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 $\to \infty$ 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) $n \to \infty$.

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].