An Information-Theoretic Primer on Complexity, Self-Organization, and Emergence

Complex Systems Science aims to understand concepts like complexity, self-organization, emergence and adaptation, among others. The inherent fuzziness in complex systems definitions is complicated by the unclear relation among these central processes: does self-organisation emerge or does it set the preconditions for emergence? Does complexity arise by adaptation or is complexity necessary for adaptation to arise? The inevitable consequence of the current impasse is miscommunication among scientists within and across disciplines. We propose a set of concepts, together with their possible information-theoretic interpretations, which can be used to facilitate the Complex Systems Science discourse. Our hope is that the suggested information-theoretic baseline may promote consistent communications among practitioners, and provide new insights into the field. Published 2008 Wiley Periodicals, Inc. Complexity 15: 11–28, 2009

MIKHAIL PROKOPENKO, FABIO BOSCHETTI, AND ALEX J. RYAN

Key Words: complexity; information theory; self-organization; emergence; predictive information; excess entropy; entropy rate; assortativeness; predictive efficiency; adaptation

Mikhail Prokopenko is affiliated with the Information and Communication Technologies Centre, Commonwealth Scientific and Industrial Research Organisation, North Ryde, New South Wales, Australia

Fabio Boschetti is affiliated with Marine and Atmospheric Research, Commonwealth Scientific and Industrial Research Organisation, Floreat, West Australia, Australia and the School of Earth and Geographical Sciences at the University of Western Australia, Crawley, Western Australia (e-mail: Fabio.Boschetti@csiro.au)

Alex J. Ryan is affiliated with the Defence Science and Technology Organisation, Edinburgh, South Australia, Australia

1. INTRODUCTION

Complex Systems Science studies general phenomena of systems comprised of many simple elements interacting in a nontrivial fashion. Currently, fuzzy quantifiers like "many" and "nontrivial" are inevitable. "Many" implies a number large enough so that no individual component/feature predominates the dynamics of the system, but not so large that features are completely irrelevant. Interactions need to be "nontrivial" so that the degrees of freedom are suitably reduced, but not constraining to the point that the arising structure possesses no further degree of freedom. Crudely put, systems with a huge number of components interacting trivially are explained by statistical mechanics, and systems with precisely defined and constrained interactions are the concern of fields like chemistry and engineering. In so far as the domain of complex systems science overlaps these fields, it contributes insights when the classical assumptions are violated.

This article was submitted as an invited paper resulting from the "Understanding Complex Systems" conference held at the University of Illinois at Urbana-Champaign, May 2007.

Vol. 15, No. 1 DOI 10.1002/cplx.20249
Published online 29 October 2008 in Wiley InterScience
(www.interscience.wiley.com)

It is unsurprising that a similar vagueness afflicts the discipline itself, which notably lacks a common formal framework for analysis. There are a number of reasons for this. Because complex systems science is broader than physics, biology, sociology, ecology, or economics, its foundations cannot be reduced to a single discipline. Furthermore, systems which lie in the gap between the "very large" and the "fairly small" cannot be easily modeled with traditional mathematical techniques.

Initially setting aside the requirement for formal definitions, we can summarize our general understanding of complex systems dynamics as follows:

(1) complex systems are "open," and receive a regular supply of energy, information, and/or matter from the environment;
(2) a large, but not too large, ensemble of individual components interact in a nontrivial fashion; in others words, studying the system via statistical mechanics would miss important properties brought about by interactions;
(3) the nontrivial interactions result in internal constraints, leading to symmetry breaking in the behavior of the individual components, from which *coordinated* global behavior arises;
(4) the system is now more organized than it was before; because neither central director nor any explicit instruction template was followed, we say that the system has *"self-organized"*;
(5) this coordination can express itself as *patterns* detectable by an external observer or as structures that convey new properties to the systems itself. New behaviors *"emerge"* from the system;
(6) coordination and emergent properties may arise from specific response to environmental pressure, in which case we can say the system displays *adaptation*;
(7) when adaptation occurs across generations at a population level we say that the system *evolved*[1];
(8) coordinated emergent properties give rise to effects at a *scale* larger than the individual components. These interdependent sets of components with emergent properties can be observed as coherent entities at lower *resolution* than is needed to observe the components. The system can be identified as a novel unit of its own and can interact with other systems/processes expressing themselves at the same scale. This becomes a building block for new iterations and the cycle can repeat from (1) mentioned above, now at a larger scale.

The process outlined earlier is not too contentious, but does not address "how" and "why" each step occurs. Consequently, we can observe the process but we cannot understand it, modify it, or engineer for it. This also prevents us from understanding what complexity is and how it should be monitored and measured; this equally applies to self-organization, emergence, evolution, and adaptation.

Even worse than the fuzziness and absence of deep understanding already described, is when the above terms are used interchangeably in the literature. The danger of not making clear distinctions in Complex Systems Science is incoherence. To have any hope of coherent communication, it is necessary to unravel the knot of assumptions and circular definitions that are often left unexamined.

Here we suggest a set of working definitions for the above mentioned concepts, essentially a dictionary for Complex Systems Science discourse. Our purpose is not to be prescriptive, but to propose a baseline for shared agreement, to facilitate communication between scientists and practitioners in the field. We would like to prevent the situation in which a scientist talks of emergence and this is understood as self-organization.

For this purpose we chose an information-theoretic framework. There are a number of reasons for this choice:

- a considerable body of work in Complex Systems Science has been cast into information theory, as pioneered by the Santa Fe Institute, and we borrow heavily from this tradition;
- it provides a well developed theoretical basis for our discussion;
- it provides definitions which can be formulated mathematically;
- it provides computational tools readily available; a number of measures can be actually computed, albeit in a limited number of cases.

Nevertheless, we believe that the concepts should also be accessible to disciplines which often operate beyond the application of such a strong mathematical and computational framework, like biology, sociology, and ecology. Consequently, for each concept we provide a "plain English" interpretation, which hopefully will enable communication across fields.

2. AN INFORMATION-THEORETICAL APPROACH

Information Theory was originally developed by Shannon [1] for reliable transmission of information from a source X to a receiver Y over noisy communication channels. Put simply, it addresses the question of "how can we achieve perfect communication over an imperfect, noisy communication channel?" [2]. When dealing with outcomes of imperfect probabilistic processes, it is useful to define the information content of an outcome x which has the probability $P(x)$, as $\log_2 \frac{1}{P(x)}$ (it is measured in bits): improbable outcomes convey more information than probable outcomes. Given a probability distribution P

[1] *This does not limit evolution to DNA/RNA based terrestrial biology—see Section 6.*

Published 2008 Wiley Periodicals, Inc.
DOI 10.1002/cplx

over the outcomes $x \in \mathcal{X}$ (a discrete random variable X representing the process, and defined by the probabilities $P(x) \equiv P(X = x)$ given for all $x \in \mathcal{X}$), the average Shannon information content of an outcome is determined by

$$H(X) = \sum_{x \in \mathcal{X}} P(x) \frac{1}{\log P(x)} = -\sum_{x \in \mathcal{X}} P(x) \log P(x), \quad (1)$$

henceforth we omit the logarithm base 2. This quantity is known as *(information) entropy*. Intuitively, it measures, also in bits, the amount of freedom of choice (or the degree of randomness) contained in the process—a process with many possible outcomes has high entropy. This measure has some unique properties that make it specifically suitable for measuring "how much "choice" is involved in the selection of the event or of how uncertain we are of the outcome?" [1]. In answering this question, Shannon required the following properties for such a measure H:

- continuity: H should be continuous in the probabilities, i.e., changing the value of one of the probabilities by a small amount changes the entropy by a small amount;
- monotony: if all the choices are equally likely, e.g. if all the probabilities $P(x_i)$ are equal to $1/n$, where n is the size of the set $\mathcal{X} = \{x_1, \ldots, x_n\}$, then H should be a monotonic increasing function of n: "with equally likely events there is more choice, or uncertainty, when there are more possible events" [1];
- recursion: H is independent of how the process is divided into parts, i.e. "if a choice be broken down into two successive choices, the original H should be the weighted sum of the individual values of H" [1],

and proved that the entropy function $-K \sum_{i=1}^{n} P(x_i) \log P(x_i)$, where a positive constant K represents a unit of measure, is the only function satisfying these three requirements.

The joint entropy of two (discrete) random variables X and Y is defined as the entropy of the joint distribution of X and Y:

$$H(X, Y) = -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P(x, y) \log P(x, y), \quad (2)$$

where $P(x, y)$ is the joint probability. The conditional entropy of Y, given random variable X, is defined as follows:

$$H(Y|X) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P(x, y) \log \frac{P(x)}{P(x, y)} = H(X, Y) - H(X). \quad (3)$$

This measures the average uncertainty that remains about $y \in \mathcal{Y}$ when $x \in \mathcal{X}$ is known [2].

Mutual information $I(X; Y)$ measures the amount of information that can be obtained about one random variable by observing another (it is symmetric in terms of these variables):

$$I(X; Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P(x, y) \log \frac{P(x, y)}{P(x)P(y)} = H(X) + H(Y) - H(X, Y). \quad (4)$$

Mutual information $I(X; Y)$ can also be expressed via the conditional entropy:

$$I(X; Y) = H(Y) - H(Y|X). \quad (5)$$

These concepts are immediately useful in quantifying qualities of communication channels. In particular, the amount of information $I(X; Y)$ shared between transmitted X and received Y signals is often maximized by designers, via choosing the best possible transmitted signal X. Channel coding establishes that reliable communication is possible over noisy channels if the rate of communication is below a certain threshold called the channel capacity. Channel capacity is defined as the maximum mutual information for the channel over all possible distributions of the transmitted signal X (the source).

The conditional entropy of Y given X, Eq. (3), is also called the equivocation of Y about X, and, rephrasing the Eq. (5) informally, we can state that

$$\text{mutual information} = \text{receiver's diversity} - \text{equivocation of receiver about source.} \quad (6)$$

Thus, the channel capacity is optimized when receiver's diversity is maximized, whereas its equivocation about the source is minimized.

Equivocation of Y about X may also be interpreted as nonassortativeness between Y and X: the degree of having no reciprocity in either positive or negative way. The term assortativeness is borrowed from studies of complex networks: the networks where highly connected nodes are more likely to make links with other highly connected links are said to mix assortatively, while the networks where the highly connected nodes are more likely to make links with more isolated, less connected, nodes are said to to mix disassortatively [3]. The conditional entropy, defined in a suitable way for a network, estimates spurious correlations in the network created by connecting the nodes with dissimilar degrees. As argued by Solé and Valverde [4], this conditional entropy represents the "assortative noise" that affects the overall diversity or the heterogeneity of the network, but does not contribute to the amount of information within it. Solé and Valverde [4] define information transfer within the network as mutual information—the difference between network's heterogeneity (entropy) and assortative noise within it (conditional entropy)—and follow with a characterization that aims to maximize such information transfer. This means that the assortative noise reflecting spurious dependencies among nonassortative components should be reduced while the network's diversity should be increased.

Although mutual information is typically used as a suitable measure for information transfer, it contains no

Published 2008 Wiley Periodicals, Inc.
DOI 10.1002/cplx

inherent directionality, and various alternatives have been proposed. For example, transfer entropy [5] measures the average information contained in the source about the next state of the destination that was not already contained in the destination's past. It can be argued that transfer entropy is the appropriate measure for *predictive* information transfer in spatiotemporal systems [6]. For example, transfer entropy has been used to characterize information flow in sensorimotor networks [7]. There also exists an alternative perturbation-based candidate that captures information flow from the perspective of causality rather than prediction [8].

Nevertheless, we believe that the two criteria fused in the maximization of mutual information $I(X;Y) = H(Y) - H(Y|X)$ (i.e., reduction of equivocation $H(Y|X)$ and increase of diversity $H(Y)$) are useful not only when dealing with channel's capacity or complex networks with varying assortativeness (Example 3.6), but in a very general context. An increase in complexity in various settings may be related to maximization of the information shared (transferred) within the system—to reiterate, this is equivalent to maximization of the system's heterogeneity (i.e. entropy $H(Y)$), and minimization of local conflicts within the system (i.e. conditional entropy $H(Y|X)$).

As pointed out by Polani et al. [9], information should not be considered simply as something that is transported from one point to another as a "bulk" quantity—instead, "looking at the intrinsic dynamics of information can provide insight into inner structure of information." This school of thought suggests that maximization of information transfer through selected channels appears to be one of the main evolutionary pressures [10–15]. We shall consider information dynamics of evolution in Section 6, noting at this stage that although the evolutionary process involves a larger number of pressures and constraints, information fidelity (i.e. preservation) is a consistent motif throughout biology [16]. For example, it was observed that evolution operates close to the error threshold [17]: Adami argued that the evolutionary process extracts relevant information, storing it in the genes. Because this process is relatively slow [18], it is a selective advantage to preserve this valuable information, once captured [19].

In the remainder of this work, we intend to point out how different concepts in Complex Systems Science can be interpreted via simple information-theoretic relationships, and illustrate the importance of the informational split between "diversity" and "equivocation" (often leading to maximization of the information transfer within the system). In particular, we shall argue that when suitable information channels are identified, the rest is often a matter of computation—the computation of "diversity" and "equivocation". In engineering, the choice of channels is typically a task for modelers, whereas in biological systems the "embodied" channels are shaped by interactions with the environment during the evolution.

There are other mathematical approaches, such as nonlinear time series analysis, Chaos Theory, etc., that also provide insights into the concepts used by Complex Systems Science. We note that these approaches are outside the scope of this article, as our intention is to point out similarities in information dynamics across multiple fields, providing a baseline for Complex Systems Science discourse rather than a competing methodology. It is possible that Information Theory has not been widely used in applied studies of complex systems because of the lack of clarity. We are proposing here to clarify the applicability and exemplify how different information channels can be identified and used.

3. COMPLEXITY

3.1. Concept

It is an intuitive notion that certain processes and systems are harder to describe than others. Complexity tries to capture this difficulty in terms of the amount of information needed for the description, the time it takes to carry out the description, the size of the system, the number of components in the system, the number of conflicting constraints, the number of dimensions needed to embed the system dynamics, etc. A large number of definitions have been proposed in the literature and because a review is beyond the scope of this work, we adopt here as definition of complexity the amount of information needed to describe a process, a system, or an object. This definition is computable (at least in one of its forms), is observer-independent (once resolution is defined), applies to both data and models [20] and provides a framework within which self-organization and emergence can also be consistently defined.

3.1.1. Algorithmic Complexity

The original formulation can be traced back to Solomonoff, Kolmogorov, and Chaitin, who developed independently what is today known as Kolmogorov–Chaitin or algorithmic complexity [21,22]. Given an entity (this could be a data set or an image, but the idea can be extended to other objects) the algorithmic complexity is defined as the length (in bits of information) of the shortest program (computer model) which can describe the entity. According to this definition a simple periodic object (a sine function for example) is not complex, because we can store a sample of the period and write a program which repeatedly outputs it, thereby reconstructing the original data set with a very small program. At the opposite end of the spectrum, an object with no internal structure cannot be described in any meaningful way but by storing every feature, because we cannot rely on any shared structure for a shorter description. It follows that a random object has maximum complexity, because the

Published 2008 Wiley Periodicals, Inc.
DOI 10.1002/cplx

shortest program able to reconstruct it needs to store the object itself.[2]

A nice property of this definition is that it does not depend on what language we use to write the program. It can be shown that descriptions using different languages differ by additive constants. However, a clear disadvantage of the algorithmic complexity is that it cannot be computed exactly but only approximated from above—see the Chaitin theorem [23].

3.1.2. Statistical Complexity

Having described algorithmic complexity, we note that associating randomness to maximum complexity seems counterintuitive. Imagine you throw a cup of rice to the floor and want to describe the spatial distribution of the grains. In most cases you do not need to be concerned with storing the position of each individual grain; the realization that the distribution is structure-less and that predicting the exact position of a specific grain is impossible is probably all you need to know. And this piece of information is very simple (and short) to store. There are applications for which our intuition suggests that both strictly periodic and totally random sequences should share low complexity.

One definition addressing this concern is the statistical complexity [24]—it attempts to measure the size of the minimum program able to statistically reproduce the patterns (configurations) contained in the data set (sequence): such a minimal program is able to statistically reproduce the configuration ensemble to which the sequence belongs. In the rice pattern mentioned earlier, there is no statistical difference in the probability of finding a grain at different positions and the resulting statistical complexity is zero.

Apart from implementation details, the conceptual difference between algorithmic and statistical complexity lies in how randomness is treated. Essentially, the algorithmic complexity implies a deterministic description of an object (it defines the information content of an individual sequence), while the statistical complexity implies a statistical description (it refers to an ensemble of sequences generated by a certain source) [25, 26]. As suggested by Boffetta et al. [26], which of these approaches is more suitable is problem-specific.

3.1.3. Excess Entropy and Predictive Information

As pointed out by Bialek et al. [27], our intuitive notion of complexity corresponds to statements about the underlying process, and not directly to Kolmogorov complexity. A dynamic process with an unpredictable and random output (large algorithmic complexity) may be as trivial as the dynamics producing predictable constant outputs (small algorithmic complexity)—while "really complex processes lie somewhere in between." Noticing that the entropy of the output strings either is a fixed constant (the extreme of small algorithmic complexity), or grows exactly linearly with the length of the strings (the extreme of large algorithmic complexity), we may conclude that the two extreme cases share one feature: corrections to the asymptotic behavior do not grow with the size of the data set. Grassberger [25] identified the slow approach of the entropy to its extensive limit as a sign of complexity. Thus, subextensive components—which grow with time less rapidly than a linear function—are of special interest. Bialek et al. [27] observe that the subextensive components of entropy identified by Grassberger determine precisely the information available for making predictions—e.g. the complexity in a time series can be related to the components which are "useful" or "meaningful" for prediction. We shall refer to this as *predictive information*. Revisiting the two extreme cases, they note that "it only takes a fixed number of bits to code either a call to a random number generator or to a constant function"—in other words, a model description *relevant to prediction* is compact in both cases.

The predictive information is also referred to as excess entropy [28, 29], stored information [30], effective measure complexity [25, 31, 32], complexity [33, 34], and has a number of interpretations.

3.2. Information-Theoretic Interpretation

3.2.1. Predictive Information

To estimate the relevance to prediction, two distributions over a stream of data with infinite past and infinite future $X = \ldots, x_{t-2}, x_{t-1}, x_t, x_{t+1}, x_{t+2}, \ldots$ are considered: a prior probability distribution for the futures, $P(x_{\text{future}})$, and a more tightly concentrated distribution of futures conditional on the past data, $P(x_{\text{future}}|x_{\text{past}})$, and their average ratio is defined

$$I_{\text{pred}}(T, T') = \left\langle \log_2 \frac{P(x_{\text{future}}|x_{\text{past}})}{P(x_{\text{future}})} \right\rangle, \tag{7}$$

where $\langle \cdots \rangle$ denotes an average over the joint distribution of the past and the future, $\mathrm{P}(x_{\text{future}}|x_{\text{past}})$, T is the length of the observed data stream in the past, and T' is the length of the data stream that will be observed in the future. This average predictive information captures the reduction of entropy, in Shannon's sense, by quantifying the information (measured in bits) that the past provides about the future:

$$I_{\text{pred}}(T, T') = H(T') - H(T'|T), \tag{8}$$

or informally,

$$\begin{aligned} &\text{predictive information} \\ &\quad = \text{total uncertainty about the future} \\ &\qquad - \text{uncertainty about the future} \\ &\qquad\qquad \text{given the past} \end{aligned} \tag{9}$$

We may point out that the total uncertainty $H(T')$ can be thought of as structural diversity of the underlying process.

[2] *This follows from the most widely used definition of randomness, as structure which can not be compressed in any meaningful way.*

Published 2008 Wiley Periodicals, Inc.
DOI 10.1002/cplx

Similarly, the conditional uncertainty $H(T'|T)$ can be related to structural nonconformity or equivocation within the process—a degree of nonassortativeness between the past and the future, or between components of the process in general:

$$\text{predictive information} = \text{diversity} \\ - \text{non-assortativeness}. \quad (10)$$

The predictive information is always positive and grows with time less rapidly than a linear function, being subextensive. It provides a universal answer to the question of how much is there to learn about the underlying pattern in a data stream: $I_{\text{pred}}(T, T')$ may either stay finite, or grow infinitely with time. If it stays finite, this means that no matter how long we observe we gain only a finite amount of information about the future: e.g. it is possible to completely predict dynamics of periodic regular processes after their period is identified. For some irregular processes the best predictions may depend only on the immediate past (e.g. a Markov process, or in general, a system far away from phase transitions and/or symmetry breaking)—and in these cases $I_{\text{pred}}(T, T')$ is also small and is bound by the logarithm of the number of accessible states: the systems with more states and longer memories have larger values of predictive information [27]. If $I_{\text{pred}}(T, T')$ diverges and optimal predictions are influenced by events in the arbitrarily distant past, then the rate of growth may be slow (logarithmic) or fast (sublinear power). If the data allows us to learn a model with a finite number of parameters or a set of underlying rules describable by a finite number of parameters, then $I_{\text{pred}}(T, T')$ grows logarithmically with the size of the observed data set, and the coefficient of this divergence counts the dimensionality of the model space (i.e. the number of parameters). Sublinear power-law growth may be associated with infinite parameter models or nonparametric models such as continuous functions with some regularization (e.g. smoothness constraints) [35].

3.2.2. Statistical Complexity

The statistical complexity is calculated by reconstructing a minimal model, which contains the collection of all situations (histories) which share a similar probabilistic future, and measuring the entropy of the probability distribution of the states.

Here we briefly sketch the approach to statistical complexity based on ϵ-machines [24, 36, 37]. Let us again consider a stream of data with infinite past and infinite future[3] $X = \ldots, x_{t-2}, x_{t-1}, x_t, x_{t+1}, x_{t+2}, \ldots$, and use $x_{\text{past}}(t)$ and $x_{\text{future}}(t)$ to denote the sequences up to x_t, and from x_{t+1} forward, respectively. Then, an equivalence relation $\sim$ over histories x_{past} of observed states is defined:

$$x_{\text{past}}(t) \sim x_{\text{past}}(t') \quad \text{if and only if} \\ P(x_{\text{future}}|x_{\text{past}}(t)) = P(x_{\text{future}}|x_{\text{past}}(t')), \\ \forall\, x_{\text{future}}. \quad (11)$$

The equivalence classes $\mathcal{S}_i$ induced by the relation $\sim$ are called causal states. For practical purposes, one considers longer and longer histories x^L_{past} up to a given length $L = L_{\text{max}}$, and obtains the partition into the classes for a fixed future horizon (e.g., for the very next observable). In principle, starting at the coarsest level which groups together those histories that have the same distribution for the very next observable, one may refine the partition by subdividing these coarse classes using the distribution of the next two observables, etc. [39]. The causal states provide an optimal description of a system's dynamics in the sense that these states make as good a prediction as the histories themselves. Different causal states "leave us in different conditions of ignorance about the future" [37]. The set of causal states $\mathcal{S}_i$ is denoted by $\mathcal{S}$.

After all causal states $\mathcal{S}_i$ are identified, one constructs an ϵ-machine—a minimal model—as an automaton with these states and the transition probabilities T_{ij} between the states. To obtain a transition probability T_{ij} between the states $\mathcal{S}_i$ and $\mathcal{S}_j$, one simply traces the data stream, identifies all the transitions from histories $x_{\text{past}}(t) \in \mathcal{S}_i$ to new histories $x_{\text{past}}(t+1) \in \mathcal{S}_j$, and calculates T_{ij} as $P(\mathcal{S}_j|\mathcal{S}_i)$. The transition probabilities of an ϵ-machine allow to calculate an invariant probability distribution $P(\mathcal{S})$ over the causal states. One can also inductively obtain the probability $P(\mathcal{S}_i)$ of finding the data stream in the causal state $\mathcal{S}_i$ by observing many configurations [29]. The statistical complexity C_μ is defined as the Shannon entropy, measured in bits, of this probability distribution $P(\mathcal{S})$:

$$C_\mu = -\sum_{\mathcal{S}_i \in \mathcal{S}} P(\mathcal{S}_i) \log P(\mathcal{S}_i). \quad (12)$$

It represents the minimum average amount of memory needed to statistically reproduce the configuration ensemble to which the sequence belongs [40]. The description of an algorithm which achieves an ϵ-machine reconstruction and calculates the statistical complexity for 1D time series can be found in [41] and for 2D time series in [38].

In general, the predictive information is bound by the statistical complexity

$$I_{\text{pred}}(T, T') \le C_\mu. \quad (13)$$

This inequality means that the memory needed to perform an optimal prediction of the future configurations cannot be lower than the mutual information between the past and future themselves [42]: this relationship reflects the fact that the causal states are a reconstruction of the hidden, effective states of the process. Specifying how the memory within a process is organized cannot

[3] *The formalism is applicable not only to time series, but also to stochastic processes, one dimensional chains of Ising spins, cellular automata, other spatial processes, e.g. time-varying random fields on networks [38], etc.*

Published 2008 Wiley Periodicals, Inc.
DOI 10.1002/cplx

be done within the framework of Information Theory, and a more structural approach based on the Theory of Computation must be used [43]—this leads (via causal states) to ϵ-machines and statistical complexity C_μ.

3.2.3. Excess Entropy

Before defining excess entropy, let us define the block-entropy $H(L)$ of length-L sequences within a data stream (an information source):

$$H(L) = -\sum_{x^L \in \mathcal{X}^L} P(x^L) \log P(x^L), \quad (14)$$

where $\mathcal{X}$ contains all possible blocks/sequences of length L. The block-entropy $H(L)$, measured in bits, is a non-decreasing function of L, and the quantity

$$h_\mu(L) = H(L) - H(L-1), \quad (15)$$

defined for $L \geq 1$, is called the entropy gain, measured in bits per symbol [43]. It is the average uncertainty about the Lth symbol, provided the $(L-1)$ previous ones are given [26]. The limit of the entropy gain

$$h_\mu = \lim_{L\to\infty} h_\mu(L) = \lim_{L\to\infty} \frac{H(L)}{L} \quad (16)$$

is the source entropy rate—also known as per-symbol entropy, the thermodynamic entropy density, Kolmogorov-Sinai entropy [44], metric entropy, etc. Interestingly, the entropy gain $h_\mu(L) = H(L) - H(L-1)$ differs in general from the estimate $\frac{H(L)}{L}$ for any given L, but converges to the same limit: the source entropy rate.

As noted by Crutchfield and Feldman [43], the length-L approximation $h_\mu(L)$ typically overestimates the entropy rate h_μ at finite L, and each difference $[h_\mu(L) - h_\mu]$ is the difference between the entropy rate conditioned on L measurements and the entropy rate conditioned on an infinite number of measurements—it estimates the information-carrying capacity in the L-blocks that is not actually random, but is due instead to correlations, and can be interpreted as the local (i.e. L-dependent) predictability [45]. The total sum of these local over-estimates is the excess entropy or intrinsic redundancy in the source:

$$E = \sum_{L=1}^{\infty} [h_\mu(L) - h_\mu]. \quad (17)$$

Thus, the excess entropy measures the amount of apparent randomness at small L values that is "explained away" by considering correlations over larger and larger blocks: it is a measure of the total apparent memory or structure in a source [43]. A finite partial-sum estimate [43] of excess entropy for length L is given by

$$E(L) = H(L) - Lh_\mu(L). \quad (18)$$

Importantly, Crutchfield and Feldman [43] demonstrated that the excess entropy E can also be seen as either:

(1) the mutual information between the source's past and the future—exactly the predictive information $I_{\text{pred}}(T, T')$, if T and T' are semi-infinite, or
(2) the subextensive part of entropy $H(L) = E + h_\mu L$, as $L \to \infty$.

It was also shown that only the first interpretation holds in two-dimensional systems [46]. This analogy, coupled with the representation (10), creates an alternative intuitive representation:

$$\text{excess entropy} = \text{diversity} - \text{nonassortativeness}. \quad (19)$$

In other words, the total structure within a system is adversely affected by nonassortative disagreements (e.g., between the past and the future) that reduce the overall heterogeneity.

3.3. Convergence

The source entropy rate h_μ captures the irreducible randomness produced by a source after all correlations are taken into account [43]:

- $h_\mu = 0$ for periodic processes and even for deterministic processes with infinite-memory (e.g. Thue-Morse process) which do not have an internal source of randomness, and
- $h_\mu > 0$ for irreducibly unpredictable processes, e.g. independent identically distributed (IID) processes which have no temporal memory and no complexity, as well as Markov processes (both deterministic and nondeterministic), and infinitary processes (e.g. positive-entropy-rate variations on the Thue-Morse process).

The excess entropy, or predictive information, increases with the amount of structure or memory within a process:

- E is finite for both periodic processes and random (e.g. it is zero for an IID process)—its value can be used as a relative measure: a larger period results in higher E, as a longer past needs to be observed before we can estimate the finite predictive information;
- finite length estimates $E(L)$ of E diverge logarithmically for complex processes because of an infinite memory (e.g. Thue-Morse process); similarly, as noted in 3.2, predictive information $I_{\text{pred}}(T, T')$ diverges logarithmically, with the size of the observed data set, for complex processes "in a known class but with unknown parameters" [35], and the coefficient of this divergence can be used as a relative measure estimating a number of parameters or rules in the underlying model;
- an even faster rate of growth is also possible, and $I_{\text{pred}}(T, T')$ exhibits a sublinear power law divergence for complex processes that "fall outside the conventional finite dimensional models" [35] (e.g. a continuous function with smoothness constraints)—typically, this happens in problems where predictability over long scales is "governed by a progressively more

Published 2008 Wiley Periodicals, Inc.
DOI 10.1002/cplx

detailed description" as more data are observed [27]; here, the relative complexity measure is the number of different parameter-estimation scales growing in proportion to the number of taken samples (e.g. the number of bins used in a histogram approximating the distribution of a random variable).

3.4. Summary

Entropy rate h_μ is a good identifier of intrinsic randomness, and is related to the Kolmogorov–Chaitin (KC) complexity. To reiterate, the KC complexity of an object is the length of the minimal Universal Turing Machine (UTM) program needed to reproduce it. The entropy rate h_μ is equal to the average length (per variable) of the minimal program that, when run, will cause an UTM to produce a typical configuration and then halt [21, 42, 47].

The relationships $I_{\text{pred}}(T, T') = E$ and $E \leq C_\mu$, suggest a very intuitive interpretation:

$$\begin{aligned} &\text{predictive information} = \text{richness of} \\ &\quad \text{structure} \leq \text{statistical complexity} \\ &\quad = \text{memory for optimal predictions.} \end{aligned} \tag{20}$$

Predictive information and statistical complexity are small at both extremes (complete order and complete randomness), and are maximal in the region somewhere between the extremes. Moreover, in some "intermediate" cases, the complexity is infinite, and may be divergent at different rates.

If one needs to maximize the total structure in a system (e.g., mutual information in the network), then a reduction of local conflicts or disagreements represented by nonassortativeness, in parallel with an increase in the overall diversity, is the preferred strategy. This is not equivalent to simply reducing the randomness of the system.

3.5. Example—Thue-Morse Process

The infinite-memory Thue-Morse sequences $\sigma^k(s)$ contain two units 0 and 1, and can be obtained by the substitution rules $\sigma^k(0) = 01$ and $\sigma^k(1) = 10$ (e.g. $\sigma^1(1) = 10$, $\sigma^2(1) = 1001$, etc.).

Despite the fact that the entropy rate $h_\mu = 0$, and the entropy gain for the process converges according to a power law, $h_\mu(L) \propto 1/L$, such a process needs an infinite amount of memory to maintain its aperiodicity [43], and hence, its past provides an ever-increasing predictive information about its future. This leads to logarithmic divergence of both block-entropy $H(L) \propto \log_2 L$, as well as partial-sum excess entropy $E(L) \propto \log_2 L$, correctly indicating an infinite-memory process [43].

The estimates of the statistical complexity $C_\mu(L)$, where L is the length of histories x^L_{past} used in defining causal states by the equivalence relation (11), also diverge for the Thue-Morse process. The exact divergence rate is still a subject of ongoing research—it is suggested [40] that the divergence may be logarithmic, i.e. $C_\mu(L) \propto \log_2 L$.

3.6. Example—Graph Connectivity

Graph connectivity can be analyzed in terms of the size of the largest connected subgraph (LCS) and its standard deviation obtained across an ensemble of graphs, as suggested by Random Graph Theory [48]. In particular, critical changes occur in connectivity of a directed graph as the number of edges increases: the size of the LCS rapidly increases as well and fills most of the graph, whereas the variance in the size of the LCS reaches a maximum at some critical point before decreasing. In other words, variability within the ensemble of graphs grows as graphs become more and more different in terms of their structure.

An information-theoretic representation can subsume this graph-theoretic model. Let us consider a network with N nodes (vertices) and M links (edges), and say that the probability of a randomly chosen node having degree k is p_k, where $1 \leq k \leq N_p$. The distribution of such probabilities is called the degree distribution of the network. However, if a node is reached by following a randomly chosen link, then the remaining number of links (the remaining degree) of this node is not distributed according to p_k—instead it is biased in favor of nodes of high degree, because more links end at a high-degree node than at a low-degree one [3]. The distribution q_k of such remaining degrees is called the remaining degree distribution, and is related to p_k as follows [3]:

$$q_k = \frac{(k+1)p_{k+1}}{\sum_j^{N_p} jp_j}, \quad 0 \leq k \leq N_p - 1. \tag{21}$$

The quantity $e_{j,k}$ can then be defined as the joint probability distribution of the remaining degrees of the two nodes at either end of a randomly chosen link [3, 49], as well as the conditional probability $\pi(j|k) = e_{j,k}/q_k$ [4, 50] defined as the probability of observing a vertex with j edges leaving it provided that the vertex at the other end of the chosen edge has k leaving edges. Following Solé and Valverde [4], we use these probability distributions in defining

- the Shannon entropy of the network, that measures the diversity of the degree distribution or the network's heterogeneity:

$$H(q_k) = -\sum_{k=0}^{N_p-1} q_k \log q_k, \tag{22}$$

- the joint entropy measuring of the average uncertainty of the network as a whole

$$H(q_j, q_k) = -\sum_{j=0}^{N_p-1} \sum_{k=0}^{N_p-1} e_{j,k} \log e_{j,k}, \tag{23}$$

Published 2008 Wiley Periodicals, Inc.
DOI 10.1002/cplx

- the conditional entropy

$$H(q_j|q_k) = -\sum_{j=0}^{N_p-1}\sum_{k=0}^{N_p-1} q_j \pi(j|k) \log \pi(j|k)$$
$$= -\sum_{j=0}^{N_p-1}\sum_{k=0}^{N_p-1} e_{j,k} \log \frac{e_{j,k}}{q_k}. \quad (24)$$

These measures are useful in analyzing how assortative, disassortative, or nonassortative is the network. Assortative mixing (AM) is the extent to which high-degree nodes connect to other high degree nodes [3]. In disassortative mixing (DM), high-degree nodes are connected to low-degree ones. Both AM and DM networks are contrasted with nonassortative mixing (NM), where one cannot establish any preferential connection between nodes. As pointed out by Solé and Valverde [4], the conditional entropy $H(q_j|q_k)$ may estimate spurious correlations in the network created by connecting the vertices with dissimilar degrees—this noise affects the overall diversity or the average uncertainty of the network, but does not contribute to the amount of information (correlation) within it. Using the joint probability of connected pairs $e_{j,k}$, one may calculate the amount of correlation between vertices in the graph via the mutual information measure, the information transfer, as

$$I(q) = I(q_j, q_k) = H(q_k) - H(q_j|q_k)$$
$$= -\sum_{j=0}^{N_p-1}\sum_{k=0}^{N_p-1} e_{j,k} \log \frac{e_{j,k}}{q_j q_k}, \quad (25)$$

Informally,

$$\begin{aligned}&\text{transfer within the network}\\ &\quad = \text{diversity in the network}\\ &\qquad - \text{assortative noise in the}\\ &\qquad\qquad \text{network structure.} \quad (26)\end{aligned}$$

This motivating interpretation is analogous to the one suggested by the Eqs. (6), (10), and (19), assuming that assortative noise is the nonassortative extent to which the preferential (either AM or DM) connections are obscured. The mutual information $I(q)$ is a better, more generic measure of dependence than the correlation functions, like the variance in the size of the LCS, that "measure linear relations whereas mutual information measures the general dependence and is thus a less biased statistic" [4].

4. SELF-ORGANIZATION

Three ideas are implied by the word self-organization: (a) the organization in terms of global implicit coordination; (b) the dynamics progressing in time from a not (or less) organized to an organized state; and (c) the spontaneous arising of such dynamics. To avoid semantic traps, it is important to notice that the word "spontaneous" should not be taken literally; we deal with open systems, exchanging energy, matter and/or information with the environment and made up of components whose properties and behaviors are defined prior to the organization itself. The "self" prefix merely states that no centralized ordering or external agent/template explicitly guides the dynamics. It is thus necessary to define what is meant by "organization" and how its arising or increase can be detected.

4.1. Concept

A commonly held view is that organization entails an increase in complexity. Unfortunately the lack of agreement of what we mean by complexity leaves such definition somehow vague. For example, De Wolf and Holvoet [51] refer to complexity as a measure of redundancy or structure in the system. The concept can be made more formal by adopting the statistical complexity described earlier as a measure of complexity, as demonstrated in Shalizi [39] and Shalizi et al. [52]. This definition offers several of the advantages of the Computational Mechanics approach; it is computable and observer independent. Also, it captures the intuitive notion that the more a system self-organizes, the more behaviors it can display, the more effort is needed to describe its dynamics. Importantly, this needs to be seen in a statistical perspective; whereas a disorganized system may potentially display a larger number of actual configurations, the distinction among several of them may not matter statistically. Adopting the statistical complexity allows us to focus on the system configurations which are statistically different (causal states) for the purpose at hand. We thus have a measure which is based only on the internal dynamics of the system (and consequently is observer-independent) but which can be tuned according to the purpose of the analysis. For an alternative definition of self-organization based on thermodynamics and the distinction between self-organization and the related concept of self-assembly we refer the reader to Halley and Winkler [53].

4.2. Information-Theoretic Interpretation

In the scientific literature the concept of self-organization refers to both living and nonliving systems, ranging from physics and chemistry to biology and sociology. Kauffman [54] suggests that the underlying principle of self-organization is the generation of constraints in the release of energy. According to this view, the constrained release allows for such energy to be controlled and channeled to perform some useful work. This work in turn can be used to build better and more efficient constraints for the release of further energy and so on; this principle is closely related to Kauffman's own definition of life [54]. It helps us to understand why an organized system with effectively less available configurations may behave and look more complex than a disorganized one to which, in principle, more configurations are available. The ability to constrain and control the release of energy may allow a system to display behaviors (reach configurations) which, although possible, would be extremely unlikely in its nonorganized state. It is surely possible that 100 parrots move independently to the same location at the same time, but this is far more likely if they fly in a flock. A limited number of coordinated behaviors become implementable because of self-organization, which would be extremely unlikely to arise in the midst of a vast number of disorganized configurations. The ability to constrain the release of energy thus provides the self-organized system with behaviors that can be selectively chosen for successful adaptation.

However, Halley and Winkler [53] correctly point out that attention should paid to how self-organization is treated if we want the concept to apply equally to both living and nonliving systems. For example, although it is temping to consider adaptation as a guiding process for self-organization, it then makes it hard to use the same definition of self-organization for nonliving systems.

Recently, Correia [55] analyzed self-organization motivated by embodied systems, i.e. physical systems situated in the real world, and established four fundamental properties of self-organization: no external control, an increase in order, *robustness*[4], and interaction. All of these properties are easily interpretable in terms of information dynamics.

First, the absence of external control may correspond to "spontaneous" arising of information dynamics without any flow of information into the self-organizing system. Second, an increase in order or complexity reflects that the predictive information is increased within the system or its specific part:

$$I_{\text{pred}}([t_1 - T, t_1], [t_1, t_1 + T']) < I_{\text{pred}}([t_2 - T, t_2], [t_2, t_2 + T']) \quad (27)$$

and

$$C_\mu^{\text{System}}(t_1) < C_\mu^{\text{System}}(t_2), \quad (28)$$

for $t_1 < t_2$ and positive T and T', where I_{pred} is the predictive information (7) estimated at different times (t_1 and t_2), and $C_\mu^{\text{System}}(t)$ is the statistical complexity (12) estimated at time t. In general, however, we believe that one may relax the distinction between these two requirements and demand only that in a self-organizing system, the change in the predictive information's gain within the system $\triangle I^{\text{system}}(t_1, t_2) = I_{\text{pred}}([t_2 - T, t_2], [t_2, t_2 + T']) - I_{\text{pred}}([t_1 - T, t_1], [t_1, t_1 + T'])$, is strictly more than the amount of information flowing from the outside $I^{\text{influence}}(t_1, t_2)$, analogously estimated for an external influence, given T and T':

$$I^{\text{influence}}(t_1, t_2) < \triangle I^{\text{system}}(t_1, t_2). \quad (29)$$

Similarly, the complexity of external influence into a self-organizing system, $C_\mu^{\text{influence}}(t_1, t_2)$ should be strictly less than the gain in internal complexity, $\triangle C_\mu^{\text{system}}(t_1, t_2) = C_\mu^{\text{system}}(t_2) - C_\mu^{\text{system}}(t_1)$:

$$C_\mu^{\text{influence}}(t_1, t_2) < \triangle C_\mu^{\text{system}}(t_1, t_2). \quad (30)$$

Third, a system is robust if it continues to function in the face of perturbations [56]. Robustness of a self-organizing system to perturbations means that it may interleave stages of an increased information transfer within some channels (dominant patterns are being exploited; assortative noise is low; $\triangle I^{\text{system}} > I^{\text{influence}}$) with periods of decreased information transfer (alternative patterns are being explored; assortative noise is high; $\triangle I^{\text{system}} < I^{\text{influence}}$)—see also Example 4.5. This flexibility provides the self-organized system with a variety of behaviors, thus informally following Ashby's Law of Requisite Variety. A more detailed information-theoretic treatment of robustness is presented by Ay and Krakauer [57].

Lastly, the interaction property is described by Correia [55] as follows: "minimization of local conflicts produces global optimal self-organization, which is evolutionarily stable"—see Example 4.4. Minimization of local conflicts, however, is only one aspect, captured in Eqs. (6), (10), (19), and (26) as equivocation or nonassortativeness, and should be generally complemented by maximizing diversity within the system. The interaction property is immediately related to the second property (robustness).

4.3. Summary

The fundamental properties of self-organization are immediately related to information dynamics, and can be studied in precise information-theoretic terms when the appropriate channels are identified. The first two properties (no external control, and an increase in order), are unified in the Eqs. (29) and (30), while the fourth, interaction, property is subsumed by the key equations of the information dynamics analyzed in this work, e.g. Eq. (10). The third, robustness, property follows from maximizing the richness-of-structure (the excess entropy), and an ensuing increase in the variety of behaviors. It manifests itself via interleaved stages of increased and decreased information transfer within certain channels.

4.4. Example—Self-Organizing Traffic

In the context of pedestrian traffic, Correia [55] argues that it can be shown that the "global efficiency of opposite pedestrian traffic is maximized when interaction rate is locally minimized for each component. When this happens two separate lanes form, one in each direction. The minimization of interactions follows directly from maximizing the average velocity in the desired direction." In other words, the division into lanes results from maximizing velocity (an overall objective or fitness), which in turn supports minimization of conflicts.

[4] *Although Correia refers to this as adaptability, according to the concepts in this paper he in fact defines robustness. This is an example of exactly the kind of issue we hope to avoid by developing this dictionary.*

Another example is provided by ants: “Food transport is done via a trail, which is an organized behavior with a certain complexity. Nevertheless, a small percentage of ants keeps exploring the surroundings and if a new food source is discovered a new trail is established, thereby dividing the workers by the trails [58] and increasing complexity” [55]. Here, the division into trails is again related to an increase in fitness and complexity.

These two examples demonstrate that when local conflicts are minimized, the degree of coupling among the components (i.e. interaction) increases and the information flows easier, thus increasing the predictive information. This means that not only the overall diversity of a system is important (more lanes or trails), but the interplay among different channels (the assortative noise within the system, the conflicts) is crucial as well.

4.5. Example—Self-organizing Locomotion

The internal channels through which information flows within the system are observer-independent, but different observers may select different channels for a specific analysis. For example, let us consider a modular robotic system modeling a multisegment snake-like (salamander) organism, with actuators (“muscles”) attached to individual segments (“vertebrae”). A particular side-winding locomotion arises as a result of individual control actions when the actuators are coupled within the system and follow specific evolved rules [14, 15].

The proposed approach [14, 15] introduced a spatial dimension across multiple Snakebot's actuators, and considered varying spatial sizes $d_s \leq D_s$ (the number of adjacent actuators) and time length $d_t \leq D_t$ (the time interval) in defining spatiotemporal patterns (blocks) $V(d_s, d_t)$ of size $d_s \times d_t$, containing values of the corresponding actuators' states from the observed multivariate time series of actuators' states. A block entropy computed over these patterns is generalized to order-2 Rényi entropy [59], resulting in the spatiotemporal generalized correlation entropy K_2:

$$K_2 = -\lim_{d_s \to \infty} \lim_{d_t \to \infty} \frac{1}{d_s} \frac{1}{d_t} \ln \sum_{V(d_s, d_t)} P^2(V(d_s, d_t)), \quad (31)$$

where the sum under the logarithm is the collision probability, defined as the probability $P_c(x)$ that two independent realizations of the random variable X show the same value $P_c(X) = \sum_{x \in \mathcal{X}} P(x)^2$. The order-$q$ Rényi entropy K_q is a generalization of the Kolmogorov–Sinai entropy: it is a measure for the rate at which information about the state of the system is lost in the course of time—see Section 3.2.3 describing the entropy rate. The finite-template (finite spatial-extent and finite time-delay) entropy rate estimates $K_2^{d_s d_t}$ converge to their asymptotic values K_2 in different ways for Snakebots with different individual control actions, and the predictive information, approximated as a generalized excess entropy:

$$E_2 = \sum_{d_s=1}^{D_s} \sum_{d_t=1}^{D_t} \left(K_2^{d_s d_t} - K_2\right) \quad (32)$$

defines a fitness landscape.

There is no global coordinator component in the evolved system, and it can be shown that the amount of predictive information between groups of actuators grows as the modular robot starts to move across the terrain. That is, the distributed actuators become more coupled when a coordinated side-winding locomotion is dominant. Faced with obstacles, the robot temporarily loses the side-winding pattern: the modules become less organized, the strength of their coupling is decreased, and rather than exploiting the dominant pattern, the robot explores various alternatives. Such exploration temporarily decreases self-organization, i.e. the predictive information within the system. When the obstacles are avoided, the modules “rediscover” the dominant side-winding pattern by themselves, recovering the previous level of predictive information and manifesting again the ability to self-organize without any global controller. Of course, the “magic” of this self-organization is explained by properties defined a priori: the rules employed by the biologically-inspired actuators have been obtained by a genetic programming algorithm, whereas the biological counterpart (the rattlesnake *Crotalus cerastes*) naturally evolved over long time. Our point is simply that we can measure the dynamics of predictive information and statistical complexity as it presents itself within the channels of interest.

5. EMERGENCE

Nature can be observed at different levels of resolution, be these intended as spatial or temporal scales or as measurement precision. For certain phenomena this affects merely the level of details we can observe. As an example, depending on the scale of observation, satellite images may highlight the shape of a continent or the make of a car; similarly, the time resolution of a temperature time series can reflect local stochastic (largely unpredictable) fluctuations or daily periodic (fairly predictable) oscillations. There are classes of phenomena though, which when observed at different levels, display behaviors which appear fundamentally different. The quantum phenomena of the “very small” and the relativistic effects of the “very large” do not seem to find obvious realizations in our everyday experience at the middle scale; similarly the macroscopic behavior of a complex organism appears to transcend the biochemistry it derives from. The apparent discontinuity between these radically different phenomena arising at different scales is usually, broadly and informally, defined as emergence.

Attempts to formally address the study of emergence have sprung at regular intervals in the last century or so (for a nice review see Corning [60]), under different names, approaches, and motivations, and is currently receiving

Published 2008 Wiley Periodicals, Inc.
DOI 10.1002/cplx

a new burst of interest [61]. Here we borrow from Crutchfield [62], who, in a particularly insightful work, proposes a distinction between two phenomena which are commonly viewed as expression of emergence: pattern formation and "intrinsic" emergence.

5.1. Concept

5.1.1. Pattern Formation

In pattern formation we imagine an observer trying to "understand" a process. If the observer detects some patterns (structures) in the system, he/she/it can then employ such patterns as tools to simplify their understanding of the system. As an example, a gazelle which learns to correlate hearing a roaring to the presence of a lion, will be able to use it as warning and flee danger. Not being able to detect the pattern "roaring = lion close by" would require the gazelle to detect more subtle signs, possibly needing to employ more attention and thus more effort. In this setting the observer (gazelle) is "external" to the system (lion) it needs to analyze.

5.1.2. Intrinsic Emergence

In intrinsic emergence, the observer is "internal" to the system. Imagine a set of traders in an economy. The traders are locally connected via their trades, but no global information exchange exists. Once the traders identify an "emergent" feature, like the stock market, they can employ it to understand and affect the functioning of the system itself. The stock market becomes a mean for global information processing, which is performed by the agents (that is, the system itself) to affect their *own* functioning.

5.2. Information-Theoretic Interpretation

Given that a system can be viewed and studied at different levels, a natural question is "what level should we choose for our analysis"? A reasonable answer could be "the level at which it is easier or more efficient to construct a workable model." This idea has been captured formally by Shalizi [39] in the definition of Efficiency of Prediction. Within a computational mechanics [37] framework, Shalizi suggests:

$$e = \frac{E}{C_\mu}, \tag{33}$$

where e is the efficiency of prediction, E is the excess entropy, and C_μ the statistical complexity discussed earlier. The excess entropy can be seen as the mutual information between the past and future of a process, that is, the amount of information observed in the past which can be used to predict the future (i.e. which can be usefully coded in the agent instructions on how to behave in the future). Recalling that the statistical complexity is defined as the amount of information needed to reconstruct a process (that is equivalent to performing an optimal prediction), we can write informally:

$$e = \frac{\text{how much can be predicted}}{\text{how difficult it is to predict}}. \tag{34}$$

Given two levels of description of the same process, the approach Shalizi suggests is to choose for analysis the level which has larger efficiency of prediction e. At this level, either:

- we can obtain better predictability (understanding) of the system (E is larger), or
- it is much easier to predict because the system is simpler (C_μ is smaller), or
- we may lose a bit of predictability (E is smaller) but at the benefit of much larger gain in simplicity (C_μ is much smaller).

We can notice that this definition applies equally to pattern formation as well as to intrinsic emergence. In the case of pattern formation, we can envisage the scientist trying to determine what level of enquiry will provide a better model. At the level of intrinsic emergence, developing an efficient representation of the environment and of *its own functioning within the environment* gives a selective advantage to the agent, either because it provides for a better model, or because it provides for a similar model at a lower cost, enabling the agent to direct resources towards other activities.

5.3. Example—The Emergence of Thermodynamics

A canonical example of emergence without self-organization is described by Shalizi [39]: thermodynamics can emerge from statistical mechanics. The example considers a cubic centimeter of argon, which is conveniently spinless and monoatomic, at standard temperature and pressure, and sample the gas every nanosecond. At the micromechanical level, and at time intervals of 10^{-9} s, the dynamics of the gas are first-order Markovian, so each microstate is a causal state. The thermodynamic entropy (calculated as 6.6×10^{20} bits) gives the statistical complexity C_μ. The entropy rate h_μ of one cubic centimeter of argon at standard temperature and pressure is quoted to be around 3.3×10^{29} bits per second, or 3.3×10^{20} bits per nanosecond. Given the range of interactions $R = 1$ for a first-order Markov process, and the relationship $E = C_\mu - Rh_\mu$ [42], it follows that the efficiency of prediction $e = E/C_\mu$ is about 0.5 at this level. Looking at the macroscopic variables uncovers a dramatically different situation. The statistical complexity C_μ is given by the entropy of the macrovariable energy which is approximately 33.28 bits, whereas the entropy rate per millisecond is 4.4 bits (i.e. $h_\mu = 4.4 \times 10^3$ bits/second). Again, the assumption that the dynamics of the macrovariables are Markovian, and the relationship $E = C_\mu - Rh_\mu$ yield $e = E/C_\mu = 1 - Rh_\mu/C_\mu = 0.87$. If the time-step is a nanosecond, like at the micromechanical level, then $e \approx 1$, i.e. the efficiency of prediction approaches maximum. This allows Shalizi to conclude that "almost all of the information needed at the statistical—mechanical level is simply irrelevant thermodynamically," and given the apparent differences

Published 2008 Wiley Periodicals, Inc.
DOI 10.1002/cplx

in the efficiencies of prediction at two levels, "thermodynamic regularities are emergent phenomena, emerging out of microscopic statistical mechanics" [39].

6. ADAPTATION AND EVOLUTION

Adaptation is a process where the behavior of the system changes such that there is an increase in the mutual information between the system and a potentially complex and nonstationary environment. The environment is treated as a black box, meaning an adaptive system does not need to understand the underlying system dynamics to adapt. Stimulus response interactions provide feedback that modifies an internal model or representation of the environment, which affects the probability of the system taking future actions.

6.1. Concept

The three essential functions for an adaptive mechanism are generating variety, observing feedback from interactions with the environment, and selection to reinforce some interactions and inhibit others. Without variation, the system cannot change its behavior, and therefore it cannot adapt. Without feedback, there is no way for changes in the system to be coupled to the structure of the environment. Without preferential selection for some interactions, changes in behavior will not be statistically different to a random walk. First-order adaptation keeps sense and response options constant and adapts by changing only the probability of future actions. However, adaptation can also be applied to the adaptive mechanism itself [63]. Second-order adaptation introduces three new adaptive cycles: one to improve the way variety is generated, another to adapt the way feedback is observed and thirdly an adaptive cycle for the way selection is executed. If an adaptive system contains multiple autonomous agents using second-order adaptation, a third-order adaptive process can use variation, feedback, and selection to change the structure of interactions between agents.

From an information-theoretic perspective, variation decreases the amount of information encoded in the system, whereas selection acts to increase information. Because adaptation is defined to increase mutual information between a system and its environment, the information loss from variation must be less than the increase in mutual information from selection.

For the case that the system is a single agent with a fixed set of available actions, the environmental feedback is a single real valued reward plus the observed change in state at each time step, and the internal model is an estimate of the future value of each state, this model of first-order adaptation reduces to reinforcement learning (see for example [64]).

For the case that the system contains a population whose generations are coupled by inheritance with variation under selective pressure, the adaptive process reduces to evolution. Evolution is not limited to DNA/RNA based terrestrial biology, because other entities, including prions and artificial life programs, also meet the criteria for evolution. Provided a population of replicating entities can make imperfect copies of themselves, and not all the entities have an equal capacity to survive, the system is evolutionary. This broader conception of evolution has been coined universal Darwinism by Dawkins [65].

6.2. Information-Theoretic Interpretation

Adami [66] advocated the view that "evolution increases the amount of information a population harbors about its niche." In particular, he proposed physical complexity—a measure of the amount of information that an organism stores in its genome about the environment in which it evolves. Importantly, physical complexity for a population X (an ensemble of sequences) is defined in relation to a specific environment Z, as mutual information:

$$I(X,Z) = H_{\max} - H(X|Z), \qquad (35)$$

where $H_{\max}$ is the entropy in the absence of selection, i.e. the unconditional entropy of a population of sequences, and $H(X|Z)$ is the conditional entropy of X given Z, i.e. the diversity tolerated by selection in the given environment. When selection does not act, no sequence has an advantage over any other, and all sequences are equally probable in ensemble X. Hence, $H_{\max}$ is equal to the sequence length. In the presence of selection, the probabilities of finding particular genotypes in the population are highly nonuniform, because most sequences do not fit the particular environment. The difference between the two terms in (35) reflects the observation that "If you do not know which system your sequence refers to, then whatever is on it cannot be considered information. Instead, it is potential information (a.k.a. entropy)". In other words, this measure captures the difference between potential and selected (filtered) information:

$$\text{physical complexity} = \text{how much data can be stored} - \text{how much data irrelevant to environment is stored.} \qquad (36)$$

Adami stated that "physical complexity *is* information about the environment that can be used to make predictions about it" [66]. There is, however, a technical difference between physical complexity and predictive information, excess entropy, and statistical complexity. Although the latter three measure correlations within a single source, physical complexity measures correlation between two sources representing the system and its environment. However, it may be possible to represent the system and its environment as a single combined system by redefining the system boundary to include the environment. Then the correlations between

Published 2008 Wiley Periodicals, Inc.
DOI 10.1002/cplx

the system and its environment can be measured in principle by predictive information and/or statistical complexity. Comparing the representation (36) with the information transfer through networks, Eq. (26), as well as analogous information dynamics Eqs. (6), (10), and (19), we can observe a strong similarity: "how much data can be stored" is related to diversity of the combined system, while "how much data irrelevant to environment is stored" (or "how much conflicting data") corresponds to assortative noise within the combined system.

6.3. Example—Perception-Action Loops

The information transfer can also be interpreted as the acquisition of information from the environment by a single adapting individual: there is evidence that pushing the information flow to the information-theoretic limit (i.e. maximization of information transfer) can give rise to intricate behavior, induce a necessary structure in the system, and ultimately adaptively reshape the system [11, 12]. The central hypothesis of Klyubin et al. is that there exists "a local and universal utility function which may help individuals survive and hence speed up evolution by making the fitness landscape smoother," while adapting to morphology and ecological niche. The proposed general utility function, *empowerment*, couples the agent's sensors and actuators via the environment. Empowerment is the perceived amount of influence or control the agent has over the world, and can be seen as the agent's potential to change the world. It can be measured via the amount of Shannon information that the agent can "inject into" its sensor through the environment, affecting future actions and future perceptions. Such a perception-action loop defines the agent's actuation channel, and technically empowerment is defined as the capacity of this actuation channel: the maximum mutual information for the channel over all possible distributions of the transmitted signal. "The more of the information can be made to appear in the sensor, the more control or influence the agent has over its sensor"—this is the main motivation for this local and universal utility function [12]. Other examples highlighting the role of information transfer in guiding selection of spatiotemporally stable multicellular patterns, well-connected network topologies, multiagent swarms, and coordinated actuators in a modular robotic system are discussed in [14, 15, 67–69].

6.4. Summary

In short, natural selection increases physical complexity by the amount of information a population contains about its environment. Adami argued that physical complexity must increase in molecular evolution of asexual organisms in a single niche if the environment does not change, due to natural selection, and that "natural selection can be viewed as a filter, a kind of semipermeable membrane that lets information flow into the genome, but prevents it from flowing out." In general, however, information *may* flow out, and it is precisely this dynamic that creates larger feedback loops in the environment. As advocated by the interactionist approach to modern evolutionary biology [70], the organism–environment relationship is dialectical and reciprocal—again highlighting the role of assortativeness.

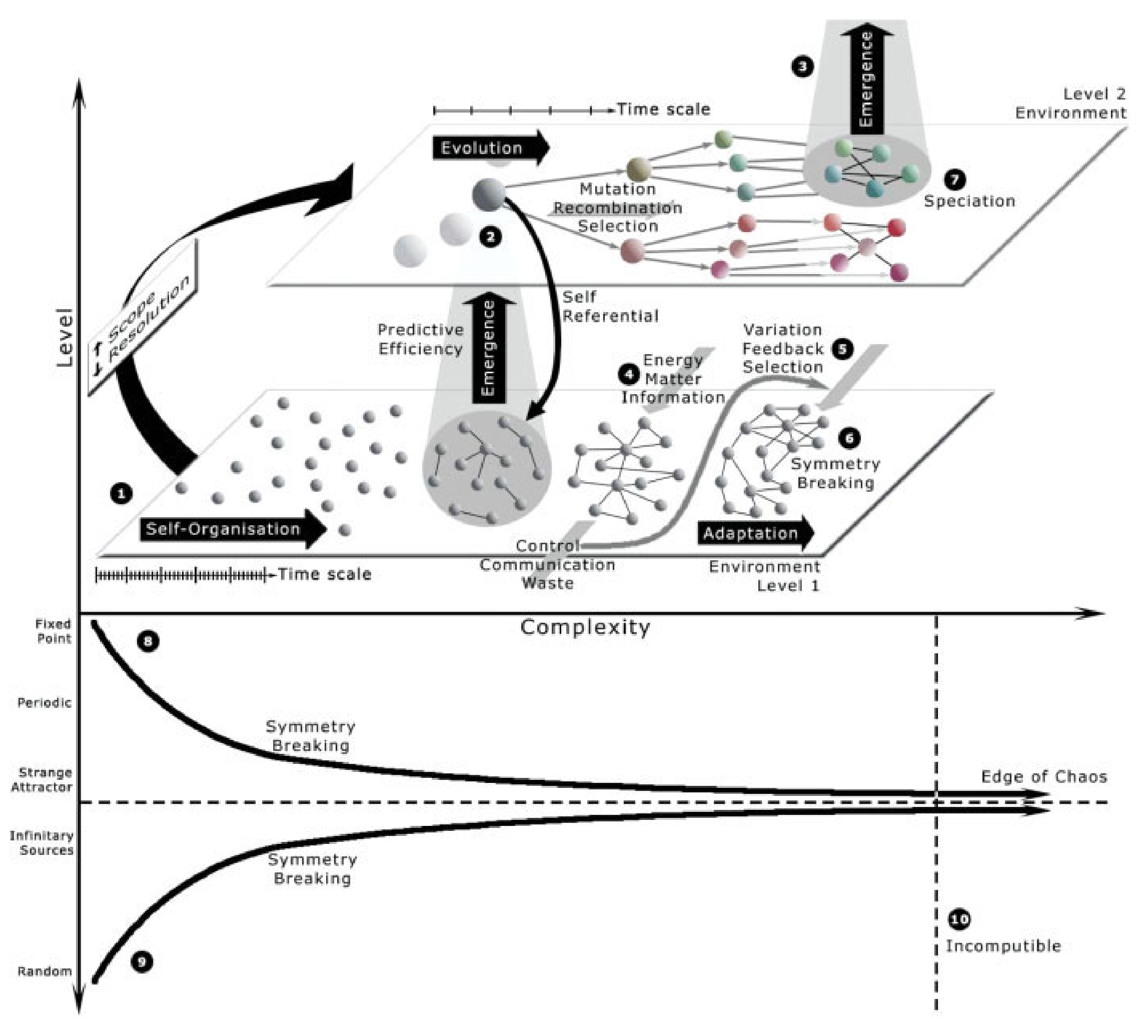
7. DISCUSSION AND CONCLUSIONS

By studying the processes which result from the local interaction of relatively simple components, Complex System Science has accepted the audacious aim of addressing problems which range from physics to biology, sociology, and ecology. It is not surprising that a common framework and language which enable practitioners of different field to communicate effectively is still lacking. As a possible contribution to this goal we have proposed a baseline using which concepts like complexity, emergence, and self-organization can be described, and most importantly, distinguished.

Figure 1 illustrates some relationships between the concepts introduced in this article. In particular, it shows two levels of an emergence hierarchy that are used to describe a complex system. The figure depicts dynamics that tend to increase complexity as arrows from left to right, and increases in the level of organization as arrows from bottom to top. The concepts can be related in numerical order as follows. (1) demonstrates self-organization, as components increase in organization over time. As the components become more organized, interdependencies arise constraining the autonomy of the components, and at some point it is more efficient to describe tightly coupled components as an emergent whole (or system). (2) depicts a lower resolution description of the whole, which may be self-referential if it causally affects the behavior of its components. Note that Level 2 has a longer time scale. The scope at this level is also increased, such that the emergent whole is seen as one component in a wider population. As new generations descend with modification through mutation and/or recombination, natural selection operates on variants and the population evolves; (3) shows that interactions between members of a population can lead to the emergence of higher levels of organisation: in this case, a species is shown. (4) emphasizes flows between the open system and the environment in the Level 1 description. Energy, matter and information enter the system, and control, communication and waste can flow back out into the environment. When the control provides feedback between the outputs and the inputs of the system in (5), its behavior can be regulated. When the feedback contains variation in the interaction between the system and its environment, and is subject to a selection pressure, the system adapts. Positive

Published 2008 Wiley Periodicals, Inc.
DOI 10.1002/cplx

FIGURE 1



A systems view of complex systems science concepts. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

feedback that reinforces variations at (6) results in symmetry breaking and/or phase transitions. (7) shows analogous symmetry breaking in Level 2 in the form of speciation.

Below the complexity axis, a complementary view of system complexity in terms of behavior, rather than organization, is provided. Fixed point behavior at (8) has low complexity, which increases for deterministic periodic and strange attractors [71, 72]. The bifurcation process is a form of symmetry breaking. Random behavior at (9) also has low complexity, which increases as the system's components become more organized into processes with "infinitary sources" [43]: e.g. positive-entropy-rate variations on the Thue-Morse process and other stochastic analogues of various context-free languages. The asymptote between (8) and (9) indicates the region where the complexity can grow without bound (it can also be interpreted as the "edge of chaos" [73]). Beyond some threshold of complexity at (10), the behavior is incomputable: it cannot be simulated in finite time on a Universal Turing Machine.

For our discussion we chose an information-theoretical framework. There are four primary reasons for this choice:

(1) it enables clear and consistent definitions and relationships between complexity, emergence, and self-organization in the physical world;
(2) the same concepts can equally be applied to biology;
(3) from a biological perspective, the basic ideas naturally extend to adaptation and evolution, which begins

Published 2008 Wiley Periodicals, Inc.
DOI 10.1002/cplx

to address the question of why complexity and self-organization are ubiquitous and apparently increasing in the biosphere; and

(4) it provides a unified setting, within which the description of relevant information channels provides significant insights of practical utility.

As noted earlier, once the information channels are identified by designers of a physical system (or naturally selected by interactions between a bio-system and its environment), the rest is mostly a matter of computation. This computation can be decomposed into "diversity" and "equivocation", as demonstrated in the discussed examples.

Information Theory is not a philosophical approach to the reading of natural processes, rather it comes with a set of tools to carry out experiments, make predictions, and computationally solve real-world problems. Like all toolboxes, its application requires a set of assumptions regarding the processes and conditions regarding data collections to be satisfied. Also, it is by definition biased towards a view of Nature as an immense information processing device. Whether this view and these tools can be successfully applied to the large variety of problems Complex Systems Science aims to address is far from obvious. Our intent, at this stage, is simply to propose it as a framework for a less ambiguous discussion among practitioners from different disciplines. The suggested interpretations of the concepts may be at best temporary place-holders in an evolving discipline—hopefully, the improved communication which can arise from sharing a common language will lead to deeper understanding, which in turn will enable our proposals to be sharpened, rethought, and even changed altogether.

ACKNOWLEDGEMENTS

This research was conducted under the CSIRO emergence interaction task (http://www.per.marine.csiro. au/ staff/ Fabio.Boschetti/CSS_emergence. htm) with support from the CSIRO Complex Systems Science Theme (http:// www.csiro.au/css). Thanks to Cosma Shalizi and Daniel Polani for their insightful contributions, and Eleanor Joel, our graphic design guru.

REFERENCES

1. Shannon, C.E. A mathematical theory of communication. The Bell Syst Tech J 1948, 27, 379–423, 623–656.
2. David, J.C.M. Information Theory, Inference, and Learning Algorithms; Cambridge University Press: Cambridge, UK, 2003.
3. Newman, M.E.J. Assortative mixing in networks. Phys Rev Lett 2002, 89, 208701.
4. Sole, R.V.; Valverde, S. Information theory of complex networks: on evolution and architectural constraints, In Complex Networks, Vol. 650: Lecture Notes in Physics; Ben-Naim, E.; Frauenfelder, H.; Toroczkai, Z., Eds.; Springer: Berlin, 2004.
5. Schreiber, T. Measuring information transfer. Phys Rev Lett 2000, 85, 461.
6. Lizier, J.T.; Prokopenko, M.; Zomaya, A.Y. Local information transfer as a spatiotemporal filter for complex systems. Phys Rev E 2008, 77, 026110.
7. Lungarella, M.; Sporns, O. Mapping information flow in sensorimotor networks. PLoS Comput Biol 2006, e144.
8. Ay, N.; Polani, D. Information flows in causal networks. Adv Complex Syst 2008, 11, 17–41.
9. Polani, D.; Nehaniv, C.; Martinetz, T.; Kim, J.T. Relevant information in optimized persistence vs. progeny strategies, In: Artificial Life X: Proceedings of the 10th International Conference on the Simulation and Synthesis of Living Systems; Rocha, L.M.; Yaeger, L.S.; Bedau, M.A.; Floreano, D.; Goldstone, R.L.; Vespignani, A., Eds.; Bloomington IN, USA, 2006.
10. Foreman, M.; Prokopenko, M.; Wang, P. Phase transitions in self-organising sensor networks, In: Advances in Artificial Life—Proceedings of the 7th European Conference on Artificial Life (ECAL), Vol. 2801: Lecture Notes in Artificial Intelligence; Banzhaf, W.; Christaller, T.; Dittrich, P.; Kim, J.T.; Ziegler, J., Eds.; Springer Verlag: Berlin, 2003, pp. 781–791.
11. Klyubin, A.S.; Polani, D.; Nehaniv, C.L. Organization of the information flow in the perception-action loop of evolved agents, In: Proceedings of 2004 NASA/DoD Conference on Evolvable Hardware; IEEE Computer Society, 2004, pp. 177–180.
12. Klyubin, A.S.; Polani, D.; Nehaniv, C.L. All else being equal be empowered. In: Advances in Artificial Life, 8th European Conference, ECAL 2005, Vol. 3630: LNCS; Capcarr'ere, M.S.; Freitas, A.A.; Bentley, P.J.; Johnson, C.G.; Timmis, J., Eds.; Springer: Berlin, 2005, pp. 744–753.
13. Nehaniv, C.L.; Polani, D.; Olsson, L.A.; Klyubin, A.S. Evolutionary information-theoretic foundations of sensory ecology: Channels of organism-specific meaningful information, In: Modeling Biology: Structures, Behaviour, Evolution, Vienna Series in Theoretical Biology; Fontoura Costa, L.; Müller, G.B., Eds.; MIT Press: Cambridge, MA, 2005.
14. Prokopenko, M.; Gerasimov, V.; Tanev, I. Evolving spatiotemporal coordination in a modular robotic system, In: From Animals to Animats 9: 9th International Conference on the Simulation of Adaptive Behavior (SAB 2006), Rome, Italy, September 25–29 2006, Vol. 4095: Lecture Notes in Computer Science, Nolfi, S., Baldassarre, G., Calabretta, R., Hallam, J., Marocco, D., Meyer, J.-A., Parisi, D., Eds.; Springer Verlag: Berlin, 2006, pp. 558–569.
15. Prokopenko, M.; Gerasimov, V.; Tanev, I. Measuring spatiotemporal coordination in a modular robotic system, In: Artificial Life X: Proceedings of The 10th International Conference on the Simulation and Synthesis of Living Systems; Rocha, L.M.; Yaeger, L.S.; Bedau, M.A.; Floreano, D.; Goldstone, R.L.; Vespignani, A., Eds.; Bloomington IN, USA, 2006, pp. 185–191.

16. Piraveenan, M.; Polani, D.; Prokopenko, M. Emergence of genetic coding: an information-theoretic model, In: Advances in Artificial Life: 9th European Conference on Artificial Life (ECAL-2007), Lisbon, Portugal, September 10–14, Vol. 4648: Lecture Notes in Artificial Intelligence; Almeida e Costa, F.; Rocha, L.M.; Costa, E.; Harvey, I.; Coutinho, A., Eds.; Springer: Berlin, 2007, pp. 42–52.
17. Adami, C. Introduction to Artificial Life; Springer: Berlin, 1998.
18. Wojciech H. Zurek, Ed. Valuable Information, Santa Fe Studies in the Sciences of Complexity; Addison-Wesley: Reading, Mass.: 1990.
19. Polani, D.; Prokopenko, M.; Chadwick, M. Modelling stigmergic gene transfer, In: Artificial Life XI: Proceedings of The 11th International Conference on the Simulation and Synthesis of Living Systems; Bullock, S.; Noble, J.; Watson, R.; Bedau, M., Eds.; MIT Press: Winchester, UK, 2008.
20. Boschetti, F. Mapping the complexity of ecological models. Ecol Complexity 2008, 5, 37–47.
21. Li, M.; Vitanyi, P.M.B. An Introduction to Kolmogorov Complexity and its Applications, 2nd ed.; Springer-Verlag: New York, 1997.
22. Chaitin, G.J. Algorithmic Information Theory; Cambridge University Press: Cambridge, UK, 1987.
23. Chaitin, G.J. Information-theoretic limitations of formal systems. J ACM 1974, 21, 403–424.
24. Crutchfield, J.P.; Young, K. Inferring statistical complexity. Phys Rev Lett 1989, 63, 105–108.
25. Grassberger, P. Toward a quantitative theory of selfgenerated complexity. Int J Theor Phys 1986, 25, 907–938.
26. Boffetta, G.; Cencini, M.; Falcioni, M.; Vulpiani, A. Predictability: A way to characterize complexity. Phys Rep 2002, 356, 367–474.
27. Bialek, W.; Nemenman, I.; Tishby, N. Complexity through nonextensivity. Physica A 2001, 302, 89–99.
28. Crutchfield, J.P.; Packard, N.H. Symbolic dynamics of noisy chaos. Physica D 1983, 7, 201–223.
29. Crutchfield, J.P.; Feldman, D.P. Statistical complexity of simple one-dimensional spin systems. Phys Rev E 1997, 55, R1239–R1243.
30. Shaw, R. The Dripping Faucet as a Model Chaotic System; Aerial Press: Santa Cruz, California, 1984.
31. Lindgren, K.; Norhdal, M.G. Complexity measures and cellular automata. Complex Syst 1988, 2, 409–440.
32. Eriksson, K-E.; Lindgren, K. Structural information in self-organizing systems. Phys Scripta 1987, 35, 388–397.
33. Li, W. On the relationship between complexity and entropy for Markov chains and regular languages. Complex Syst 1991, 5, 381–399.
34. Arnold, D. Information-theoretic analysis of phase transitions. Complex Systems 1996, 10, 143–155.
35. Bialek, W.; Nemenman, I.; Tishby, N. Predictability, complexity, and learning. Neur Comp 2001, 13, 2409–2463.
36. James P. Crutchfield; Cosma Rohilla Shalizi Thermodynamic depth of causal states: Objective complexity via minimal representations. Phys Rev E 1999, 59, 275–283.
37. Shalizi, C.R.; Crutchfield, J.P. Computational mechanics: Pattern and prediction, structure and simplicity. J Stat Phys 2001, 104, 819–881.
38. Shalizi, C.R.; Shalizi, K.L. Optimal nonlinear prediction of random fields on networks. Discrete Math Theor Comput Sci 2003, AB(DMCS), 11–30.
39. Shalizi, C. Causal architecture, complexity and self-organization in time series and cellular automata. PhD thesis, University of Michigan, 2001.
40. Varn, D.P. Language extraction from ZnS. Phd thesis, University of Tennessee, 2001.
41. Shalizi, C.R.; Shalizi, K.L. Blind construction of optimal nonlinear recursive predictors for discrete sequences, In: Uncertainty in Artificial Intelligence: Proceedings of the Twentieth Conference; Chickering, M.; Joseph Halpern, J., Eds.; AUAI Press: Arlington, Virginia, 2004, pp. 504–511.
42. Feldman, D.P.; Crutchfield, J.P. Discovering noncritical organization: Statistical mechanical, information theoretic, and computational views of patterns in one-dimensional spin systems, Technical Report 98-04-026, SFI Working Paper 1998.
43. Crutchfield, J.P.; Feldman, D.P. Regularities unseen, randomness observed: Levels of entropy convergence. Chaos 2003, 13, 25–54.
44. Kolmogorov, A.N. Entropy per unit time as a metric invariant of automorphisms. Doklady Akademii Nauk SSSR 1959, 124, 754–755.
45. Ebeling, W. Prediction and entropy of nonlinear dynamical systems and symbolic sequences with lro. Physica D 1997, 109, 42–52.
46. Feldman, D.P.; Crutchfield, J.P. Structural information in two-dimensional patterns: Entropy convergence and excess entropy. Phys Rev E 2003, 67, 051104.
47. Cover, T.M.; Thomas, J.A. Elements of Information Theory; Wiley: New York, 1991.
48. Erdos, P.; Renyi, A. On the strength of connectedness of random graphs. Acta Math Sci Hungary 1961, 12, 261–267.
49. Callaway, D.S.; Hopcroft, J.E.; Kleinberg, J.M.; Newman, M.E.; Strogatz, S.H. Are randomly grown graphs really random? Phys Rev E October 2001, 64(4 Pt 1), 041902.
50. Ash, R.B. Information theory; Dover: London, 1965.
51. T De Wolf; Holvoet, T. Emergence versus self-organisation: Different concepts but promising when combined. In: Engineering Self-Organising Systems; Brueckner, S.; Serugendo, G.D.M.; Karageorgos, A.; Nagpal, R., Eds.; Springer: Berlin, 2005, pp. 1–15.
52. Shalizi, C.R.; Shalizi, K.L.; Haslinger, R. Quantifying self-organization with optimal predictors. Phys Rev Lett 2004, 93, 11870114.
53. Halley, J.; Winkler, D. Consistent concepts of self-organization and self-assembly. Complexity, in press.
54. Kauffman, S.A. Investigations; Oxford University Press: Oxford, 2000.
55. Correia, L. Self-organisation: a case for embodiment, In: Proceedings of The Evolution of Complexity Workshop at Artificial Life X: The 10th International Conference on the Simulation and Synthesis of Living Systems, 2006, pp. 111–116.
56. Wagner, A. Robustness and Evolvability in Living Systems; Princeton University Press: Princeton, NJ, 2005.
57. Ay, N.; Flack, J.; Krakauer, D. Robustness and complexity co-constructed in multimodal signalling networks. Philos Trans R Soc B 2007, 362, 441–447.
58. Hubbell, S.P.; Johnson, L.K.; Stanislav, E.; Wilson, B.; Fowler, H. Foraging by bucket-brigade in leafcutter ants. Biotropica 1980, 12, 210–213.
59. Rényi, A. Probability theory; North-Holland: Amsterdam, 1970.
60. Corning, P.A. The re-emergence of "emergence": A venerable concept in search of a theory. Complexity 2002, 7, 18–30.
61. Boschetti, F.; Prokopenko, M.; Macreadie, I.; Grisogono, A.-M. Defining and detecting emergence in complex networks, In: Knowledge-Based Intelligent Information and Engineering Systems, 9th International Conference, KES 2005, Melbourne, Australia, September 14–16, 2005; Proceedings, Part IV; Khosla, R.; Howlett, R.J.; Jain, L.C., Eds.; Vol. 3684: Lecture Notes in Computer Science; Springer: Berlin, 2005, pp. 573–580.
62. Crutchfield, J. The calculi of emergence: Computation, dynamics, and induction. Physica D 1994, 75, 11–54.
63. Grisogono, A.-M. Co-adaptation, In: SPIE Symposium on Microelectronics, MEMS and Nanotechnology, Vol. Paper 6039-1; Brisbane, Australia, 2005.
64. Sutton, R.S.; Barto, A.G. Reinforcement Learning: An Introduction, A Bradford Book; The MIT Press: Cambridge, MA, 1998.
65. Dawkins, R. Universal Darwinism, In: Evolution from Molecules to Men, Bendall, D.S., Ed.; Cambridge University Press, 1983.
66. Adami, C. What is complexity? Bioessays 2002, 24, 1085–1094.
67. Prokopenko, M.; Wang, P.; Price, D.C.; Valencia, P.; Foreman, M.; Farmer, A.J. Self-organizing hierarchies in sensor and communication networks. Artificial Life, Special Issue on Dynamic Hierarchies 2005, 11, 407–426.

Published 2008 Wiley Periodicals, Inc.
DOI 10.1002/cplx

68. Prokopenko, M.; Wang, P.; Foreman, M.; Valencia, P.; Price, D.C.; Poulton, G.T. On connectivity of reconfigurable impact networks in ageless aerospace vehicles. Journal of Robotics and Autonomous Systems 2005, 53, 36–58.
69. Mathews, G.; Durrant-Whyte, H.; Prokopenko, M. Measuring global behaviour of multi-agent systems from pair-wise mutual information. In: Knowledge-Based Intelligent Information and Engineering Systems, 9th International Conference, KES 2005, Melbourne, Australia, September 14–16, 2005; Proceedings, Part IV; Khosla, R.; Howlett, R.J.; Jain, L.C., Eds.; Vol. 3684: Lecture Notes in Computer Science; Springer: Berlin, 2005, pp. 587–594.
70. Ridley, M. Nature Via Nurture: Genes, Experience and What Makes Us Human; Fourth Estate, 2003.
71. Wolfram, S. Universality and complexity in cellular automata. Physica D 1984, 10, 1–35.
72. Casti, J.L. Chaos, godel and truth, In: Beyond Belief: Randomness, Prediction, and Explanation in Science; Casti, J.L.; Karlqvist, A., Eds.; CRC Press: London, 1991.
73. Langton, C. Computation at the edge of chaos: Phase transitions and emergent computation, In Emergent Computation; Forest, S. Ed.; MIT: Cambridge, MA, 1991.

Published 2008 Wiley Periodicals, Inc.
DOI 10.1002/cplx