set reflect a target architecture, which can be built following a set-specific list of instructions. In other component systems, instead, the underlying functional design and constraints are not obvious *a priori*, and their detection is often a challenge of both scientific and practical importance, requiring a clear understanding of component statistics. Importantly, some quantitative invariants appear to be common to many component systems, most notably a common broad distribution of component abundances, which often resembles the well-known Zipf's law. Such "laws" affect in a general and nontrivial way the component statistics, potentially hindering the identification of system-specific functional constraints or generative processes. Here, we specifically focus on the statistics of shared components, i.e., the distribution of the number of components shared by different system realizations, such as the common bricks found in different LEGO sets. To account for the effects of component heterogeneity, we consider a simple null model, which builds system realizations by random draws from a universe of possible components. Under general assumptions on abundance heterogeneity, we provide analytical estimates of component occurrence, which quantify exhaustively the statistics of shared components. Surprisingly, this simple null model can positively explain important features of empirical component-occurrence distributions obtained from large-scale data on bacterial genomes, LEGO sets, and book chapters. Specific architectural features and functional constraints can be detected from occurrence patterns as deviations from these null predictions, as we show for the illustrative case of the "core" genome in bacteria.

DOI: 10.1103/PhysRevX.8.021023

Subject Areas: Biological Physics, Complex Systems,
Interdisciplinary Physics

I. INTRODUCTION

A large number of complex systems in very different contexts—ranging from biology to linguistics, social sciences, and technology—can be broken down to clearly defined basic building blocks or components. For example, books are composed of words, genomes of genes, and many technological systems are assemblies of simple modules. Once components are identified, a specific realization of a system (e.g., a specific book, a LEGO set, a genome) can be represented by its parts list, which is the subset of the possible elementary components (e.g., words, bricks, genes), with their abundances, present in the realization. We use the term "component systems" for empirical systems to which this general representation can be applied.

Occurrence patterns of components across realizations are expected to reveal relevant architectural constraints. For example, the bricks present in each LEGO set clearly reflect a target architecture that can be built with them following the instruction booklet. While for LEGO sets the assembly instructions are provided by the seller, in most component systems the architectural constraints are not obvious. Inferring such constraints from the statistics of components may answer important questions about the nature of a system. For example, it could reveal new clues about the complex combination of selective pressure and random events that shaped the functional composition of extant genomes. Even in those cases where the architecture is partially or even fully known and the instruction manual is available, the statistics of components may help us distill

[*]To whom correspondence should be addressed.
mosella@to.infn.it

some general principles characterizing a given class of component systems, in some cases revealing basic features of the underlying generative processes.

In order to perform detection of system-dependent features from patterns of shared components, we need to have a clear idea of the general behavior of component systems even in the absence of functional constraints on the presence or absence of specific classes of components. This is by itself a challenging task, as such systems show a large degree of nontrivial universal properties [1–3] that could in principle affect the occurrence statistics. Indeed, several notable quantitative laws can be identified in the composition of component systems of very different nature. This is well known, e.g., in linguistics, where the notorious "Zipf's law" [4] describing the word frequency distribution (or its equivalent rank plot) in a linguistic corpus has been the subject of extensive investigations [5–9]. In this context, the existence of quantitative "universal" laws may in principle provide insights on the cognitive mechanisms of text production, and can have practical applications in data mining and data search techniques [1]. Analogously, for genomes across the whole tree of life, the number of genes in different evolutionary families is power-law distributed, a discovery that represents one of the first examples of "laws" of the genome sequencing era [2,10]. Such heterogeneous usage of the different basic components, often resulting in an approximately power-law distribution of their frequencies, can be seen as a hallmark of the complexity of component systems [6].

A large body of theoretical work addresses the origins of this heterogeneity. Several models have emerged in different areas of science, with context-specific ingredients. For example, stochastic processes based on gene duplication, deletion, and innovation have been proposed as simple evolutionary models of genome evolution at the basis of the observed heterogeneous component usage [3,11–13]. On the other hand, specific communication optimization principles [14,15] and stochastic models for text generation [5,16,17] have been invoked to explain the emergence of Zipf's law in natural language. In many, but not all, of these models a preferential attachment principle is at the origin of the emergence of the power-law distribution of component frequencies. More importantly, the ubiquity of this emergent behavior raises the question of whether (and to what extent) empirical laws like Zipf's law are pervasive statistical patterns that transcend system-specific mechanisms [2,18]. In this spirit, the analysis of radically different systems can help the discovery of patterns that descend from pure statistical effects or general principles [18,19].

Here, we analyze empirical data from three very different component systems from linguistics (book chapters), genomics (protein domain families in sequenced genomes), and technology (LEGO toys) and we look for general statistical *consequences* of their heterogeneous frequency distributions. The different data sources considered here reasonably do not share any generative mechanisms, nor are they expected to share the same type of constraints, selection criteria, or optimization principles. However, the frequency of their components is heterogeneous and they all obey laws that are similar to Zipf's.

The marginal statistics that we concentrate on is the fraction of components that are shared among a certain number of realizations, for example, the fraction of LEGO bricks with the same shape found in a given fraction of unequal LEGO boxes. In genomics, this is the so-called "gene-frequency distribution," which was shown to follow a U shape at several taxonomic levels [20–22]. A U shape of this distribution of shared components indicates that there is a set of "core" components that are common to most realizations, as well as an enriched set of realization-specific components. This histogram also decays approximately as a power law for rare components, both in genomic data and in technological systems [19]. In evolutionary genomics, the origins of this pattern are the focus of a lively debate. The pattern has been rationalized theoretically by neutral or selective population dynamics models [22–25], or as a consequence of functional dependencies among different components [19]. For component systems outside of genomics, the distribution of shared components remains underexplored, and is typically neglected by the current debate, for example, in linguistics [1].

Using theoretical calculations based on random sampling of components (with replacement) from their overall frequencies (estimated by their total abundance across empirical realizations), we show that a distribution of shared components with a power-law behavior is a general feature of component systems not only with Zipf-like component frequency distributions, but also for general power laws and exponential decay of the overall component frequencies. In other words, a U-shaped distribution of shared components can naturally emerge in component systems with a heterogeneous component usage (which is often the case empirically). Importantly, we quantitatively identify the general features of the system leading to a U-shaped distribution of shared components, a given core size, and a specific decay of the realization-specific bulk of this distribution.

II. DATA

A. Data sources

1. Genomes

We use the superfamily classification of protein domains from the SUPERFAMILY database [26] considering a set of $R = 1061$ prokaryotic genomes ("realizations") and a total number of different families $N = 1531$ ("components"). Protein domain families are the basic modular topologies of folded proteins [27]. Different domains of the same family can be found in each genome in the same or different proteins. As a functional annotation of protein domains in SUPERFAMILY, we considered the SCOP annotations mapped into 7 general function categories, as developed by Vogel and Chothia [28].

2. LEGO sets

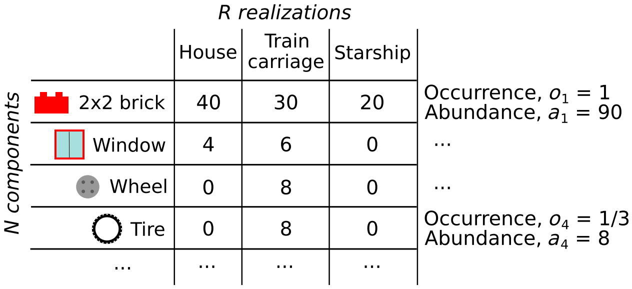
The composition in bricks of several LEGO sets ($R = 2820$) can be freely downloaded [29]. We exclude from the analysis LEGO sets belonging to the category of "LEGO Technic" since, by construction, they share a very small number of bricks with the classic LEGO toys. Similarly, we do not consider LEGO sets with less than 80 components or belonging to the categories "Educational and Dacta" and "Supplemental," in order to exclude sets that are actually collections of spare parts or additional bricks for other sets.

3. Texts

The analyzed linguistic corpus is composed by $R = 1721$ book chapters (realizations) of several English books randomly chosen from the most popular ones in the Project Gutenberg database [30]. We define chapters as realizations, instead of entire books, to obtain a corpus with a range of sizes (total number of components per realization) comparable to the one of genomes and LEGO toys [Fig. S1 of Supplemental Material (SM) [31]]. The complete list of books considered is reported in Table S1 of SM [31]. The elementary components are defined as the words regardless of capitalization (e.g., "We" and "we" are considered as the same component).

B. Data structure: Matrix representation of component systems

A set of empirical realizations of a component system can be naturally described as a matrix $\{n_{ij}\}$ defined such that the entry n_{ij} represents the abundance of the component i ($i = 1, \ldots N$) in the realization j ($j = 1, \ldots, R$). Thus, each realization (a literary text, a LEGO set, or a prokaryotic genome) is represented as a matrix column (Fig. 1). Some key observables can be easily defined using this representation. First, the total abundance a_i of the component i in the whole ensemble is defined by summing over all realizations $a_i = \sum_j n_{ij}$. The normalized abundance represents the component frequency $f_i = (a_i / \sum_i a_i)$. The "component occurrence" o_i is instead defined as the fraction of realizations in which the component is found; thus, $o_i = (1/R) \sum_j (1 - \delta_{n_{ij},0})$. Two other crucial quantities are the total number N of different components in the system, which is essentially the number of bricks of different shape or the vocabulary, and the size of a realization j, defined as the total number of its components $M_j = \sum_i n_{ij}$.



	House	Train carriage	Starship
🟥 2x2 brick	40	30	20
▨ Window	4	6	0
● Wheel	0	8	0
○ Tire	0	8	0
…	…	…	…

FIG. 1. Matrix representation of complex component systems. (a) Each column is a realization (e.g., a LEGO set, a genome, or a book chapter) and each row is a component type (e.g., a LEGO brick, a protein domain family, a word). The element n_{ij} represents the abundance of component i in realization j. The frequency f_i of component i is given by its total abundance (90 for the red brick, 8 for the tire) divided by the total number of components in the system. The occurrence o_i of component i is the fraction of realizations (toys in the example) in which there is at least one token of i (1 for the red brick, $1/3$ for the tire).

III. RESULTS

A. Component frequency distribution and distribution of shared components show general features across systems

This section illustrates two empirical laws in the analyzed datasets (LEGO toys, bacterial genomes, and literary texts). We first consider the component frequencies in the whole universe of available realizations of a given system, which is essentially the generalized Zipf's law [6] for the three systems. Figure 2 shows the rank plots of these component frequencies. The three data sets share a power-law behavior for components with high frequencies (low rank), with an exponent close to 1 as in the classic Zipf's law [4], and a faster decay at higher ranks (components with low frequency). This double-scaling behavior has been recently observed in the context of linguistics [17]. In evolutionary genomics, the gene frequency was previously analyzed over single genomes and shown to be approximately power-law distributed with an exponent dependent on genome size [3,10]. Figure 2 shows that the same distribution calculated over thousands of prokaryotic genomes has a double scaling, with an exponential-like decay for low ranks in its rank plot. We tested that the shape of these component frequency distributions do not strongly depend on the specific size or number of realizations analyzed. The rank plots in Fig. 2 do not vary when evaluated on different subsamples of the whole data sets (Fig. S2 of the SM [31]). This suggests that the frequency distributions evaluated using the available finite empirical data sets estimate reliably the global heterogeneity of the component usage in the systems.

We aim to also evaluate the distribution of shared components, $\{o_i\}$, and how much of its features can be explained from other measurable quantities, namely, the

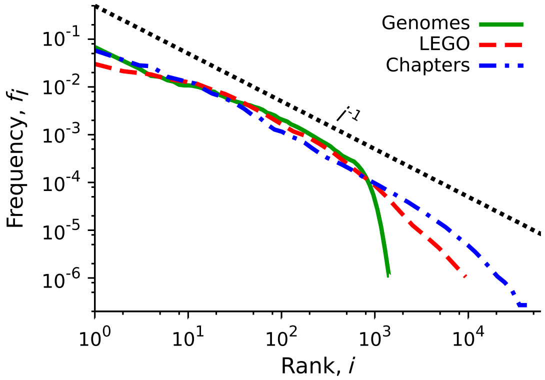


FIG. 2. Different empirical component systems show similar component frequency distributions. The rank plot of the component frequencies is reported for the three data sets (book chapters, genomes, LEGO sets). The frequency of a component is defined as the abundance of that component in the whole data set normalized by the total number of components (Fig. 1). The three curves follow similar behavior, which can be described qualitatively as a power-law-like decay with exponent close to 1 for low ranks (high frequency), and a faster data-set-specific decay for higher ranks.

component frequencies, the realization sizes $\{M_j\}$, and the number of different components in the universe N. Figure 3 shows this distribution for the three data sets considered here. For small occurrences, the plots are compatible with a power-law decay, with a data-set-specific exponent. Only for genomes this curve is clearly U shaped(see also Fig. S3 of SM [31]) and shows a "core" of shared components, i.e., protein domains shared by almost all the genomes, together with a rich group of rare components. Book chapters do not show this marked behavior, due to the fact that the ubiquitous words (e.g., articles, pronouns, prepositions) are much less than the chapter-specific words. Finally, LEGO sets display no core of shared components, and this is probably due to the wide range of themes using poorly overlapping brick types.

B. Random-sampling model as a minimal model for component systems with defined component frequencies

In order to identify the statistical consequences of a heterogeneous usage of components on the statistics of shared components, a suitable model is needed. In particular, we would like to generate system realizations starting from a fixed component frequency distribution without any additional functional information or constraint. To this end, we employ a random-sampling procedure [8,17,32,33] that builds artificial realizations through an iterative random extraction (with replacement) of components from their frequencies $\{f_i\}$ in the whole system. Each realization size M is specified by the number of random extractions.

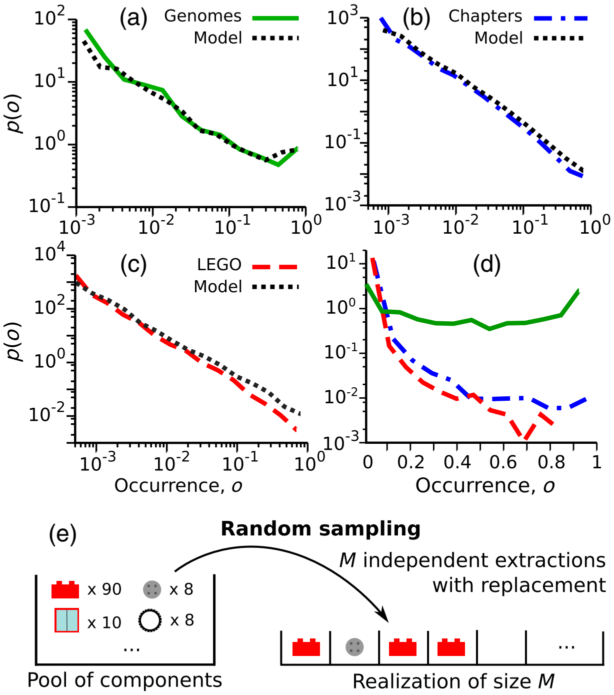


FIG. 3. The random-sampling model captures the main features of the empirical statistics of shared components. The plots show the normalized distribution $p(o)$ of component occurrences, quantifying the statistics of shared components for the three data sets: genomes (a), book chapters (b), and LEGO sets (c). The log-log scale highlights the power-law-like decay. The black dashed lines represent the prediction of the random-sampling model assuming the empirical component frequencies and realization sizes. The model reproduces very well the power-law decay, but may differ quantitatively from the empirical laws in the high-occurrence region. Panel (d) plots the same quantities in log-lin scale, to highlight the quantitative differences between systems and the presence or absence of a peak of core components. Note that the different range of the y-axis values with respect to previous panels is due to the different binning procedures, logarithmic vs linear. (e) Scheme of the random-sampling process: samples of size M are generated from independent draws from the "universe" of all possible components with their specific abundances. Therefore, the probability of a component extraction is proportional to its global abundance, i.e., the sum of its abundances over all realizations of the systems.

More precisely, the following prescriptions [Fig. 3(e)] define the random-sampling model that is used in the following. (i) The component abundance rank distribution is assumed to be a universal property of the component system and well represented by the empirical overall abundances (see Fig. S2 and the SM for a discussion of this assumption [31]). (ii) The extraction probability of a component is proportional to its overall abundance. (iii) A realization of size M is generated by M independent extractions from the pool of components. Statements (ii) and (iii) define a multinomial process. Given a normalized list of component frequencies $\{f_i\}$, $i = 1, \ldots N$ (where N is the size of the available "vocabulary"), and the size M of the realization, the probability of a specific configuration $\{n_1, n_2, \ldots, n_N\}$, where n_i is the number of the components with frequency f_i, is

$$P(n_1, n_2, \ldots, n_N; M) = \frac{M!}{\prod_{i=1}^{N} n_i!} \prod_{i=1}^{N} f_i^{n_i} \tag{1}$$

under the constraint that $\sum_{i=1}^{N} n_i = M$. Note that the expected value of n_i is Mf_i. Therefore, on average the global abundance distribution is conserved in each realization. In other words, the component composition in each realization is a sampled copy of the universe, without any of the possible complex correlations which may follow from architectural and functional properties of an empirical system.

For example, in the context of bacterial genome evolution, the random-sampling model translates into a scenario in which there is continuous and completely random horizontal gene transfer (exchange of genetic material) between species [34]. Thus, genome composition would simply reflect the pan-genome abundances of protein domains. While horizontal gene transfer is indeed a major force in bacterial evolution [21,35,36], several additional genome-specific functional constraints are clearly in place in evolution [35,37–40], and these are neglected by the model. Therefore, the random sampling can be considered as a null model useful to disentangle the consequences of the observed global heterogeneity in the component usage from actual hallmarks of more complex functional constraints.

C. Distribution of shared components is mainly a consequence of component frequencies, number of available components, and realization sizes

The fact that the distribution of shared components is qualitatively very similar in systems that are so different triggers the question of whether it may be an emergent statistical consequence of other system properties. In particular, we asked to what extent the statistics of shared components could be a direct consequence of component frequencies. As explained above, this question can be addressed quantitatively using a random-sampling model that generates an artificial copy of the empirical system by drawing realizations (whose sizes are fixed by the empirical ones) from the component frequency distribution. Figure 3 compares the empirical occurrence distributions with simulations of a random sampling. The null-model curves (dashed lines) provide very good approximations of the empirical laws, particularly for low component occurrences. Additionally, the model matches well the power-law decay with the system-specific exponent. Finally, the model predicts also the qualitative behavior of core components, and specifically that only genomes show a clear U-shaped distribution of shared components. The relative core sizes of the three systems are also well approximated, although there are some quantitative deviations from the empirical values that are addressed in detail in Sec. III F. These results suggest that the shape of the distribution of shared components in the three widely different empirical systems considered here is well described by a random-sampling model that conserves only the empirical component frequencies, the vocabulary (i.e., the set of possible components), and the realization sizes. The next section provides an analytical understanding of this observation.

D. A wide range of component frequency patterns lead to occurrence distributions with power-law decay and U shape

Thus far we have used the model only to address the specific statistics of component sharing of the empirical systems under consideration. To this end, we have simulated the random-sampling model fixing the component frequencies and realization sizes as in the empirical cases. More in general, one can ask whether a power-law decaying and/or U-shaped distribution of component occurrences is expected for a given distribution of component frequencies. To address this question, we have computed analytically the distribution of shared components under general prescriptions for the component frequency distributions within the random-sampling model.

For the sampling procedure explained in Sec. III B, the probability q_i that a component of rank i is present in a realization of size M_j is $q_i(M_j) = 1 - (1 - f_i)^{M_j}$, where f_i is the component probability of extraction. Therefore, the expectation value for the occurrence of component i over a set of R realizations is

$$o_i = \frac{1}{R} \sum_{j=1}^{R} q_i(M_j) = 1 - \frac{1}{R} \sum_{j=1}^{R} (1 - f_i)^{M_j}. \tag{2}$$

In order to obtain the probability distribution associated to this rank representation, one can use the fact that the rank of a component with occurrence o is the number of components with occurrence higher than o. In fact, these naturally correspond to components with higher frequency and thus lower rank. Therefore, we can write the rank $i(o)$ as

$$i(o) = \text{rank}(o) = \sum_{o'=o}^{o_1} N p(o') \simeq N \int_{o}^{o_1} p(o') do', \tag{3}$$

where o_1 is the highest possible occurrence, which corresponds to the component of rank 1. The function $i(o)$ is simply the inverse function of Eq. (2). From the approximate integral representation of $i(o)$, the occurrence probability distribution $p(o)$ is defined by the simple relation $(di(o)/do) = -Np(o)$.

Equation (3) provides a general relation between the representation of the frequency distribution as a rank plot and the representation as a probability distribution. Indeed, the arguments we present here to introduce Eqs. (2) and (3) have been used previously to establish the connection between Zipf's law as a rank plot and Zipf's law as a frequency distribution [41].

1. Observed versus possible vocabulary of components and Heaps' law

When a set of R realizations of size M is generated through a random-sampling procedure from a pool of $\tilde{N}$ possible different components with their probabilities of extraction $\{f_i\}$, the expected size N of the vocabulary that is actually sampled can be expressed as [32]

$$N = \tilde{N} - \sum_{i=1}^{\tilde{N}} (1 - f_i)^{MR}. \tag{4}$$

Thus, in general, $N \leq \tilde{N}$.

If the system size, defined by the total number of extractions MR, is large enough, essentially all possible components are expected to be sampled at least once, thus leading to the simplification $N \simeq \tilde{N}$ that we implicitly assume in Eq. (1). However, in general, the observed vocabulary in an ensemble of realizations is an increasing function of the system size, i.e., $N(MR)$. This functional dependence is equivalent to Heaps' law, which is the empirical power-law growth of the number of distinct components with the system size observed in linguistics [1,17] and in genomics [3]. This distinction between the observed and the possible vocabulary of components is discussed in more detail in the SM [31] and is relevant in the following sections.

2. Analytical distribution of shared components for component frequencies with a power-law or an exponential distribution

Explicit expressions for the occurrence distribution can be derived assuming a simple scenario, in which all realizations have the same size M, and the component frequency statistics follows a prescribed function. We first consider the empirically relevant case of a power-law frequency rank plot (Fig. 4, left-hand panel) defined by

$$f_i = \frac{1}{\alpha} i^{-\gamma}, \qquad \alpha = \sum_{i=1}^{\tilde{N}} i^{-\gamma}. \tag{5}$$

Under these assumptions and using Eqs. (2) and (3), the exact expression of the occurrence distribution can be calculated:

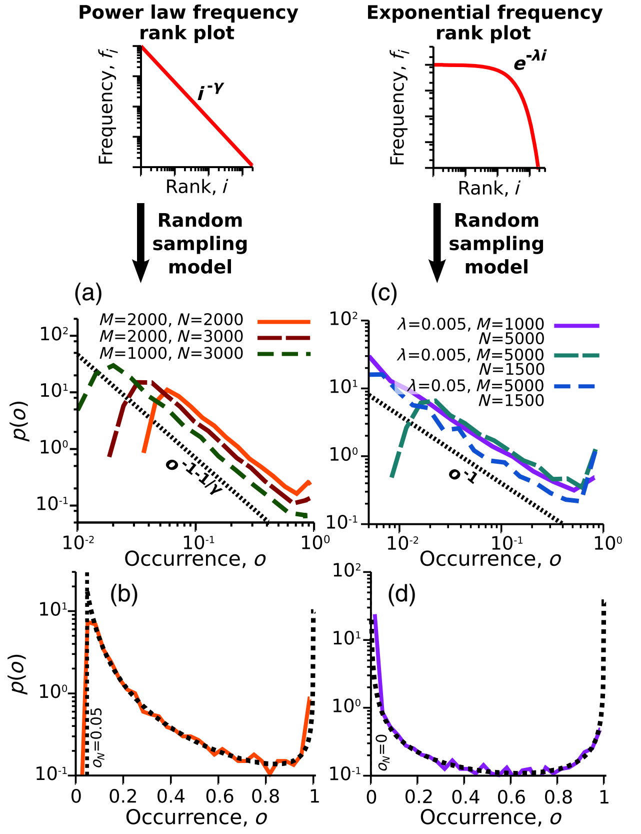


FIG. 4. Power-law decaying and U-shaped component occurrence distributions may descend from both power-law and exponential distributed universe component frequencies. (a) A power-law rank plot for the frequency (and thus for the abundance), whose exponent is $-\gamma$ ($\gamma = 1.2$ in the plot), produces a power-law decay of the component occurrence distribution with exponent $-1 - (1/\gamma)$, independently of the realization size M and the number of components N (for sufficiently large values of these parameters). (b) Agreement between the theoretical prediction of Eq. (6) (black line) and a simulated random sampling with parameters $R = 1000$, $N = 2000$, $\gamma = 1.2$, $M = 2000$ [the black vertical dashed line is the left boundary of the $p(o)$ domain]. Panels (c) and (d) are the counterpart of (a) and (b) for an exponential frequency rank plot. In this case $p(o)$ always decreases with exponent -1, for every value of λ, M, and N (sufficiently large). Parameter values are $R = 1000$, $N = 2000$, $\lambda = 0.005$, $M = 5000$. Given the system sizes MR in these examples, the number of possible different components essentially coincides with the vocabulary actually sampled; i.e., $\tilde{N} \simeq N$.

$$p(o) = \frac{(1-o)^{1/M-1}}{\gamma M N \alpha^{1/\gamma} [1 - (1-o)^{1/M}]^{1/\gamma+1}}. \tag{6}$$

The distribution is defined in the interval of occurrences $[o_N; o_1]$, where o_i is computed by Eq. (2) and N is the effective or observed component vocabulary, which can be a function of the system size, i.e., $N(MR)$, as described by Eq. (4)). Considering the limit of small occurrences and large sizes, i.e., $o \ll 1$ and $M \gg 1$, one finds precisely the empirically observed power-law decay. Specifically, in this limit the occurrence distribution takes the form

$$p(o) \simeq \frac{M^{1/\gamma}}{\alpha^{1/\gamma}\gamma N} o^{-1/\gamma-1}, \tag{7}$$

where the power-law exponent depends only on the exponent γ of the frequency rank plot.

Analogous calculations (details in the SM [31]) can be performed assuming a frequency distribution described by an exponential rank plot $f_i \sim e^{-\lambda i}$ (right-hand panel of Fig. 4). In this case, the distribution of shared components, for large enough realizations $M \gg 1$, has the expression

$$p(o) \simeq \frac{(1-o)^{-1}}{N\lambda \log[(1-o)^{-1}]}. \tag{8}$$

Interestingly, for rare families the above expression further simplifies to a power-law decay,

$$p(o) \simeq \frac{1}{N\lambda} o^{-1}, \tag{9}$$

with a "universal" exponent -1. This indicates that also systems with a heterogeneous but more compact frequency distribution are expected to show a power-law decay in the occurrence distribution. Figure 4 shows the agreement between these predictions and simulations of the random-sampling model for the two illustrative examples of a power law and of an exponential distribution of component frequencies. These analytical predictions have a dependence on the sampled vocabulary N and are expected to hold even if this is actually smaller than the total number of possible components $\tilde{N}$ (Fig. S5 of SM [31]). The effects of a dependence of the observed dictionary on system size [i.e., Heaps' law $N(MR)$] become relevant and has to be taken into account when comparing statistical features of ensembles of realizations with different sizes MR.

3. Shape of the distribution of shared components and rescaling properties

We now turn our attention to the conditions for a U-shaped distribution of shared components in the random-sampling model. Figures 4(a) and 4(c) already show that the decay of the occurrence of rare components is set only by the exponent γ as described by Eq. (7), but for different values of M and N the distribution may or may not display a significant fraction of core components. Additionally, Figs. 4(b) and 4(d) prove that Eqs. (6) and (8) can capture quantitatively the occurrence distributions and thus can well describe the relative proportion of core and specific components. In order to understand under what conditions this distribution becomes clearly U shaped for an underlying power-law frequency distribution, it is useful to note a rescaling property of Eq. (6). Taking the limit of large realizations $M \gg 1$, Eq. (6) becomes

$$p(o) = k(\gamma, M, N) \frac{(1-o)^{-1}}{\gamma(-\log(1-o))^{1+1/\gamma}}, \tag{10}$$

which depends only on two parameters, γ, and the rescaling parameter,

$$k(\gamma, M, N) = \frac{M^{1/\gamma}}{\alpha^{1/\gamma} N}. \tag{11}$$

This rescaling property shows that the statistics of component sharing is actually a function of a specific combination of realization sizes (e.g., text lengths) and of the range of possible components (e.g., the observed vocabulary). Specifically, the functional form of the distribution is purely defined by the exponent γ, while the rescaling parameter k sets the normalization factor and the range of possible occurrences. In fact, the analytical expression of the occurrence corresponding to the distribution minimum, i.e., $o_{\min} = 1 - e^{-1-1/\gamma}$, is only a function of γ, while the minimum possible occurrence value $o_N \simeq 1 - e^{-k^\gamma}$ scales with k. Therefore, a U-shaped occurrence distribution should be generally expected for component systems with highly heterogeneous component frequencies since the power-law decay and the presence of a minimum before the core are robust features with respect to system parameters. This is confirmed by the analysis of component systems with different values of k and γ (illustrative examples in Fig. S7 of SM [31]): the system specificities set the power-law decay of the left part of the distribution, its support, and the relative proportion of core and rare components, but the U shape is conserved. However, this shape can be more or less symmetric and more or less clearly evident depending on the actual size of the core fraction. The following section discusses in detail the nontrivial dependences of the core size on system parameters.

For the case of component frequency distributions with an exponential rank plot, the statistics of shared components [Eq. (8)] is a function of a single effective parameter λN, and does not depend on the realization sizes M. In other words, the shape of the distribution, and whether it is clearly U shaped, depends only on the decay of component frequencies and on the total number of components. In fact, occurrence distributions corresponding to different exponential frequency rank plots collapse if λN is constant, even if the realizations have widely different size. This is shown in Fig. S4 of the SM [31].

4. Core size

We can estimate the "core size" by computing the fraction of components with occurrence greater than a given arbitrary occurrence threshold θ_c as a function of the only two effective parameters γ and k. Integrating Eq. (6) between θ_c and the maximum occurrence o_1, and then taking the limit $M \gg 1$, this quantity reads

$$\begin{aligned} c &= 1 && \text{if } o_N \geq \theta_c \\ c &= k[-\log(1-\theta_c)]^{-1/\gamma} && \text{otherwise,} \end{aligned} \qquad (12)$$

where o_N is the left boundary of the occurrence distribution, corresponding to the component with lowest frequency.

Starting from this estimate of the core size, Figs. 5(a) and 5(b) show how the scaling property is verified in simulations.

Figure 5(c) compares the analytical predictions for the core size with simulations for different values of γ, showing perfect agreement. Equally, one can obtain analytical estimates for the fraction of rare components (occurrence below a fixed threshold), which are tested in Fig. 5(d). Thus, with increasing k, core families increase linearly with a γ-dependent slope until all components are shared, and concurrently rare components decrease linearly until they hit zero (when the lower cutoff of occurrence exceeds the chosen threshold value). Component number and realization size enter only through the combination defined by the rescaling parameter k. This phenomenology fully characterizes the distribution of shared components with varying parameters.

The general relation [Eq. (12)] between the core size and the rescaling parameter k translates into different dependences of the core size on the typical realization size M, depending on the relation between the system size MR and the total number of accessible components $\tilde{N}$.

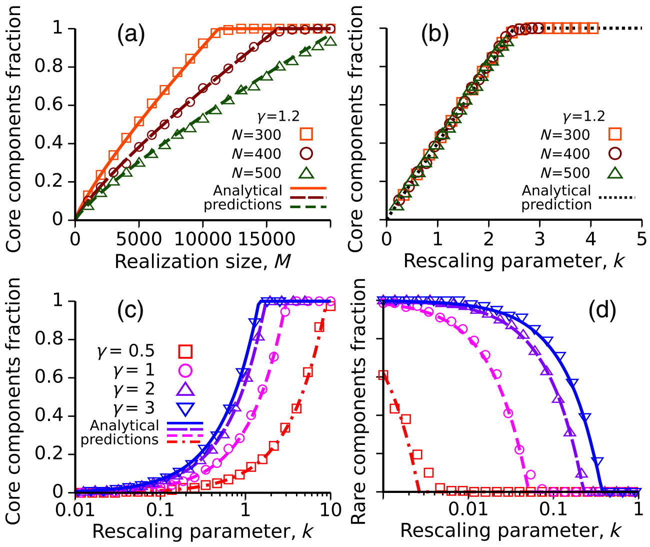


FIG. 5. Scaling of the distribution of shared components and fraction of rare and core components. (a) The fraction of core components (defined by the occurrence threshold $o > \theta_c = 0.95$) for a power-law component frequency distribution with exponent $\gamma = 1.2$, plotted as a function of component size M for three values of realization number N. (b) Collapse of the curves shown in (a) when plotted as a function of the rescaled parameter k, defined in Eq. (11). (c),(d) Fraction of core and rare ($o < 0.05$) components plotted as a function of k for different values of γ. For sufficiently large k (i.e., typically when M dominates over N), the fraction of core components saturates to 1. Conversely, the fraction of rare components drops to zero for increasing k. Symbols refer to numerical simulations of the random-sampling model, while the lines are the theoretical predictions of Eq. (12).

While this issue is discussed in more detail in the SM [31], it is easy to intuitively understand the different regimes. For large enough systems, all possible components $\tilde{N}$ are expected to be sampled at least once, thus making the observed vocabulary $N \simeq \tilde{N}$ a constant parameter. This is the regime considered in Fig. 5(a). In this regime, Eq. (12) simplifies to the simple scaling $c \sim (M^{1/\gamma}/\tilde{N})$. On the other hand, in several empirical systems the observed vocabulary is a function of the system size, and typically with the power-law dependence $N(MR) \sim (MR)^{\beta}$ (with $\beta < 1$) called Heaps' law. Thus, in general, the core fraction is expected to show the more complex dependences $c \sim M^{1/\gamma - \beta} R^{-\beta}$. However, a random-sampling procedure starting from a Zipf's law described by Eq. (5) leads to the approximate relation $\beta \simeq 1/\gamma$ between the exponents of Zipf's and Heaps' laws [7,8,32]. Therefore, in this regime the core fraction becomes only a function of the number of realizations as $c \sim R^{1/\gamma}$. These different scaling relations in different regimes are tested in Figure S6 [31].

Note that the absolute number of core components cN, as estimated from Eqs. (11) and (12), is instead always independent from the number of realizations, even in the regime where Heaps' law is expected to hold (Fig. S6 [31]).

For component frequency distributions with an exponential rank plot, the sampling procedure leads to an occurrence distribution that is independent from the realization size M [Eq. (8)]. However, the exact analytical prediction for the core size [the counterpart of Eq. (12)] still has a dependence on M. But this is due to the residual dependence of the maximum occurrence values (o_1) on M and does not affect the shape of the distribution. This last technical point is discussed in more detail in the SM [31].

E. Empirical distributions of shared components satisfy the relations predicted by the random sampling

One can ask whether the general analytical predictions discussed in the previous section can be applied to empirical data. In particular, we first asked how the power-law decay exponent of the distribution of shared components relates to the component frequency rank plot in empirical systems, and if this relation follows our analytical prediction. An analytical mapping would give a more synthetic and powerful description than the direct simulations discussed in Fig. 3. Importantly, the analytical formulas for the distribution of shared components are derived under the hypothesis of a pure power-law or exponential component frequency rank plot. However, the three empirical data sets (as previously discussed) show a double-scaling frequency distribution. To override this issue, we restrict the frequency rank-plot range in which the predictions are applicable. The procedure to perform this comparison is described in Fig. 6.

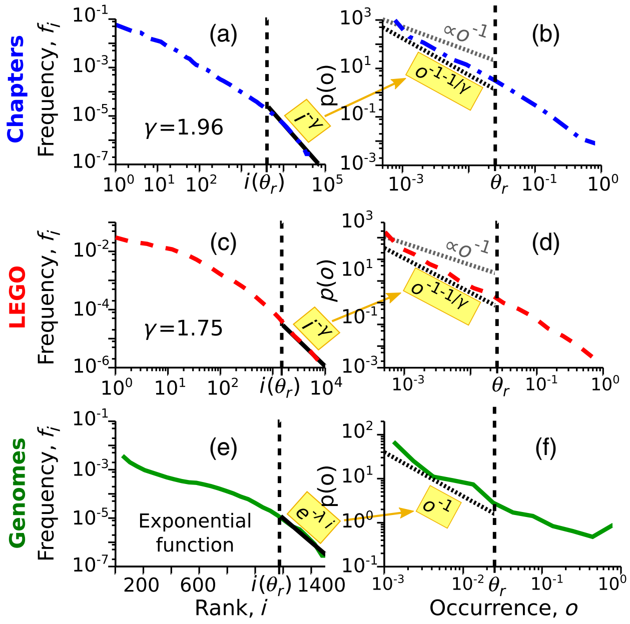


FIG. 6. The relation between the exponents of frequency rank plot and occurrence distribution is satisfied in all the three data sets. The plots consider the low occurrence region, below the arbitrary threshold $\theta_r = 0.025$, which corresponds to the high-rank region above $i(\theta_r)$ in the frequency rank plot (see main text). Panels (a) and (b) refer to book chapters, for which the tail of rank plot is a power law with exponent $\gamma = 1.96$, which implies a power-law decay of $p(o)$ with exponent $1 + (1/\gamma) = 1.51$. Panels (c) and (d) show the LEGO data set ($\gamma = 2.8$, $1 + (1/\gamma) = 1.36$). Panels (e) and (f) correspond to protein domains in genomes, where the best fit of the tail region is an exponential function [note that (e) is in linear-logarithmic scale], which implies a power-law decay with exponent -1.

First, we choose an arbitrary threshold θ_r defining the rare components and we map it to the frequency rank plot (assuming the model), by using the inverse function of Eq. (2). The frequency rank associated to the occurrence threshold θ_r, $i(\theta_r)$ in the figure, is the rank above which the model prediction for the decay of the distribution of shared components should apply as long as $i(\theta_r)$ does not cross the position of the change in scaling. In other words, since in the model there is a monotonic relation between occurrence and frequency [Eq. (2)], all components with rank greater than $i(\theta_r)$ (and frequency smaller that $f_{i(\theta_r)}$) are assumed to be the components with occurrence lower than θ_r. We then estimate the behavior of the frequency rank plot in the high-rank region [after $i(\theta_r)$] as the best fit with a power-law function or an exponential. This leads to a prediction for the decay exponent of the distribution of shared components [using Eq. (7) or Eq. (9) for the exponential case] in the range $[o_N, \theta_r]$. Figure 6 shows that the predicted decay exponents correspond well with the data.

The random-sampling model also gives qualitative analytical predictions for the expected fraction of core components, and thus for the expected shape of the distribution of shared components for a given empirical system. While the analytical relations between exponents applied in Fig. 6 do not depend on the realization sizes, the analytical formulas for the fraction of core components [see, e.g., Eq. (12)] were derived assuming realizations of fixed size M. The actual size distributions for the three empirical systems are quite broad (Fig. S1 [31]), but we can still use the analytical framework to get an estimate of the core fraction considering the average realization size of each empirical system. Following the same line of reasoning as for the low-occurrence tail of the distribution of shared components, we can use a restricted region of the frequency rank plot. In this case, the low-rank region (with exponent around 1 for all the data sets; see Fig. 2) is expected to contain the core components. Therefore, the parameter γ can be fixed to 1, implying that the fraction of core components, given by Eq. (12), should be simply proportional to the rescaling parameter k [Eq. (11)]. However, the normalization factor α, which is present in the definition of k and defined in Eq. (5), takes an approximately constant value with respect to $\tilde{N}$ for large values of $\tilde{N}$, as it is the case for the empirical examples considered. As a consequence, the core fraction should be simply proportional to (M/N). This estimate can be used to explain why the core fraction is much larger in genomes than in the other two empirical systems [see Fig. 2(d)]. In fact, genome sizes are typically of the same order as the total number of families ($M \simeq 3000$, $N = 1531$; see Fig. S1 [31]) leading to a large expected core. By comparison, book chapters have similar realization sizes but a much larger vocabulary ($N \simeq 50\,000$), and LEGO sets have very small sizes ($M \simeq 100$) compared to vocabulary size ($N \simeq 13\,000$).

More in general, Eqs. (11) and (12) lead to a scaling estimate (dependent on the decay of the frequency rank plot) as a function of the system parameters M and N, which can be applied to data, in order to generate expectations for the core components. For example, for Zipf-like (exponent -1) frequency distributions, we expect the absolute number of core components to be linearly dependent on the average size of realizations M, and essentially insensitive to the vocabulary size N and the total number of realizations R. In genomics language, this would imply that the number of core protein domains does not directly depend on the number of sequenced genomes but only on their sizes and on the total number of different protein domains discovered. Note that adding new genomes to the data set is not expected to alter the power-law exponent $\gamma \simeq 1$ of the global frequency distribution for high-frequency components, since it does not change if the distribution is evaluated on subsamples of the empirical data set (Fig. S2 [31]).

As previously discussed, the core fraction, instead of the absolute number of core components, is expected to have a more complex dependence on the typical realization size M and on the number of realizations R. Moreover, in empirical

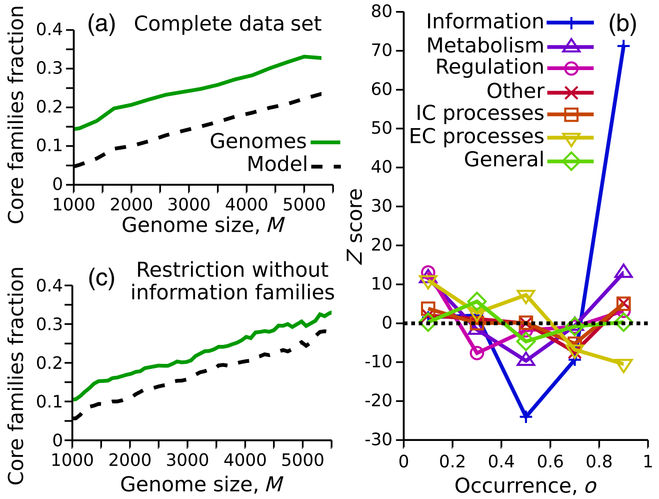


FIG. 7. Specific functional constraints can be detected by deviations from the predictions of a random sampling. (a) Fraction of common protein domain families as a function of the genome sizes. Each point of the curves corresponds to the core families ($o > \theta_c = 0.95$) given the occurrence distribution of a genome's subset whose sizes are inside a certain window. The average size of the genomes within the size window defines the x axis. (b) Enrichment analysis in the occurrence distribution for specific functional categories. Considering domain families relative to a single functional category, their relative component occurrence distribution was evaluated for an ensemble of systems built with a random sampling. From this, the average value and the standard deviation for the expected fraction of components at each occurrence value o can be calculated. This provides a measure (Z score) of over- or underrepresentation of domain families belonging to each functional category in the empirical data set. IC and EC denote intra-cellular and extra-cellular processes respectively. (c) Excluding from the analysis the domain families associated to information processes (i.e., DNA replication, transcription, and translation) significantly reduces the offset between the random-sampling prediction and the empirical trend.

systems these relations are further complicated by the fact that the frequency distributions cannot be described by simple power laws (Fig. 2). Nevertheless, the relation between the core fraction and the average realization size predicted by a random-sampling model can be tested numerically, as Fig. 7(a) shows for prokaryotic genomes, and seems accurately verified and roughly linear in the tested range of sizes. However, the predicted fraction of core components is actually much smaller than the empirical one. This highlights the presence of additional functional constraints and/or specific correlations in the empirical system that the model cannot capture. The next section addresses this point in more detail.

F. Deviations from the random-sampling predictions can highlight system-specific properties

Beyond the striking agreement with null predictions for shared components, the deviations from sampling can be used to quantify specific functional and architectural features of a component system. While the scope of this work is to highlight the common trends and their origins, we discuss a specific example, in order to show the feasibility of this procedure. Of the three data sets considered here, the case where the clearest deviations emerge are genomes. For example, Fig. 7(a) illustrates how the random sampling underestimates the empirical core size by a constant offset, for genomes of increasing size. Generally speaking, this larger core of components is due to the components that tend to occur in most realizations, but in few copies. The natural explanation is that there are specific basic functions that are essential for all (or most) genomes, but the domains involved in these functions are not necessarily needed in many copies per genome, and thus their presence in all realizations does not simply correlate with high global abundances as the random sampling would entail [42].

To test this hypothesis, we divided the domain families in functional categories (see Sec. II for the functional annotation), and tested if most of the deviations from the random-sampling prediction can be ascribed to the statistics of domains belonging to specific categories. The result of this analysis is reported in Fig. 7(b). Different parts of the distribution of shared components are indeed enriched in components of different biological functions with respect to the random-sampling expectation. In particular, protein domains that play a functional role in information processes—such as DNA translation, DNA transcription, and DNA replication—are clearly enriched in the core. At the same time, they seem statistically underrepresented at occurrences around 0.6. These two deviations can be explained as two sides of the same coin if this category contains domain families that empirically occur in all genomes but in a single copy per genome. Indeed, the global frequency (i.e, across all genomes) of families that are both single copy and ubiquitous is $f = (R/RM) = 1/M$. Therefore, their occurrence predicted by the random-sampling model is $o = 1 - [1 - (1/M)]^M = 1 - e^{M \log(1-1/M)} \simeq 1 - e^{-1} \simeq 0.6$ (where the rough approximation holds for large enough M), thus naturally leading to an excess of those families in the core and to a depletion around $o \simeq 0.6$.

The observation of a strong presence of protein domains related to basic cellular function in the core genome is not new [21,42]. However, the random-sampling model allows us in principle to distinguish families whose presence in the core could be simply explained by their high abundance in the pan-genome and thus it would be expected also in a simple scenario of random gene exchange. Finally, the observed correlation between biological functions and deviations from random-sampling predictions seems coherent with a picture, recently proposed [23], in which natural selection and functional constraints have played an important role in defining the empirical U-shaped distribution of gene occurrences.

IV. DISCUSSION AND CONCLUSIONS

This work employs a simple statistical model based on random sampling to describe the distribution of shared components in complex component systems. A similar approach was employed in quantitative linguistics to explain how the dictionary used in a text scales with text size as measured in number of words (the so-called "Heaps' law") while assuming Zipf's law for component frequencies [7–9,32,43]. We extend the model to show that there is a general link between the heterogeneity in component frequency and the statistics of shared components, regardless of the mechanisms that generate heterogeneity. Consequently, models or generative processes able to explain the heterogeneity in component frequency implicitly carry predictions for the statistics of shared components.

The striking similarities of laws governing both component abundance and occurrence found in empirical systems of very different origins (LEGO sets, genomes, book chapters) support the idea that the concept of "component system" defined in this work can capture in a unified framework a large class of complex systems with some common global properties. Different component systems, besides having specific architectural constraints, may show convergent phenomena in terms of global statistics. Such "universal" phenomena may be regarded as emergent properties due to system heterogeneity, which transcend the specific design, generative process, or selection criteria at the origin of a system. Analogous phenomena occur, for example, in ecosystems, where emergent species-abundance distributions appear for forests, birds, or insects [44].

Beyond the examples considered here, modular systems in a wide range of disciplines can be represented as component systems. Developing a common theoretical language for such systems can help the exchange of ideas, models, and data-analysis techniques between distant communities of researchers [45]. For example, the statistics of component sharing considered here plays a central role in genomics [2,23,46] but is relatively unexplored in the context of natural languages [1]. Conversely, the random-sampling approach used here was developed in quantitative linguistics [8], and this work shows that it is applicable to other systems, including the detection of functional constraints in prokaryotic genome evolution.

An important result of this work is a proof of the clear link between the heterogeneity of component abundance in a system and the statistics of shared components. This link is consistent with data from three very different empirical systems and well captured by the random-sampling model. The fact that emergent patterns can be explained by largely null models resembles again the case of biodiversity, where neutral theories ignoring species interactions and competitive exclusion appear to capture many of the emerging trends of species abundance [44,47].

If the trends of component sharing of generic component systems are to be regarded as largely null and due to the heterogeneity in component usage, system-specific investigations should be informed of this general trend. Quantitative null models, such as the one provided here, may be crucial for identifying data-set-specific deviations that are related to functional reasons or constraints. In the data considered in this work, the patterns of shared components show differences between empirical data and the null model in some cases. This is particularly true in the genomic context, where the differences can indeed be traced back to functional constraints in genome composition. Therefore, the framework can be useful to pinpoint hallmarks of functional design and distinguish them from statistical effects, particularly for the detection of causality, dependency, and correlation structures between components from occurrence patterns.

Once a null model is defined, these features can emerge as significant deviations from the null behavior, for example, as violations of the constraints linking different global statistics such as the abundance rank plot, the distribution of shared components, and Heaps' law. We have considered here a specific example for the case of shared protein domain families in genomes (Fig. 7), but this question still needs to be approached systematically. In this specific case, core components are particularly enriched by specific functional classes of components with respect to the random-sampling prediction. In evolutionary terms, the random sampling defines a scenario in which the pangenome fully determines the overall abundance of the gene families in each genome, while in empirical bacterial genomes genome-specific functional constraints are clearly in place [38,39,48]. Deviations from the null scenario can thus highlight the role of selection for specific functions, supporting from a different perspective the idea that the empirical U-shaped gene occurrence distribution is affected by selective rather than neutral processes [22–25].

ACKNOWLEDGMENTS

We thank Erik van Nimwegen for useful discussions. This work is supported by the "Departments of Excellence" Grant awarded by the Italian Ministry of Education, University and Research (MIUR) (L. 232/2016).

[1] E. G. Altmann and M. Gerlach, in *Creativity and Universality in Language*, edited by M. D. Esposti, E. G. Altmann, and F. Pachet (Springer International Publishing, Cham, Switzerland, 2016), pp. 7–26.

[2] E. V. Koonin, *Are There Laws of Genome Evolution?*, PLoS Comput. Biol. **7**, e1002173 (2011).

[3] M. C. Lagomarsino, A. L. Sellerio, P. D. Heijning, and B. Bassetti, *Universal Features in the Genome-Level Evolution of Protein Domains*, Genome Biol. **10**, R12 (2009).

[4] G. K. Zipf, *The Psychobiology of Language* (Houghton-Mifflin, New York, 1935).

[5] D. Zanette and M. Montemurro, *Dynamics of Text Generation with Realistic Zipf's Distribution*, J. Quant. Linguist. **12**, 29 (2005).

[6] M. E. J. Newman, *Power Laws, Pareto Distributions and Zipf's Law*, Contemp. Phys. **46**, 323 (2005).

[7] L. Lü, Z.-K. Zhang, and T. Zhou, *Zipf's Law Leads to Heaps' Law: Analyzing Their Relation in Finite-Size Systems*, PLoS One **5**, e14139 (2010).

[8] I. Eliazar, *The Growth Statistics of Zipfian Ensembles: Beyond Heaps Law*, Physica (Amsterdam) **390A**, 3189 (2011).

[9] F. Font-Clos and Á. Corral, *Log-Log Convexity of Type-Token Growth in Zipfs Systems*, Phys. Rev. Lett. **114**, 238701 (2015).

[10] M. A. Huynen and E. van Nimwegen, *The Frequency Distribution of Gene Family Sizes in Complete Genomes.* Mol. Biol. Evol. **15**, 583 (1998).

[11] J. Qian, N. M. Luscombe, and M. Gerstein, *Protein Family and Fold Occurrence in Genomes: Power-Law Behaviour and Evolutionary Model.* J. Mol. Biol. **313**, 673 (2001).

[12] G. P. Karev, Y. I. Wolf, A. Y. Rzhetsky, F. S. Berezovskaya, and E. V. Koonin, *Birth and Death of Protein Domains: A Simple Model of Evolution Explains Power Law Behavior*, BMC Evol. Biol. **2**, 18 (2002).

[13] G. P. Karev, Y. I. Wolf, and E. V. Koonin, *Simple Stochastic Birth and Death Models of Genome Evolution: Was There Enough Time for Us to Evolve?*, Bioinformatics **19**, 1889 (2003).

[14] B. Mandelbrot, *An Informational Theory of the Statistical Structure of Language*, Communication Theory **84**, 486 (1953).

[15] R. F. I Cancho and R. V. Solé, *Least Effort and the Origins of Scaling in Human Language*, Proc. Natl. Acad. Sci. U.S.A. **100**, 788 (2003).

[16] H. A. Simon, *On a Class of Skew Distribution Functions*, Biometrika **42**, 425 (1955).

[17] M. Gerlach and E. G. Altmann, *Stochastic Model for the Vocabulary Growth in Natural Languages*, Phys. Rev. X **3**, 021006 (2013).

[18] S. K. Baek, S. Bernhardsson, and P. Minnhagen, *Zipf's Law Unzipped*, New J. Phys. **13**, 043004 (2011).

[19] T. Y. Pang and S. Maslov, *Universal Distribution of Component Frequencies in Biological and Technological Systems*, Proc. Natl. Acad. Sci. U.S.A. **110**, 6235 (2013).

[20] M. Touchon, C. Hoede, O. Tenaillon, V. Barbe, S. Baeriswyl, P. Bidet, E. Bingen, S. Bonacorsi, C. Bouchier, O. Bouvet *et al.*, *Organised Genome Dynamics in the Escherichia coli Species Results in Highly Diverse Adaptive Paths*, PLoS Genet. **5**, e1000344 (2009).

[21] E. V. Koonin and Y. I. Wolf, *Genomics of Bacteria and Archaea: The Emerging Dynamic View of the Prokaryotic World*, Nucleic Acids Res. **36**, 6688 (2008).

[22] B. Haegeman and J. S. Weitz, *A Neutral Theory of Genome Evolution and the Frequency Distribution of Genes*, BMC Genomics **13**, 196 (2012).

[23] A. E. Lobkovsky, Y. I. Wolf, and E. V. Koonin, *Gene Frequency Distributions Reject a Neutral Model of Genome Evolution*, Genome Biol. Evol. **5**, 233 (2013).

[24] F. Baumdicker, W. R. Hess, and P. Pfaffelhuber, *The Infinitely Many Genes Model for the Distributed Genome of Bacteria*, Genome Biol. Evol. **4**, 443 (2012).

[25] R. E. Collins and P. G. Higgs, *Testing the Infinitely Many Genes Model for the Evolution of the Bacterial Core Genome and Pangenome*, Mol. Biol. Evol. **29**, 3413 (2012).

[26] D. Wilson, M. Madera, C. Vogel, C. Chothia, and J. Gough, *The Superfamily Database in 2007: Families and Functions*, Nucleic Acids Res. **35**, D308 (2007).

[27] C. A. Orengo and J. M. Thornton, *Protein Families and Their Evolution—A Structural Perspective.* Annu. Rev. Biochem. **74**, 867 (2005).

[28] C. Vogel and C. Chothia, *Protein Family Expansions and Biological Complexity*, PLoS Comput. Biol. **2**, e48 (2006).

[29] https://rebrickable.com.

[30] http://www.gutenberg.org.

[31] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevX.8.021023 for further details.

[32] D. C. van Leijenhorst and T. P. Van der Weide, *A Formal Derivation of Heaps' Law*, Information Sciences (NY) **170**, 263 (2005).

[33] A. M. Petersen, J. N. Tenenbaum, S. Havlin, H. E. Stanley, and M. Perc, *Languages Cool as They Wxpand: Allometric Scaling and the Decreasing Need for New Words*, Sci. Rep. **2**, 943 (2012).

[34] A. E. Lobkovsky, Y. I. Wolf, and E. V. Koonin, *Estimation of Prokaryotic Supergenome Size and Composition from Gene Frequency Distributions*, BMC Genomics **15**, S14 (2014).

[35] S. M. Soucy, J. Huang, and J. P. Gogarten, *Horizontal Gene Transfer: Building the Web of Life*, Nat. Rev. Genet. **16**, 472 (2015).

[36] P. D. Dixit, T. Y. Pang, F. W. Studier, and S. Maslov, *Recombinant Transfer in the Basic Genome of Escherichia coli*, Proc. Natl. Acad. Sci. U.S.A. **112**, 9070 (2015).

[37] E. van Nimwegen, *Scaling Laws in the Functional Content of Genomes*, Trends Genet. **19**, 479 (2003).

[38] N. Molina and E. van Nimwegen, *Scaling Laws in Functional Genome Content across Prokaryotic Clades and Lifestyles*, Trends Genet. **25**, 243 (2009).

[39] S. Maslov, S. Krishna, T. Y. Pang, and K. Sneppen, *Toolbox Model of Evolution of Prokaryotic Metabolic Networks and Their Regulation*, Proc. Natl. Acad. Sci. U.S.A. **106**, 9743 (2009).

[40] J. Grilli, B. Bassetti, S. Maslov, and M. C. Lagomarsino, *Joint Scaling Laws in Functional and Evolutionary Categories in Prokaryotic Genomes*, Nucleic Acids Res. **40**, 530 (2012).

[41] M. Mitzenmacher, *A Brief History of Generative Models for Power Law and Lognormal Distributions*, Internet Math. **1**, 226 (2004).

[42] E. V. Koonin, *Comparative Genomics, Minimal Gene-Sets and the Last Universal Common Ancestor*, Nat. Rev. Microbiol. **1**, 127 (2003).

[43] H. S. Heaps, *Information Retrieval: Computational and Theoretical Aspects* (Academic Press, Inc., New York, 1978).

[44] S. P. Hubbell, *The Unified Neutral Theory of Biodiversity and Biogeography (MPB-32)* (Princeton University Press, Princeton, NJ, 2001).

[45] Y. Holovatch, R. Kenna, and S. Thurner, *Complex Systems: Physics beyond Physics*, Eur. J. Phys. **38**, 023002 (2017).

[46] P. Lapierre and J. P. Gogarten, *Estimating the Size of the Bacterial Pan-Genome*, Trends Genet. **25**, 107 (2009).

[47] S. Azaele, S. Suweis, J. Grilli, I. Volkov, J. R. Banavar, and A. Maritan, *Statistical Mechanics of Ecological Systems: Neutral Theory and Beyond*, Rev. Mod. Phys. **88**, 035003 (2016).

[48] J. Grilli, M. Romano, F. Bassetti, and M. C. Lagomarsino, *Cross-Species Gene-Family Fluctuations Reveal the Dynamics of Horizontal Transfers*, Nucleic Acids Res. **42**, 6850 (2014).