Physica A 453 (2016) 35–43

Contents lists available at ScienceDirect

Physica A

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

An entropy based measure for comparing distributions of complexity

R. Rajaram [a,*], B. Castellani [b]

[a] Department of Math. Sci., Kent State University, USA
[b] Department of Sociology, 3300, Lake Rd West, Kent State University, USA

HIGHLIGHTS

- An entropy based measure for comparison of diversity of complexity is proposed.
- The measure allows for comparison of diversity both within and across distributions.
- The measure is multiplicative i.e., a doubling of value implies a doubling of diversity.

ARTICLE INFO

Article history:
Received 21 August 2015
Received in revised form 6 January 2016
Available online 23 February 2016

Keywords:
Probability distributions
Complex systems
Shannon entropy
Measures of complexity

ABSTRACT

This paper is part of a series addressing the empirical/statistical distribution of the diversity of complexity within and amongst complex systems. Here, we consider the problem of measuring the diversity of complexity in a system, given its ordered range of complexity types i and their probability of occurrence p_i, with the understanding that larger values of i mean a higher degree of complexity. To address this problem, we introduce a new complexity measure called *case-based entropy* C_c — a modification of the Shannon–Wiener entropy measure H. The utility of this measure is that, unlike current complexity measures – which focus on the macroscopic complexity of a single system – C_c can be used to empirically identify and measure the *distribution of the diversity of complexity* within and across multiple natural and human-made systems, as well as the diversity contribution of complexity of any part of a system, relative to the total range of ordered complexity types.

1. Measuring complexity: A non-exhaustive list

Over the past several decades, scholars have given considerable attention to measuring the complexity of systems [1–7]. The result has been a proliferation of a wide array of approaches, which vary considerably in mathematical form and focus. In 2001, for example, Lloyd [8] counted roughly forty measures, all differing as a function of the type of question asked, such as (1) how hard is it to describe the complex system of study? (2) How hard is it to create? (3) And, what is its degree of organization? Examples of the first question include Shannon entropy, algorithmic complexity and Renyi entropy; examples of the second include computational complexity, thermodynamic depth and information-based complexity; and examples of the third include stochastic complexity, true-measure complexity and tree subgraph diversity.

Still, despite this considerable variation, the purpose of these measures – as the above questions suggest – has been generally the same: they were created, for the most part, to measure the relative complexity of a single system, *at the macroscopic level*, with or without links to empirical data [2,5,9,10].

* Corresponding author. Tel.: +1 440 964 4537.
E-mail address: rrajaram@kent.edu (R. Rajaram).

http://dx.doi.org/10.1016/j.physa.2016.02.007
0378-4371/

One unintended consequence of this focus is that, to date, little attention has been given to the empirical distribution of the diversity of complexity within or across multiple natural or human-made systems; or, more specifically, the diversity contribution of complexity of any part of a system. Hence, the purpose of the current study.

1.1. Purpose of current study

This paper is part of a series of studies addressing the empirical/statistical distribution of the diversity of complexity within and amongst complex systems [11,12]. Our goal here is to introduce and mathematically validate a measure we developed for our research, called *case-based entropy* C_c.

Grounded in a case-based approach to complexity, [13–16] the purpose of C_c is to effectively measure the true diversity of complexity within systems. It is based on a modification of the Shannon–Wiener entropy measure H.

To date, we have used C_c to empirically explore the distribution of the diversity of complexity in a wide variety of systems. In Paper 2 of our series, we used C_c to examine eight empirical systems: (1) a segment of the World-Wide-Web, (2) household income in the United States for 2013, (3) the body mass of Late Quaternary mammals, (4) the human diseasome map, (5) Hubble's classic data on the velocity of galaxies, (6) USA cities by population size for 2011, (7) the Financial Times 2014 biggest 500 companies, and (8) the twelve-month prevalence of mental disorders in the United States for a given year [11]. We examined such a wide variety of systems in order to search for universal properties regarding the diversity of complexity in systems. In Paper 3 of our series, we extended our search by using C_c to examine the Maxwell–Boltzmann distribution for kinetic energy of an ideal gas at thermodynamic equilibrium, in both one and three dimensions [12].

Across these two studies, C_c proved highly useful, as it allowed us to do three things. First, it allowed us to statistically calculate the distribution of the diversity of complexity within and across multiple systems. Second, it allowed us to compute the diversity contribution of complexity of any part of a system, given its ordered complexity types i and their probability of occurrence p_i. Third, it allowed us to do these calculations despite differences in the indices of complexity used — which, generally speaking, vary significantly in the literature (i.e., descriptive, organizational, structural, behavioral, informational, etc. [2,5,9,10]) — as well as differences in the scale or the degree of macroscopic complexity of these systems. (For details see Refs. [11,12].)

Still, given their empirical focus, neither Paper 2 or Paper 3 had time to provide a formal theoretical overview or mathematical validation of C_c — hence the purpose of the current paper, which is organized as follows. We begin with an introduction to case-based entropy (and, more generally, case-based complexity) and its utility for studying the probability distributions of complex systems.

Next, we validate the utility of C_c by calculating the diversity contribution D_c of some set of complexity types k up to a cumulative probability c. For our validation we examine two different distributions: the uniform distribution and the geometric distribution. We chose the former because it is the simplest case; and we chose the latter because it takes the characteristic form of a positively skewed distribution, which seems to be, based on our initial studies, the 'signature shape' of the diversity of complexity in most systems [11,12]. That is, we find that, for the complex systems we have studied, as complexity $\rightarrow \infty$ on the x-axis, a skewed-right histogram emerges, decreasing asymptotically as complexity increases (with or without long-tail). With our validation complete, we provide a formula for the general case of C_c, which works for all distributions.

In the final section, we use our general formula to compute the distribution of the diversity of complexity for two empirical examples, culled from our research: the body mass of Late Quaternary mammals and a segment of the World-Wide-Web. We end with the implications of our approach for advancing the study of complex systems.

2. Case-based complexity and probability distributions

We begin by considering a probability distribution (or relative frequency) for some imaginary complex system of study — as shown in Fig. 1. Grounding ourselves in the field of statistical mechanics, we begin here for two reasons:

First, as identified by Sornette [17] and others [18,19], probability distributions are the "first quantitative characteristics of complex systems", [17] providing researchers an effective tool for identifying and describing macroscopic regularities that, otherwise, would be difficult to detect, as in the case of measuring the complexity of a system. Second, as demonstrated by the central limit theorem, Boltzmann distribution, power laws, etc. — measurements on a wide range of physical, biological, psychological, sociological and ecological systems are well approximated by the shape of these probability distributions, particularly as the sample size (or number of samples or trials) $n \rightarrow \infty$.

In regard to such distributions, however, an important dimension has been missed: how they illustrate, in compressed two-dimensional form, the distribution of the diversity of complexity in a system. To explain, we turn to the tools of *case-based complexity* [13–16].

Case-based complexity is distinct in the complexity sciences and statistics in that it treats the elements in a complex system (i.e., gas molecules, diseases, mammals, cities, countries, galaxies, etc.) as a set of *qualitatively* distinct cases [13,14]. Following Weaver and his notion of organized complexity [20], by 'qualitative' we mean that each row in a study's database D constitutes a complex case c_i, where each c_i is a k dimensional row vector $c_i = [x_{i1}, \ldots, x_{ik}]$ and where each x_{ij} represents a measurement on the profile of intersecting and interconnected empirical variables for D – what case – comparative researchers call the *case-based profile* [21,15,16].

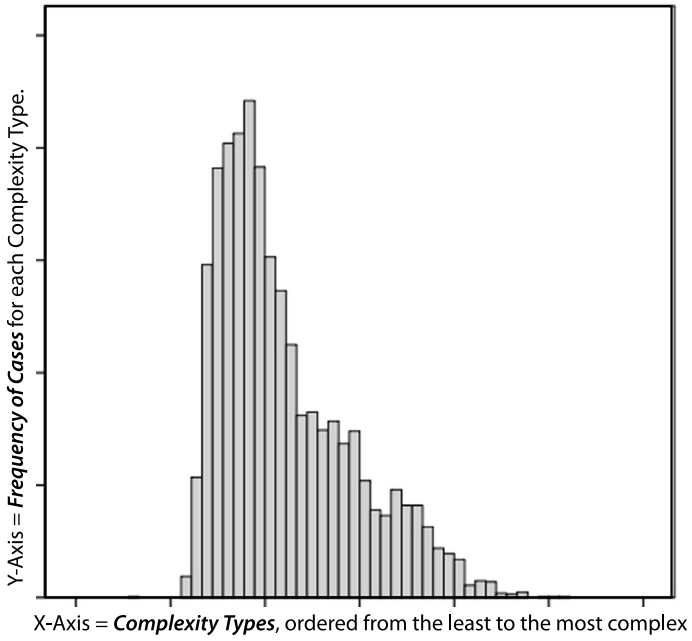


Fig. 1. Histogram for the distribution of the diversity of complexity.

With this profile in hand, cases can be empirically ordered, ranked or grouped according to their degree of complexity; which, when combined, constitute the system's range of complexity types — discrete or continuous. In other words, in terms of case-based complexity, we can think of the probability distribution of a complex system as a set of cases, with respect to some ordered range of complexity types — as shown in Fig. 1.

For such a distribution as Fig. 1, the x-axis is empirically defined according to a chosen metric of complexity, which (based on the current complexity measurement literature [1–7]), can take a variety of forms: descriptive, organizational, structural, behavioral, informational. Some easy examples of such complexity-based measures/indices include (a) velocity dispersion of galaxies, (b) degree distribution of a network, (c) body mass of late quaternary mammals, (d) household income, (e) gene count for different diseases, (f) population density of cities, and the (g) market value of global companies [11,12].

Still, while differences in metric are possible, in terms of defining the x-axis, the key is that each complexity type (be it continuous or discrete) is treated as an empirically derived probability state for some given vector space of cases of a complex system of study. In this way, each type is an empirically derived measure of one of the major or minor complexity trajectories within a given state/phase space, based on the empirical observation of the set of cases representing the complex system of study — for details on this approach, see Refs. [21,15,16].

Also, for each complexity type (whether measured continuously or discretely), an empirically-driven theoretical distinction has to be made, such that the range of types is ordered from the simplest to the most complex. For example, if we were studying galaxies using Hubble's original 1929 data, then the velocity dispersion of a galaxy would be an indication of its degree of complexity, as it is well known that galaxies with larger mass spin faster and hence are inherently more complex. On the x-axis, then, the range would go from the lowest to the highest levels of velocity dispersion. Another easy example is the phylogenetic tree: moving from bacteria and simple cells to primates and, more specifically, humans, life evolves toward increasing levels of complexity [22,23]. If one constructed a histogram of this evolution, grouping all of life according to some measure of complexity (i.e., cellular, morphological, cognitive, behavioral or informational), the result would be a probability distribution moving along the x-axis from the least to the most complex forms of life. A final example would be a degree distribution for some segment of nodes in the World Wide Web. In such a network, taking a walk along the x-axis, we would go from the least connected to the most connected nodes, with the hubs being far along the tail of the x-axis — see Fig. 3. In turn, for each of the case-based complexity distributions used as examples above, the y-axis would constitute the frequency of cases (as in galaxies or species or nodes, for example) associated with each complexity type (for example, velocity spin, species complexity, or degree connectedness), based on empirically derived differences in their respective *case-based profiles* [21,15,16].

With all of this in mind, we come to our first definition, which outlines what we mean by the distribution of case-based complexity:

Definition 1. We define a distribution of case-based complexity as the probability distribution of cases with probability p_i on the y-axis and an ordered measure of complexity types i on the x-axis, with the understanding that larger values of i mean a higher degree of complexity.

It is crucial to remember, however – again, following Weaver's notion of organized complexity [20] – that a case-based probability distribution does not reduce-away the complexity of a case to a single dimension. Instead, as with a discrete one-dimensional Boltzmann distribution for an ideal gas, it uses this single dimension (i.e., its degree of complexity) to identify macroscopic regularities that would be otherwise difficult to model across a range of cases.

With our notion of case-based complexity defined, we turn next to a more in-depth discussion of the distribution of the diversity of complexity and its corresponding measurement.

2.1. Measuring the diversity of complexity via case-based entropy

An effective way to define and measure the diversity of complexity D in a case-based probability distribution (as shown in Fig. 1) is through a specific form of Shannon entropy H [24]. With such a modified measure (i.e., complexity index), researchers can determine the overall information (complexity) contributed by the range of complexity types in a system, relative to their different frequencies of occurrence.

Our impetus for using the Shannon entropy index H comes from evolutionary biology and ecology, where it is employed to measure the true diversity of species (types) in a given ecological system of study — think back, for example, to our above discussion of the phylogenetic tree [25–28]. More specifically, our impetus comes from Jost's paper on *Entropy and Diversity* [26], in which he states that: "In physics, economics, information theory, and other sciences, the distinction between the entropy of a system and the effective number of elements of a system is fundamental. It is this latter number, not the entropy, that is at the core of the concept of diversity in biology" (p. 363).

And so we come to our second definition – on the distribution of the diversity of complexity – which we will use in the remainder of this paper to explicate:

Definition 2. Given an ordered set of complexity types numbered as $i \in \mathbf{N}$ and their corresponding probabilities p_i, with the understanding that larger values of i denote a higher degree of complexity, the diversity of complexity of the entire distribution (or a part of it) is defined as the number of equi-probable types of complexity that are needed to yield the same value of Shannon entropy H.

Following this definition, given the probability p_i of the occurrence of a particular type of complexity i, the Shannon–Weiner entropy index H is given by:

$$H = -\sum_{i=1}^{N} p_i \ln(p_i). \tag{1}$$

The problem, however, with the Shannon entropy index H is that, while useful for studying the diversity of a single system, it cannot be used to compare the diversity of complexity between or across systems. In other words, H is not multiplicative i.e., a doubling of value for H does not mean that the actual diversity of complexity has doubled. For example, one could not compare the H for the velocity spin of galaxies with the H of the World Wide Web using such a measure.

To address this problem, we turn to the *true diversity measure* D in biology and ecology [29–31]. The utility of this measure is that, given the value of H, it can compute the number of complexity types that have the same probability of occurrence and gives the same value of H, as given by the following formula:

$$D = \mathrm{e}^{H} = \Pi_{i=1}^{N} \frac{1}{p_i^{p_i}}. \tag{2}$$

In other words, the utility of D for comparing the diversity of complexity across systems is that in D a doubling of the value means that the number of equi-probable types of complexity in a complex system of study has doubled as well, a property that is not true of H. D calculates the number of such equi-probable types of complexity that will give the same value of Shannon entropy H as observed in the complex system.

But that is not where our novel usage of H ends. In order to define and measure the distribution of the diversity of complexity, we next need to determine how to compute the percentage contribution to overall diversity from any part of a complex system — say, for example the first K set of types. In other words, we need to be able to compute the diversity contribution D_c up to a certain cumulative probability c of the first K set of types.

To do so, D is still employed, but H is replaced with H_c — which is the conditional entropy, given that only the first K types are observed with conditional probability of occurrence $\hat{p}_i$ of the first K set of types within the given cumulative probability c, as given below:

$$\hat{p}_i = \frac{p_i}{c} \tag{3}$$

$$H_c = -\sum_{i=1}^{N} \hat{p}_i \ln(\hat{p}_i) \tag{4}$$

Table 1
General dataset with complexity types x_i each having a probability p_i.

X	P
x_1	p_1
x_2	p_2
x_3	p_3
⋮	⋮
x_K	p_K
⋮	⋮
x_N	p_N

$$D_c = e^{H_c} = \Pi_{i=1}^{K} \frac{1}{\hat{p}_i^{\hat{p}_i}} = \frac{c}{\Pi_{i=1}^{K} p_i^{\left(\frac{p_i}{c}\right)}} \tag{5}$$

$$C_c = \frac{D_c * 100}{D}. \tag{6}$$

In other words, if, for each complex system studied, D stands for the true diversity of the entire dataset, and D_c stands for the true diversity of the dataset up to a cumulative probability c, then we can plot the percentage diversity contribution (given by $\frac{D_c * 100}{D}$) versus the cumulative probability c. For example, we could plot the percentage diversity contribution for some K set of nodes for some segment of the World Wide Web versus their cumulative probability c. The same could be done for the velocity spin of galaxies, income distributions, and so forth, across all the examples our series has currently explored [11,12]. And, because 100% of the diversity contribution comes from the entire database ($c = 1$ or a cumulative frequency of a hundred percent), (100, 100) will be the end point of the graph. We call this novel measure *case-based entropy* C_c.

2.2. Mathematical validation

To validate the utility of our new measure, we will calculate the diversity contribution D_c of some complexity type up to a cumulative probability c for two different distributions. The first distribution is a uniform distribution; and the second is the geometric distribution. We chose the former because it is the simplest case; and we chose the latter because it takes the characteristic form of a positively skewed distribution, which seems to be the 'signature shape' of the diversity of complexity for all the systems we have so far examined [11,12]. In other words, as the diversity of complexity $\to \infty$ (primarily in terms of the number of types; but also, secondarily, in terms of the frequency of cases), a skewed-right distribution emerges for the histograms of complex systems, which decreases asymptotically as complexity increases (with or without long-tail) — see, for example, Fig. 3.

For our validation, let X denote the random variable measuring the probability of occurrence of some complexity type x_i i.e. $P(X = x_i) = p_i$ as shown in Table 1. In this table, N stands for the total number of complexity types in all.

We also note that $c = \sum_{i=1}^{K} p_i$, $\hat{p}_i = \frac{p_i}{c}$ and $\sum_{i=1}^{N} p_i = 1$, where c is the cumulative probability up to K complexity types, $\hat{p}_i$ is the conditional probability of observing complexity type x_i where $i \leq K$ given that the first K types have been observed.

2.2.1. Uniform distribution

The simplest case is a dataset where all complexity types are equi-probable i.e. $p_i = \frac{1}{N}$. In this case, the diversity of the entire dataset is N, and the diversity contribution of the first K types is K itself i.e., $D_c = K$. In the equi-probable case, $c = \frac{K}{N}$ and $\hat{p}_i = \frac{1}{K}$.

To compute the entropy contribution of the first K types up to a cumulative probability c, denoted by H_c, we essentially replace the pure probability term p_i with the conditional probability $\hat{p}_i$ in the formula for Shannon entropy ($H = -\sum_{i=1}^{N} p_i \ln(p_i)$) and sum only the terms up to K i.e.

$$H_c = -\sum_{i=1}^{K} \hat{p}_i \ln(\hat{p}_i).$$

For the equi-probable case mentioned above, the above expression amounts to the following:

$$H_c = -\sum_{i=1}^{K} \hat{p}_i \ln(\hat{p}_i) = -\sum_{i=1}^{K} \frac{1}{K} \ln\left(\frac{1}{K}\right) = \ln(K).$$

Next, as outlined in Refs. [29–31], we exponentiate H_c to compute the true number of equi-probable complexity types D_c as follows:

$$D_c = e^{H_c} = e^{\ln(K)} = K.$$

Hence, we recover the required diversity contributed by the first K complexity types for the equi-probable case and this is in a sense, reassuring as the trivial answer is captured by our method.

We also note that the same expression for H_c in Eq. (4) can be derived by defining a new variable Y which keeps track of whether or not the first K types are observed (say $Y = 0$ stands for the case when the first K types are observed and $Y = 1$ for the case where the first K types are not observed), and then computing the specific conditional entropy $H(X/Y = 0)$, which is by definition, the entropy given that the first K types have been observed. Then, it is true that $H_c = H(X/Y = 0)$, and hence this is an alternative route to deriving an expression for H_c.

2.2.2. The geometric distribution

A more complex case is one where all complexity types are not equi-probable. For this case, we consider the geometric distribution given by $p_i = p \cdot q^{i-1}\ \forall i = 1, 2, 3, \ldots, \infty$, which is followed by the random variable X denoting the number of Bernoulli trials until success, with p being the probability of success and $q = 1 - p$ being the probability of failure. If we assume that x_i refers to an ordered discrete complexity type whose probability of occurrence p_i decays exponentially, then the geometric distribution will serve as an abstract probability model for such an example.

It is well known that the geometric distribution is skewed-right with geometric (or exponential) decay of probability. One can also show that the total entropy H is given by

$$H = \frac{-p\ln(p) - q\ln(q)}{p} = \ln\left(\frac{1}{p \cdot q^{\frac{q}{p}}}\right), \tag{7}$$

and hence the total diversity D is given by

$$D = \mathrm{e}^{H} = \mathrm{e}^{\ln\left(\frac{1}{p \cdot q^{\frac{q}{p}}}\right)} = \frac{1}{p \cdot q^{\frac{q}{p}}}. \tag{8}$$

In other words, it will take $\frac{1}{p \cdot q^{\frac{q}{p}}}$ number of equi-probable discrete complexity types to cover the overall diversity of the geometric random variable X. It is not surprising that the diversity is finite since the higher complexity types x_i have a geometrically decaying probability of occurrence p_i.

Next, we compute the diversity contribution D_c up to a cumulative probability of $c = 1 - q^K$ which includes up to $i = K$ types. We first compute the partial entropy H_c up to a cumulative probability c.

$$\hat{p}_i = \frac{p \cdot q^{i-1}}{1 - q^K} \quad \forall i = 1, 2, \ldots, K \tag{9}$$

$$\begin{aligned} H_c &= -\sum_{i=1}^{K}\left(\frac{p \cdot q^{i-1}}{1 - q^K}\right)\ln\left(\frac{p \cdot q^{i-1}}{1 - q^K}\right) \\ &= -\frac{1}{(1 - q^K)}\sum_{i=1}^{K}(p \cdot q^{i-1})\ln(p \cdot q^{i-1}) + \ln(1 - q^K) \\ &= -\frac{p\ln(p)}{(1-q)} - \frac{pq\ln(q)}{(1-q)^2} + \frac{(K-1)pq^K\ln(q)}{(1-q)(1-q^K)} + \ln(1 - q^K) \\ &= H + \frac{(K-1)q^K\ln(q)}{(1-q^K)} + \ln(1 - q^K) = H + \ln\underbrace{\left(q^{\frac{(K-1)q^K}{(1-q^K)}}(1 - q^K)\right)}_{<1} \end{aligned} \tag{10}$$

$$D_c = \mathrm{e}^{H_c} = \frac{\left(q^{\frac{(K-1)q^K}{(1-q^K)}}(1 - q^K)\right)}{(p \cdot q^{\frac{q}{p}})} = D \cdot \underbrace{\left(q^{\frac{(K-1)q^K}{(1-q^K)}}(1 - q^K)\right)}_{<1}. \tag{11}$$

$$C_c = \frac{D_c \cdot 100}{D} = \left(q^{\frac{(K-1)q^K}{(1-q^K)}}(1 - q^K)\right) \cdot 100. \tag{12}$$

In regard to this last set of computations, several observations are in order. First as $K \to \infty$, we have that $c \to 1$, $H_c \to H$ and $D_c \to D$ as expected.

Second, we also have that $H_c < H$ and $D_c < D\ \forall\ 0 < c < 1$, indicating that the partial entropy H_c from the truncated geometric distribution and its corresponding partial diversity contribution D_c both monotonically increase to

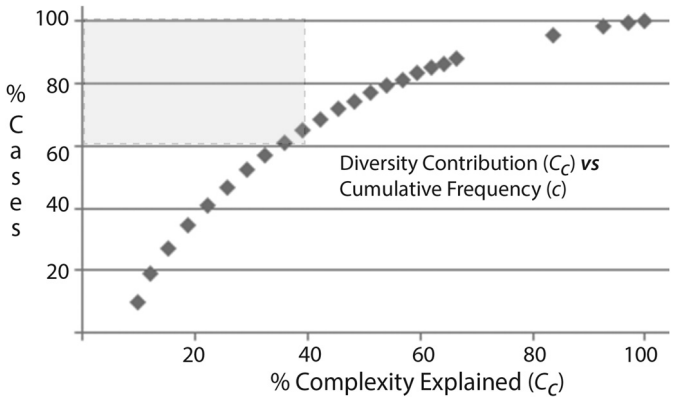


Fig. 2. C_c for geometric distribution.

their corresponding total entropy and total diversity H and D respectively, as $c \to 1$ (or alternatively as $H \to \infty$). These results indicate, as shown in Fig. 2, that diversity D_c increases monotonically as we include more and more parts of the infinite support of the geometric distribution.

Third, as shown in Fig. 2, we can plot and calculate the diversity contribution C_c versus the cumulative frequency c. For example, the gray area in Fig. 2 shows that the curve passes through a region where more than 60% of cases (probability tries) account for less than 40% of the total distribution of the diversity for this system.

Finally, as also shown in Fig. 2, we also see that as $c \to 1$, we have that $K \to \infty$ and accordingly, $D_c \to 100$ indicating that (100, 100) is a point on the C_c versus c curve. That (0, 0) is a point is a trivial observation.

2.2.3. The general case

Based on our validation of the uniform and geometric distribution, we can now write the following formula for the general case, for all such systems where the complexity types are not equi-probable:

$$D_c = e^{H_c} = e^{-\sum_{i=1}^{K} \hat{p}_i \ln(\hat{p}_i)} = e^{-\sum_{i=1}^{K} \ln(\hat{p}_i)^{\hat{p}_i}} = e^{-\ln\left(\Pi_{i=1}^{K} \hat{p}_i\right)^{\hat{p}_i}}$$

$$= \frac{1}{\Pi_{i=1}^{K} \hat{p}_i^{\hat{p}_i}} = \frac{1}{\Pi_{i=1}^{K} \left(\frac{p_i}{c}\right)^{\left(\frac{p_i}{c}\right)}} = \frac{\Pi_{i=1}^{K} c^{\left(\frac{p_i}{c}\right)}}{\Pi_{i=1}^{K} p_i^{\left(\frac{p_i}{c}\right)}} = \frac{c^{\left(\frac{\sum_{i=1}^{K} p_i}{c}\right)}}{\Pi_{i=1}^{K} p_i^{\left(\frac{p_i}{c}\right)}} = \frac{c}{\Pi_{i=1}^{K} p_i^{\left(\frac{p_i}{c}\right)}}.$$

$$D = \lim_{K \to \infty} \frac{c}{\Pi_{i=1}^{K} p_i^{\left(\frac{p_i}{c}\right)}}. \tag{13}$$

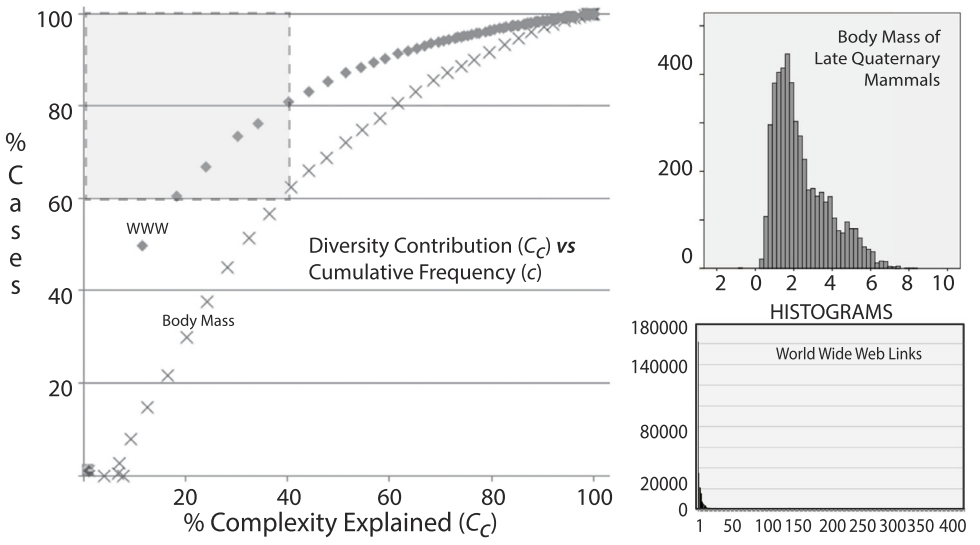
In Eq. (13), we note that as $K \to \infty$ we automatically have that $c \to 1$ since c is that cumulative probability of cases up to complexity type $i = K$.

Finally, we find the percentage of the total diversity contribution D that D_c covers by (a) calculating $\frac{D_c * 100}{D}$ and (b) plotting that versus the cumulative probability c, by repeating the same calculations above for $c = 0$ through $c = 1$, which gives us the type of distributions shown in Fig. 3.

3. Two empirical examples

Now that our general formula is defined, we can use it to calculate the distribution of the diversity of complexity in two physical systems, culled from our recent research. As stated earlier, the current paper is part of a series of studies addressing the empirical/statistical distribution of the diversity of complexity within and amongst complex systems [11,12]. To date, we have used C_c to empirically explore the distribution of the diversity of complexity in a wide variety of systems, from the Maxwell–Boltzmann distribution for kinetic energy of an ideal gas at thermodynamic equilibrium and the human diseasome map to income distributions and Hubble's classic data on the velocity of galaxies. For the current paper, we chose two examples: one natural and one human-made.

The natural system is the *Body Mass of Late Quaternary Mammals* (MOM v4.1). It is part of a taxonomic list of all known mammals of the world ($N = 4{,}629$ species) to which the researchers added status, distribution, and body mass estimates compiled from the literature (see http://biology.unm.edu/fasmith/Datasets/) [32,33]. Here, following the literature on allometry, complexity and scaling-laws [34], complexity was defined as a function of body mass, moving from the smallest body mass (least complex) to the largest body mass (most complex).



Complex System	Coordinates for Each System in the 60/40 Region		
World Wide Web	***80.86***	***40.16***	325,729
Body Mass of Late Quaternary Mammals	62.30	42.56	4,916
	% Cases	% Complexity	N= total cases

Fig. 3. C_c for mammalian body mass and segment of world wide web.

The human-made system is a segment of the *World Wide Web*. This segment is a directed network comprised of 325,729 vertices and 1,497,135 arcs (27,455 loops). It is part of the *Pajek* complex network software platform and its online collection of datasets; and was originally part of the Notre Dame Self-Organizing Networks database (see http://vlado.fmf.uni-lj.si/pub/networks/data/ND/NDnets.htm). For this system, following the literature on complex networks, [35,18] complexity was numerically defined as the number of links for each document, with one link (arc) being the least complex and the hubs (the most densely connected nodes in the network) being the most complex.

As concerns our usage of C_c to examine these two systems, several comments are in order: First, as shown in Fig. 3, one can easily see that the distribution of the diversity of complexity in both systems, when plotted as a histogram, takes the characteristic shape of a skewed-right distribution.

Second, the percentage diversity of complexity in both systems, up to some chosen point, can be easily computed. In Fig. 3, for example, we see that both curves pass through our designated 60/40 region, such that, at minimum, 60% of the cases in both systems are constrained to the lower 40% or less of the diversity of complexity in both systems.

Finally, the distribution of the diversity of complexity in these two systems can be compared. In Fig. 3, for example, we see that the diversity of complexity is much more highly constrained for the World Wide Web in the 60/40 region (over 80.86% of cases) than it is for the body mass of Quaternary Mammals.

4. Conclusion

We have modified the traditional Shannon–Wiener entropy to compute (a) the distribution of the diversity of complexity and (b) the diversity contribution of complexity of any part of a system. We have shown that for the simple case of an equi-probable set of complexity types, we recover the obvious answer i.e. the diversity is simply equal to the number of complexity types. We have also computed an analytical formula for C_c in the case of the geometric distribution, which serves as an abstract model for most complexity measures i.e., a skewed right distribution. And to complete the cycle, we have also computed an analytical formula for C_c in the most general case of an arbitrary distribution of complexity types x_i, with probability of occurrence p_i. We also showed that, because C_c is based on information coming from a part of the system, up to a cumulative probability of c, it can be used to compare the diversity of complexity within and across real physical systems. Furthermore, since the measure is derived purely from information theory, we can use it to compare the distribution of any property of the system and not just complexity i.e., instead of complexity we can use some other property of the system for which we have a histogram (or a probability distribution) for a large collection of systems. This makes C_c a new statistical measure of comparing whole or parts of probability distributions from the viewpoint of information theory.

However, while the current paper has mathematically validated and initially demonstrated the utility of C_c, and while we have used our measure in two subsequent studies to examine a series of different empirical and mathematical systems,

[11,12] the necessary next step is to continue to examine the universal macroscopic properties of the diversity of complexity across an even wider range of mathematical, natural and human-made systems. We hypothesize that C_c will allow for such a comparison because the diversity contribution D_c mentioned above is computed by calculating the equivalent number of equi-probable cases that are required to give the same Shannon entropy value H_c, and thereby makes the measure multiplicative, as stated in the paper. We will also be pursuing the applicability of C_c in the field of probability theory, to a variety of well known discrete and continuous probability distributions with an emphasis on comparing the distribution of diversity of the random variable in our future studies.

Acknowledgments

The authors want to thank the following colleagues at Kent State University: (1) Dean Susan Stocker, (2) Kevin Acierno and Michael Ball (Computer Services), and (3) the Complexity in Health and Infrastructure Group for their support. We also wish to thank Emma Uprichard and David Byrne and the ESRC Seminar Series on Complexity and Method in the Social Sciences (Centre for Interdisciplinary Methodologies, University of Warwick, UK) for the chance to work through the initial framing of these ideas.

References

[1] C. Adami, What is complexity? BioEssays 24 (12) (2002) 1085–1094.
[2] J.E. Contreras-Reyes, Rényi entropy and complexity measure for skew-gaussian distributions and related families, Physica A 433 (2015) 84–91.
[3] R. Lopez-Ruiz, H. Mancini, X. Calbet, A statistical measure of complexity, arXiv preprint.
[4] T.J. McCabe, A complexity measure, IEEE Trans. Softw. Eng. (4) (1976) 308–320.
[5] M. Mitchell, Complexity: A Guided Tour, Oxford University Press, 2009.
[6] I. Rojdestvenski, M. Cottam, G. Oquist, N. Huner, Thermodynamics of complexity, Physica A 320 (2003) 318–328.
[7] T. Yamano, A statistical measure of complexity with nonextensive entropy, Physica A 340 (1) (2004) 131–137.
[8] S. Lloyd, Measures of complexity: a nonexhaustive list, IEEE Control Syst. Mag. 21 (4) (2001) 7–8.
[9] L.T. Lui, G. Terrazas, H. Zenil, C. Alexander, N. Krasnogor, Complexity measurement based on information theory and kolmogorov complexity, Artificial life.
[10] C.R. Shalizi, Methods and techniques of complex systems science: An overview, in: Complex Systems Science in Biomedicine, Springer, 2006, pp. 33–114.
[11] B. Castellani, R. Rajaram, Complex systems and the limiting law of restricted diversity, Complexity.
[12] R. Rajaram, B. Castellani, Kinetic energy and the limiting law of restricted diversity, Eur. Phys. Lett.
[13] D. Byrne, G. Callaghan, Complexity Theory and the Social Sciences: The State of the Art, Routledge, UK, 2013.
[14] D. Byrne, C. Ragin, Using cluster analysis, qualitative comparative analysis and nvivo in relation to the establishment of causal configurations with pre-existing large-N datasets: Machining hermeneutics, in: The Sage Handbook of Case-Based Methods, Sage Press, CA, 2009, pp. 260–268.
[15] R. Rajaram, B. Castellani, Modeling complex systems macroscopically: Case/agent-based modeling, synergetics and the continuity equation, Complexity (2012) http://dx.doi.org/10.1002/cplx.21412.
[16] R. Rajaram, B. Castellani, The utility of non-equilibrium statistical mechanics, specifically transport theory, for modeling cohort data, Complexity (2014) http://dx.doi.org/10.1002/cplx.21512.
[17] D. Sornette, Probability distributions in complex systems, in: R. Meyers (Ed.), Encyclopedia of Complexity and Systems Science, Springer, 2009, pp. 7009–7024.
[18] A.L. Barabasi, Scale-free networks: a decade and beyond, Science 325 (5939) (2009) 412.
[19] M. Newman, Power laws, pareto distributions and zipf's law, Contemp. Phys. 46 (5) (2005) 323–351.
[20] W. Weaver, Science and complexity, Am. Sci. 36 (1948) 536–544.
[21] B. Castellani, R. Rajaram, Case-based modeling and the sacs toolkit: A mathematical outline, Comput. Math. Organ. Theory 18 (2) (2012) 153–174.
[22] L.S. Yaeger, How evolution guides complexity, HFSP J. 3 (5) (2009) 328–339.
[23] S.B. Carroll, Chance and necessity: the evolution of morphological complexity and diversity, Nature 409 (6823) (2001) 1102–1109.
[24] C. Shannon, A mathematical theory of communication, Bell Syst. Tech. J., 27 (3).
[25] M.O. Hill, Diversity and evenness: a unifying notation and its consequences, Ecology 54 (2) (1973) 427–432.
[26] L. Jost, Entropy and diversity, Oikos 113 (2) (2006) 363–375.
[27] T. Leinster, C.A. Cobbold, Measuring diversity: the importance of species similarity, Ecology 93 (3) (2012) 477–489.
[28] J. Beck, W. Schwanghart, Comparing measures of species diversity from incomplete inventories: an update, Methods Ecol. Evol. 1 (1) (2010) 38–44.
[29] R. MacArthur, Patterns of species diversity, Biol. Rev. 40 (1965) 510–533.
[30] M. Hill, Diversity and evenness: a unifying notation and its consequences, Ecology 54 (1973) 427–432. http://dx.doi.org/10.2307/1934352.
[31] R. Peet, The measurement of species diversity, Ann. Rev. Ecol. Syst. 5 (1974) 285–307. http://dx.doi.org/10.1146/annurev.es.05.110174.001441.
[32] F.A. Smith, J.H. Brown, J.P. Haskell, J. Alroy, E.L. Charnov, T. Dayan, B.J. E, et al., Similarity of mammalian body size across the taxonomic hierarchy and across space and time, Am. Nat. 163 (2004) 672–691.
[33] F.A. Smith, S. Lyons, S. Ernest, K. Jones, D. Kaufman, T. Dayan, P. Marquet, J. Brown, J. Haskell, Body mass of late quaternary mammals (updated version of data obtained from senior author), Ecology 84 (2003) 3402.
[34] G.B. West, B.J. H, B.J. Enquist, A general model for the origin of allometric scaling laws in biology, Science 276 (5309) (1997) 122–126.
[35] R. Albert, A.L. Barabasi, Statistical mechanics of complex networks, Rev. Modern Phys. 74 (1) (2002) 47–97.