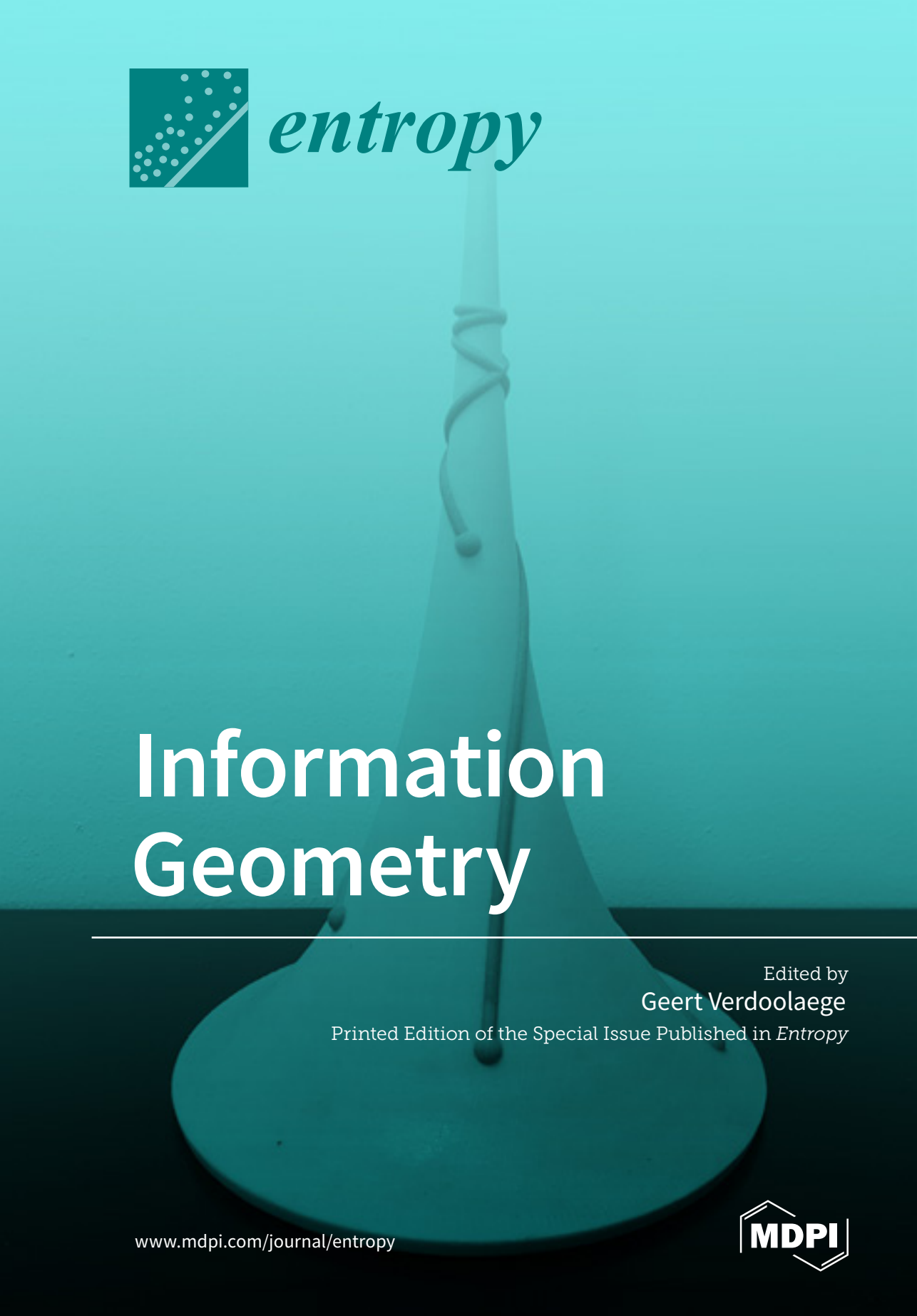
entropy



Information Geometry

Edited by
Geert Verdoolaege

Printed Edition of the Special Issue Published in *Entropy*

www.mdpi.com/journal/entropy

MDPI

Information Geometry

Information Geometry

Special Issue Editor

Geert Verdoolaege

MDPI • Basel • Beijing • Wuhan • Barcelona • Belgrade

Special Issue Editor
Geert Verdoolaege
Ghent University
Belgium

Editorial Office
MDPI
St. Alban-Anlage 66
4052 Basel, Switzerland

This is a reprint of articles from the Special Issue published online in the open access journal *Entropy* (ISSN 1099-4300) in 2014 (available at: https://www.mdpi.com/journal/entropy/special_issues/information-geometry)

For citation purposes, cite each article independently as indicated on the article page online and as indicated below:

LastName, A.A.; LastName, B.B.; LastName, C.C. Article Title. *Journal Name* **Year**, *Article Number*, Page Range.

ISBN 978-3-03897-632-5 (Pbk)
ISBN 978-3-03897-633-2 (PDF)

Cover image courtesy of Geert Verdoolaege.

Contents

About the Special Issue Editor . . . . . **vii**

Preface to "Information Geometry" . . . . . **ix**

Shun-ichi Amari
Information Geometry of Positive Measures and Positive-DefiniteMatrices: Decomposable Dually Flat Structure
Reprinted from: *Entropy* **2014**, *16*, 2131–2145, doi:10.3390/e16042131 . . . . . **1**

Harsha K. V. and Subrahamanian Moosath K S
F-Geometry and Amari's $\alpha-$Geometry on a Statistical Manifold
Reprinted from: *Entropy* **2014**, *16*, 2472–2487, doi:10.3390/e16052472 . . . . . **14**

Frank Critchley and Paul Marriott
Computational Information Geometry in Statistics: Theory and Practice
Reprinted from: *Entropy* **2014**, *16*, 2454–2471, doi:10.3390/e16052454 . . . . . **29**

Paul Vos and Karim Anaya-Izquierdo
Using Geometry to Select One Dimensional Exponential Families That Are Monotone Likelihood Ratio in the Sample Space, Are Weakly Unimodal and Can Be Parametrized by a Measure of Central Tendency
Reprinted from: *Entropy* **2014**, *16*, 4088–4100, doi:10.3390/e16074088 . . . . . **44**

Guido Montúfar, Johannes Rauh and Nihat Ay
On the Fisher Metric of Conditional Probability Polytopes
Reprinted from: *Entropy* **2014**, *16*, 3207–3233, doi:10.3390/e16063207 . . . . . **56**

André Klein
Matrix Algebraic Properties of the Fisher Information Matrix of Stationary Processes
Reprinted from: *Entropy* **2014**, *16*, 2023–2055, doi:10.3390/e16042023 . . . . . **80**

Keisuke Yano and Fumiyasu Komaki
Asymptotically Constant-Risk Predictive Densities When the Distributions of Data and Target Variables Are Different
Reprinted from: *Entropy* **2014**, *16*, 3026–3048, doi:10.3390/e16063026 . . . . . **110**

Samuel Livingstone and Mark Girolami
Information-Geometric Markov Chain Monte Carlo Methods Using Diffusions
Reprinted from: *Entropy* **2014**, *16*, 3074–3102, doi:10.3390/e16063074 . . . . . **131**

Hui Zhao and Paul Marriott
Variational Bayes for Regime-Switching Log-Normal Models
Reprinted from: *Entropy* **2014**, *16*, 3832–3847, doi:10.3390/e16073832 . . . . . **155**

Frank Nielsen, Richard Nock and Shun-ichi Amari
On Clustering Histograms with k-Means by Using Mixed α-Divergences
Reprinted from: *Entropy* **2014**, *16*, 3273–3301, doi:10.3390/e16063273 . . . . . **169**

Salem Said, Lionel Bombrun and Yannick Berthoumieu
New Riemannian Priors on the Univariate Normal Model
Reprinted from: *Entropy* **2014**, *16*, 4015–4031, doi:10.3390/e16074015 . . . . . **194**

Luigi Malagò and Giovanni Pistone
Combinatorial Optimization with Information Geometry: The Newton Method
Reprinted from: *Entropy* **2014**, *16*, 4260–4289, doi:10.3390/e16084260 . . . . . **209**

Domenico Felice, Carlo Cafaro and Stefano Mancini
Information Geometric Complexity of a Trivariate Gaussian Statistical Model
Reprinted from: *Entropy* **2014**, *16*, 2944–2958, doi:10.3390/e16062944 . . . . . **237**

Alexandre Levada
Learning from Complex Systems: On the Roles of Entropy and Fisher Information in Pairwise Isotropic Gaussian Markov Random Fields
Reprinted from: *Entropy* **2014**, *16*, 1002–1036, doi:10.3390/e16021002 . . . . . **250**

Masatoshi Funabashi
Network Decomposition and Complexity Measures: An Information Geometrical Approach
Reprinted from: *Entropy* **2014**, *16*, 4132–4167, doi:10.3390/e16074132 . . . . . **283**

Roger Balian
The Entropy-Based Quantum Metric
Reprinted from: *Entropy* **2014**, *16*, 3878–3888, doi:10.3390/e16073878 . . . . . **315**

Xiaozhao Zhao, Yuexian Hou, Dawei Song and Wenjie Li
Extending the Extreme Physical Information to Universal Cognitive Models via a Confident Information First Principle
Reprinted from: *Entropy* **2014**, *16*, 3670–3688, doi:10.3390/e16073670 . . . . . **324**

About the Special Issue Editor

Geert Verdoolaege obtained an M.Sc. degree in Theoretical Physics in 1999 and the Ph.D. in Engineering Physics in 2006, both at Ghent University (UGent, Belgium). His Ph.D. work concerned applications of Bayesian probability theory to plasma spectroscopy in fusion devices. He was a postdoctoral researcher in the field of computer vision at the University of Antwerp (2007–2008), working on probabilistic modeling of image textures using information geometry. From 2008 to 2010, he was with the Department of Data Analysis at UGent, where he worked on modeling and estimation of brain activity, based on functional magnetic resonance imaging. In 2010, he returned to the Department of Applied Physics at UGent, first as a postdoctoral assistant and from 2014 onwards, as a part-time assistant professor. Since 2013, he has held a cross-appointment as a researcher at the Laboratory for Plasma Physics of the Royal Military Academy (LPP-ERM/KMS) in Brussels. His research activities comprise development of data analysis techniques using methods from probability theory, machine learning and information geometry, and their application to nuclear fusion experiments. He also teaches a Master course on Continuum Mechanics at Ghent University. He serves on the editorial board of the multidisciplinary journal Entropy and is a member of the scientific committees of several conferences (IAEA Technical Meeting on Fusion Data Processing, Validation and Analysis; International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering; Conference on Geometric Science of Information). In addition, he is a consulting expert in the International Tokamak Physics Activity (ITPA) Transport and Confinement Topical Group and member of the General Assembly of the European Fusion Education Network (FuseNet).

Preface to "Information Geometry"

The mathematical field of information geometry originated from the observation the Fisher information can be used to define a Riemannian metric on manifolds of probability distributions. This led to a geometrical description of probability theory and statistics, which over the years has developed into a rich mathematical field with a broad range of applications in the data sciences. Moreover, similar to the concept of entropy, there are various connections to and applications of information geometry in statistical mechanics, quantum mechanics, and neuroscience.

It has been a pleasure to act as a guest editor for this first Special Issue on information geometry in the journal Entropy. For me, as a physicist working on the development and application of data science techniques in the context of nuclear fusion experiments, the interdisciplinary character of information geometry has always been one of the main reasons for its appeal. There are, of course, many other domains in physics where geometrical notions play a key role, including classical mechanics, continuum mechanics (which I have been teaching at Ghent University for several years now), general relativity, and much of modern physics. This interplay between the beautiful and elegant formalism of differential geometry on the one hand and physics and data science on the other hand is both fascinating and inspiring. The variety of topics covered by this Special Issue is a reflection of this cross-fertilization between disciplines.

"Information Geometry I" has been a great success, and although the papers were published already several years ago, it was decided that it was worthwhile to reprint the collection of papers in book form. Indeed, even though all papers present original research, many have a strong tutorial character, and we were honored to receive multiple contributions by authorities in the field. The papers have been structured according to their main subject area, or field of application, and we briefly discuss each of them in the following.

We start with two papers related to the foundations of information geometry. We were very pleased to receive a contribution by one of the founders of the field of information geometry, prof. Shun-ichi Amari. In his paper, the dually flat structure of the manifold of positive measures is discussed, derived from a class of Bregman divergences. These so-called (ρ,τ)-divergences, originally proposed by J. Zhang, are defined in terms of two monotone, scalar functions (ρ and τ) and form a unique class of dually flat, decomposable divergences. This is extended to the set of positive-definite matrices, additionally requiring invariance of the divergence under matrix transformations. It is well known that such dually flat manifolds have computationally desirable properties in applications to classification and information retrieval.

Harsha K. V. and Subrahamanian Moosath K. S. introduce F-geometry as a generalization of α-geometry, based on a representation of a probability density function through a function F. They then combine this with another function G to define a weighted expectation, from which an (F,G)-metric and connection are deduced. A condition for two of such structures to lead to dual connections is also derived. However, it was shown by Zhang (J. Zhang, Entropy 17, pp. 4485–4499, 2015) that this framework is equivalent to the (ρ,τ)-geometry introduced earlier by him. Although the present paper is slightly different in perspective, it should be read with this equivalence in mind.

The next four papers deal with applications of information geometry in statistics. The paper by Frank Critchley and Paul Marriott presents an important research program aimed at rendering some of the most useful results of information geometry more accessible to statisticians in

practical applications. Indeed, the formalism of differential geometry and tensor algebra can appear daunting at first sight and may present an obstacle to adoption of many useful results by practitioners. The paper describes a computational framework that facilitates implementation of results from information geometry, based on an embedding of various important statistical models in a (sufficiently large) simplex. Challenges related to extension of the framework to the infinite–dimensional case are touched upon as well.

In the paper by Paul Vos and Karim Anaya-Izquierdo, the goal is to identify one-dimensional exponential families enjoying a number of properties that are convenient for statistical modeling, i.e., parametrization by a measure of central tendency, unimodality, and monotone likelihood ratio. The basis for the framework is the multinomial distribution, modeled geometrically by the simplex. The selection of exponential families with desirable properties is then based on a partitioning of the natural parameter space of the family of multinomial distributions by means of convex cones.

Guido Montúfar and co-workers consider various possibilities to define natural Riemannian metrics on polytopes of stochastic matrices, which describe the conditional probability distribution of two categorical random variables. Inspired by the classical result regarding the uniqueness of the Fisher metric by requiring invariance under Markov morphisms, they define metrics derived from a natural class of stochastic maps between such polytopes, or, alternatively, through embeddings in various possible model spaces. They provide recommendations as to which metric to use, depending on the application.

André Klein, in his article, provides a survey of several matrix algebraic properties of the Fisher information matrix corresponding to weakly stationary time series. The link with various structured matrices arising from a number of time series models is demonstrated. A statistical distance measure is built using the Fisher information matrix in the context of classical and quantum information. Finally, conditions are obtained for the Fisher information of a stationary process to obey certain forms of the Stein equation.

We continue with three papers concerning applications of information geometry in Bayesian inference and simulation. Keisuke Yano and Fumiyasu Komaki, in their paper, construct constant-risk Bayesian predictive densities using the Kullback-Leibler loss function when the distributions of the data and the target variable to be predicted are different but have a common unknown parameter. Specifically, the issue of prior selection is investigated, and several applications are given.

Samuel Livingstone and Mark Girolami provide an introduction to recent enhancements of Markov chain Monte Carlo simulation techniques inspired by information geometry. They apply this to the Metropolis–Hastings algorithm driven by Langevin diffusion, gradually transforming the ingredients to the setting of a Riemannian manifold equipped with a metric similar to the Fisher information metric. Pointers to various applications are given. The paper is written in a way that also makes it accessible to practitioners with little background in stochastic processes and geometry.

The paper by Hui Zhao and Paul Marriott concerns Bayesian inference making use of variational methods for approximating the posterior distribution. In the context of inference for time series models that switch between different regimes, variational Bayes is shown to be a computationally attractive alternative to Markov chain Monte Carlo simulations. The geometry related to the projection of the posterior onto a computationally tractable family of distributions is elucidated by means of a simple example. This is followed by an application wherein it is shown that variational Bayes is successful in estimating the regime-switching model, including the number of regimes.

Applications of information geometry in machine learning are represented by the following

three papers. The article by Frank Nielsen and colleagues considers κ-means histogram clustering, with applications to, e.g., information retrieval. Based on the α-divergences as similarity measures, clustering is performed using either the sided (asymmetric) or symmetrized divergence, or by means of the interesting notion of a mixed divergence. An important computational advantage is that the centroids based on the sided and mixed divergences have a closed-form expression. Next, the scheme is extended to algorithms with optimized initialization of cluster centroids, as well as soft clustering.

Salem Said and co-workers present a class of distributions on the manifold of the univariate normal model equipped with the Fisher information metric. Expressed in terms of the Fisher-Rao distance, the distributions are used as priors for modeling the classes in Bayesian classification problems of normal distributions. Characteristics of this "Gaussian" distribution on the manifold are discussed, as well as estimation of its parameters and the posterior using the Laplace approximation. In an application to classification of image textures, the improved performance of these priors over conjugate priors is demonstrated.

Luigi Malagò and Giovanni Pistoni address optimization on manifolds of exponential distributions on a discrete state space using Newton's method, which is based on second-order calculus. In particular, the goal is to find maxima of the expectation of a function with respect to the distribution (stochastic relaxation). Details of the computation are provided, including calculation of the Riemannian Hessian. A nonparametric formalism is used, with a view to extension to the infinite–dimensional case.

The next three papers are related to the role of information geometry in complex systems research. Domenico Felice and colleagues consider the time-averaged volume explored by geodesics on a statistical manifold as an indicator of complexity of the entropic dynamics of a system. The parameters of the model play the role of macrovariables conveying information on the system's microstate. Examples are given for the case of univariate, bivariate, and trivariate normal distributions, providing interesting results depending on correlations between the microvariables.

Alexandre Levada investigates the role of entropy and Fisher information in pairwise isotropic Gaussian Markov random fields, acting as models for complex systems. Expressions for these quantities are derived and the evolution of the Fisher information, and entropy is studied as a function of the inverse temperature of the system. An interesting interpretation is given of asymmetries between these curves in terms of the arrow of time.

Masatoshi Funabashi presents a framework for measuring statistical dependence between subsystems of a stochastic model, based on the model's graph representation. A description in terms of the mixed coordinates of the system is used to quantify the complexity loss incurred by cutting an edge of the graph. In addition, a complexity measure is defined as a geometric mean of Kullback–Leibler divergences between decompositions of the system in terms of subsystems with fewer statistical dependencies. This quantifies the degree to which the system can be decomposed.

The following paper concerns an application to physics, specifically quantum mechanics. Roger Balian gives an overview of a geometrical framework for measuring information loss in quantum systems resulting from the mixing of states. A Riemannian metric is defined, based on the von Neumann entropy, generating a mapping between states and observables. The metric is compared to other quantum metrics, as well as the Fisher–Rao metric, and various geometrical properties are derived. Applications are given to quantum information, as well as equilibrium and non-equilibrium quantum statistical mechanics.

The final paper in the Special Issue is situated at the interface between physics and neuroscience.

Xiaozhao Zhao and colleagues consider the principle of extreme physical information based on the Fisher information, which has been used before in an attempt to establish an information-theoretical basis for physical laws. They extend the idea to cognitive systems and aim at narrowing the gap between the information bound and the data information for such complex systems, by transforming the model to a simpler one. This is done by means of a dimensionality reduction technique, also based on the Fisher information. The approach is applied to derive the model for single-layer Boltzmann machines and interpret their learning algorithms.

We are convinced that the varied collection of papers in this Special Issue will be useful for scientists who are new to the field, while providing an excellent reference for the more seasoned researcher. Furthermore, it is worth mentioning that the second *Entropy* Special Issue in this series, "Information Geometry II", will also be published as a book, and that a third Special Issue is being prepared. We hope that the reader will enjoy browsing and reading through this collection of papers as much as we enjoyed guest editing this Special Issue "Information Geometry I".

Finally, I would like to thank the Editor-in-Chief of *Entropy*, Prof. Dr. Kevin H. Knuth, for suggesting the opportunity to guest-edit a Special Issue on information geometry. Furthermore, I wish to thank the editorial staff at MDPI for their great help with contacting authors, organizing paper reviews, and editing the original Special Issue in *Entropy*, as well as the reprinted version in the present book.

Geert Verdoolaege
Special Issue Editor

entropy

MDPI

Article

Network Decomposition and Complexity Measures: An Information Geometrical Approach

Masatoshi Funabashi

Sony Computer Science Laboratories, inc. Takanawa muse bldg. 3F, 3-14-13, Higashi Gotanda, Shinagawa-ku, Tokyo 141-0022, Japan; E-Mail: masa_funabashi@csl.sony.co.jp; Tel.: +81-3-5448-4380; Fax: +81-3-5448-4273

Received: 28 March 2014; in revised form: 24 June 2014 / Accepted: 14 July 2014 / Published: 23 July 2014

Abstract: We consider the graph representation of the stochastic model with n binary variables, and develop an information theoretical framework to measure the degree of statistical association existing between subsystems as well as the ones represented by each edge of the graph representation. Besides, we consider the novel measures of complexity with respect to the system decompositionability, by introducing the geometric product of Kullback–Leibler (KL-) divergence. The novel complexity measures satisfy the boundary condition of vanishing at the limit of completely random and ordered state, and also with the existence of independent subsystem of any size. Such complexity measures based on the geometric means are relevant to the heterogeneity of dependencies between subsystems, and the amount of information propagation shared entirely in the system.

Keywords: information geometry; complexity measure; complex network; system decompositionability; geometric mean

1. Introduction

Complex systems sciences emphasize on the importance of non-linear interactions that can not be easily approximated linearly. In other word, the degrees of non-linear interactions are the source of complexity. The classical reductionism approach generally decomposes a system into its components with linear interactions, and tries to evaluate whether the whole property of the system can still be reproduced. If this decomposition of a system destroys too much information to reproduce the system's whole property, the plausibility of such reductionism is lost. Inversely, if we can evaluate how much information is ignored by the decomposition, we can assume how much complexity of the whole system is lost. This gives us a way to measure the complexity of a system with respect to the system decomposition.

In stochastic systems described as a set of joint distributions, the interaction can basically be expressed as the statistical association between the variables. The simplest reductionism approach is to separate the whole system into some subsets of variables, and assume the independence between them. If such decomposition does not affect the system's property, the isolated subsystem is independent from the rest. On the other hand, if the decomposition loses too much information, then the subsystem is inside of a larger subsystem with strong internal dependencies and can not be easily separated.

The stochastic models have often been represented with the use of graph representation, and treated with the name of complex network [1–3]. Generally, the nodes represent the variables and the weights on the edges are the statistical association between them. However, if we consider the information contained in the different orders of dependencies among variables, the graph with a single kind of edges is not sufficient to express the whole information of the system [4]. An edge of a graph with n nodes contains the information of statistical association up to the n-th order dependencies among n variables. If we try to decompose the system independently by cutting these information, we have to consider what it means to cut the edge of the graph from the information theoretical point of view.

Indeed, analysis on the degree of dependencies existing between variables derived many definition of complexity in stochastic model [5], which have been mostly studied with information theoretical perspective. Beginning with seminal works of Lempel and Ziv (e.g., [6]), computation-oriented definition of complexity takes deterministic formalization and measures the necessary information to reproduce a given symbolic sequence exactly, which is classified with the name of *algorithmic complexity* [7–9].

On the other hand, statistical approach to complexity, namely *statistical complexity*, assumes some stochastic model as theoretical basis, and refers to the structure of information source on it in measure-theoretic way [10–12].

One of the most classical statistical complexities is the mutual information between two stochastic variables, and its generalized form to measure dependence between n variables is proposed (e.g., [13]) and explored in relevance to statistical models and theories by several authors [14–16].

We should also recall that complexity is not necessary conditioned only by information theory, but rather motivated from the organization of living system such as brain activity. The TSE complexity shows further extension of generalized mutual information into biological context, where complexity exists as the heterogeneity between different system hierarchies [17]. These statistical complexities are all based on the boundary condition of vanishing at the limit of completely random and ordered state [18].

The complexity measure is usually the projection from system's variables to one-dimensional quantity, which is composed to express the degree of characteristic that we define to be important in what means "*complexity*". Since the complexity measure is always a many-to-one association, it has both aspects of compressing information to classify the system from simple to complex, and losing resolution of the system's phase space. If the system has n variables, we generally need n independent complexity measures to completely characterize the system with real-value resolution. The problematics of defining a complexity measure is situated on the edge of balancing the information compression on system's complexity with theoretical support, and the resolution of the system identification to be maintained high enough to avoid trivial classification. The latter criterion increases its importance as the system size becomes larger. The better complexity measure is therefore a set of indices, with as less number as possible, which characterizes major features related to the complexity of the system. In this sense, the ensemble of complexity measures is also analogous to the feature space of support vector machine. A non-trivial set of complexity measures need to be complementary to each other in parameter space for the possible best discrimination of different systems.

In this paper, we first consider the stochastic system with binary variables and theoretically develop a way to measure the information between subsystems, which is consistent to the information represented by the edges of the graph representation.

Next, we particularly focus on the generalized mutual information as a start point of the argument, and further consider to incorporate network heterogeneity into novel measures of complexity with respect to the system's decompositionability. This approach will be revealed to be complementary to TSE complexity as the difference between arithmetic and geometric means of information.

2. System Decomposition

Let us consider the stochastic system with n binary variables $\mathbf{x} = (x_1, \cdots, x_n)$ where $x_i \in \{0, 1\}$ $(1 \leq i \leq n)$. We denote the joint distribution of $\mathbf{x}$ by $p(\mathbf{x})$. We define the decomposition $p^{dec}(\mathbf{x})$ of $p(\mathbf{x})$ into two subsystems $\mathbf{y^1} = (x^1_1, \cdots, x^1_{n_1})$ and $\mathbf{y^2} = (x^2_1, \cdots, x^2_{n_2})$ ($n_1 + n_2 = n$, $\mathbf{y^1} \cup \mathbf{y^2} = \mathbf{x}$, $\mathbf{y^1} \cap \mathbf{y^2} = \phi$) as follows:

$$p^{dec}(\mathbf{x}) = p(\mathbf{y^1})p(\mathbf{y^2}), \tag{1}$$

where $p(\mathbf{y^1})$ and $p(\mathbf{y^2})$ are the joint distributions of $\mathbf{y^1}$ and $\mathbf{y^2}$, respectively. For simplicity, hereafter we denote the system decomposition using the smallest subscript of variables in each subsystem. For example, in case $n = 4$, $\mathbf{y^1} = (x_1, x_3)$ and $\mathbf{y^2} = (x_2, x_4)$, we describe the decomposed system $p^{dec}(\mathbf{x})$

as $< 1212 >$. The system decomposition means to cut all statistical association between the two subsystems, which is expressed as setting the independent relation between them.

We will further consider the Equation (1) in terms of the graph representation. We define the undirected graph $\Gamma := (V, E)$ of the system $p(\mathbf{x})$, whose vertices $V = \{x_1, \cdots, x_n\}$ and edges $E = V \times V$ represent the variables and the statistical association, respectively. To express the system, we set the value of each vertex as the value of the corresponding variable, and the weight of each edge as the degree of dependency between the connected variables.

There is however a problem considering the representation with a single kind of edge. The statistical association among variables is not only between two variables, but can be independently defined among plural variables up to the n-th order. Therefore, the exact definition of the weight of the edges remains unclear. To clarify these problematics, we consider the hierarchical marginal distributions $\mathbf{\jmath}$ as another coordinates of the system $p(\mathbf{x})$ as follows:

$$\mathbf{\jmath} = (\mathbf{\jmath^1}; \mathbf{\jmath^2}; \cdots; \mathbf{\jmath^n}), \tag{2}$$

where

$$\begin{aligned} \mathbf{\jmath^1} &= (\eta_1, \cdots, \eta_i, \cdots, \eta_n), (1 < i < n), \\ \mathbf{\jmath^2} &= (\eta_{1,2}, \cdots, \eta_{i,j}, \cdots, \eta_{n-1,n}), (1 < i < j < n), \\ &\vdots \\ \mathbf{\jmath^n} &= \eta_{1,2,\cdots,n}, \end{aligned} \tag{3}$$

and

$$\begin{aligned} \eta_1 &= \sum_{i_2,\cdots,i_n \in \{0,1\}} p(1, i_2, \cdots, i_n), \\ &\vdots \\ \eta_n &= \sum_{i_1,\cdots,i_{n-1} \in \{0,1\}} p(i_1, \cdots, i_{n-1}, 1), \\ \eta_{1,2} &= \sum_{i_3,\cdots,i_n \in \{0,1\}} p(1, 1, i_3, \cdots, i_n), \\ &\vdots \\ \eta_{n-1,n} &= \sum_{i_1,\cdots,i_{n-2} \in \{0,1\}} p(i_1, \cdots, i_{n-2}, 1, 1), \\ &\vdots \\ \eta_{1,2,\cdots,n} &= p(1, 1, \cdots, 1). \end{aligned} \tag{4}$$

Since the definition of $\mathbf{\jmath}$ is a linear transformation of $p(\mathbf{x})$, both coordinates have the degrees of freedom $\sum_{k=1}^{n} {}_n\mathrm{C}_k$.

The subcoordinates $\mathbf{\jmath^1}$ are simply the set of marginal distributions of each variable. The subcoordinates $\mathbf{\jmath^k}$ $(1 < k \leq n)$ include the statistical association among k variables, that can not be expressed with the coordinates less than the k-th order. This means that the different statistical associations exist independently in each order among the corresponding sets of the variables. The statistical association represented by the weight of a graph edge $\{x_i, x_j\}$ is therefore the superposition of the different dependencies defined on every subset of $\mathbf{x}$ including x_i and x_j.

To measure the degree of statistical association in each order, the information geometry established the following setting [19]. We first define another coordinates $\grave{} = (\grave{}^{\mathbf{1}}; \grave{}^{\mathbf{2}}; \cdots; \grave{}^{\mathbf{n}})$ that are the dual

coordinates of $\mathbf{\jmath}$ with respect to the Legendre transformation of the exponential family's potential function $\psi(\grave{})$ to its conjugate potential $\phi(\mathbf{\jmath})$ as follows:

$$\begin{aligned} \grave{}^{\mathbf{1}} &= (\theta_1, \cdots, \theta_n), \\ \grave{}^{\mathbf{2}} &= (\theta_{1,2}, \cdots, \theta_{n-1,n}), \\ &\vdots \\ \grave{}^{\mathbf{n}} &= \theta_{1,2,\cdots,n}, \end{aligned} \tag{5}$$

where

$$\begin{aligned} \psi(\grave{}) &= \log \frac{1}{p(0, \cdots, 0)}, \\ \phi(\mathbf{\jmath}) &= \sum_i \theta_i \eta_i + \sum_{i<j} \theta_{i,j} \eta_{i,j} + \cdots + \theta_{1,2,\cdots,n} \eta_{1,2,\cdots,n} - \psi(\grave{}), \\ \theta_i &= \frac{\partial \phi(\mathbf{\jmath})}{\partial \eta_i}, (1 \leq i \leq n), \\ \theta_{i,j} &= \frac{\partial \phi(\mathbf{\jmath})}{\partial \eta_{i,j}}, (1 \leq i < j \leq n), \\ &\vdots \\ \theta_{1,2,\cdots,n} &= \frac{\partial \phi(\mathbf{\jmath})}{\partial \eta_{1,2,\cdots,n}}. \end{aligned} \tag{6}$$

Note that $\mathbf{\jmath}$ can be inversely derived from $\grave{}$, following Legendre transformation between $\phi(\mathbf{\jmath})$ and $\psi(\grave{})$:

$$\begin{aligned} \eta_i &= \frac{\partial \psi(\grave{})}{\partial \theta_i}, (1 \leq i \leq n), \\ \eta_{i,j} &= \frac{\partial \psi(\grave{})}{\partial \theta_{i,j}}, (1 \leq i < j \leq n), \\ &\vdots \\ \eta_{1,2,\cdots,n} &= \frac{\partial \psi(\grave{})}{\partial \theta_{1,2,\cdots,n}}. \end{aligned} \tag{7}$$

Using the coordinates $\grave{}$, the system is described in the form of the exponential family as follows:

$$p(\mathbf{x}) = \sum_i \theta_i x_i + \sum_{i<j} \theta_{i,j} x_i x_j + \cdots + \theta_{1,2,\cdots,n} x_1 x_2 \cdots x_n - \psi(\grave{}). \tag{8}$$

The information geometry revealed that the exponential family of probability distribution forms a manifold with a dual-flat structure. More precisely, the coordinates $\grave{}$ form a flat manifold with respect to the Fisher information matrix as the Riemannian metric, and α-connection with $\alpha = 1$. Dually to $\grave{}$, the coordinates $\mathbf{\jmath}$ are flat with respect to the same metric but α-connection with $\alpha = -1$. It is known that $\grave{}$ and $\mathbf{\jmath}$ are orthogonal to each other with respect to the Fisher information matrix. This structure give us a way to decompose the degree of statistical association among variables into separated elements of arbitrary orders. We define the so-called k-cut mixture coordinates $\mathbf{\imath}^{\mathbf{k}}$ as follows [14].

$$\mathbf{\imath}^{\mathbf{k}} = (\mathbf{\jmath}^{\mathbf{k}-}; \grave{}^{\mathbf{k}+}), \tag{9}$$

$$\mathbf{\jmath}^{\mathbf{k}-} = (\mathbf{\jmath}^{\mathbf{1}}, \cdots, \mathbf{\jmath}^{\mathbf{k}}), \tag{10}$$

$$\grave{}^{\mathbf{k}+} = (\grave{}^{\mathbf{k+1}}, \cdots, \grave{}^{\mathbf{n}}). \tag{11}$$

We also define the k-cut mixture coordinates $\mathbf{\imath}_{\mathbf{0}}^{\mathbf{k}} = (\mathbf{\jmath}^{\mathbf{k}-}; 0, \cdots, 0)$ with no dependency above the k-th order. We denote the system specified with $\mathbf{\imath}^{\mathbf{k}}$ and $\mathbf{\imath}_{\mathbf{0}}^{\mathbf{k}}$ as $p(\mathbf{x}, \mathbf{\imath}^{\mathbf{k}})$ and $p(\mathbf{x}, \mathbf{\imath}_{\mathbf{0}}^{\mathbf{k}})$, respectively.

Then the degree of the statistical association more than the k-th order in the system can be measured by the Kullback-Leibler (KL-) divergence $D[p(\mathbf{x}, \mathbf{\imath}) : p(\mathbf{x}, \mathbf{\imath}_{\mathbf{0}}^{\mathbf{k}})]$.

$$2N \cdot D[p(\mathbf{x}, \mathbf{\imath}) : p(\mathbf{x}, \mathbf{\imath}_{\mathbf{0}}^{\mathbf{k}})] \sim \chi^2(\sum_{i=k+1}^{n} {}_nC_i), \tag{12}$$

where $D[\cdot : \cdot]$ is the KL-divergence from the first system to the second one.

Here, the decomposition is performed according to the orders of statistical association, which does not spatially distinguish the vertices. If we define the weight of an edge $\{x_i, x_j\}$ with the KL-divergence, the above k-cut coordinates $\mathbf{\imath}^{\mathbf{k}}$ are not appropriate to measure the information represented in each edge. We need to set another mixture coordinates so that to separate only the existing information between x_i and x_j regardless of its order.

Let us return to the definition of the system decomposition and consider on the dual-flat coordinates $\grave{\ }$ and $\mathbf{\jmath}$.

Proposition 1. *The independence between the two decomposed systems* $\mathbf{y^1} = (x_1^1, \cdots, x_{n_1}^1)$ *and* $\mathbf{y^2} = (x_1^2, \cdots, x_{n_2}^2)$ *can be expressed on the new coordinates* $\mathbf{\jmath}^{\mathbf{dec}}$ *as follows:*

$$\begin{aligned}
\eta_i^{dec} &= \eta_i, (1 \le i \le n), \\
\eta_{i,j}^{dec} &= \begin{cases} \eta_{i,j}, (1 \le i < j \le n), & \text{if } \{x_i, x_j\} \subseteq \mathbf{y^1} \text{ or } \subseteq \mathbf{y^2} \\ \eta_i \eta_j, (1 \le i < j \le n), & \text{else} \end{cases}, \\
\eta_{i,j,k}^{dec} &= \begin{cases} \eta_{i,j,k}, (1 \le i < j < k \le n), & \text{if } \{x_i, x_j, x_k\} \subseteq \mathbf{y^1} \text{ or } \subseteq \mathbf{y^2} \\ \eta_{i,j}\eta_k, (1 \le i < j < k \le n), & \text{else if } \{x_i, x_j\} \subseteq \mathbf{y^1} \text{ or } \subseteq \mathbf{y^2} \\ \eta_i\eta_{j,k}, (1 \le i < j < k \le n), & \text{else if } \{x_j, x_k\} \subseteq \mathbf{y^1} \text{ or } \subseteq \mathbf{y^2} \\ \eta_j\eta_{i,k}, (1 \le i < j < k \le n), & \text{else (if } \{x_i, x_k\} \subseteq \mathbf{y^1} \text{ or } \subseteq \mathbf{y^2}) \end{cases}, \\
&\vdots \\
\eta_{1,2,\cdots,n}^{dec} &= \eta_{s[i,k_1,\cdots,k_{n_1-1}]}\eta_{s[j,l_1,\cdots,l_{n_2-1}]}, (x_i \in \mathbf{y^1}, x_j \in \mathbf{y^2}),
\end{aligned} \tag{13}$$

where $s[\cdots]$ *is the ascending sort of the internal sequence.*

Then the corresponding dual coordinates $\grave{\ }^{\mathbf{dec}}$ *take* 0 *elements as follows:*

$$\begin{aligned}
\theta_{i,j}^{dec} &= 0, \quad (1 \le i < j <\le n), \quad \text{if } \{x_i, x_j\} \cap \mathbf{y^1} \neq \phi \text{ and } \{x_i, x_j\} \cap \mathbf{y^2} \neq \phi \\
\theta_{i,j,k}^{dec} &= 0, \quad (1 \le i < j < k \le n), \quad \text{if } \{x_i, x_j, x_k\} \cap \mathbf{y^1} \neq \phi \text{ and } \{x_i, x_j, x_k\} \cap \mathbf{y^2} \neq \phi \\
&\vdots \\
\theta_{1,2,\cdots,n}^{dec} &= 0.
\end{aligned} \tag{14}$$

Proof. For simplicity, we show the cases of $n = 2$ and $n = 3$ for the first node separation.

For $n = 2$, the above defined $\mathbf{J}^{\mathbf{dec}}$ for the system decomposition $< 12 >$ give its dual coordinates $\grave{}^{\mathbf{dec}}$ as follows:

$$\begin{aligned}
\theta_1^{dec} &= \log \frac{\eta_1^{dec} - \eta_{1,2}^{dec}}{1 - \eta_1^{dec} - \eta_2^{dec} + \eta_{1,2}^{dec}} &= \log \frac{\eta_1}{1 - \eta_1}, \\
\theta_2^{dec} &= \log \frac{\eta_2^{dec} - \eta_{1,2}^{dec}}{1 - \eta_1^{dec} - \eta_2^{dec} + \eta_{1,2}^{dec}} &= \log \frac{\eta_2}{1 - \eta_2}, \\
\theta_{1,2}^{dec} &= \log \frac{\eta_{1,2}^{dec}(1 - \eta_1^{dec} - \eta_2^{dec} + \eta_{1,2}^{dec})}{(\eta_1^{dec} - \eta_{1,2}^{dec})(\eta_2^{dec} - \eta_{1,2}^{dec})} &= 0,
\end{aligned} \tag{15}$$

which means the first and second node is independent.

For $n = 3$, the above defined $\mathbf{J}^{\mathbf{dec}}$ for the system decomposition $< 122 >$ give its dual coordinates $\grave{}^{\mathbf{dec}}$ as follows:

$$\begin{aligned}
\theta_1^{dec} &= \log \frac{\eta_1^{dec} - \eta_{1,2}^{dec} - \eta_{1,3}^{dec} + \eta_{1,2,3}^{dec}}{1 - \eta_1^{dec} - \eta_2^{dec} - \eta_3^{dec} + \eta_{1,2}^{dec} + \eta_{1,3}^{dec} + \eta_{2,3}^{dec} - \eta_{1,2,3}^{dec}} = \log \frac{\eta_1}{1 - \eta_1}, \\
\theta_2^{dec} &= \log \frac{\eta_2^{dec} - \eta_{1,2}^{dec} - \eta_{1,3}^{dec} + \eta_{1,2,3}^{dec}}{1 - \eta_1^{dec} - \eta_2^{dec} - \eta_3^{dec} + \eta_{1,2}^{dec} + \eta_{1,3}^{dec} + \eta_{2,3}^{dec} - \eta_{1,2,3}^{dec}} = \log \frac{\eta_2 - \eta_{2,3}}{1 - \eta_2 - \eta_3 + \eta_{2,3}}, \\
\theta_3^{dec} &= \log \frac{\eta_3^{dec} - \eta_{1,3}^{dec} - \eta_{2,3}^{dec} + \eta_{1,2,3}^{dec}}{1 - \eta_1^{dec} - \eta_2^{dec} - \eta_3^{dec} + \eta_{1,2}^{dec} + \eta_{1,3}^{dec} + \eta_{2,3}^{dec} - \eta_{1,2,3}^{dec}} = \log \frac{\eta_3 - \eta_{2,3}}{1 - \eta_2 - \eta_3 + \eta_{2,3}},
\end{aligned} \tag{16}$$

$$\begin{aligned}
\theta_{1,2}^{dec} &= \log \frac{(\eta_{1,2}^{dec} - \eta_{1,2,3}^{dec})(1 - \eta_1^{dec} - \eta_2^{dec} - \eta_3^{dec} + \eta_{1,2}^{dec} + \eta_{1,3}^{dec} + \eta_{2,3}^{dec} - \eta_{1,2,3}^{dec})}{(\eta_1^{dec} - \eta_{1,2}^{dec} - \eta_{1,3}^{dec} + \eta_{1,2,3}^{dec})(\eta_2^{dec} - \eta_{1,2}^{dec} - \eta_{2,3}^{dec} + \eta_{1,2,3}^{dec})} \\
&= 0, \\
\theta_{1,3}^{dec} &= \log \frac{(\eta_{1,3}^{dec} - \eta_{1,2,3}^{dec})(1 - \eta_1^{dec} - \eta_2^{dec} - \eta_3^{dec} + \eta_{1,2}^{dec} + \eta_{1,3}^{dec} + \eta_{2,3}^{dec} - \eta_{1,2,3}^{dec})}{(\eta_1^{dec} - \eta_{1,2}^{dec} - \eta_{1,3}^{dec} + \eta_{1,2,3}^{dec})(\eta_3^{dec} - \eta_{1,3}^{dec} - \eta_{2,3}^{dec} + \eta_{1,2,3}^{dec})} \\
&= 0, \\
\theta_{2,3}^{dec} &= \log \frac{(\eta_{2,3}^{dec} - \eta_{1,2,3}^{dec})(1 - \eta_1^{dec} - \eta_2^{dec} - \eta_3^{dec} + \eta_{1,2}^{dec} + \eta_{1,3}^{dec} + \eta_{2,3}^{dec} - \eta_{1,2,3}^{dec})}{(\eta_2^{dec} - \eta_{1,2}^{dec} - \eta_{2,3}^{dec} + \eta_{1,2,3}^{dec})(\eta_3^{dec} - \eta_{1,3}^{dec} - \eta_{2,3}^{dec} + \eta_{1,2,3}^{dec})} \\
&= \log \frac{\eta_{2,3}(1 - \eta_2 - \eta_3 + \eta_{2,3})}{(\eta_2 - \eta_{2,3})(\eta_3 - \eta_{2,3})},
\end{aligned} \tag{17}$$

$$\begin{aligned}
\theta_{1,2,3}^{dec} &= \log \left[\frac{\eta_{1,2,3}^{dec}}{(\eta_{1,2}^{dec} - \eta_{1,2,3}^{dec})(\eta_{1,3}^{dec} - \eta_{1,2,3}^{dec})(\eta_{2,3}^{dec} - \eta_{1,2,3}^{dec})} \right. \\
&\quad \left. \times \frac{(\eta_1^{dec} - \eta_{1,2}^{dec} - \eta_{1,3}^{dec} + \eta_{1,2,3}^{dec})(\eta_2^{dec} - \eta_{1,2}^{dec} - \eta_{2,3}^{dec} + \eta_{1,2,3}^{dec})(\eta_3^{dec} - \eta_{1,3}^{dec} - \eta_{2,3}^{dec} + \eta_{1,2,3}^{dec})}{(1 - \eta_1^{dec} - \eta_2^{dec} - \eta_3^{dec} + \eta_{1,2}^{dec} + \eta_{1,3}^{dec} + \eta_{2,3}^{dec} - \eta_{1,2,3}^{dec})} \right] \\
&= 0,
\end{aligned} \tag{18}$$

which means the first node is independent from the other nodes.

The generalization is possible with the use of recurrence formula between system size n and $n + 1$, according to the symmetry of the model and Legendre transformation between $\mathbf{J}^{\mathbf{dec}}$ and $\grave{}^{\mathbf{dec}}$ coordinates.

Numerical proof can be obtained by computing directly 0 elements of $\grave{}^{\mathbf{dec}}$ from $\mathbf{J}^{\mathbf{dec}}$. □

The definition of $\mathfrak{J}^{\mathbf{dec}}$ means to decompose the hierarchical marginal distributions $\mathfrak{J}$ into the products of the subsystems' marginal distributions, in case the subscripts traverse the two subsystems. Therefore, only the statistical associations between two subsystems are set to be independent, while the internal dependencies of each subsystem remain unchanged. This is analytically equivalent to compose another mixture coordinates $\mathfrak{z}$, namely the $< \cdots >$-cut coordinates, with proper description of the system decomposition with $< \cdots >$. The $\mathfrak{z}$ consists of the $\mathfrak{J}$ coordinates with the subscripts that do not traverse between the decomposed subsystems, and the $\mathfrak{H}$ coordinates whose subscripts traverse between them.

For simplicity, we only describe here the case $n = 4$ and the decomposition $< 1133 >$ (the set of the first, second, and the third, fourth nodes each form a subsystem). The system $p(\mathbf{x})$ is expressed with the $< 1133 >$-cut coordinates $\mathfrak{z}$ as

$$\begin{aligned}
\mathfrak{z}_1 &= \eta_1, \\
&\vdots \\
\mathfrak{z}_4 &= \eta_4, \\
\mathfrak{z}_{1,2} &= \eta_{1,2}, \\
\mathfrak{z}_{1,3} &= \theta_{1,3}, \\
\mathfrak{z}_{1,4} &= \theta_{1,4}, \\
\mathfrak{z}_{2,3} &= \theta_{2,3}, \\
\mathfrak{z}_{2,4} &= \theta_{2,4}, \\
\mathfrak{z}_{3,4} &= \eta_{3,4}, \\
\mathfrak{z}_{1,2,3} &= \theta_{1,2,3}, \\
&\vdots \\
\mathfrak{z}_{2,3,4} &= \theta_{2,3,4}, \\
\mathfrak{z}_{1,2,3,4} &= \theta_{1,2,3,4}.
\end{aligned} \tag{19}$$

The decomposed system with no statistical association between two subsystems have the following coordinates $\mathfrak{z}^{\mathbf{dec}}$, which is, in any decomposition, equivalent to set all θ in $\mathfrak{z}$ as 0:

$$\begin{aligned}
\mathfrak{z}^{dec}_1 &= \eta_1, \\
&\vdots \\
\mathfrak{z}^{dec}_4 &= \eta_4, \\
\mathfrak{z}^{dec}_{1,2} &= \eta_{1,2}, \\
\mathfrak{z}^{dec}_{1,3} &= 0, \\
\mathfrak{z}^{dec}_{1,4} &= 0, \\
\mathfrak{z}^{dec}_{2,3} &= 0, \\
\mathfrak{z}^{dec}_{2,4} &= 0, \\
\mathfrak{z}^{dec}_{3,4} &= \eta_{3,4}, \\
\mathfrak{z}^{dec}_{1,2,3} &= 0, \\
&\vdots \\
\mathfrak{z}^{dec}_{2,3,4} &= 0, \\
\mathfrak{z}^{dec}_{1,2,3,4} &= 0.
\end{aligned} \tag{20}$$

This is analytically equivalent to the definition of the decomposition (13)–(14) in case of $< 1133 >$. Therefore, the KL-divergence $D[p(\mathbf{x}, \lambda) : p(\mathbf{x}, \lambda^{\mathbf{dec}})]$ measures the information lost by the system decomposition. The following asymptotic agreement to χ^2 test also holds.

Proposition 2.

$$2N \cdot D[p(\mathbf{x}, \lambda) : p(\mathbf{x}, \lambda^{\mathbf{dec}})] \sim \chi^2(\sharp_\theta(\lambda)), \tag{21}$$

where $\sharp_\theta(\lambda)$ is the number of θ coordinates appearing in the λ coordinates.

3. Edge Cutting

We further expand the concept of system decomposition to eventually quantify the total amount of information expressed by an edge of the graph. Let us consider to cut an edge $\{x_i, x_j\}$ $(1 \leq i < j \leq n)$ of the graph with n vertices. Hereafter we call this operation as the edge cutting $i - j$. In the same way as the system decomposition, the edge cutting corresponds to modify the η coordinates to produce $\eta^{\mathbf{ec}}$ coordinates as follows:

$$\begin{aligned}
\eta^{ec}_{i,j} &= \eta_i \eta_j, \\
\eta^{ec}_{s[i,j,k_1]} &= \eta_{s[i,k_1]} \eta_{s[j,k_1]}, \\
\eta^{ec}_{s[i,j,k_1,k_2]} &= \eta_{s[i,k_1,k_2]} \eta_{s[j,k_1,k_2]}, \\
&\vdots \\
\eta^{ec}_{s[i,j,k_1,\cdots,k_{n-2}]} &= \eta_{s[i,k_1,\cdots,k_{n-2}]} \eta_{s[j,k_1,\cdots,k_{n-2}]}, \\
(\{i, j, k_1, \cdots, k_{n-2}\} &= \{1, \cdots, n\}),
\end{aligned} \tag{22}$$

and the rest of $\eta^{\mathbf{ec}}$ remains the same as those of η.

The formation of η^{ec} from η consists of replacing the k-th order elements ($k \geq 3$) of η including both i and j in its subscripts, with the product of the $k-1$-th order η in maximum subgraphs ($k-1$ vertices) each including i or j. This means that all orders of statistical association including the variables x_i and x_j are set to be independent only between them. Other relations that do not include simultaneously x_i and x_j remain unchanged.

Certain combinations of edge cuttings coincide with system decompositions. For example, in case $n = 4$, the edge cuttings $1-2$, $1-3$, and $1-4$ are equivalent to the system decomposition $< 1222 >$.

We define the $i - j$-cut mixture coordinates λ for orthogonal decomposition of the statistical association represented by the edge $\{x_i, x_j\}$. Although actual calculation can be performed only with η coordinates, this generalization is necessary to have a geometrical definition of the orthogonality. For simplicity, we only describe the λ in the case of $n = 4$:

$$\begin{aligned}
\xi_1 &= \eta_1, \\
&\vdots \\
\xi_4 &= \eta_4, \\
\xi_{1,2} &= \theta_{1,2}, \\
\xi_{1,3} &= \eta_{1,3}, \\
\xi_{1,4} &= \eta_{1,4}, \\
\xi_{2,3} &= \eta_{2,3}, \\
\xi_{2,4} &= \eta_{2,4}, \\
\xi_{3,4} &= \eta_{3,4}, \\
\xi_{1,2,3} &= \theta_{1,2,3}, \\
\xi_{1,2,4} &= \theta_{1,2,4}, \\
\xi_{1,3,4} &= \eta_{1,3,4}, \\
\xi_{2,3,4} &= \eta_{2,3,4}, \\
\xi_{1,2,3,4} &= \theta_{1,2,3,4},
\end{aligned} \tag{23}$$

where orthogonality between the elements of $\mathbf{J}$ and $\mathbf{L}$ holds with respect to the Fisher information matrix.

Calculating the dual coordinates $\mathbf{L}^{\mathbf{ec}}$ of $\mathbf{J}^{\mathbf{ec}}$, we can define the coordinates $\mathbf{H}^{\mathbf{ec}}$ of the system after the edge cutting $1-2$ as follows:

$$\begin{aligned}
\xi_1^{ec} &= \eta_1, \\
&\vdots \\
\xi_4^{ec} &= \eta_4, \\
\xi_{1,2}^{ec} &= \theta_{1,2}^{ec}, \\
\xi_{1,3}^{ec} &= \eta_{1,3}, \\
\xi_{1,4}^{ec} &= \eta_{1,4}, \\
\xi_{2,3}^{ec} &= \eta_{2,3}, \\
\xi_{2,4}^{ec} &= \eta_{2,4}, \\
\xi_{3,4}^{ec} &= \eta_{3,4}, \\
\xi_{1,2,3}^{ec} &= \theta_{1,2,3}^{ec}, \\
\xi_{1,2,4}^{ec} &= \theta_{1,2,4}^{ec}, \\
\xi_{1,3,4}^{ec} &= \eta_{1,3,4}, \\
\xi_{2,3,4}^{ec} &= \eta_{2,3,4}, \\
\xi_{1,2,3,4}^{ec} &= \theta_{1,2,3,4}^{ec}.
\end{aligned}$$

Note that the edge cutting can not be defined simply by setting the corresponding elements of $\mathbf{L}^{\mathbf{ec}}$ as 0.

Then the KL-divergence $D[p(\mathbf{x}, \mathbf{H}) : p(\mathbf{x}, \mathbf{H}^{\mathbf{ec}})]$ represent the total amount of information represented by the edge $1-2$.

The following asymptotic agreement to χ^2 test also holds:

Proposition 3.

$$2N \cdot D[p(\mathbf{x}, \_) : p(\mathbf{x}, \_^{\mathbf{ec}})] \sim \chi^2(1 + \sum_{k=1}^{n-2} {}_{n-2}C_k). \tag{24}$$

We call this χ^2 value or the KL-divergence itself as edge information of edge $1-2$.

4. Generalized Mutual Information as Complexity with Respect to the Total System Decomposition

In previous sections, we have introduced a measure of complexity in terms of system decomposition, by measuring the KL-divergence between a given system and its independently decomposed subsystems. We consider here the total system decomposition, and measure the informational distance I between the system and the totally decomposed system where each element are independent.

$$I := \sum_{i=1}^{n} H(x_i) - H(x_1, \cdots, x_n), \tag{25}$$

where

$$H(\mathbf{x}) := -\sum_{\mathbf{x}} p(\mathbf{x}) \log(\mathbf{x}). \tag{26}$$

This quantity is the generalization of mutual information, and is named in various ways such as generalized mutual information, integration, complexity, multi-information, *etc*. according to different authors. For simplicity, we call the I as "*multi-information* taking after [15]. This quantity can be interpreted as a measure of complexity that sums up the order-wise statistical association existing in each subset of components with information geometrical formalization [14]

For simplicity, we denote the multi-information I of n-dimensional stochastic binary variables as follows, using the notation of the system decomposition:

$$I = D[< 111 \cdots 1 >:< 123 \cdots n >]. \tag{27}$$

5. Rectangle-Bias Complexity

The multi-information contains some degrees of freedom in case $n > 2$. That is, we can define a set of distributions $\{p(\mathbf{x}) | I = const.\}$ with different parameters but the same I value. This fact can be clearly explained with the use of information geometry. From the Pythagorean relation, we obtain the followings in case of $n = 3$:

$$\begin{aligned} D[< 111 >:< 113 >] + D[< 113 >:< 123 >] &= D[< 111 >:< 123 >], \\ D[< 111 >:< 121 >] + D[< 121 >:< 123 >] &= D[< 111 >:< 123 >], \\ D[< 111 >:< 122 >] + D[< 122 >:< 123 >] &= D[< 111 >:< 123 >]. \end{aligned} \tag{28}$$

Using these relations, we can schematically represent the decomposed systems on a circle diagram with diameter $\sqrt{I}$. This representation is based on the analogous algebra between Pythagorean relation of KL-divergence, and that of Euclidian geometry where the circumferential angle of a semi-circular arc is always $\frac{\pi}{2}$.

Figure1 represents two different cases with the same I value in case $n = 3$. The distance between two systems in the same diagram corresponds to the square root value of KL-divergence between them. Clearly the left and right figures represent different dependencies between nodes, although they both have the same I value. Such geometrical variation is possible by the abundance of degree of freedom in dual coordinates compared to the given constraint. There exist 7 degrees of freedom in η or θ coordinates for $n = 3$, while the only constraint is the invariance of I value, which only reduce 1

degree of freedom. The remaining 6 degrees of freedom can then be deployed to produce geometrical variation in the circle diagram. As for considering system decomposition, the left figure is difficult to obtain decomposed systems without losing much information. While in the right figure there exists relatively easy decomposition $< 122 >$, which loses less information than any decomposition in the left figure. We call such degree of facility of decomposition with respect to the losing information as *system decompositionability*. In this sense, the left system is more complex although the 2 systems both have the same I value. Especially, in case $D[< 111 >:< 122 >] = D[< 111 >:< 113 >] = D[< 111 >:< 121 >]$, the system does not have any easiest way of decomposition, and any isolation of a node loses significant amount of information.

To further incorporate such geometrical structure reflecting system decompositionability into a measure of complexity, we consider a mathematical way to distinguish between these two figures. Although the total sum of KL-divergence along the sequence of system decomposition is always identical to I by Pythagorean relation, their product can vary according to the geometrical composition in the circle diagram. This is analogous to the isoperimetric inequality of rectangle, where regular tetragon gives the maximum dimensions amongst constant perimeter rectangles.

We propose provisionary a new measure of complexity as follows, namely *rectangle-bias complexity* C_r:

$$C_r = \frac{1}{|SD| - 2} \sum_{< \cdots > \in SD} D[< 11 \cdots 1 >:< \cdots >] \cdot D[< \cdots >:< 12 \cdots n >], \tag{29}$$

where SD is the set of possible system decomposition in n binary variables, and $|SD|$ is the element number of SD. For example, $SD = \{< 111 >, < 122 >, < 121 >, < 113 >, < 123 >\}$ and $|SD| = 5$ for $n = 3$. This measure distinguishes between the two systems in Figure 1, and gives larger value for the left figure. It also gives maximum value in case $D[< 111 >:< 122 >] = D[< 111 >:< 113 >] = D[< 111 >:< 121 >]$. We propose provisionary a new measure of complexity as follows, namely *rectangle-bias complexity* C_r:

$$C_r = \frac{1}{|SD| - 2} \sum_{< \cdots > \in SD} D[< 11 \cdots 1 >:< \cdots >] \cdot D[< \cdots >:< 12 \cdots n >], \tag{30}$$

where SD is the set of possible system decomposition in n binary variables, and $|SD|$ is the element number of SD. For example, $SD = \{< 111 >, < 122 >, < 121 >, < 113 >, < 123 >\}$ and $|SD| = 5$ for $n = 3$. This measure distinguishes between the two systems in Figure 1, and gives larger value for the left figure. It also gives maximum value in case $D[< 111 >:< 122 >] = D[< 111 >:< 113 >] = D[< 111 >:< 121 >]$.

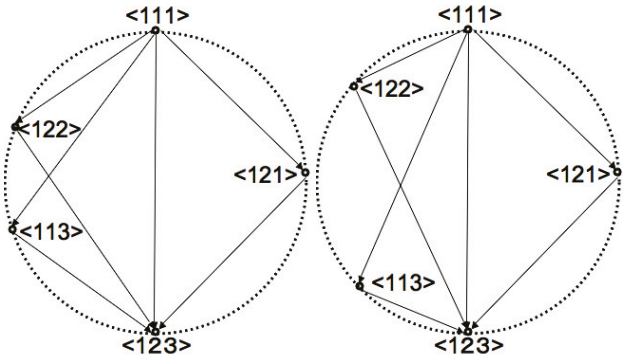


Figure 1. *Circle diagrams of system decomposition in 3-node network.* Both systems have the same value of multi-information I that is expressed as the identical diameter length of the circles. 2 variations are shown, where the left system is more complex (C_r high) in a sense any system decomposition requires to lose more information than the easiest one ($< 122 >$) in the right figure (C_r low).

6. Complementarity between Complexities Defined with Arithmetic and Geometric Means

We evaluate the possibility and the limit of rectangle-bias complexity C_r comparing with other proposed measures of complexity.

The Interests in measuring network heterogeneity have been developed toward the incorporation of multi-scale characteristics into complexity measures. The TSE complexity is motivated from the structure of the functional differentiation of brain activity, which measures the difference of neural integration between all sizes of subsystems and the whole system [17]. Biologically motivated TSE complexity is also investigated from theoretical point of view, to further attribute desirable property as an universal complexity measure independent of system size [20]. The hierarchical structure of the exponential family in information geometry also leads to the order-wise description of statistical association, which can be regarded as a multi-scale complexity measure [14]. The relation between the order-wise dependencies and the TSE complexity is theoretically investigated to establish the order-wise component correspondence between them [15].

These indices of network heterogeneity, however, all depend on the arithmetic mean of the component-wise information theoretical measure. We show that these arithmetic means still miss to measure certain modularity based on the statistical independence between subsystems.

Figure 2 present the simplified cases where complexity measures with arithmetic means fail to distinguish. We consider the two systems with different heterogeneity but identical multi-information I. Here, the multi-information can not reflect the network heterogeneity. The TSE complexity and its information geometrical correspondence in [15] has a sensitivity to measure the network heterogeneity, but since the arithmetic mean is taken over all subsystems, they do not distinguish the component-wise break of symmetry between different scales. The renormalized TSE complexity with respect to the multi-information I still has the same insensitivity. Even by incorporating the information of each subsystem scale, the arithmetic mean can balance out between the scale-wise variations, and a large range of the heterogeneity in different scale can realize the same value of these complexities. For the application in neuroscience, the assumption of a model with simple parametric heterogeneity and the comparison of TSE complexity between different I values alleviate this limitation [17].

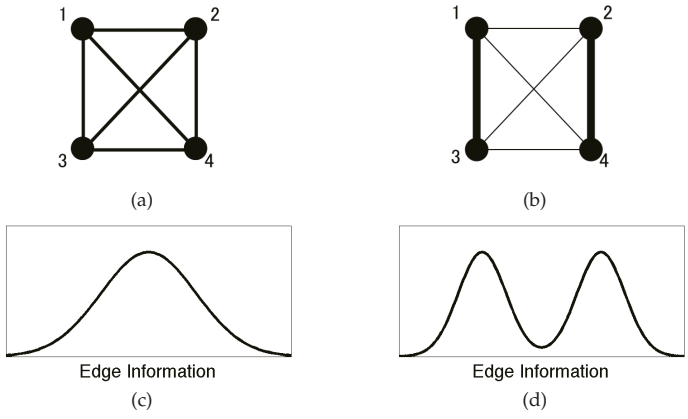


Figure 2. *Schematic examples of stochastic systems with identical multi-information I where complexity measures with arithmetic mean fail to distinguish.* (**a**): Example 1 of stochastic system with homogeneous mean of edge information and symmetric fluctuation of its heterogeneity; (**b**): Example 2 of heterogeneous stochastic system with bimodal edge information distribution and identical multi-information I and complexity based on arithmetic mean as example 1; (**c**): schematic representation of the distribution of statistical association (edge information) in upper network; (**d**): schematic representation of the distribution of statistical association (edge information) in upper network.

In contrast to complexities with arithmetic mean, the rectangle-bias complexity C_r is related to the geometrical mean. The C_r can distinguish the two systems in Figure 2, giving relatively high C_r value to the left system and low value to the right one.

This does not mean , however, that the C_r has a finer resolution than other complexity measures. The constant conditions of complexity measures are the constraints on $\sum_{k=1}^{n} {}_nC_k$ degrees of freedom in model parameter space, which define different geometrical composition of corresponding submanifolds. We basically need $\sum_{k=1}^{n} {}_nC_k$ independent measures to assure the real-value resolution of network feature characterization. Complexities with arithmetic and geometric means are just giving complementary information on network heterogeneity, or different constant-complexity submanifolds structure in statistical manifold as depicted in Figure 3. Therefore, it is also possible to construct a class of systems that has identical I and C_r values but different TSE complexity. Complexity measures should be utilized in combination, with respect to the non-linear separation capacity of network features of interest.

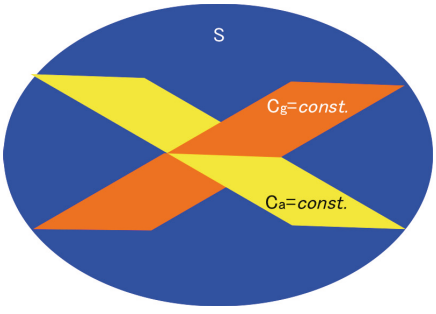


Figure 3. *Schematic representation of complementarity between complexity measures based on arithmetic mean (C_a) and geometric mean (C_g) of informational distance.* An example of the $n-1$ dimensional constant-complexity submanifolds with respect to $C_a = const.$ and $C_g = const.$ conditions are depicted with yellow and orange surface, respectively. The dimension of the whole statistical manifold S is the parameter number n.

7. Cuboid-Bias Complexity with Respect to System Decompositionability

We consider the expansion of C_r into general system size n. The $n \geq 4$ situation is different from $n = 3$ and less in the existence of a hierarchical structure between system decompositions.

Figure 4 shows the hierarchy of the system decompositions in case $n = 4$. Such hierarchical structure between system decompositions is not homogeneous with respect to the subsystems number, and depends on the isomorphic types of decomposed systems. This fact produces certain difference of meaning in complexity between each KL-divergences when considering the system decompositionability.

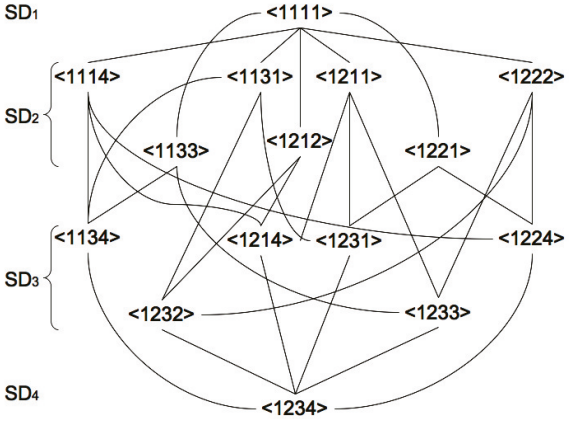


Figure 4. *Hierarchy of system decomposition for 4 nodes network ($n = 4$).* Possible sequences of $Seq = \{SD_1(i_s) \to SD_2(i_s) \to SD_3(i_s) \to SD_4(i_s) | 1 \leq i_s \leq |Seq| = 18\}$ are connected with the lines.

A simple example in 4 nodes network is shown in Figure 5.

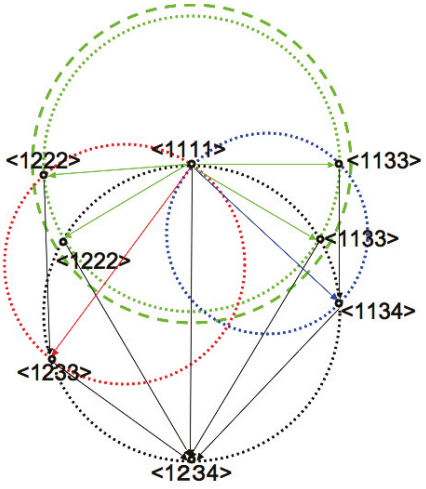


Figure 5. *Hierarchical effect of sequential system decomposition on cuboid volume and rectangle surface on circle graph.* We consider to increase the diameter of the green circle from dotted to dashed one without changing those of the red and blue circles, which gives different effect on the change of $D[< 1222 >:< 1233 >]$ and $D[< 1133 >:< 1134 >]$ according to the hierarchical structure of the decomposition sequences.

We consider the modification of 2 KL-divergences in the figure, $D[< 1111 >:< 1222 >]$ and $D[< 1111 >:< 1133 >]$ from the diameter of green dotted circle to the dashed one.

The joint distribution $P(x_1, x_2, x_3, x_4)$ of a discrete distribution with 4 binary variables (x_1, x_2, x_3, x_4) $(x_1, x_2, x_3, x_4 \in \{0, 1\})$ have $2^4 - 1 = 15$ parameters, which define the dual-flat coordinates of statistical manifold in information geometry.

On the other hand, the possible system decompositions exist as the followings in $n = 4$:

$$\begin{aligned} SD := \{&< 1111 >, < 1114 >, < 1131 >, < 1211 >, < 1222 >, \\ &< 1133 >, < 1212 >, < 1221 >, < 1134 >, < 1214 >, \\ &< 1231 >, < 1224 >, < 1232 >, < 1233 >, < 1234 >\}. \end{aligned} \tag{31}$$

Since the number of possible system decompositions is 15, and each is associated with the modification of different sets of $P(x_1, x_2, x_3, x_4)$ parameters, the system decompositions and KL-divergences between them can be defined independently. This also holds even under the constant condition of I value or other complexity measures except the ones imposing dependency between system decompositions.

This means that we can independently modify the diameter of green dotted circle in Figure 5, without changing the diameters of the red and blue circles, which define the system decompositions $< 1233 >$ and $< 1134 >$ in the sub-hierarchy of $< 1222 >$ and $< 1133 >$, respectively. Other KL-divergences can also be maintained as given constant values for the same reason.

The rectangle-biased complexity C_r increases its value with such modification, but does not reflect the heterogeneity of KL-divergences according to the hierarchy of system decompositions. If we consider the system decompositionability as the mean facility to decompose the given system into its finest components with respect to the "all" possible system decompositions, such hierarchical difference also has a meaning in the definition of complexity.

The effect of modifying the diameter of the green dotted circle is different between the decomposition sequences $< 1111 > \rightarrow < 1222 > \rightarrow < 1233 > \rightarrow < 1234 >$ and $< 1111 > \rightarrow <$

$1133 >\rightarrow< 1134 >\rightarrow< 1234 >$. The decrease of the KL-divergence $D[< 1222 >:< 1233 >]$ is less than $D[< 1133 >:< 1134 >]$ since the diameter of the red dotted circle is larger than the blue one in Figure 5. This means that the effect of changing the same amount of KL-divergences in $D[< 1111 >:< 1222 >]$ and $D[< 1111 >:< 1133 >]$ produces larger effect on the sequence $< 1111 >\rightarrow< 1133 >\rightarrow< 1134 >\rightarrow< 1234 >$ than $< 1111 >\rightarrow< 1222 >\rightarrow< 1233 >\rightarrow< 1234 >$, if compared at the sequence level. The rectangle-biased complexity C_r does not reflect such characteristics since it does not distinguish between the hierarchical structure between the diameters of the green, red and blue dotted circles.

To incorporate such hierarchical effect in a complexity measure with geometric mean, we have the natural expansion of the rectangle-biased complexity C_r as the *cuboid-bias complexity* C_c, which is defined as follows:

$$C_c := \frac{1}{|Seq|}\sum_{i_s=1}^{|Seq|}\prod_{i=1}^{n-1} D[SD_i(i_s) : SD_{i+1}(i_s)], \tag{32}$$

where Seq represents the possible sequences of hierarchical system decompositions as follows:

$$Seq = \{SD_1(i_s) \rightarrow SD_2(i_s) \rightarrow \cdots SD_i(i_s) \cdots \rightarrow SD_n(i_s) | 1 \leq i_s \leq |Seq|\}. \tag{33}$$

The elements $SD_i(i_s)$ of Seq corresponds to the system decomposition, which is aligned according to the hierarchy with the following algorithmic procedure (based on [15]):

(1) Initialization: Set the initial sets of system decomposition of all sequences in Seq as the whole system $SD_1(i_s) :=< 111 \cdots 1 > (1 \leq i_s \leq |Seq|)$.
(2) Step $i \rightarrow i+1$: If the system decomposition is the total system decomposition ($SD_i(i_s) :=< 123 \cdots n >$), then stop. Otherwise, choose a non-decomposed subsystem $SS_i(i_s)$ of the system decomposition $SD_i(i_s)$, and further divide it into two independent subsystems $SS_i^1(i_s)$ and $SS_i^2(i_s)$ different for each i_s. $SD_{i+1}(i_s)$ is then defined as a system decomposition of total system that further separates independently subsystems $SS_i^1(i_s)$ and $SS_i^2(i_s)$, in addition to the previous decomposition $SD_i(i_s)$.
(3) Go to the next step $i+1 \rightarrow i+2$.

The value of $|Seq|$ corresponds to the number of different sequences generated by this algorithm. For example, $|Seq| = 3$ and $|Seq| = 18$ holds for $n = 3$ and $n = 4$, respectively. The general analytical form $|Seq|_n$ of $|Seq|$ with system size n is obtained as the following recurrence formula:

$$|Seq|_n = \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} {}_nC_i |Seq|_{n-i} |Seq|_i, \tag{34}$$

where $\lfloor \cdot \rfloor$ is a floor function and with formal definition of $|Seq|_1 := 1$.

The products of KL-divergences according to the hierarchical sequences of system decompositions in Equation (32) is related to the volume of $n-1$-dimensional cuboids in the circle diagram. An example in case of $n = 4$ is presented in Figure 5, where two cuboids with 3 orthogonal edges of the different decomposition sequences $< 1111 >\rightarrow< 1222 >\rightarrow< 1233 >\rightarrow< 1234 >$ and $< 1111 >\rightarrow< 1133 >\rightarrow< 1134 >\rightarrow< 1234 >$ are depicted, whose cuboid volumes are

$$\sqrt{D[< 1111 >:< 1222 >]D[< 1222 >:< 1233 >]D[< 1233 >:< 1234 >]}, \tag{35}$$

and

$$\sqrt{D[< 1111 >:< 1133 >]D[< 1133 >:< 1134 >]D[< 1134 >:< 1234 >]}, \tag{36}$$

respectively.

In the same way as C_r, we took in the definition of C_c the arithmetic average of cuboid volumes so that to renormalize the combinatorial increase of the decomposition paths ($|Seq|$) according to the system size n.

Note that on the other hand we did not renormalize the rectangle-bias complexity C_r and the cuboid-bias complexity C_c by taking the exact geometrical mean of each product of KL-divergences such as $\sqrt[n-1]{\prod_{i=1}^{n-1} D[SD_i(i_s) : SD_{i+1}(i_s)]}$. This is for further accessibility to theoretical analysis such as variational method (see "Further Consideration" section), and does not change qualitative behavior of C_r and C_c since the power root is a monotonically increasing function. This treatment can be interpreted as taking the $(n-1)$-th power of the geometric means for the hierarchical sequences of KL-divergences.

A more comprehensive example on the utility of the cuboid-bias complexity C_c with respect to the rectangle-biased one C_r is shown in Figure 6. We consider the 6 nodes networks ($n = 6$) with the same I and C_r values but different heterogeneity. The system in the top left figure has a circularly connected structure with medium intensity, while that of the top right figure has strongly connected 3 subsystems. These systems have qualitatively five different ways of system decomposition that are the basic generators of all hierarchical sequences $Seq = \{SD_1(i_s) \to \cdots \to SD_5(i_s) | 1 \leq i_s \leq |Seq|\}$ for these networks. The five basial system decompositions are shown with the number ①, ②, ②′, ③ and ④ in top figures.

The circle diagrams of these systems are depicted in the middle figures. To suppose the same constant value of C_r in both systems, the following condition is satisfied in the middle right figure: $D[< 111111 >: ②] < D[< 111111 >: ①\text{in Middle Left figure}] < D[< 111111 >: ①] < D[< 111111 >: ②\text{in Middle Left figure}] < D[< 111111 >: ③] < D[< 111111 >: ④]$. Furthermore, the total surface of right triangles sharing the circle diameter as hypotenuse in the middle left and the middle right figures are conditioned to be identical, therefore the rectangle-bias complexity C_r fails to distinguish.

On the other hand, under the same condition, the cuboid-bias complexity C_c distinguishes between these two systems and gives higher value to the left one. The volume of 5-dimensional cuboids of the decomposition sequence $< 111111 > \xrightarrow{①②②'③④} < 123456 >$ are schematically shown in the bottom figures, maintaining the quantitative difference between KL-divergences. Since the multi-information I is identical between the two systems, so is the values of KL-divergence $D[< 111111 >:< 123456 >]$, which is the sum of the KL-divergences along the sequence $< 111111 > \xrightarrow{①②②'③④} < 123456 >$ from the Pythagorean theorem. This means that the inequality between the cuboid volumes can be represented as the isoperimetric inequality of high-dimensional cuboid. As a consequence, the left system has quantitatively higher value of C_c than the right one. The cuboid-bias complexity C_c is also sensitive to such heterogeneity.

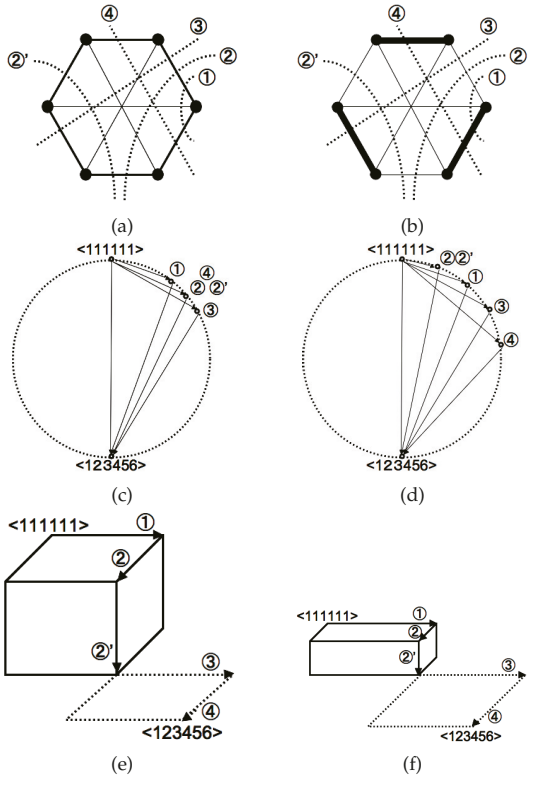


Figure 6. *Meaning of taking geometric mean over the sequence of system decomposition in cuboid-bias complexity* C_c. (**a**): Example of 6-node network with circularly connected structure with medium intensity. Edge width is proportional to edge information; (**b**): Example of 6-node network with strongly connected 3 subsystems. Edge width is proportional to edge information. The multi-information I of the two systems in Top figures are conditioned to be identical; The dotted lines schematically represent possible system decompositions. (**c,d**): Circle diagrams of each system decomposition in upper networks; The total surface of right triangles sharing the circle diameter as hypotenuse in (c) and (d) are conditioned to be identical, therefore the rectangle-bias complexity C_r fails to distinguish. (**e,f**): 5-dimensional cuboids of upper networks (Figure 6a,b) whose edges are the root of KL-divergences for the strain of system decomposition $< 111111 > \xrightarrow{①②②'③④} < 123456 >$. Only the first 3-dimensional part is shown with solid line, and the remaining 2-dimensional part is represented with dotted line. The volume of cuboid in (e) is larger than the one in (f), according to the isoperimetric inequality of high-dimensional cuboid. The total squared length of each side is identical between two cuboids, which represents multi-information $I = D[< 111111 > :< 123456 >]$.

8. Regularized Cuboid-Bias Complexity with Respect to Generalized Mutual Information

We further consider the geometrical composition of system decompositions in the circle diagram and insist the necessity of renormalizing the cuboid-bias complexity C_c with the multi-information I, which gives another measure of complexity namely "*regularized cuboid-bias complexity* C_c^R."

We consider the situation in actual data where the multi-information I varies. Figure 7 shows the $n = 3$ cases where the C_c fails to distinguish. Both the blue and red systems are supposed to have the same C_c value by adjusting the red system to have relatively smaller values of KL-divergences

$D[< 111 >:< 122 >]$ and $D[< 113 >:< 123 >]$ than the blue one. Such conditioning is possible since the KL-divergences are independent parameters with each other.

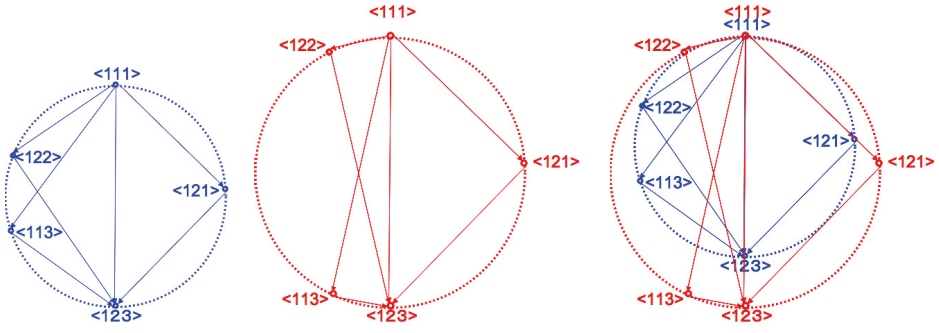


Figure 7. *Examples of the 3-node systems with identical cuboid-bias complexity C_c but different multi-information I on circle graph.* (**a**): System with smaller I but larger C_c^R; (**b**): System with larger I but smaller C_c^R; (**c**): Superposition of the above two systems. The regularized cuboid-bias complexity C_c^R distinguishes between the blue and red systems.

Although the C_c value is identical, the two systems have different geometrical composition of system decompositions in the circle diagram. The red system has relatively easier way of decomposition $< 111 > \rightarrow < 122 >$ if renormalized with the total system decomposition $< 111 > \rightarrow < 123 >$. This relative decompositionability with respect to the renormalization with the multi-information I can be clearly understood by superimposing the circle diagram of the two systems and comparing the angles between each and total decomposition paths (bottom figure). The red system has larger angle between the decomposition paths $< 111 > \rightarrow < 122 >$ and $< 111 > \rightarrow < 123 >$ than any others in the blue system, which represents the relative facility of the decomposition under renormalization with I. In this term, the paths $< 111 > \rightarrow < 121 >$ in the red and blue system do not change its relative facility, and the paths $< 111 > \rightarrow < 113 >$ are easier in the blue system.

To express the system decompositionability based on these geometrical compositions in a comprehensive manner, we define the *regularized cuboid-bias complexity* C_c^R as follows:

$$\begin{aligned} C_c^R &:= \frac{1}{|Seq|} \sum_{i_s=1}^{|Seq|} \prod_{i=1}^{n-1} \frac{D[SD_i(i_s) : SD_{i+1}(i_s)]}{D[< 11\cdots1 >:< 12\cdots n >]} \\ &:= \frac{C_c}{D[< 11\cdots1 >:< 12\cdots n >]^{n-1}} \\ &:= \frac{C_c}{I^{n-1}}. \end{aligned} \tag{37}$$

The red system then has quantitatively smaller C_c^R value than the blue system in Figure 7.

9. Modular Complexity with Respect to the Easiest System Decomposition Path

We have considered so far the system decompositionability with respect to the all possible decomposition sequences. This was also a way to avoid the local fluctuation of the network heterogeneity to be reflected in some specific decomposition paths. On the other hand, the easiest decomposition is particularly important when considering the modularity of the system. If there exists hierarchical structure of modularity in different scales with different coherence of the system, the KL-divergence and the sequence of the easiest decomposition gives much information.

Figure 8 schematically shows a typical example where there exist two levels of modularity. Such structure with different scales of statistical coherence appears as functional segregation in neural systems [17], and is expected to be observed widely in complex systems.

The hierarchical topology of the easiest decomposition path reflects these structures. For example, in the system of Figure 8, the decompositions between $< 1\ 1\ \cdots\ 1 >$ and $<$ 1 1 1 1 5 5 5 5 9 9 9 9 13 13 13 13 $>$ are easier than those inside of the 4-node subsystems. The values of KL-divergence also reflect the hierarchy, giving relatively low values for the decomposition between the 4-node subsystems, and high values inside of them. By examining the shortest decomposition path and associated KL-divergences in possible *Seq*, one can project the hierarchical structure of the modularity existing in the system.

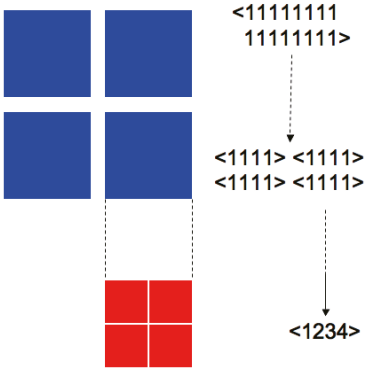


Figure 8. *Example of* 16*-node system* $< 11\cdots1 >$ *that has different levels of modularity.* The four 4-node subsystems $< 1111 >$ (blue blocks) are loosely connected and easy to be decomposed, while inside each component (red blocks) is tightly connected. The degree of connection represents statistical dependency or edge information between subsystems. Such hierarchical structure can be detected by observing the decomposition path of the modular complexity C_m.

For this reason, we define the *modular complexity* C_m as follows, which is the shortest path component of the cuboid-bias complexity C_c:

$$C_m := \prod_{i=1}^{n-1} D[SD_i(i_{min}) : SD_{i+1}(i_{min})], \tag{38}$$

where the index i_{min} of the sequence $SD_1(i_{min}) \to SD_2(i_{min}) \to \cdots \to SD_n(i_{min})$ is chosen as follows:

$$i_{min} \quad = \quad \{i_1\} \cap \{i_2\} \cap \cdots \cap \{i_{n-1}\}, \tag{39}$$

where

$$\begin{aligned} \{i_1\} &= \underset{i_s}{\text{argmin}}\{D[SD_1(i_s) : SD_2(i_s)] | 1 \le i_s \le |Seq|\}, \\ \{i_2\} &= \underset{i_1}{\text{argmin}}\{D[SD_2(i_1) : SD_3(i_1)] | i_1 \in \{i_1\}\}, \\ &\vdots \\ \{i_{n-1}\} &= \underset{i_{n-2}}{\text{argmin}}\{D[SD_{n-1}(i_{n-2}) : SD_n(i_{n-2})] | i_{n-1} \in \{i_{n-1}\}\}, \end{aligned} \tag{40}$$

which gives eventually

$$i_{min} = i_{n-1}. \tag{41}$$

This means that beginning from the undecomposed state $< 11\cdots1 >$, we continue to choose the shortest decomposition path in the next hierarchy of system decomposition. The minimization of the path length is guaranteed by the sequential minimization since the geometric mean of isometric path division is bounded below by its minimum component. i_{min} is unique if the system is completely heterogenous (*i.e.*, $D[SD_1(i_k) : SD_2(i_k)] \neq D[SD_1(i_l) : SD_2(i_l)]$, $1 \leq i_k < i_l \leq |Seq|$), otherwise plural decomposition paths that give the same C_m value are possible according to the homogeneity of the system. Besides its value, the modular complexity C_m should be utilized with the sequence information of the shortest decomposition path to evaluate the modularity structure of a system.

The cases where C_m are identical but C_c are different can be composed by varying the system decompositions other than in the shortest path $SD_1(i_{min}) \to SD_2(i_{min}) \to \cdots \to SD_n(i_{min})$ without modifying the index i_min. There exist also inverse examples with identical C_c and different C_m, due to the complementarity between C_m and C_c.

We finally define the *regularized modular complexity* C_m^R as follows, for the same reason as defining C_c^R from C_c;

$$\begin{aligned} C_m^R &:= \prod_{i=1}^{n-1} \frac{D[SD_i(i_{min}) : SD_{i+1}(i_{min})]}{D[< 11\cdots1 >:< 12\cdots n >]} \\ &:= \frac{C_m}{D[< 11\cdots1 >:< 12\cdots n >]^{n-1}} \\ &:= \frac{C_m}{I^{n-1}}. \end{aligned} \tag{42}$$

Proposition 4. *The cuboid-bias complexities C_c and C_c^R are bounded by the modular complexities C_m and C_m^R respectively:*

$$C_c \leq C_m, \tag{43}$$
$$C_c^R \leq C_m^R. \tag{44}$$

And they coincide at the maximum values under the given multi-information I:

$$\max\{C_m | I = const.\} = \max\{C_c | I = const.\}, \tag{45}$$
$$\max\{C_m^R\} = \max\{C_c^R\}. \tag{46}$$

These relations (43)–(46) are numerically shown in the "Numerical Comparison" section.

The superiority of the modular complexities is due to the hierarchical dependency of KL-divergence value in decomposition paths. In the shortest decomposition path defining modular complexities, the easier system decomposition relatively increase its value since they incorporate more number of edge cutting. Since we eventually cut all edges to obtain $< 12\cdots n >$ at the end of the decomposition sequence, collecting the edges with relatively weak edge information and cutting them together augment the value of the product of KL-divergences. The modular complexities are then the maximum value components among the possible decomposition paths calculated in cuboid-bias complexities:

$$C_m = \max\left\{ \prod_{i=1}^{n-1} D[SD_i(i_s) : SD_{i+1}(i_s)] \middle| 1 \leq i_s \leq |Seq| \right\}, \tag{47}$$

$$C_m^R = \max\left\{ \prod_{i=1}^{n-1} \frac{D[SD_i(i_s) : SD_{i+1}(i_s)]}{D[< 11\cdots1 >:< 12\cdots n >]^{n-1}} \middle| 1 \leq i_s \leq |Seq| \right\}. \tag{48}$$

The difference between the cuboid-bias complexities and the modular complexities is an index of the geometrical variation of decomposed systems in the circle graph, which reflects the fluctuation of the sequence-wise system decompositionability. If the variation of the system decompositionability for each system decomposition is large, accordingly the modular complexities tend to give higher values than the cuboid-bias complexities.

10. Numerical Comparison

We numerically investigate the complementarity between the proposed complexities, C_c, C_c^R, C_m, and C_m^R. Since the minimum node number giving non-trivial meaning to these measures is $n = 4$, the corresponding dimension of parameter space is $\sum_{k=1}^{n} {}_nC_k = 15$. The constant-complexity submanifolds are therefore difficult to visualize due to the high dimensionality. For simplicity, we focus on the 2-dimensional subspace of this parameter space whose first axis ranging from random to maximum dependencies of the system, and the second one representing the system decompositionability of $< 1133 >$.

For this purpose, we introduce the following parameters α and β ($0 \leq \alpha, \beta \leq 1$) in the **ɲ**-coordinates of the discrete distribution with 4-dimensional binary stochastic variable:

$$\begin{aligned}
\eta_1 &= \eta_0, \\
\eta_2 &= \eta_0, \\
\eta_3 &= \eta_0, \\
\eta_4 &= \eta_0, \\
\eta_{1,2} &= \eta_1\eta_2 + \alpha(\eta_0 - \epsilon - \eta_1\eta_2), \\
\eta_{3,4} &= \eta_3\eta_4 + \alpha(\eta_0 - \epsilon - \eta_3\eta_4), \\
\eta_{1,3} &= \eta_1\eta_3 + \alpha\beta(\eta_0 - \epsilon - \eta_1\eta_3), \\
\eta_{1,4} &= \eta_1\eta_4 + \alpha\beta(\eta_0 - \epsilon - \eta_1\eta_4), \\
\eta_{2,3} &= \eta_2\eta_3 + \alpha\beta(\eta_0 - \epsilon - \eta_2\eta_3), \\
\eta_{2,4} &= \eta_2\eta_4 + \alpha\beta(\eta_0 - \epsilon - \eta_2\eta_4), \\
\eta_{1,2,3} &= \eta_{1,2}\eta_3 + \alpha\beta(\eta_0 - 2\epsilon - \eta_{1,2}\eta_3), \\
\eta_{1,2,4} &= \eta_{1,2}\eta_4 + \alpha\beta(\eta_0 - 2\epsilon - \eta_{1,2}\eta_4), \\
\eta_{1,3,4} &= \eta_1\eta_{3,4} + \alpha\beta(\eta_0 - 2\epsilon - \eta_1\eta_{3,4}), \\
\eta_{2,3,4} &= \eta_2\eta_{3,4} + \alpha\beta(\eta_0 - 2\epsilon - \eta_2\eta_{3,4}), \\
\eta_{1,2,3,4} &= \eta_{1,2}\eta_{3,4} + \alpha\beta(\eta_0 - 3\epsilon - \eta_{1,2}\eta_{3,4}).
\end{aligned} \tag{49}$$

Where α represents the degree of statistical association from random ($\alpha = 0$) to maximum ($\alpha = 1$), and β control the system decompositionability of $< 1133 >$. If $\beta = 1$, the system has the maximum KL-divergence $D[< 1111 > :< 1133 >]$ under the constraint of α parameter, and $\beta = 0$ gives $D[< 1111 > :< 1133 >] = 0$.

ϵ is the minimum value of the joint distribution of 4-dimensional variable, which is defined to be more than 0 to avoid singularity in the dual-flat coordinates of statistical manifold. $\epsilon = 1.0 \times 10^{-10}$ and $\eta_0 = 0.5$ was chosen for the calculation.

The system with maximum statistical association under given η_0 corresponds to the $\alpha = \beta = 1$ condition in given parameters, whose J-coordinates become as follows:

$$\begin{aligned}
\eta_1 &= \eta_0, \\
&\vdots \\
\eta_4 &= \eta_0, \\
\eta_{1,2} &= \eta_0 - \epsilon, \\
&\vdots \\
\eta_{3,4} &= \eta_0 - \epsilon, \\
\eta_{1,2,3} &= \eta_0 - 2\epsilon, \\
&\vdots \\
\eta_{2,3,4} &= \eta_0 - 2\epsilon, \\
\eta_{1,2,3,4} &= \eta_0 - 3\epsilon, .
\end{aligned} \tag{50}$$

On the other hand, the totally decomposed system corresponds to the $\alpha = 0$ condition, and the J-coordinates are:

$$\begin{aligned}
\eta_1 &= \eta_0, \\
&\vdots \\
\eta_4 &= \eta_0, \\
\eta_{1,2} &= \eta_0\eta_0, \\
&\vdots \\
\eta_{3,4} &= \eta_0\eta_0, \\
\eta_{1,2,3} &= \eta_0\eta_0\eta_0, \\
&\vdots \\
\eta_{2,3,4} &= \eta_0\eta_0\eta_0, \\
\eta_{1,2,3,4} &= \eta_0\eta_0\eta_0\eta_0.
\end{aligned} \tag{51}$$

Note that the completely deterministic case $\eta_0 = 1.0$ and $\alpha = \beta = 1$ gives $I = 0$.

The intuitive meaning of these parameters α and β are also schematically depicted in Figure 9 bottom right.

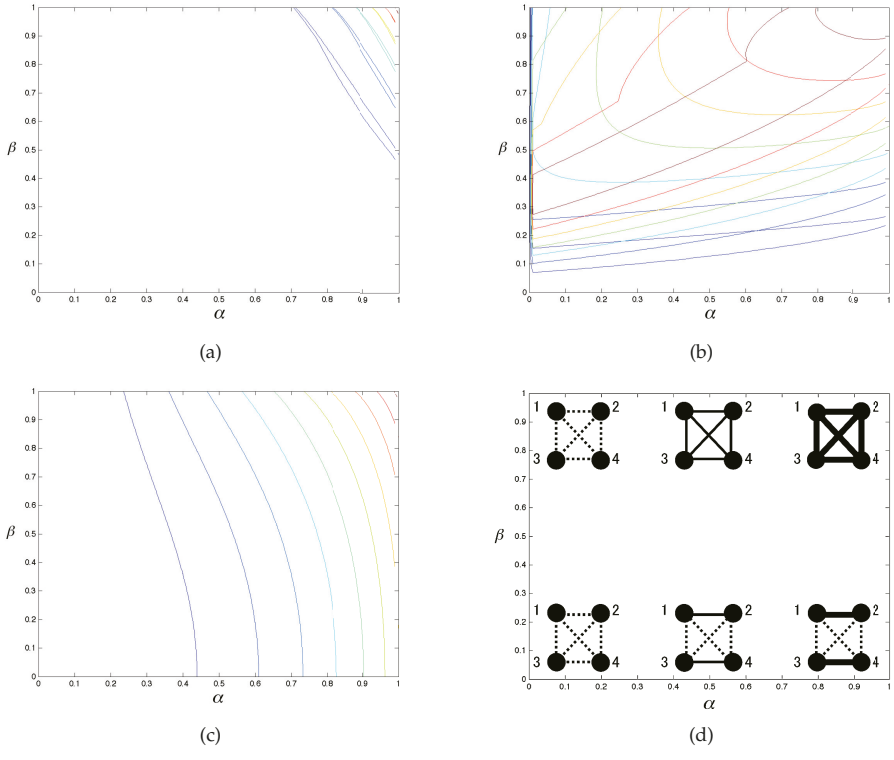


Figure 9. *Contour plot of the complexity landscape of I, C_c, C_m, C_c^R, and C_m^R on α-β plane.* (**a**): Contour plot superposition of C_c and C_m. (**b**): Contour plot superposition of C_c^R and C_m^R. (**c**): Contour plot of I. The color of contour plots corresponds to the color gradient of 3D plots in Figure 10; (**d**): Schematic representation of the system in different regions of α-β plane. Edge width represents the degree of edge information, and independence is depicted with dotted line.

Figure 10 shows the landscape of the proposed complexities on the α-β plane. Their contour plots are depicted in Figure 9. The proposed complexities each differs from others in almost everywhere points on α-β plane except at the intersection lines. Therefore, these measures serve as the independent features of the system, each has its specific meaning with respect to the system decompositionability. The α-β plane shows a section of the actual structure of the complementarity expressed in Figure 3 between the proposed complexity measures.

The relations between the cuboid-bias complexities and modular complexities in Equations (43)–(46) are also numerically confirmed. The modular complexities are superior than the corresponding cuboid-bias complexities, and coincide at the parameter $\alpha = \beta = 1$ giving maximum values and dependencies in this parameterization.

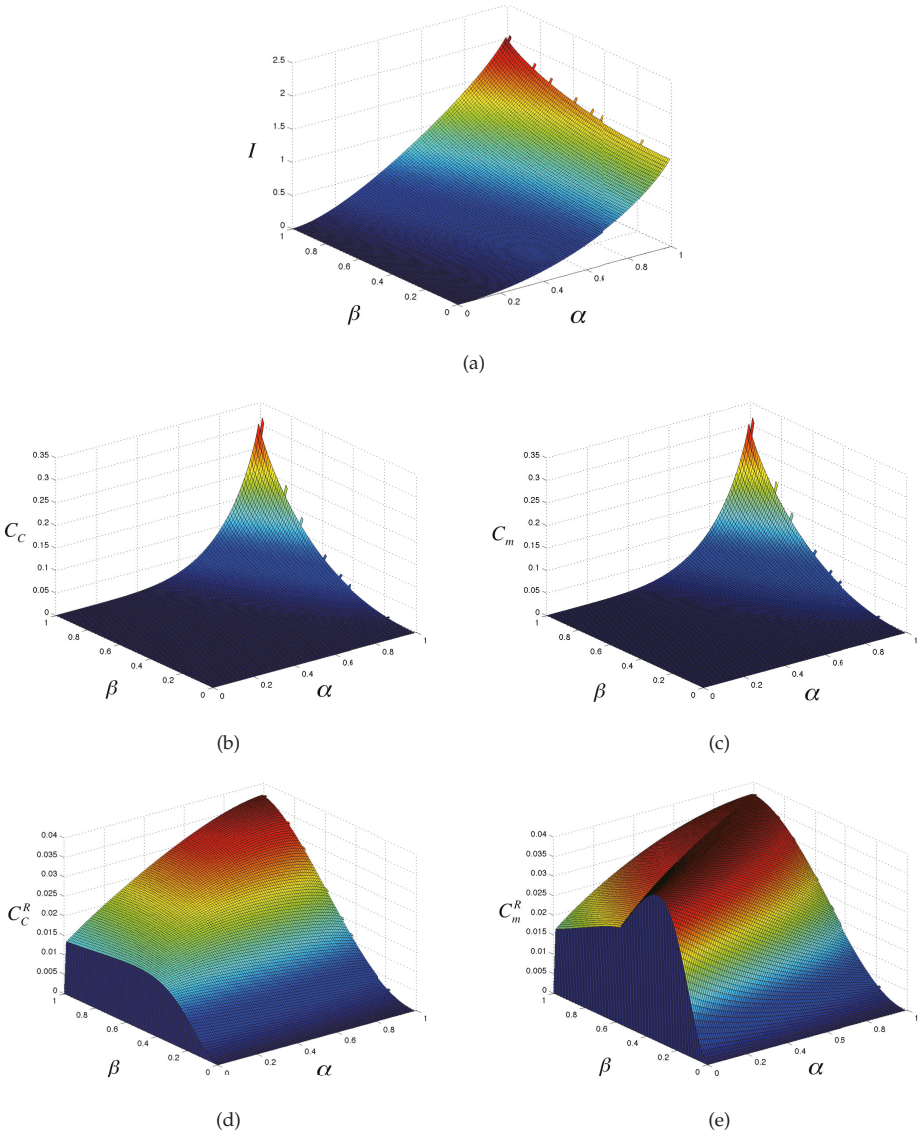


Figure 10. *Landscape of complexities I, C_c, C_m, C_c^R, and C_m^R on α-β plane.* (**a**): Multi-information I; (**b**): Cuboid-bias complexity C_c. (**c**): Modular complexity C_m;(**d**): Regularized cuboid-bias complexity C_c^R; (**e**): Regularized modular complexity C_m^R. All complexity measures show the complementarity intersecting with each other, satisfying the boundary conditions vanishing at $\alpha = 0$ and $\beta = 0$ except the multi-information I. Note that regularized complexities C_c^R and C_m^R show singularity of convergence at $\alpha \to 0$ due to the regularization of infinitesimal value.

In general case without the parameterization with α, β and η_0, the boundary conditions of C_c, C_c^R, C_m and C_m^R include that of the multi-information I, which vanish at the completely random or ordered state. This is common to other complexity measures such as the LMC complexity, and fit to the basic

intuition on the concept of complexity situated equivalently far from the completely predictable and disordered states [21,22].

The proposed complexities further incorporate boundary conditions that vanish with the existence of a completely independent subsystem of any size. This means that the C_c, C_c^R, C_m and C_m^R of a system become 0 if we add another independent variable. This property does not reflect the intuition of complexity defined by the arithmetic average of statistical measures. The proposed complexity can better find its meaning in comparison to other complexity measures such as the multi-information I, and by interactively changing the system scale to avoid trivial results with small independent subsystem. For example, the proposed complexities could be utilized as the information criteria for the model selection problems, especially with an approximative modular structure based on the statistical independency of data between subsystems. We insist that the complementarity principle between plural complexity measures of different foundation is the key to understand the complexity in a comprehensive manner.

To characterize the property of C_c, C_c^R, C_m and C_m^R in relation to the diverse composition of each system decomposition, it is useful to consider the geometry of their contour structure, as compared in Figure 9. The contour can be formalized as $C_c, C_c^R, C_m, C_m^R = const.$ for each complexity measure, and $D[< 11 \cdots 1 >: SD_i(i_s)] = const.$ $(1 \leq i \leq n-1, 1 \leq i_s \leq |Seq|)$ for each system decomposition. For that purpose, analysis with algebraic geometry can be considered as a prominent tool. Algebraic geometry investigates the geometrical property of polynomial equations [23]. The complexities C_c, C_c^R, C_m and C_m^R can be interpreted as polynomial functions by taking each system decomposition as novel coordinates, therefore directly accessible to algebraic geometry. However, if we want to investigate the contour of the complexities on the $\mathbf{p}$ parameter space, logarithmic function appears as the definition of KL-divergence, which is a transcendental function and outreach the analytical requirement of algebraic geometry. To introduce compatibility between the $\mathbf{p}$ parameter space of information geometry and algebraic geometry, it suffices to describe the model by replacing the logarithmic functions as another n variables such as $\mathbf{q} = \log \mathbf{p}$, and reconsider the intersection between the result from algebraic geometry on the coordinates $(\mathbf{p}, \mathbf{q})$ and $\mathbf{q} = \log \mathbf{p}$ condition. The contour of C_c, C_c^R, C_m and C_m^R is also important to seek for the utility of these measures as a potential to interpret the dynamics of statistical association as geodesics.

11. Further Consideration

11.1. Pythagorean Relations in System Decomposition and Edge Cutting

We further look back at the system decomposition and edge cutting in terms of the Pythagorean relation between KL-divergences, which is based on the orthogonality between $\grave{}$ and $\jmath$ coordinates.

In system decomposition, the distribution of decomposed system is analytically obtained from the product of subsystems' η coordinates, which is equivalent to set all θ^{dec} parameters as 0 in mixture coordinate ξ^{dec}. From the consistency of θ^{dec} parameters in ξ^{dec} being 0 in all system decompositions, we have the Pythagorean relation according to the inclusion relation of system decomposition. For example, the following holds:

$$\begin{aligned} D[< 1111 >:< 1234 >] &= D[< 1111 >:< 1222 >] \\ &+ D[< 1222 >:< 1233 >] \\ &+ D[< 1233 >:< 1234 >]. \end{aligned} \tag{52}$$

The proof is in the same way as k-cut coordinates isolating k-tuple statistical association between variables [14].

On the other hand, the edge cutting previously defined using the product of remaining maximum cliques' η coordinates does not coincides with the $\theta^{ec} = 0$ condition in mixture coordinates ξ^{ec}. We have defined the η^{ec} values of edge cutting based only on the orthogonal relation between η and θ

coordinates, by generalizing the rule of system decomposition in η^{ec} coordinates, and did not consider the Pythagorean relation between different edge cuttings.

It is then possible to define another way of edge cutting using $\theta^{ec} = 0$ condition in ξ^{ec}. Indeed, in k-cut mixture coordinates, $\theta^{k+} = 0$ condition is derived from the independent condition of the variables in all orders, and k-tuple statistical association is measured by reestablishing the η parameters for the statistical association up to $k-1$-tuple order. In the same way, we can set $\theta^{dec} = 0$ condition for ξ^{dec} of a system decomposition, and reestablish edges with respect to the η parameters, except the one in focus for edge cutting.

As a simple example, consider the system decomposition $< 1222 >$ and edge cutting $1-2$ in 4-node graph. We have the mixture coordinate ξ^{dec} for the system decomposition as follows:

$$\begin{aligned} \xi^{dec}_{1,2} &= \theta^{dec}_{1,2} = 0, \\ \xi^{dec}_{1,3} &= \theta^{dec}_{1,3} = 0, \\ \xi^{dec}_{1,4} &= \theta^{dec}_{1,4} = 0, \\ \xi^{dec}_{1,2,3} &= \theta^{dec}_{1,2,3} = 0, \\ \xi^{dec}_{1,3,4} &= \theta^{dec}_{1,3,4} = 0, \\ \xi^{dec}_{1,2,3,4} &= \theta^{dec}_{1,2,3,4} = 0, \end{aligned} \tag{53}$$

where all the rest of ξ^{dec} coordinates is equivalent to that of η coordinates.

We then consider the new way of edge cutting $1-2$ by recovering the statistical association in edges $1-3$ and $1-4$ from system decomposition $< 1222 >$, orthogonally to that of edge $1-2$. The new mixture coordinate ξ^{EC} changes to the following:

$$\begin{aligned} \xi^{EC}_{1,2} = \theta^{EC}_{1,2} &= 0, \\ \xi^{EC}_{1,3} &= \eta_{1,3}, \\ \xi^{EC}_{1,4} &= \eta_{1,4}, \\ \xi^{EC}_{1,2,3} = \theta^{EC}_{1,2,3} &= 0, \\ \xi^{EC}_{1,3,4} &= \eta_{1,3,4}, \\ \xi^{EC}_{1,2,3,4} = \theta^{EC}_{1,2,3,4} &= 0, \end{aligned} \tag{54}$$

and the rest is equivalent to that of η coordinates.

This new ξ^{EC} is also compatible with k-cut coordinates formalization for its simple $\theta^{EC} = 0$ conditions. To obtain ξ^{EC} for arbitrary edge cutting $i-j$, one should take θ^{EC} containing i and j in its subscript, set them to 0, and combine with η coordinates for the rest of the subscript. For plural edge cuttings $i-j, \cdots, k-l$ $(1 \leq i, j, k, l \leq n)$, it suffices to take θ^{EC} containing i and j, ... , k and l in its subscript respectively, then set them to 0.

We finally obtain the Pythagorean relation between edge cuttings. Denoting the general edge cutting(s) coordinates as $\xi^{i-j,\cdots,k-l}$, the following holds for the example of system decomposition $< 1222 >$:

$$\begin{aligned} D[< 1111 >:< 1222 >] &= D[< 1111 >: p(\xi^{1-2})] \\ &+ D[p(\xi^{1-2}) : p(\xi^{1-2,1-3})] \\ &+ D[p(\xi^{1-2,1-3}) : p(\xi^{1-2,1-3,1-4})]. \end{aligned} \tag{55}$$

Despite the consistency with the dual structure between θ and η, we do not generally have analytical solution to determine η^{EC} values from $\theta^{EC} = 0$ conditions. We should call for some numerical algorithm to solve $\theta^{EC} = 0$ conditions with respect to η^{EC} values, which are in general high-degree simultaneous polynomials. Furthermore, numerical convergence of the solution has to be

very strict, since tiny deviation from the conditions can become non-negligible by passing fractional function and logarithmic function of θ coordinates.

On the other hand, the previously defined edge cutting with ξ^{ec} using the product between subgraphs' η coordinates is analytically simple and does not need to consider the other edges' recovery from system decomposition or independence hypothesis. We then chose the previous way of edge cutting for both calculability and clarity of the concept.

There have been many attempts to approximate complex network by low-dimensional system with the use of statistical physics and network theory. As a contemporary example, moment-closure approximation provides a various way to abstract essential dynamics e.g., in discrete adaptive network [24]. Although the approximation takes several theoretical assumptions such as random graph approximation, it is difficult to quantitatively reproduce the dynamics even in some simplest model. This is partly due to homogeneous treatment of statistics such as truncation into pair-wise order. The edge cutting can offer a complementary view on the evaluation of moment-closure approximations. Using orthogonal decomposition between edge information, one can evaluate which part of network link and which order of statistics contain essential information, which does not necessary conform to top-down theoretical treatment.

11.2. Complexity of the Systems with Continuous Phase Space

We have developed the concept of system decompositionability based on discrete binary variables. One can also apply the same principle to continuous variable.

For an ergodic map $G : X \to X$ in continuous space X, KS entropy $h(\mu, G)$ is defined as the maximum of entropy rate with respect to all possible system decomposition A, when the invariant measure μ exists:

$$h(\mu, G) = \sup_A h(\mu, G, A). \tag{56}$$

where A is the disjoint decomposition of X that consists of non-trivial sets a_i, whose total number is $n(A)$, defined as

$$X = \bigcup_{i=1}^{n(A)} a_i, \tag{57}$$

$$a_i \cap a_j = \phi,\ i \neq j,\ 1 \leq i, j \leq n(A), \tag{58}$$

meaning the natural expansion of system decomposition into continuous space.

The entropy rate $h(\mu, G, A)$ in Equation (56) is defined as

$$h(\mu, G, A) = \lim_{n\to\infty} \frac{1}{n} H(\mu, A \vee G^{-1}(A) \vee \cdots \vee G^{-n+1}(A)), \tag{59}$$

according to the entropy $H(\mu, A)$ based on the decomposition $A = \{a_i\}$

$$H(\mu, A) = -\sum_{i=1}^{n(A)} \mu(a_i) \ln \mu(a_i), \tag{60}$$

and the product $C = A \vee B$ as

$$\begin{aligned} C &= A \vee B \\ &= \{c_i = a_j \cap b_k | 1 \leq j \leq n(A), 1 \leq k \leq n(B)\}. \end{aligned} \tag{61}$$

In a more general case, topological entropy $h_T(G)$ is defined simply with the number of decomposed subsystem elements by preimages as follows, without requiring ergodicity, therefore neither the existence of invariant measure μ:

$$h_T(G) = \sup_A \lim_{n\to\infty} \frac{1}{n} \ln n(A \vee G^{-1}(A) \vee \cdots \vee G^{-n+1}(A)). \tag{62}$$

Topological entropy takes the maximum value of the possible preimage divisions, in order to measure the complexity in terms of the mixing degree of the orbits. For example, if the KS entropy is positive as $h(\mu, G) > 0$, the dynamics of G on an invariant set of invariant measure μ is chaotic for almost everywhere initial conditions. As for the positive topological entropy $h_T(G) > 0$, the dynamics of G contain chaotic orbits, but not necessary as attractive chaotic invariant set, since $h_T(G) \geq h(\mu, G)$ and the KS entropy can be negative.

Although these definitions are useful to characterize the existence of chaotic dynamics, the system decompositionability is another property representing different aspect of the system complexity. It is rather the matter of the existence of independent dynamics components, or the degree of orbit localization between arbitrary system decompositions. We propose the following "*geometric topological entropy*" $h_g(G)$ applying the same principle of taking geometric product between all hierarchical structure of the system decomposition A.

$$h_g(G) := \prod_{\sigma(A)>0} \lim_{n\to\infty} \frac{1}{n} \ln n(A \vee G^{-1}(A) \vee \cdots \vee G^{-n+1}(A)), \tag{63}$$

where $\sigma(A) > 0$ means to take all components of A having positive Lebesgue measure on X.

This gives 0 if the preimage of certain $a_i \in A$ is a_i itself, meaning there exist a subsystem a_i whose range is invariant under G, closed by itself. The system X can be completely divided into a_i and the rest. This corresponds to the existence of an independent subsystem in cuboid-bias and modular complexities. In case such independent components do not exist, it still reflects the degree of orbit localization for all possible system decompositions in multiplicative manner. The condition $\sigma(A) > 0$ is to avoid trivial case such as the existence of unstable limit cycle, whose Lebesgue measure is 0.

Typical example giving $h_g(G) = 0$ is the function having independent ergodic components, such as the Chirikov-Taylor map with appropriate parameter [25].

12. Conclusions and Discussion

We have theoretically developed a framework to measure the degree of statistical association existing between subsystems as well as the ones represented by each edge of the graph representation. We then reconsidered the problem of how to define complexity measures in terms of the construction of non-linear feature space. We defined new type of complexity based on the geometrical product of KL-divergence representing the degree of system decompositionability. Different complexity measures as well as newly proposed ones are compared on a complementarity basis on statistical manifold.

Application of presented theory can encompass a large field of complex systems and data science, such as social network, genetic expression network, neural activities, ecological database, and any kind of complex networks with binary co-occurrence matrix data e.g., [26–29], databases: [30–34]. Continuous variables are also accessible by appropriate discretization of information source with e.g., entropy maximization principle.

In contrast to arithmetic mean of information over the whole system, geometric mean has not been investigated sufficiently in the analysis of complex network. However in different fields, theoretical ecology has already pointed out the importance of geometric mean when considering the long-term fitness of a species population in a randomly varying environment [35,36]. Long-term fitness refers to the ecological complexity of its survival strategy under large stochastic fluctuation. Here, we can find useful analogy between the growth rate of a population in ecology and the spatio-temporal

propagation rate of information between subsystems in general. If we take an arbitrary subsystem and consider the amount of information it can exchange with all other subsystems, the proposed complexity measures with geometric mean reflect the minimum amount with amongst all possible other subsystems, which can not be distinguished with arithmetic mean. The propagation rate of a population in ecology and the information transmission in complex network hold mathematically analogous structure. In population ecology, the variance of growth rate is crucial to evaluate the long-term survival of the population. Even if the arithmetic mean of growth rate is high, large variance will lead to low geometric mean even with a small amount of exceptionally small fitness situation, which ecologically means extinction of an entire species. In stochastic network, the variance of system decompositionability is essential to evaluate the amount of information shared between subsystems, or information persistence in the entire network. Even the multi-information I is high, large heterogeneity of edge information can lead to informational isolation of certain subsystem, which means extinction of its information. If such subsystem is situated on the transmission pathway, information cannot propagate across these nodes. Therefore, the proposed complexity measures C_C, C_C^R, C_m and C_m^R generally reflect the minimum amount of information propagation rate spread entirely on the system without exception of isolated division.

Some recent studies on adaptive network focus on the evolution of network topology in response to node activity, such as game-theoretic evolution of strategies [37], opinion dynamics on an evolving network [38], epidemic spreading on an adaptive network [39], *etc*. Analysis of coevolution network between variables and interactions can capture important dynamical feature of complex systems. In contrast to topological network analysis, the newly proposed complexity measures can complement its statistical dynamics analysis. In addition to the topological change of network model, (e.g., linking dynamics of game theory, opinion community network structure, contact network of epidemics transmission), one can evaluate the emerged statistical association between the variables that does not necessary coincide with the network topology. Interesting feature of non-linear dynamics is the unexpected correlation between distant variables, which is quantified as Tsallis entropy [40]. The complementary relation between concrete interaction and resulting statistical association can provide a twofold methodology to characterize the coevolutionary dynamics of adaptive network. Such strategy can promote integrated science from laboratory experiments to open-field *in natura* situation, where actual multi-scale problematics remain to be solved [41].

Arithmetic and geometric means can be integrated in a mutual formula called generalized mean [42]. Therefore, the proposed complexity measures with geometric mean of KL-divergence is an expansion of preexisting complexity measures with mixture coordinates. Table 1 summarizes the generalization of complexity measure in this article. Based on the k-cut coordinates $\mathbf{1}$, the weighted sum of KL-divergence representing k-tuple order of statistical association derived complexity measures with (weighted) arithmetic mean such as multi-information I and TSE complexity. On the other hand, we showed that subsystem-wise correlation can also be isolated with the use of mixture coordinates, namely $< \cdots >$-cut coordinates $\boldsymbol{\lambda}$. To quantify the heterogeneity of system decompositionability, we generally took a weighted geometric mean of KL-divergence in C_C, C_C^R, C_m and C_m^R. Here, the shortest path selection of C_m and C_m^R, and regularization of C_C^R and C_m^R with respect to multi-information I can be interpreted as the weight function of geometric mean. This perspective brings a definition of a generalized class of complexity measures based on the mixture coordinates and generalized mean of KL-divergence. Information discrepancy can also be generalized from KL-divergence to Bregman divergence, providing access to the concept of multiple centroids in large stochastic data analysis such as image processing [43]. The blank columns of the Table 1 imply the possibility of other complexity measures in this class. For example, the weighted geometric mean of KL-divergence defined between k-cut coordinates is expected to yield complexity measures that are sensitive to the heterogeneity of correlation orders. The weighted arithmetic mean of KL-divergence defined between $< \cdots >$-cut coordinates should be sensitive to the mean decompositionability of arbitrary subsystem. Since these measures take analytically different form on mixture coordinates and/or mean

functions, their derivatives do not coincide, which give independent information of the system on the complementary basis on statistical manifold, as long as the number of complexity measures are inferior to the freedom degree of the system.

Table 1. Classification of complexity measures with KL-divergence on mixture coordinates.

		Generalized Mean of KL-Divergence	
		Arithmetic Mean	**Geometric Mean**
Mixture Coordinates	k-cut $\mathbf{1}$	TSE complexity, I	
	$< \cdots >$-cut $\mathbf{\lambda}$		C_C, C_C^R, C_m, C_m^R

Acknowledgments: This study was partially supported by CNRS, the long term study abroad support program of the university of Tokyo, and the French government (Promotion Simone de Beauvoir).

Conflicts of Interest: Conflicts of Interest

The author declares no conflict of interest.

References

1. Boccalettia, S.; Latorab, V.; Morenod, Y.; Chavezf, M.; Hwang, D.U. Complex Networks: Structure and Dynamics. *Phys. Rep.* **2006**, *424*, 175–308.
2. Strogatz, S.H. Exploring Complex Networks. *Nature* **2001**, *410*, 268–276.
3. Wasserman, S.; Faust, K. *Social Network Analysis*; Cambridge University Press: Cambridge, UK, 1994.
4. Funabashi, M.; Cointet, J.P.; Chavalarias, D. Complex Network. In *Studies in Computational Intelligence*; Springer: Berlin/Heidelberg, Germany, 2009; Volume 207, pp. 161–172.
5. Badii, R.; Politi, A. *Complexity: Hierarchical Structures and Scaling in Physics*; Cambridge University Press: Cambridge, UK, 2008.
6. Lempel, A.; Ziv, J. On the Complexity of Finite Sequences. *IEEE Trans. Inf. Theory* **1976**, *22*, 75–81.
7. Li, M.; Vitanyi, P. Texts in Computer Science. In *An Introduction to Kolmogorov Complexity and Its Applications*, 2nd ed.; Springer: Berlin/Heidelberg, Germany, 1997.
8. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*; Wiley: New York, NY, USA, 2006.
9. Bennett, C. On the Nature and Origin of Complexity in Discrete, Homogeneous, Locally-Interacting Systems. *Found. Phys.* **1986**, *16*, 585–592.
10. Grassberger, P. Toward a Quantitative Theory of Self-Generated Complexity. *Int. J. Theor. Phys.* **1986**, *25*, 907–938.
11. Crutchfield, J.P.; Feldman, D.P. Regularities Unseen, Randomness Observed: The Entropy Convergence Hierarchy. *Chaos* **2003**, *15*, 25–54.
12. Crutchfield, J.P. Inferring Statistical Complexity. *Phys. Rev. Lett.* **1989**, *63*, 105–108.
13. Prichard, D.; Theiler, J. Generalized Redundancies for Time Series Analysis. *Physica D* **1995**, *84*, 476–493.
14. Amari, S. Information Geometry on Hierarchy of Probability Distributions. *IEEE Trans. Inf. Theory* **2001**, *47*, 1701–1711.
15. Ay, N.; Olbrich, E.; Bertschinger, N.; Jost, J. *A Unifying Framework for Complexity Measures of Finite Systems*; Report 06-08-028; Santa Fe Institute: Santa Fe, NM, USA, 2006.
16. MacKay, R.S. Nonlinearity in Complexity Science. *Nonlinearity* **2008**, *21*, T273–T281.
17. Tononi, G.; Sporns, O.; Edelman, M. A Measure for Brain Complexity: Relating Functional Segregation and Integration in the Nervous System. *Proc. Natl. Acad. Sci. USA* **1994**, *91*, 5033.
18. Feldman, D.P.; Crutchfield, J.P. Measures of statistical complexity: Why? *Phys. Lett. A* **1998**, *238*, 244–252.
19. Nakahara, H.; Amari, S. Information-Geometric Measure for Neural Spikes. *Neural Comput.* **2002**, *14*, 2269–2316.
20. Olbrich, E.; Bertschinger, N.; Ay, N.; Jost, J. How Should Complexity Scale with System Size? *Eur. Phys. J. B* **2008**, *63*, 407–415.
21. Feldman, D.P.; Crutchfield, J.P. Measures of Statistical Complexity: Why? *Phys. Lett. A* **1998**, *238*, 244–252.
22. Lopez-Ruiz, R.; Mancini, H.; Calbet, X. A Statistical Measure of Complexity. *Phys. Lett. A* **1995**, *209*, 321–326.

23. Hodge, W.; Pedoe, D. *Methods of Algebraic Geometry*; Cambridge Mathematical Library, Cambridge University Press: Cambridge, UK, 1994; Volume 1–3.
24. Demirel, G.; Vazquez, F.; Bohme, G.; Gross, T. Moment-closure Approximations for Discrete Adaptive Networks. *Physica D* **2014**, *267*, 68–80.
25. Fraser, G., Ed. *The New Physics for the Twenty-First Century*; Cambridge University Press: Cambridge, UK, 2006; p. 335.
26. Scott, J. *Social Network Analysis: A Handbook*; SAGE Publications Ltd.: London, UK, 2000.
27. Geier, F.; Timmer, J.; Fleck, C. Reconstructing Gene-Regulatory Networks from Time Series, Knock-Out Data, and Prior Knowledge. *BMC Syst. Biol.* **2007**, *1*, doi:10.1186/1752-0509-1-11.
28. Brown, E.N.; Kass, R.E.; Mitra, P.P. Multiple Neural Spike Train Data Analysis: State-of-the-Art and Future Challenges. *Nat. Neurosci.* **2004**, *7*, 456–461.
29. Yee, T.W. The Analysis of Binary Data in Quantitative Plant Ecology. Ph.D. Thesis, The University of Auckland, New Zealand, 1993.
30. Stanford Large Network Dataset Collection. Available online: http://snap.stanford.edu/data/ (accessed on 19 July 2014).
31. BioGRID. Available online: http://thebiogrid.org/ (accessed on 19 July 2014).
32. Neuroscience Information Framework. Available online: http://www.neuinfo.org/ (accessed on 19 July 2014).
33. Global Biodiversity Information Facility. Available online: http://www.gbif.org/ (accessed on 19 July 2014).
34. UCI Network Data Repository. Available online: http://networkdata.ics.uci.edu/index.php (accessed on 19 July 2014).
35. Lewontin, R.C.; Cohen, D. On Population Growth in a Randomly Varying Environment. *Proc. Natl. Acad. Sci. USA* **1969**, *62*, 1056–1060.
36. Yoshimura, J.; Clark, C.W. Individual Adaptations in Stochastic Environments. *Evol. Ecol.* **1969**, *5*, 173–192.
37. Wu, B.; Zhou, D.; Wang, L. Evolutionary Dynamics on Stochastic Evolving Networks for Multiple-Strategy Games. *Phys. Rev. E* **2011**, *84*, 046111.
38. Fu, F.; Wang, L. Coevolutionary Dynamics of Opinions and Networks: From Diversity to Uniformity. *Phys. Rev. E* **2008**, *78*, 016104.
39. Gross, T.; D'Lima, C.J.D.; Blasius, B. Epidemic Dynamics on an Adaptive Network. *Phys. Rev. Lett.* **2006**, *96*, 208701.
40. Tsallis, C. Possible Generalization of Boltzmann-Gibbs Statistics. *J. Stat. Phys.* **1988**, *52*, 479–487.
41. Quintana-Murci, L.; Alcais, A.; Abel, L.; Casanova, J.L. Immunology in natura: Clinical, Epidemiological and Evolutionary Genetics of Infectious Diseases. *Nat. Immunol.* **2007**, *8*, 1165–1171.
42. Hardy, G.; Littlewood, J.; Polya, G. *Inequalities*; Cambridge University Press: Cambridge, UK, 1967; Chapter 3.
43. Nielsen, F.; Nock, R. Sided and symmetrized Bregman centroids. *IEEE Trans. Inf. Theory* **2009**, *55*, 2882–2904.

MDPI
St. Alban-Anlage 66
4052 Basel
Switzerland
Tel. +41 61 683 77 34
Fax +41 61 302 89 18
www.mdpi.com

Entropy Editorial Office
E-mail: entropy@mdpi.com
www.mdpi.com/journal/entropy

MDPI
St. Alban-Anlage 66
4052 Basel
Switzerland

Tel: +41 61 683 77 34
Fax: +41 61 302 89 18

www.mdpi.com

ISBN 978-3-03897-633-2