entropy

MDPI

Article

Multiscale Information Theory and the Marginal Utility of Information

Benjamin Allen [1,2,*], Blake C. Stacey [3,4] and Yaneer Bar-Yam [4]

[1] Department of Mathematics, Emmanuel College, Boston, MA 02115, USA
[2] Program for Evolutionary Dynamics, Harvard University, Cambridge, MA 02138, USA
[3] Department of Physics, University of Massachusetts-Boston, Boston, MA 02125, USA; bstacey@sunclipse.org
[4] New England Complex Systems Institute, Cambridge, MA 02139, USA; yaneer@necsi.edu
[*] Correspondence: benjcallen@gmail.com

Received: 28 February 2017; Accepted: 9 June 2017; Published: 13 June 2017

Abstract: Complex systems display behavior at a range of scales. Large-scale behaviors can emerge from the correlated or dependent behavior of individual small-scale components. To capture this observation in a rigorous and general way, we introduce a formalism for multiscale information theory. Dependent behavior among system components results in overlapping or shared information. A system's structure is revealed in the sharing of information across the system's dependencies, each of which has an associated scale. Counting information according to its scale yields the quantity of scale-weighted information, which is conserved when a system is reorganized. In the interest of flexibility we allow information to be quantified using any function that satisfies two basic axioms. Shannon information and vector space dimension are examples. We discuss two quantitative indices that summarize system structure: an existing index, the complexity profile, and a new index, the marginal utility of information. Using simple examples, we show how these indices capture the multiscale structure of complex systems in a quantitative way.

Keywords: complexity; complex systems; entropy; information; scale

1. Introduction

The field of complex systems seeks to identify, understand and predict common patterns of behavior across the physical, biological and social sciences [1–7]. It succeeds by tracing these behavior patterns to the structures of the systems in question. We use the term "structure" to mean the totality of quantifiable relationships, or *dependencies*, among the components comprising a system. Systems from different domains and contexts can share key structural properties, causing them to behave in similar ways. For example, the central limit theorem tells us that the sum over many independent random variables yields an aggregate value whose probability distribution is well-approximated as a Gaussian. This helps us understand systems composed of statistically independent components, whether those components are molecules, microbes or human beings. Likewise, different chemical elements and compounds display essentially the same behavior near their respective critical points. The critical exponents which encapsulate the thermodynamic properties of a substance are the same for all substances in the same universality class, and membership in a universality class depends upon structural features such as dimensionality and symmetry properties, rather than on details of chemical composition [8].

Outside of known universality classes, identifying the key structural features that dictate the behavior of a system or class of systems often relies upon an *ad hoc* leap of intuition. This becomes particularly challenging for complex systems, where the set of system components is not only large, but also interwoven and resistant to decomposition. Information theory [9,10] holds promise as a general tool for quantifying the dependencies that comprise a system's structure [11]. We can consider the amount of information that would be obtained from observations of any component or set of components. Dependencies mean that one observation can be fully or partially inferred from another, thereby reducing the amount of joint information present in a set of components, compared to the amount that would be present without those dependencies. Information theory allows one to quantify not only fixed or rigid relationships among components, but also "soft" relationships that are not fully determinate, e.g., statistical or probabilistic relationships.

However, traditional information theory is primarily concerned with amounts of independent bits of information. Consequently, each bit of non-redundant information is regarded as equally significant, and redundant information is typically considered irrelevant, except insofar as it provides error correction [10,12]. These features of information theory are natural in applications to communication, but present a limitation when characterizing the structure of a physical, biological, or social system. In a system of celestial bodies, the same amount of information might describe the position of a moon, a planet, or a star. Likewise, the same amount of information might describe the velocity of a solitary grasshopper, or the mean velocity of a locust swarm. A purely information-theoretic treatment has no mechanism to represent the fact that these observables, despite containing the same amount of information, differ greatly in their significance.

Overcoming this limitation requires a multiscale approach to information theory [13–18]—one that identifies not only the *amount* of information in a given observable but also its *scale*, defined as the number or volume of components to which it applies. Information describing a star's position applies at a much larger scale than information describing a moon's position. In shifting from traditional information theory to a multiscale approach, redundant information becomes not irrelevant but crucial: Redundancy among smaller-scale behaviors gives rise to larger-scale behaviors. In a locust swarm, measurements of individual velocity are highly redundant, in that the velocity of all individuals can be inferred with reasonable accuracy by measuring the velocity of just one individual. The multiscale approach, rather than collapsing this redundant information into a raw number of independent bits, identifies this information as large-scale and significant precisely because it is redundant across many individuals.

The multiscale approach to information theory also sheds light on a classic difficulty in the field of complex systems: to clarify what a "complex system" actually is. Naïvely, one might think to define complex systems as those that display the highest complexity, as quantified using Shannon information or other standard measures. However, the systems deemed the most "complex" by these measures are those in which the components behave independently of each other, such as ideal gases. Such systems lack the multiscale regularities and interdependencies that characterize the systems typically studied by complex systems researchers. Some theorists have argued that true complexity is best viewed as occupying a position between order and randomness [19–21]. Music is, so the argument goes, intermediate between still air and white noise. But although complex systems contain both order and randomness, they do not appear to be mere blends of the two. A more satisfying answer is that complex systems display behavior across a wide range of scales. For example, stock markets can exhibit small-scale behavior, as when an individual investor sells a small number of shares for reasons unrelated to overall market activity. They can also exhibit large-scale behavior, e.g., a large institutional investor sells many shares [22], or many individual investors sell shares simultaneously in a market panic [23].

Formalizing these ideas requires a synthesis of statistical physics and information theory. Statistical physics [24–27]—in particular the renormalization group of phase transitions [28,29]—provides a notion of scale in which individual components acting in concert can be considered equivalent to larger-scale units. Information theory provides the tool of *multivariate mutual information*: the information shared among an arbitrary number of variables (also called *interaction information* or *co-information*) [13–17,30–39]. These threads were combined in the *complexity profile* [13–18], a quantitative index of structure that characterizes the amount of information applying at a given scale or higher. In the context of the complexity profile, the multivariate mutual information of a set of n

variables is considered to have scale n. In this way, information is understood to have scale equal to the multiplicity (or redundancy) at which it arises—an idea which is also implicit in other works on multivariate mutual information [40,41].

Here we present a mathematical formalism for multiscale information theory, for use in quantifying the structure of complex systems. Our starting point is the idea discussed in the previous paragraph, that information has scale equal to the multiplicity at which it arises. We formalize this idea mathematically and generalize it in two directions: First, we allow each system component to have an arbitrary intrinsic scale, reflecting its inherent size, volume, or multiplicity. For example, the mammalian muscular system includes both large and small muscles, corresponding to different scales of environmental challenge (e.g., pursuing prey and escaping from predators versus chewing food) [42]. Scales are additive in our formalism, in the sense that a set of components acting in perfect coordination is formally equivalent to a single component with scale equal to the sum of the scales of the individual components. This equivalence can greatly simplify the representation of a system. Consider, for example, an avalanche consisting of differently-sized rocks. To represent this avalanche within the framework of traditional (single-scale) information theory, one must either neglect the differences in size (thereby diminishing the utility of the representation) or else model each rock by a collection of myriad statistical variables, each corresponding to a equally-sized portion. Our formalism, by incorporating scale as a fundamental quantity, allows each rock to be represented in a direct and physically meaningful way.

Second, in the interest of generality, we use a new axiomatized definition of information, which encompasses traditional measures such as Shannon information as well as other quantifications of freedom or indeterminacy. In this way, our formalism is applicable to system representations for which traditional information measures cannot be used.

Using these concepts of information and scale, we identify how a system's joint information is distributed across each of its *irreducible dependencies*—relationships among some components conditional on all others. Each irreducible dependency has an associated scale, equal to the sum of the scales of the components included in this dependency. This formalizes the idea that any information pertaining to a system applies at a particular scale or combination of scales. Multiplying quantities of information by the scales at which they apply yields the *scale-weighted information*, a quantity that is conserved when a system is reorganized or restructured.

We use this multiscale formalism to develop quantitative indices that summarize important aspects of a system's structure. We generalize the complexity profile to allow for arbitrary intrinsic scales and a variety of information measures. We also introduce a new index, the *marginal utility of information* (MUI), which characterizes the extent to which a system can be described using limited amounts of information. The complexity profile and the MUI both capture a tradeoff of complexity versus scale that is present in all systems.

Our basic definitions of information, scale, and systems are presented in Sections 2–4, respectively. Sections 5 formalizes the multiscale approach to information theory by defining the information and scale of each of a system's dependencies. Sections 6 and 7 discuss our two indices of structure. Section 8 establishes a mathematical relationship between these two indices for a special class of systems. Section 9 applies our indices of structure to the noisy voter model [43]. We conclude in Sections 10 and 11 by discussing connections between our formalism and other work in information theory and complex systems science.

2. Information

We begin by introducing a generalized, axiomatic notion of information. Conceptually, information specifies a particular entity out of a set of possibilities and thus enables us to describe or characterize that entity. Information measures such as Shannon information quantify the amount of resources needed in this specification. Rather than adopting a specific information measure, we consider that the amount of information may be quantified in different ways, each appropriate to different contexts.

Let A be the set of components in a system. An *information function*, H, assigns a nonnegative real number to each subset $U \subset A$, representing the amount of information needed to describe the components in U. (Throughout, the subset notation $U \subset A$ includes the possibility that $U = A$.) We require that an information function satisfy two axioms:

- *Monotonicity:* The information in a subset U that is contained in a subset V cannot have more information than V, that is, $U \subset V \Rightarrow H(U) \leq H(V)$.
- *Strong subadditivity:* Given two subsets, the information contained in both cannot exceed the information in each of them separately minus the information in their intersection:

$$H(U \cup V) \leq H(U) + H(V) - H(U \cap V). \tag{1}$$

Strong subadditvity expresses how information combines when parts of a system (U and V) are regarded as a whole ($U \cup V$). Information regarding U may overlap with information regarding V for two reasons. First, U and V may share components; this is corrected for by subtracting $H(U \cap V)$. Second, constraints in the behavior of non-shared components may reduce the information needed to describe the whole. Thus, information describing the whole may be reduced due to overlaps or redundancies in the information applying to different parts, but it cannot be increased.

In contrast to other axiomatizations of information, which uniquely specify the Shannon information [9,44–46] or a particular family of measures [47–50], the two axioms above are compatible with a variety of different measures that quantify information or complexity:

- *Microcanonical or Hartley entropy*: For a system with a finite number of joint states, $H_0(U) = \log m$ is an information function, where m is the number of joint states available to the subset U of components. Here, information content measures the number of yes-or-no questions which must be answered to identify one joint state out of m possibilities.
- *Shannon entropy*: For a system characterized by a probability distribution over all possible joint states, $H(U) = -\sum_{i=1}^{m} p_i \log p_i$ is an information function, where $p_1, \ldots, p_m$ are the probabilities of the joint states available to the components in U [9]. Here, information content measures the number of yes-or-no questions which must be answered to identify one joint state out of all the joint states available to U, where more probable states can be identified more concisely.
- *Tsallis entropy*: The Tsallis entropy [51,52] is a generalization of the Shannon entropy with applications to nonextensive statistical mechanics. For the same setting as in Shannon entropy, Tsallis entropy is defined as $H_q(U) = -\sum_{i=1}^{m} p_i^q (p_i^{1-q} - 1)/(1-q)$ for some parameter $q \geq 0$. Shannon entropy is recovered in the limit $q \to 1$. Tsallis entropy is an information function for $q \geq 1$ (but not for $q < 1$); this follows from Proposition 2.1 and Theorem 3.4 of [53].
- *Logarithm of period*: For a deterministic dynamic system with periodic behavior, an information function $L(U)$ can be defined as the logarithm of the period of a set U of components (i.e., the time it takes for the joint state of these components to return to an initial joint state) [54]. This information function measures the number of questions which one should expect to answer in order to locate the position of those components in their cycle.
- *Vector space dimension*: Consider a system of n components, each of whose state is described by a real number. Then the joint states of any subset U of $m \leq n$ components can be described by points in some linear subspace of $\mathbb{R}^m$. The minimal dimension $d(U)$ of such a subspace is an information function, equal to the number of coordinates one must specify in order to identify the joint state of U.
- *Matroid rank*: A matroid consists of a set of elements called the *ground set*, together with a *rank function* that takes values on subsets of the ground set. Rank functions are defined to include the monotonicity and strong subadditivity properties [55], and generalize the notion of vector

subspace dimension. Consequently, the rank function of a matroid is an information function, with the ground set identified as the set of system components.

In principle, measurements of algorithmic complexity may also be regarded as information functions. For example, when a subset U can be encoded as a binary string, the algorithmic complexity $H(U)$ can be quantified as the length of the shortest self-delimiting program producing this string, with respect to some universal Turing machine [56]. Information content then measures the number of machine-language instructions which must be given to reconstruct U. Algorithmic complexity—at least under certain formulations—obeys versions of the monotonicity and strong subadditivity axioms [56,57]. However, while conceptually clean, this definition is difficult to apply quantitatively. First, the algorithmic complexity is only defined up to a constant which depends on the choice of universal Turing machine. Second, as a consequence of the halting problem, algorithmic complexity can only be bounded, not computed exactly.

3. Scale

A defining feature of complex systems is that they exhibit nontrivial behavior on multiple scales [1,13,14]. While the term "scale" has different meanings in different scientific contexts, we use the term scale here in the sense of the number of entities or units acting in concert, with each involved entity potentially weighted according to a measure of importance.

For many systems, it is reasonable to regard all components as having *a priori* equal scale. In this case we may choose the units of scale so that each component has scale equal to 1. This convention was used in previous work on the complexity profile [13–17]. However, it is in many cases necessary to represent the components of a system as having different intrinsic scales, reflecting their built-in size, multiplicity or redundancy. For example, in a system of many physical bodies, it may be natural to identify the scale of each body as a function of its mass, reflecting the fact that each body comprises many molecules moving in concert. In a system of investment banks [58–60], it may be desirable to assign weight to each bank according to its volume of assets. In these cases, we denote the *a priori* scale of a system component a by a positive real number $\sigma(a)$, defined in terms of some meaningful scale unit.

Scales are additive, in the sense that a set of completely interdependent components can be replaced by a single component whose scale is equal to the sum of the scales of the individual components. We describe this property formally in Section 5.4 and in Appendix B.

4. Systems

We formally define a *system* $\mathcal{A}$ to comprise three elements:

- A finite set A of components,
- An information function $H_{\mathcal{A}}$, giving the information in each subset $U \subset A$,
- A scale function $\sigma_{\mathcal{A}}$, giving the intrinsic scale of each compoent $a \in A$.

The choice of information and scale functions will reflect how the system is modeled mathematically, and the kind of statements we can make about its structure. We omit the subscripts from H and σ when only one system is under consideration.

In this work, we treat the three elements of a system as unchanging, even though the system itself may be dynamic (existing in a sequence of states through time). A dynamic system can be represented as a set of time histories, or—using the approach of ergodic theory—by defining a probability distribution over states with probabilities corresponding to frequencies of occupancy over extended periods of time. The methods outlined here could also be used to explore the dynamics or histories of a system's structure, using information and scale functions whose values vary as relationships change within a system over time. However, our current work focuses only on the static or time-averaged properties of a system.

In requiring that the set A of components be finite, we exclude, for example, systems represented as continuum fields, in which each point in a continuous space might be regarded as a component. While the concepts of multiscale information theory may still be useful in thinking about such systems, the mathematical representation of these concepts presents challenges that are beyond the scope of this work.

We shall use four simple systems as running examples. Each consists of three binary random variables, each having intrinsic scale one.

- *Example* ***A****: Three independent components.* Each component is equally likely to be in state 0 or state 1, and the system as a whole is equally likely to be in any of its eight possible states.
- *Example* ***B****: Three completely interdependent components.* Each component is equally likely to be in state 0 or state 1, but all three components are always in the same state.
- *Example* ***C****: Independent blocks of dependent components.* Each component is equally likely to take the value 0 or 1; however, the first two components always take the same value, while the third can take either value independently of the coupled pair.
- *Example* ***D****: The* $2+1$ *parity bit system.* The components can exist in the states 110, 101, 011, or 000 with equal probability. In each state, each component is equal to the parity (0 if even; 1 if odd) of the sum of the other two. Any two of the component are statistically independent of each other, but the three as a whole are constrained to have an even sum.

We define a *subsystem* of $\mathcal{A} = (A, H_{\mathcal{A}}, \sigma_{\mathcal{A}})$ as a triple $\mathcal{B} = (B, H_{\mathcal{B}}, \sigma_{\mathcal{B}})$, where B is a subset of A, $H_{\mathcal{B}}$ is the restriction of $H_{\mathcal{A}}$ to subsets of B, and $\sigma_{\mathcal{B}}$ is the restriction of $\sigma_{\mathcal{A}}$ to elements of B.

5. Multiscale Information Theory

Here we formalize the multiscale approach to information theory. We begin by introducing notation for dependencies. We then identify how information is shared across irreducible dependencies, generalizing the notion of multivariate mutual information [13–17,30–39] to an arbitrary information function. We then define the scale of a dependency and introduce the quantity of scale-weighted information. Finally, we formalize the key concepts of independence and complete interdependence.

5.1. Dependencies

A *dependency* among a collection of components $a_1, \ldots, a_m$ is the relationship (if any) among these components such that the behavior of some of the components is in part obtainable from the behavior of others. We denote this dependency by the expression $a_1; \ldots; a_m$. This expression represents a relationship, rather than a number or quantity. We use a semicolon to keep our notation consistent with information theory (in particular, with multivariate mutual information; see Section 5.2).

We can identify a more general concept of *conditional dependencies*. Consider two disjoint sets of components $a_1, \ldots, a_m$ and $b_1, \ldots, b_k$. The conditional dependency $a_1; \ldots; a_m | b_1, \ldots, b_k$ represents the relationship (if any) between $a_1, \ldots, a_m$ such that the behavior of some of these components can yield improved inferences about the behavior of others, relative to what could be inferred from the behavior of $b_1, \ldots, b_k$. We call this the dependency of $a_1, \ldots, a_m$ *given* $b_1, \ldots, b_k$, and we say $a_1, \ldots, a_m$ are *included* in this dependency, while $b_1, \ldots, b_k$ are *excluded*.

We call a dependency *irreducible* if every system component is either included or excluded. We denote the set of all irreducible dependencies of a system $\mathcal{A}$ by $\mathfrak{D}_{\mathcal{A}}$. A system's dependencies can be organized in a Venn diagram, which we call a *dependency diagram* (Figure 1).

The relationship between the components and dependencies of $\mathcal{A}$ can be captured by a mapping, which we denote by δ, from A to subsets of $\mathfrak{D}_{\mathcal{A}}$. A component $a \in A$ maps to the set of irreducible dependencies that include a (or in visual terms, the region of the dependency diagram that corresponds to component a). For example, in a system of three components a, b, c, we have

$$\delta(a) = \{a;b;c, \quad a;b|c, \quad a;c|b, \quad a|b,c\}. \tag{2}$$

We extend the domain of δ to subsets of components, by mapping each subset $U \subset A$ onto to the set of all irreducible dependencies that include at least one element of U; for example,

$$\delta(\{a,b\}) = \{a;b;c, \quad a;b|c, \quad a;c|b, \quad b;c|a, \quad a|b,c, \quad b|a,c\}. \tag{3}$$

Visually, $\delta(\{a,b\})$ is the union of the circles representing a and b in the dependency diagram. Finally, we extend the domain of δ to dependencies, by mapping the dependency $a_1;\ldots;a_m|b_1,\ldots,b_k$ onto the set of all irreducible dependencies that include $a_1,\ldots,a_m$ and exclude $b_1,\ldots,b_k$; for example,

$$\delta(a|c) = \{a;b|c, \quad a|b,c\}. \tag{4}$$

Visually, $\delta(a|c)$ consists of the regions corresponding to a but not to c.

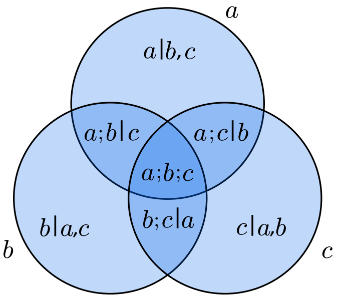


Figure 1. The dependency diagram of a system with three components, a, b and c, represented by the interiors of the three circles. The seven irreducible dependencies shown above correspond to the seven interior regions of the Venn diagram encompassed by the boundaries of the three circles. Irreducible dependencies are shaded according to their scale, assuming that each component has scale one. Reducible dependencies such as $a|b$ are not shown.

5.2. Information Quantity in Dependencies

Here we define the shared information, $I_{\mathcal{A}}(x)$, of a dependency x in a system $\mathcal{A}$. I generalizes the multivariate mutual information [13–17,30–39] to an arbitrary information function H. We note that H and I characterize the same quantity—information—but are applied to different kinds of arguments: H is applied to subsets of components of $\mathcal{A}$, while I is applied to dependencies.

The values of I on irreducible dependencies x of $\mathcal{A}$ are uniquely determined by the system of equations

$$\sum_{x \in \delta(U)} I(x) = H(U) \qquad \text{for all subsets } U \subset A \tag{5}$$

As U runs over all subsets of A, the resulting system of equations determines the values $I(x)$, $x \in \mathfrak{D}_{\mathcal{A}}$, in terms of the values $H(U)$, $U \subset A$. The solution is an instance of the inclusion-exclusion principle [61], and can also be obtained by Gaussian elimination. An explicit formula obtained in the context of Shannon information [32] applies as well to any information function. Figure 2 shows the information in each irreducible dependecy for our four running examples.

We extend I to dependencies that are not irreducible by defining the shared information $I(x)$ to be equal to the sum of the values of $I(y)$ for all irreducible dependencies y encompassed by a dependency x:

$$I(x) = \sum_{y \in \delta(x)} I(y). \tag{6}$$

Our notation corresponds to that of Shannon information theory. For example, in a system of two components a and b, solving (5) yields

$$I(a;b) = H(a) + H(b) - H(a,b). \tag{7}$$

Above, $H(a,b)$ is shorthand for $H(\{a,b\})$; we use similar shorthand throughout. Equation (7) coincides with the classical definition of mutual information [9,10], with H representing joint Shannon information. Similarly, $I(a_1|b_1,\ldots,b_k)$ is the conditional entropy of a_1 given $b_1,\ldots,b_k$, and $I(a_1;a_2|b_1,\ldots,b_k)$ is the conditional mutual information of a_1 and a_2 given $b_1,\ldots,b_k$. For a dependency including more than two components, $I(x)$ is the multivariate mutual information (also called interaction information or co-information) of the dependency x [30–33,36–39].

For any information function H, we observe that the information of one component conditioned on others, the conditional information $I(a_1|b_1,\ldots,b_k)$ is nonnegative due to the monotonicity axiom. Likewise, the mutual information of two components conditioned on others, $I(a_1;a_2|b_1,\ldots,b_k)$, is nonnegative due to the strong subadditivity axiom. However, the information shared among three or more components can be negative. This is illustrated in running Example **D**, for which the tertiary shared information $I(a;b;c)$ is negative (Figure 2D). Such negative values appear to capture an important property of dependencies, but their interpretation is the subject of continuing discussion [34,36,37,62,63].

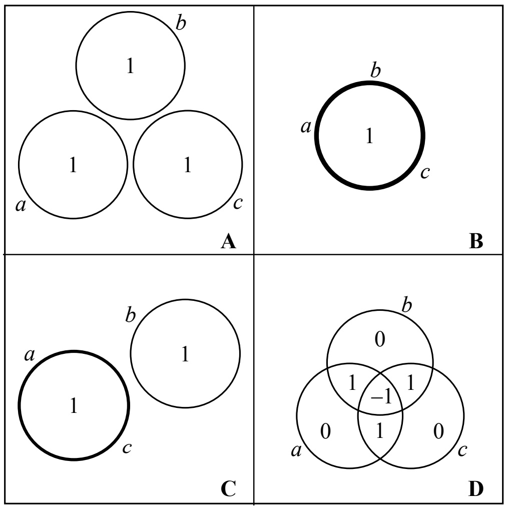


Figure 2. Dependency diagrams for our running example systems: (**A**) three independent bits; (**B**) three completely interdependent bits; (**C**) independent blocks of dependent bits; and (**D**) the $2+1$ parity bit system. Regions of information zero in (**A**–**C**) are not shown.

5.3. Scale-Weighted Information

Multiscale information theory is based on the principle that any information about a system should be understood as applying at a specific scale. Information shared among a set of components—arising from dependent behavior among these components—has scale equal to the sum of the scales of these components. This principle was first discussed in the context of the complexity profile [13–17], and is also implicit in other works on multivariate information theory [40,41], as we discuss in Section 10.2.

To formalize this principle, we define the scale $s(x)$ of an irreducible dependency $x \in \mathfrak{D}_{\mathcal{A}}$ to be equal to the total scale of all components included in x:

$$s(x) = \sum_{\substack{a \in A \\ x \text{ includes } a}} \sigma(a). \tag{8}$$

The information in an irreducible dependency $x \in \mathfrak{D}_{\mathcal{A}}$ is understood to apply at scale $s(x)$. Large-scale information pertains to many components and/or to components of large intrinsic scale; whereas small-scale information pertains to few components, and/or components of small intrinsic scale. In running Example **C** (Figure 2C), the bit of information that applies to components a and c has scale 2, while the bit applying to component b has scale 1.

The overall significance of a dependency in a system depends on both its information and its scale. It therefore natural to weight quantities of information by their scale. We define the *scale-weighted information* $S(x)$ of an irreducible dependency x to be the scale of x times its information quantity

$$S(x) = s(x)I(x). \tag{9}$$

Extending this definition, we define the scale-weighted information of any subset $U \subset \mathfrak{D}_{\mathcal{A}}$ of the dependence space to be the sum of the scale-weighted information of each irreducible-dependency in this subset:

$$S(U) = \sum_{x \in U} S(x) = \sum_{x \in U} s(x)I(x). \tag{10}$$

The scale-weighted information of the entire dependency space $\mathfrak{D}_{\mathcal{A}}$—that is, the scale-weighted information of the system $\mathcal{A}$—is equal to the sum of the scale-weighted information of each component:

$$S(\mathfrak{D}_{\mathcal{A}}) = \sum_{a \in A} \sigma(a)H(a). \tag{11}$$

As we show in Appendix A, this property arises directly from the fact that scale-weighted information counts redundant information according to its multiplicity or total scale.

According to Equation (11), the total scale-weighted information $S(\mathfrak{D}_{\mathcal{A}})$ does not change if the system is reorganized or restructured, as long as the information $H(a)$ and scale $\sigma(a)$ of each individual component $a \in A$ is maintained. The value $S(\mathfrak{D}_{\mathcal{A}})$ can therefore be considered a conserved quantity. The existence of this conserved quantity implies a tradeoff of information versus scale, which can be illustrated using the example of a stock market. If investors act largely independently of each other, information overlaps are minimal. The total amount of information is large, but most of this information is small-scale—applying only to a single investor at a time. On the other hand, in a market panic, there is much overlapping or redundant information in their actions—the behavior of one can be largely inferred from the behavior of others [23]. Because of this redundancy, the amount of information needed to describe their collective behavior is low. This redundancy also makes this collective behavior large-scale and highly significant.

5.4. Independence and Complete Interdependence

Components $a_1, \ldots, a_k \in A$ are *independent* if their joint information is equal to the sum of the information in each separately:

$$H(a_1, \ldots, a_k) = H(a_1) + \ldots + H(a_k). \tag{12}$$

In running Example **C**, components a, b, and c are independent. This definition generalizes standard notions of independence in information theory, linear algebra, and matroid theory.

We extend the notion of independence to subsystems: subsystems $\mathcal{B}_i = (B_i, H_{\mathcal{B}_i}, \sigma_{\mathcal{B}_i})$ of $\mathcal{A}$, for $i = 1, \ldots, k$, are defined to be independent of one another if

$$H_{\mathcal{A}}(B_1 \cup \ldots \cup B_k) = H_{\mathcal{B}_1}(B_1) + \ldots + H_{\mathcal{B}_k}(B_k). \tag{13}$$

We recall from Section 4 that $H_{\mathcal{B}_i}$ is the restriction of $H_{\mathcal{A}}$ to subsets of B_i. In running Example **C**, the subsystem comprised of components a and c is independent of the subsystem comprised of component b.

Independence has the following *hereditary property* [64]: if subsystems $\mathcal{B}_1, \ldots, \mathcal{B}_k$ are independent, then all components and subsystems of $\mathcal{B}_i$ are independent of all components and subsystems of $\mathcal{B}_j$, for all $j \neq i$. We prove the hereditary property of independence from our axioms in Appendix C.

At the opposite extreme, we define a set of components $U \subset A$ to be *completely interdependent* if $H(a) = H(U)$ for any component $a \in U$. In words, any information applying to any component in U applies to all components in U.

A set $U \subset A$ of completely interdependent components can be replaced by a single component of scale $\sum_{a \in U} \sigma(a)$ to obtain an equivalent, reduced representation of the system. Thus, in running Example **C**, the set $\{a, c\}$ is completely interdependent, and can be replaced by a single component of scale two. We show in Appendix B that replacements of this kind preserve all relevant quantities of information and scale.

6. Complexity Profile

We now turn to quantitative indices that summarize a system's structure. One such index is the complexity profile [13–17], which concretizes the observation that a complex system exhibits structure at multiple scales. We define the complexity profile of a system $\mathcal{A}$ to be a real-valued function $C_{\mathcal{A}}(y)$ of a positive real number y, equal to the total amount of information at scale y or higher in $\mathcal{A}$:

$$C_{\mathcal{A}}(y) = I\big(\{x \in \mathfrak{D}_{\mathcal{A}} \,:\, s(x) \geq y\}\big). \tag{14}$$

Equation (14) generalizes previous definitions of the complexity profile [13–17], which use Shannon information as the information function and consider all components to have scale one.

The complexity profile reveals the levels of interdependence in a system. For systems where components are highly independent, $C(0)$ is large and $C(y)$ decreases sharply in y, since only small amounts of information apply at large scales in such a system. Conversely, in rigid or strongly interdependent systems, $C(0)$ is small and the decrease in $C(y)$ is shallower, reflecting the prevalence of large-scale information, as shown in in Figure 3. We plot the complexity profiles of our four running examples in Figure 4.

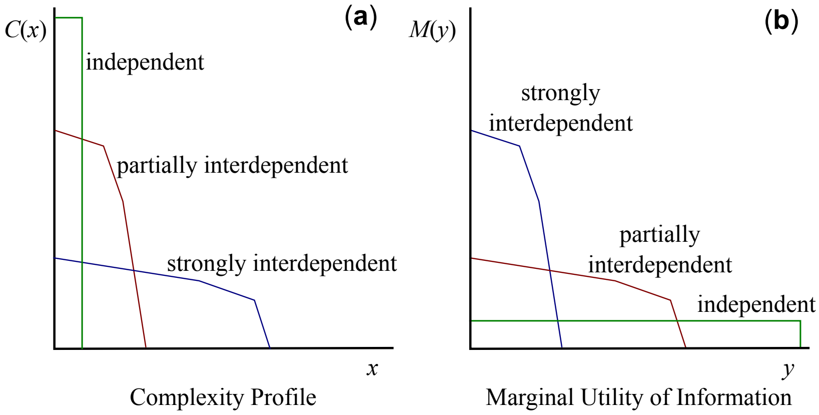


Figure 3. Schematic illustration of the (**a**) complexity profile (CP) and (**b**) marginal utility of information (MUI) for systems with varying degrees of interdependence among components. If the components are independent, all information applies at scale 1, so the complexity profile has $C(1)$ equal to the number of components and $C(x) = 0$ for $x > 1$. As the system becomes more interdependent, information applies at successively larger scales, resulting in a shallower decrease of $C(x)$. For the MUI, if components are independent, the optimal description scheme describes only a single component at a time, with marginal utility 1. As the system becomes more interdependent, information overlaps allow for more efficient descriptions that achieve greater marginal utility. For both the CP and MUI, the total area under the curve is equal to the total scale-weighted information $S(\mathfrak{D})$, which is preserved under reorganizations of the system. The CP and MUI are not reflections of each other in general, but they are for an important class of systems (see Section 8).

Previous works have developed and applied an explicit formula for the complexity profile [13,14,16,17,35] for cases where all components have equal intrinsic scales, $\sigma(a) = 1$ for all $a \in A$. To construct this formula, we first define the quantity $Q(j)$ as the sum of the joint information of all collections of j components:

$$Q(j) = \sum_{i_1,\ldots,i_j} H(a_{i_1},\ldots,a_{i_j}). \tag{15}$$

The complexity profile can then be expressed as

$$C(k) = \sum_{j=N-k}^{N-1} (-1)^{j+k-N} \binom{j}{j+k-N} Q(j+1), \tag{16}$$

where $N = |A|$ is the number of components in $\mathcal{A}$ [13,14]. The coefficients in this formula can be inferred from the inclusion-exclusion principle [61]; see [13] for a derivation.

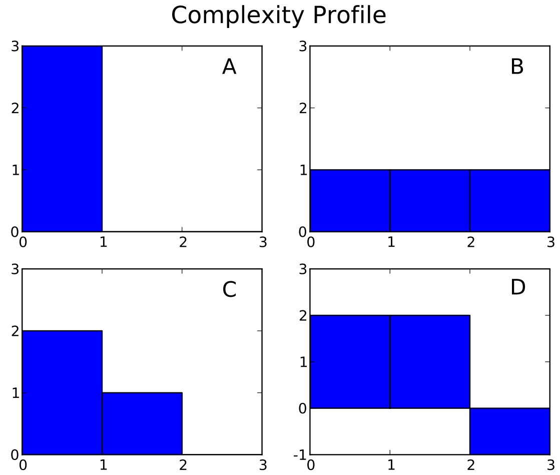


Figure 4. (**A**–**D**) Complexity profile $C(k)$ for Examples **A** through **D**. Note that the total (signed) area bounded by each curve equals $S(\mathfrak{D}_{\mathcal{A}}) = \sum_{a\in A} H(a) = 3$. For Example **D** (the parity bit), the information at scale 3 is negative.

The complexity profile has the following properties:

1. *Conservation law:* The area under $C(y)$ is equal to the total scale-weighted information of the system, and is therefore independent of the way the components depend on each other [13]:

$$\int_0^\infty C(y)\, dy = S(\mathfrak{D}_{\mathcal{A}}). \tag{17}$$

   This result follows from the conservation law for scale-weighted information, Equation (11), as shown in Appendix A.
2. *Total system information:* At the lowest scale $y = 0$, $C(y)$ corresponds to the overall joint information: $C(0) = H(A)$. For physical systems with the Shannon information function, this is the total entropy of the system, in units of information rather than the usual thermodynamic units.
3. *Additivity:* If a system $\mathcal{A}$ is the union of two independent subsystems $\mathcal{B}$ and $\mathcal{C}$, the complexity profile of the full system is the sum of the profiles for the two subsystems, $C_{\mathcal{A}}(y) = C_{\mathcal{B}}(y) + C_{\mathcal{C}}(y)$. We prove this property Appendix D.

Due to the combinatorial number of dependencies for an arbitrary system, calculation of the complexity profile may be computationally prohibitive; however, computationally tractable approximations to the complexity profile have been developed [15].

The complexity profile has connections to a number of other information-theoretic characterizations of structure and dependencies among sets of random variables [38,40,41,65–67], as we discuss in

Section 10.2. What distinguishes the complexity profile from these other approaches is the explicit inclusion of scale as an axis complementary to information.

7. Marginal Utility of Information

Here we introduce an new index characterizing multiscale structure: the *marginal utility of information* (MUI), denoted $M(y)$. The MUI quantifies how well a system can be characterized using a limited amount of information.

To obtain this index, we first ask how much scale-weighted information (as defined in Section 5.3) can be represented using y or fewer units of information. We call this quantity the *maximal utility of information*, denoted $U(y)$. For small values of y, an optimal characterization will convey only large-scale features of the system. As y increases, smaller-scale features will be progressively included. For a given system $\mathcal{A}$, the maximal amount of scale-weighted information that can be represented, $U(y)$, is constrained not only by the information limit y, but also by the pattern of information overlaps in $\mathcal{A}$—that is, the structure of $\mathcal{A}$. More strongly interdependent systems allow for larger amounts of scale-weighted information to be described using the same amount of information y.

We define the marginal utility of information as the derivative of maximal utility: $M(y) = U'(y)$. $M(y)$ quantifies how much scale-weighted information each additional unit of information can impart. The value of $M(y)$, being the derivative of scale-weighted information with respect to information, has units of scale. $M(y)$ declines steeply for rigid or strongly interdependent systems, and shallowly for weakly interdependent systems.

We now develop the formal definition of $U(y)$. We call any entity d that imparts information about system $\mathcal{A}$ a *descriptor* of $\mathcal{A}$. The utility of a descriptor will be defined as a quantity of the form

$$u = \sum_{a \in A} \sigma(a) I(d; a). \tag{18}$$

For this to be a meaningful expression, we consider each descriptor d to be an element of an augmented system $\mathcal{A}^\dagger = (A^\dagger, H_{\mathcal{A}^\dagger})$, whose components include d as well as the original components of $\mathcal{A}$, which is a subsystem of $\mathcal{A}^\dagger$. The amount of information that d conveys about any subset $V \subset A$ of components is given by

$$\begin{aligned} I(d; V) &= I_{\mathcal{A}^\dagger}(d; V) \\ &= H_{\mathcal{A}^\dagger}(d) + H_{\mathcal{A}^\dagger}(V) - H_{\mathcal{A}^\dagger}(\{d\} \cup V). \end{aligned} \tag{19}$$

For example, the amount that d conveys about a component $a \in A$ can be written $I(d; a) = H(d) + H(a) - H(d, a)$. $I(d; A)$ is the total information d imparts about the system. Because the original system $\mathcal{A}$ is a subsystem of $\mathcal{A}^\dagger$, the augmented information function $H_{\mathcal{A}^\dagger}$ coincides with $H_{\mathcal{A}}$ on subsets of A.

The quantities $I(d; V)$ are constrained by the structure of $\mathcal{A}$ and the axioms of information functions. Applying these axioms, we arrive at the following constraints on $I(d; V)$:

(i) $0 \leq I(d; V) \leq H(V)$ for all subsets $V \subset A$.

(ii) For any pair of nested subsets $W \subset V \subset A$, $0 \leq I(d; V) - I(d; W) \leq H(V) - H(W)$.

(iii) For any pair of subsets $V, W \subset A$,

$$I(d; V) + I(d; W) - I(d; V \cup W) - I(d; V \cap W) \leq H(V) + H(W) - H(V \cup W) - H(V \cap W).$$

To obtain the maximum utility of information, we interpret the values $I(d; V)$ as variables subject to the above constraints. We define $U(y)$ as the maximum value of the utility expression, Equation (18), as $I(d; V)$ vary subject to constraints (i)–(iii) and that the total information d imparts about $\mathcal{A}$ is less than or equal to y: $I(d; A) \leq y$.

$U(y)$ characterizes the maximal amount of scale-weighted information that could in principle be conveyed about $\mathcal{A}$ using y or less units of information, taking into account the information-sharing in $\mathcal{A}$ and the fundamental constraints on how information can be shared. $U(y)$ is well-defined since it is the maximal value of a linear function on a bounded set. Moreover, elementary results in linear programming theory [68] imply that $U(y)$ is piecewise linear, increasing and concave in y.

The above results imply that the marginal utility of information, $M(y) = U'(y)$, is piecewise constant, positive and nonincreasing. The MUI thus avoids the issue of counterintuitive negative values. The value of $M(y)$ can be understood as the additional scale units that can be described by an additional bit of information, given that the first y bits have been optimally utilized. Code for computing the MUI has been developed and is available online [69].

The marginal utility of information has the following additional properties:

1. *Conservation law:* The total area under the curve $M(y)$ equals the total scale-weighted information of the system:
$$\int_0^\infty M(y)\, dy = S(\mathfrak{D}_A). \tag{20}$$
This property follows from the observation that, since $M(y)$ is the derivative of $U(y)$, the area under this curve is equal to the maximal utility of any descriptor, which is equal to $S(\mathfrak{D}_A)$ since utility is defined in terms of scale-weighted information.
2. *Total system information:* The marginal utility vanishes for information values larger than the total system information, $M(y) = 0$ for $y > H(A)$, since, for higher values, the system has already been fully described.
3. *Additivity:* If $\mathcal{A}$ separates into independent subsystems $\mathcal{B}$ and $\mathcal{C}$, then
$$U_{\mathcal{A}}(y) = \max_{\substack{y_1+y_2=y \\ y_1,y_2\geq 0}} \left(U_{\mathcal{B}}(y_1) + U_{\mathcal{C}}(y_2)\right). \tag{21}$$
The proof follows from recognizing that, since information can apply either to $\mathcal{B}$ or to $\mathcal{C}$ but not both, an optimal description allots some amount y_1 of information to subsystem $\mathcal{B}$, and the rest, $y_2 = y - y_1$, to subsystem $\mathcal{C}$. The optimum is achieved when the total maximal utility over these two subsystems is maximized. Taking the derivative of both sides and invoking the concavity of U yields a corresponding formula for the marginal utility M:
$$M_{\mathcal{A}}(y) = \min_{\substack{y_1+y_2=y \\ y_1,y_2\geq 0}} \max\left\{M_{\mathcal{B}}(y_1), M_{\mathcal{C}}(y_2)\right\}. \tag{22}$$
Equations (21) and (22) are proven in Appendix E. This additivity property can also be expressed as the reflection (generalized inverse) of M. For any piecewise-constant, nonincreasing function f, we define the reflection $\tilde{f}$ as
$$\tilde{f}(x) = \max\{y : f(y) \leq x\}. \tag{23}$$
A generalized inverse [15] is needed since, for piecewise constant functions, there exist x-values for which there is no y such that $f(y) = x$. For such values, $\tilde{f}(x)$ is the largest y such that $f(y)$ does not exceed x. This operation is a reflection about the line $f(y) = y$, and applying it twice recovers the original function. If $\mathcal{A}$ comprises independent subsystems $\mathcal{B}$ and $\mathcal{C}$, the additivity property, Equation (22), can be written in terms of the reflection as
$$\tilde{M}_{\mathcal{A}}(x) = \tilde{M}_{\mathcal{B}}(x) + \tilde{M}_{\mathcal{C}}(x). \tag{24}$$
Equation (24) is proven in Appendix E.

Plots of the MUI for our four running examples are shown in Figure 5. The most interesting case is the parity bit system, Example **D**, for which the marginal utility is

$$M(y) = \begin{cases} \frac{3}{2} & 0 \leq y \leq 2 \\ 0 & y > 2. \end{cases} \tag{25}$$

The optimal description scheme leading to this marginal utility is shown in Figure 6. The marginal utility of information $M(y)$ captures the intermediate level of interdependency among components in Example **D**, in contrast to the maximal independence and maximal interdependence in Examples **A** and **B**, respectively (Figure 5). For an N-component generalization of Example **D**, in which each component acts as a parity bit for all others, we show in Appendix F that the MUI is given by

$$M(y) = \begin{cases} \frac{N}{N-1} & 0 \leq y \leq N-1 \\ 0 & y > N-1. \end{cases} \tag{26}$$

The MUI is similar in spirit to, and can be approximated by, principal components analysis, Information Bottleneck methods [70–73], and other methods that characterize the best possible description of a system using limited resources [66,74–79]. We discuss these connections in Section 10.3.

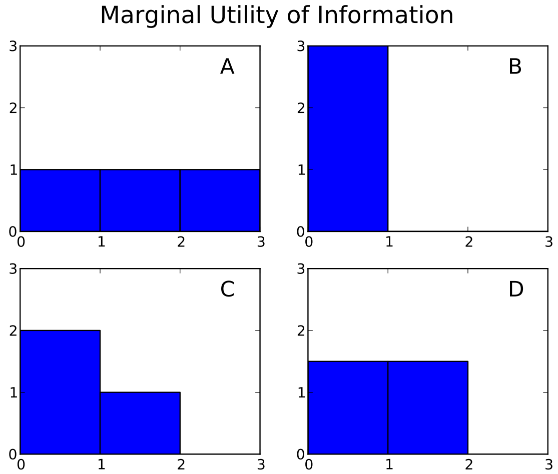


Figure 5. (**A**–**D**) Marginal Utility of Information for Examples **A** through **D**. The total area under each curve is $\int_0^\infty M(y)\,dy = S(\mathfrak{D}) = 3$. For Example **A**, all components are independent, and there is no more efficient description scheme than to describe one component at a time, with marginal utility 1. In Example **B**, the system state can be communicated with a single bit, with marginal utility 3. For Example **C**, the most efficient description scheme describes the fully correlated pair first (marginal utility 2), followed by the third component (marginal utility 1). The MUI for Example **C** can also be deduced from the additivity property, Equation (22). Examples **A**–**C** are all independent block systems; it follows from the results of Section 8 that their MUI functions are reflections (generalized inverses) of the corresponding complexity profiles shown in Figure 4. For Example **D**, the optimal description scheme is illustrated in Figure 6, leading to a marginal utility of $M(y) = 3/2$ for $0 \leq y \leq 2$ and $M(y) = 0$ for $y > 2$.

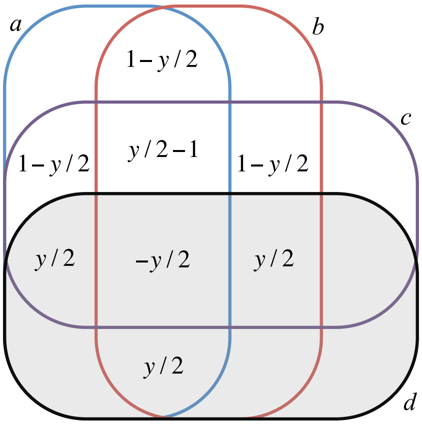


Figure 6. Information overlaps in the parity bit system, Example **D** of Figure 2, augmented with a descriptor d having information content $y \leq 2$ and maximal utility. Symmetry considerations imply that such a descriptor must convey an equal amount of information about each of the three components a, b and c. Constraints (i)–(iv) then yield that the amount described about each component must equal $y/2$ for $0 \leq y \leq 2$, and 1 for $y > 2$. Thus the maximal utility is $U(y) = 3y/2$ for $0 \leq y \leq 2$, and 3 for $y > 2$, leading to the marginal utility given in Equation (25) and shown in Figure 5**D**.

8. Reflection Principle for Systems of Independent Blocks

For systems with a particularly simple structure, the complexity profile and the MUI turn out to be reflections (generalized inverses) of each other. The simplest case is a system consisting of a single component a. In this case, according to Equation (14), the complexity profile $C(x)$ has value $H(a)$ for $0 \leq x \leq \sigma(a)$ and zero for $x > \sigma(a)$:

$$C(x) = H(a)\Theta\big(\sigma(a) - x\big). \tag{27}$$

Above, $\Theta(y)$ is a step function with value 1 for $y \geq 0$ and 0 otherwise. To compute the marginal utility of information for this system, we observe that a descriptor with maximal utility has $I(d;a) = \min\{y, H(a)\}$ for each value of the informational constraint y, and it follows that

$$M(y) = \sigma(a)\Theta\big(H(a) - y\big). \tag{28}$$

We observe that $C(x)$ and $M(x)$ are reflections of each other: $C(x) = \tilde{M}(x)$, where $\tilde{M}(x)$ is defined in Equation (23).

We next consider a system whose components are all independent of each other. Additivity over independent subsystems (Property 3 in Sections 6 and 7), together with Equations (27) and (28), implies

$$C(x) = \tilde{M}(x) = \sum_{a \in A} H(a)\Theta\big(\sigma(a) - x\big). \tag{29}$$

Thus the reflection principle holds for systems of independent components.

More generally, one can consider a system of "independent blocks"—that is, a system that can be partitioned into independent subsystems, where the components of each subsystem are completely interdependent (see Section 5.4 for definitions.) Running example **C** is such a system, in that it can be partitioned into independent subsystems with component sets $\{a, c\}$ and $\{b\}$, and each of these sets is completely interdependent. We show in Appendix B that any set of completely interdependent components can be replaced by a single component, with scale equal to the sum of the scales of the replaced components, without altering the complexity profile or MUI. Thus, for systems of independent blocks, each block can be collapsed into a single component, whereupon Equation (29) applies and the reflection principle holds.

We have thus established that for any system of independent blocks, the complexity profile and the MUI are reflections of each other $C(x) = \tilde{M}(x)$. However, this relationship does not hold for every system. $C(x)$ and $M(y)$ are not reflections of each other in the case of Example **D**, and, more generally, for a class of systems that exhibit negative information, as shown in Equation (26).

9. Application to Noisy Voter Model

As an application of our framework, we compute the marginal utility of information for the noisy voter model [43] on a complete graph. This model is a stochastic process with N "voters". Each voter, $i = 1, \ldots, N$, can exist in one of two states, which we label $\eta_i \in \{-1, 1\}$. Each voter updates its state at Poisson rate 1. With probability u it chooses ±1 with equal probability; otherwise, with probability $1 - u$, it copies a random other individual. Here $u \in [0, 1]$ is noise (or mutation) parameter that mediates the level of interdependence among voters. It follows that voter i flips its state (from $+1$ to -1 or vice versa) at Poisson rate

$$\lambda_i = \frac{u}{2} + \frac{1-u}{N-1} \sum_{j=1}^{N} \frac{|\eta_j - \eta_i|}{2}. \tag{30}$$

For $u \ll 1$, voters are typically in consensus; for $u = 1$, all voters behave independently. The noisy voter model is mathematically equivalent to the Moran model [80] of neutral drift with mutation, and to a model of financial behavior on networks [23,81].

This model has a stationary distribution in which the number m of voters in the $+1$ state has a beta-binomial distribution (Figure 7) [18,82]:

$$P(m) \propto \frac{\Gamma(N + M - m)\Gamma(M + m)}{\Gamma(N + 1 - m)\Gamma(1 + m)}, \qquad M = \frac{(N-1)u}{2(1-u)}. \tag{31}$$

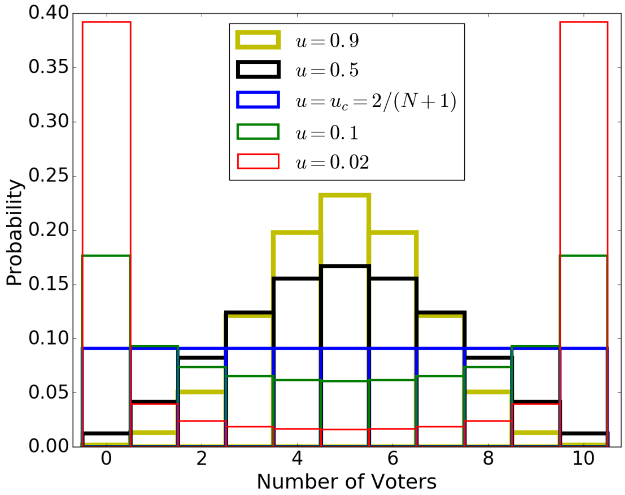


Figure 7. Stationary probability distribution for the noisy voter model [43] on a complete graph of size 10. Plot shows the probability of finding a given number i of voters in the $+1$ state, for different values of the noise (mutation) parameter u, according to Equation (31). For small values of u, voters are typically in consensus ($m = 0$ or $m = N$); as u increases their behavior becomes more independent.

For small u, P is concentrated on the "consensus" states 0 and N, converging to the uniform distribution on these states as $u \to 0$. For large u, P exhibits a central tendency around $m = N/2$, and converges to the binomial distribution with $p = 1/2$ as $u \to 1$. These two modes are separated by the critical value of $u_c = 2/(N+1)$, at which $M = 1$ and P becomes the uniform distribution on $\{0, \ldots, N\}$. This is the scenario in which mutation exactly balances the uniformizing effect of faithful copying. As $u \to 0$, P converges to the uniform distribution on the states 0 and N.

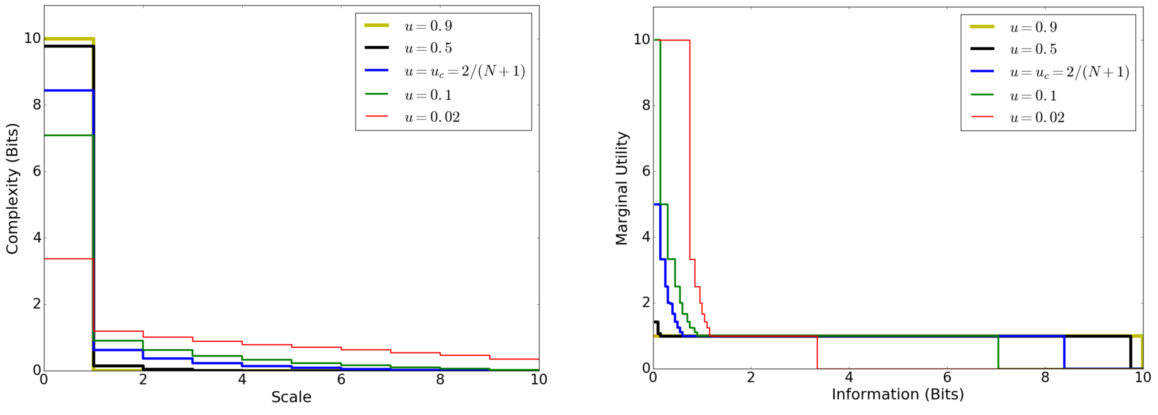
The noisy voter model on a complete graph possesses *exchange symmetry*, meaning its behavior is preserved under permutation of its components. As a consequence, if subsets U and V have the same cardinality, $|U| = |V|$, then they have the same information, $H(U) = H(V)$. The information function is therefore fully characterized by the quantities $H_1, \ldots, H_N$, where H_n is the information in each subset with $n \leq N$ components.

To calculate the MUI for systems with exchange symmetry, it suffices to consider descriptors that also possess exchange symmetry, so that $I(d;U) = I(d;V)$ whenever $|U| = |V|$. Denoting by I_n the information that a descriptor imparts about a subset of size n, constraints (i)–(iii) of Section 7 reduce to

(i) $0 \leq I_n \leq H_n$ for all $n \in \{1, \ldots, N\}$,
(ii) $0 \leq I_n - I_{n-1} \leq H_n - H_{n-1}$ for all $n \in \{1, \ldots, N\}$,
(iii) $I_n + I_m - I_{n+m-\ell} - I_\ell \leq H_n + H_m - H_{n+m-\ell} - H_\ell$ for all $n, m, \ell \in \{1, \ldots, N\}$.

The maximum utility of information $U(y)$ is the maximum value of NI_1, subject to (i)–(iii) above and $I_N \leq y$. Since the number of constraints is polynomial in N, the maximum utility—and therefore the MUI—are readily computable.

The complexity profile and MUI for this model (Figure 8) both capture how the interdependence of voters is mediated by the noise parameter u. For small u, conformity among voters leads to large-scale information (positive $C(x)$ for large x) and enables efficient partial descriptions of the system (large $M(y)$ for small y). For large u, weak dependence among voters means that most information applies at small scale ($C(x)$ decreases rapidly) and optimal descriptions cannot do much better than to describe each component singly ($M(y) = 1$ for most values of y). Unlike the case of independent block systems (Section 8), the complexity profile and MUI are not reflections of each other for this model, but the reflection of each is qualitatively similar to the other.



(**a**) (**b**)

Figure 8. (**a**) Complexity profile and (**b**) marginal utility of information (MUI) for the noisy voter model [43] on a complete graph of size 10. The MUI is approximated by computing the (exact) maximal utility of information $U(y)$ at a discrete set of points, and then approximating $M(y) = U'(y) \approx \Delta M/\Delta y$. For small u, since voters are largely coordinated, much of their collective behavior can be described using small amounts of large-scale information. This leads to positive values of $C(x)$ for large x, and large values of $M(y)$ for small y. For large u, voters are largely independent, and therefore most information applies to one voter at a time. In this case it follows that $C(x)$ decreases rapidly to zero, and $M(y) = 1$ for most values of y. For both indices, the area under each curve is 10, which is the sum of the Shannon information of each voter (i.e., the total scale-weighted information), as guaranteed by Equation (20). For all values of u, the MUI appears to take a subset of the values $10/n$ for $n = 1, \ldots, 10$.

10. Discussion

10.1. Potential Applications

Here we have presented a multiscale extension of information theory, along with two quantitative indices, as tools the analysis of structure in complex systems. Our generalized, axiomatic definition of information enables this framework to be applied to a variety of physical, biological, social, and economic systems. In particular, we envision applications to spin systems [1,17,83,84], gene regulatory systems [85–91], neural systems [92–94], biological swarming [95–97], spatial evolutionary dynamics [82,98,99], and financial markets [22,59,81,100–105].

In each of these systems, multiscale information theory enables the analysis of such questions as: Do components (genes, neurons, investors, etc.) behave largely independently, or are their behaviors strongly interdependent? How significant are intermediate scales of organization such as the genetic pathway, the cerebral lobe, or the financial sector? Can other "hidden" scales of organization be identified? Do the scales of behavior vary across different instances of a particular kind of system (e.g., gene regulation in stem versus differentiated cells; neural systems across taxa)? And how do the scales of organization in these systems compare to the scales of challenges they face?

The realization of these applications faces a computational hurdle: a full characterization of a system's structure requires the computation of 2^N informational quantities. Thus for large real-world systems, efficient approximations must be developed. Fortunately, there already exist efficient approximations to the complexity profile [15], and approximations to the MUI can be obtained using restricted description schemes as we discuss in Section 10.3 below.

10.2. Multivariate and Multiscale Information

The idea of using entropy or information to quantify a system's structure has deep roots. One of the earliest and most influential attempts was Schrödinger's concept of *negative entropy* or *negentropy* [106,107], which he introduced to quantify the extent to which a living system deviates from maximum entropy. Negentropy can be expressed in our formalism as $J = \sum_{a \in A} H(a) - H(A)$. This same quantity is known in other contexts as *multi-information* [38,65,66], *integration* [40] or *intrinsic information* [67], and is used to characterize the extent of statistical dependence among a set of random variables. Supposing that each component of our system has scale one, we have the following equivalent characterizations of this quantity:

$$J = S(\mathfrak{D}_{\mathcal{A}}) - I(\mathfrak{D}_{\mathcal{A}}) = \sum_{k=1}^{\infty} C(k) - C(1) = \sum_{k=2}^{\infty} C(k). \tag{32}$$

Equation (32) makes clear that while J quantifies the deviation of a system from full independence, it does not identify the scale at which this deviation arises: all information at scales 2 and higher is subsumed into a single quantity.

Other proposed information-theoretic measures of structure also aggregate over scales in various ways. For example, the excess entropy, as defined by Ay et al. [41], is equal to $C(2)$, the amount of information that applies at scales 2 and higher. Another example is the complexity measure of Tononi et al. [40], which is defined for an N-component system as

$$T = \sum_{k=1}^{N} \left(\frac{1}{\binom{N}{k}} \sum_{\substack{U \subset A \\ |U|=k}} H(U) - \frac{k}{N} H(A) \right). \tag{33}$$

Using Equation (5), we can re-express T as a weighted sum of the information in each irreducible dependency, where each dependency x is weighted by a particular combinatorial function of its scale $s(x)$:

$$T = \sum_{x \in \mathfrak{D}_A} I(x) \left(\sum_{k=1}^{N-s(x)} \frac{\binom{N-s(x)}{k}}{\binom{N}{k}} - \frac{N-1}{2} \right). \tag{34}$$

In this equation it is assumed that each it is assumed that each component has scale 1, so that $s(x)$ equals the number of components included in dependency x.

These previous measures can be understood as attempts to capture the idea that complex systems are not merely the sums of their parts—that is, they exhibit multiple scales of organization. We argue that this idea is best captured by making the notion of scale explicit, as a complementary axis to information. Doing so provides a formal basis for ideas that are implicit in earlier approaches.

The amount of information I that we assign to each of a system's dependencies is known in the context of Shannon information as the multivariate mutual information or interaction information [30–33,37–39]. The use of multivariate mutual information is sometimes criticized [36,62], in part because it yields negative values that may be difficult to interpret. Such negative values arise for the complexity profile, but are avoided by the MUI. The value of $M(y)$ is always nonnegative and has a consistent interpretation as the additional scale units describable by an additional bit of information. The notion of descriptors also avoids negative values, since the information that a descriptor d provides about a subset U of components is always nonnegative: $I(d; U) \geq 0$.

Finally, some recent work raises the important question of whether measures based on shared information suffice to characterize the structure of a system. James and Crutchfield [63] exhibit pairs of systems of random variables that have qualitatively different probabilistic relationships, but the same joint Shannon information $H(U)$ for each subset U of variables. As a consequence, the shared information I, complexity profile C, and marginal utility of information M, are the same for the two systems in such a pair. These examples demonstrate that probabilistic relationships among variables need not be determined by their information-theoretic measures, raising the question as to whether structure can be defined in terms of those measures. This conclusion, however, is dependent on the definition of the variables that are used to identify system states. To illustrate, if we take a particular system and combine variables into fewer ones with more states, the fewer information-theoretic measures that are obtained in the usual way become progressively less distinguishing of system probabilities. In the extreme case of a single variable having all of the possible states of the system, there is only one information measure. In the reverse direction, we have found [108] that a given system of random variables can be augmented with additional variables, the values of which are completely determined by the original system, in such a way that the probabilistic structure of the original system is uniquely determined by information overlaps in the augmented system. Thus, in this sense, moving to a more detailed representation can reveal relationships that are obscured in higher-level representations, and information theory may be sufficient to define structure in a general way.

10.3. Relation of the MUI to Other Measures

Our new index of structure, the MUI, is philosophically similar to data-reduction or dimensionality reduction techniques like principal component analysis, multidimensional scaling and detrended fluctuation analysis [74,76]; to the Information Bottleneck methods of Shannon information theory [70–73]; to Kolmogorov structure functions and algorithmic statistics in Turing-machine-based complexity theory [77–79]; to Gell-Mann and Lloyd's "effective complexity" [75]; and to Schneiderman et al.'s "connected information" [66]. All of these methods are mathematical techniques for characterizing the most important behaviors of the system under study. Each is an implementation of the idea of finding the best possible partial description of a system, where the resources available for this description (bits, coordinates, etc.) are constrained.

The essential difference from these previous measures is that the MUI is not tied to any particular method for generating partial descriptions. Rather, the MUI is defined in terms of optimally effective descriptors: for each possible amount of information invested in describing the system, the MUI considers the descriptor that provides the best possible theoretical return (in terms of scale-weighted information) on that investment. These returns are limited only by the structure of the system being described and the the fundamental constraints on information as encapsulated by our axioms.

In some applied contexts, it may be difficult or impossible to realize these theoretical maxima, due to constraints beyond those imposed by the axioms of information functions. It is often useful in these contexts to consider a particular "description scheme", in which descriptors are restricted to be of a particular form. Many of the data reduction and dimensionality reduction techniques described above can be understood as finding an optimal description of limited information using a specified description scheme. In these cases, the maximal utility found using the specified description scheme is in general less than the theoretical optimum. Calculating the marginal utility under a particular description scheme yields an approximation to the MUI.

10.4. Multiscale Requisite Variety

The discipline of cybernetics, an ancestor to modern control theory, used Shannon's information theory to quantify the difficulty of performing tasks, a topic of relevance both to organismal survival in biology and to system regulation in engineering. Ashby [109] considered scenarios in which a regulator device must protect some important entity from the outside environment and its disruptive influences. Successful regulation implies that if one knows only the state of the protected component, one cannot deduce the environmental influences; i.e., the job of the regulator is to minimize mutual information between the protected component and the environment. This is an information-theoretic statement of the idea of homeostasis. Ashby's "Law of Requisite Variety" states that the regulator's effectiveness is limited by its own information content, or *variety* in cybernetic terminology. An insufficiently flexible regulator will not be able to cope with the environmental variability.

Multiscale information theory enables us to overcome a key limitation of the requisite variety concept. In the framework of traditional cybernetics [109], each action of the environment requires a specific, unique reaction on the part of the regulator. This framework neglects the important difference between large-scale and fine-scale impacts. Systems may be able to absorb fine-scale impacts without any specific response, whereas responses to large-scale impacts are potentially critical to survival. For example, a human being can afford to be indifferent to the impact of a single molecule, whereas a falling rock (which may be regarded as the collective motion of many molecules) cannot be neglected. Ashby's Law does not make this distinction; indeed, there is no framework for this distinction in traditional information theory, since the molecule and the rock can be specified using the same amount of information.

This limitation can be overcome by a multiscale generalization of Ashby's Law [14], in which the responses of the system must occur at a scale appropriate to the environmental challenge. To protect against infection, for example, organisms have physical barriers (e.g., skin), generic physiological responses (e.g., clotting, inflammation) and highly specific adaptive immune responses, involving interactions among many cell types, evolved to identify pathogens at the molecular level. The evolution of immune systems is the evolution of separate large- and small-scale countermeasures to threats, enabled by biological mechanisms for information transmission and preservation [110]. By allowing for arbitrary intrinsic scales of components, and a range of different information functions, our work provides an expanded mathematical foundation for the multiscale generalization of Ashby's Law.

10.5. Mechanistic versus Informational Dependencies

Information-theoretic measures of a system's structure are essentially descriptive in nature. The tools we have proposed are aimed at identifying the scales of behavior of a system, but not necessarily the causes of this behavior. Importantly, causal influences at one scale can produce correlations at another. For example, the interactions in an Ising spin system are pairwise in character: the energy of a state depends only on the relative spins of neighboring pairs. These pairwise couplings can, however, give rise to long-range patterns [27]. Similarly, in models of coupled oscillators, dyadic physical interactions can lead to global synchronization [111]. Thus local interactions can create large-scale collective behavior.

11. Conclusions

Information theory has made, and will continue to make, formidable contributions to all areas of science. We argue that, in applying information theory to the study of complex systems, it is crucial to identify the scales at which information applies, rather than collapsing redundant or overlapping information into a raw number of independent bits. The multiscale approach to information theory falls squarely within the tradition of statistical physics—itself born of a marriage between probability theory and classical mechanics. By providing a general axiomatic framework for multiscale information theory, along with quantitative indices, we hope to deepen, clarify, and expand the mathematical foundations of complex systems theory.

Acknowledgments: We are grateful to Ryan G. James for developing code to compute the MUI.

Author Contributions: Benjamin Allen, Blake C. Stacey, and Yaneer Bar-Yam conceived the project; Benjamin Allen and Blake C. Stacey performed mathematical analysis; Benjamin Allen, Blake C. Stacey, and Yaneer Bar-Yam wrote the paper. All authors have read and approved the final version of the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Total Scale-Weighted Information

Here we prove two results regarding the total scale-weighted information of a system, $S(\mathfrak{D}_\mathcal{A})$. First we prove Equation (11), which shows that $S(\mathfrak{D}_\mathcal{A})$ depends only on the information and scale of each individual component:

Theorem A1. *For any system $\mathcal{A}$,*

$$S(\mathfrak{D}_\mathcal{A}) = \sum_{a \in A} \sigma(a) H(a). \tag{A1}$$

Proof. The proof amounts to a rearrangement of summations. We begin with the definition of scale-weighted information,

$$S(\mathfrak{D}_\mathcal{A}) = \sum_{x \in \mathfrak{D}_\mathcal{A}} s(x) I(x). \tag{A2}$$

Substituting the definition of $s(x)$, Equation (8), and rearranging yields

$$\begin{aligned} S(\mathfrak{D}_\mathcal{A}) &= \sum_{x \in \mathfrak{D}_\mathcal{A}} \left(\sum_{\substack{a \in A \\ x \text{ includes } a}} \sigma(a) \right) I(x) \\ &= \sum_{a \in A} \sigma(a) \sum_{\substack{x \in \mathfrak{D}_\mathcal{A} \\ x \text{ includes } a}} I(x) \\ &= \sum_{a \in A} \sigma(a) I(\delta_a) \\ &= \sum_{a \in A} \sigma(a) H(a). \quad \square \end{aligned}$$

Next we prove Equation (17) showing that the area under the complexity profile is equal to $S(\mathfrak{D}_\mathcal{A})$:

Theorem A2. *For any system $\mathcal{A}$,*

$$\int_0^\infty C(y)\, dy = S(\mathfrak{D}_\mathcal{A}). \tag{A3}$$

Proof. We begin by substituting the definition of $C(y)$:

$$\begin{aligned}\int_0^\infty C(y)\,dy &= \int_0^\infty I\big(\{x \in \mathfrak{D}_\mathcal{A} \;:\; \sigma(x) \geq y\}\big)\,dy \\ &= \int_0^\infty \left(\sum_{\substack{x \in \mathfrak{D}_\mathcal{A} \\ y \leq \sigma(x)}} I(x) \right) dy.\end{aligned}$$

We then interchange the sum and integral on the right-hand side and apply Theorem A1:

$$\begin{aligned}\int_0^\infty C(y)\,dy &= \sum_{x \in \mathfrak{D}_\mathcal{A}} \left(I(x) \int_0^{\sigma(x)} dy \right) \\ &= \sum_{x \in \mathfrak{D}_\mathcal{A}} \sigma(x) I(x) \\ &= S(\mathfrak{D}_\mathcal{A}). \quad \square\end{aligned}$$

Appendix B. Complete Interdependence and Reduced Representations

We mentioned in Section 5.4 that if a set of components is completely interdependent, they can be replaced by a single component, with scale equal to the sum of the scales of the replaced components. Here we define this replacement formally, and show that it preserves all quantities of shared information and scale-weighted information.

Let $\mathcal{A} = (A, H_\mathcal{A}, \sigma_\mathcal{A})$ be a system. We begin by recalling that a set of components $U \subset A$ is *completely interdependent* if $H(a) = H(U)$ for each $a \in U$. It follows from the monotonicity axiom that $H(V) = H(U)$ for any nonempty subset $V \subset U$. The following lemma shows that, in evaluating the information function H, the entire set U can be replaced by any subset thereof:

Lemma A1. *Let $U \subset A$ be a set of completely interdependent components. For any nonempty $V \subset U$ and any $W \subset A$,*

$$H(V \cup W) = H(U \cup W).$$

Proof. Applying the strong subadditivity axiom to the sets U and $V \cup (W \setminus U)$, and invoking the fact that $H(V) = H(U)$, we obtain

$$H(U \cup W) \leq H(U) + H\big(V \cup (W \setminus U)\big) - H(V) = H\big(V \cup (W \setminus U)\big). \tag{A4}$$

But the monotonicity axiom implies the inequalities $H\big(V \cup (W \setminus U)\big) \leq H(V \cup W) \leq H(U \cup W)$. Combining with (A4) yields $H\big(V \cup (W \setminus U)\big) = H(V \cup W) = H(U \cup W)$. $\square$

Now, for a system $\mathcal{A} = (A, H_\mathcal{A}, \sigma_\mathcal{A})$ with a set U of completely interdependent components, let us define a reduced system $\mathcal{A}^* = (A^*, H_{\mathcal{A}^*}, \sigma_{\mathcal{A}^*})$ in which the set U has been replaced by a single component u. The reduced set of components is $A^* = (A \setminus U) \cup \{u\}$. The information $H_{\mathcal{A}^*}(V)$, for $V \subset A^*$, is defined by

$$H_{\mathcal{A}^*}(V) = \begin{cases} H_\mathcal{A}\big(U \cup (V \setminus \{u\})\big) & \text{if } u \in V \\ H_\mathcal{A}(V) & \text{if } u \notin V. \end{cases} \tag{A5}$$

Component u of $\mathcal{A}^*$ has scale equal to the sum of the scales of components in U, while all other components maintain their scale:

$$\sigma_{\mathcal{A}^*}(a) = \begin{cases} \sum_{b \in U} \sigma_\mathcal{A}(b) & \text{if } a = u \\ \sigma_\mathcal{A}(a) & \text{if } a \neq u. \end{cases} \tag{A6}$$

The following theorem shows that shared information and scale-weighted information are preserved in moving from $\mathcal{A}$ to $\mathcal{A}^*$:

Theorem A3. *Let $U = \{u_1, \dots, u_k\} \subset A$ be a set of completely interdependent components of $\mathcal{A} = (A, H_\mathcal{A}, \sigma_\mathcal{A})$, with $A \setminus U = \{a_1, \dots, a_m\}$. Let $\mathcal{A}^* = (A^*, H_{\mathcal{A}^*}, \sigma_{\mathcal{A}^*})$ be the reduced system described above. Then the shared information $I_\mathcal{A}$ and $I_{\mathcal{A}^*}$ of the original and reduced systems, respectively, are related by*

$$\begin{aligned} I_\mathcal{A}(u_1; \dots; u_k; a_1; \dots; a_\ell | a_{\ell+1}, \dots, a_m) &= I_{\mathcal{A}^*}(u; a_1; \dots, a_\ell | a_{\ell+1}, \dots, a_m) \\ I_\mathcal{A}(a_1; \dots; a_\ell | u_1, \dots, u_k, a_{\ell+1}, \dots, a_m) &= I_{\mathcal{A}^*}(a_1; \dots; a_\ell | u, a_{\ell+1}, \dots, a_m) \\ I_\mathcal{A}(u_1; \dots; u_p; a_1; \dots; a_\ell | u_{p+1}, \dots, u_k, a_{\ell+1}, \dots, a_m) &= 0 \qquad \text{for } 1 \leq p \leq k-1. \end{aligned} \tag{A7}$$

The above equations also hold with the shared information $I_\mathcal{A}$ and $I_{\mathcal{A}^*}$ replaced by the scale-weighted information $S_\mathcal{A}$ and $S_{\mathcal{A}^*}$, respectively.

In other words, if the irreducible dependency x of $\mathcal{A}$ includes either all elements of U or no elements of U, then, upon collapsing the elements of U to the single component u to obtain the dependency x^* of $\mathcal{A}^*$, one has $I_{\mathcal{A}^*}(x^*) = I_\mathcal{A}(x)$ and $S_{\mathcal{A}^*}(x^*) = S_\mathcal{A}(x)$. If x includes some elements of U and excludes others, then $I_\mathcal{A}(x) = S_\mathcal{A}(x) = 0$. Thus all nonzero quantities of shared information and scale-weighted information are preserved upon collapsing the set U to the single component u.

Proof. Define the function J on the irreducible dependencies of $\mathcal{A}$ by

$$\begin{aligned} J(u_1; \dots; u_k; a_1; \dots; a_\ell | a_{\ell+1}, \dots, a_m) &= I_\mathcal{A}(u_1; \dots; u_k; a_1; \dots; a_\ell | a_{\ell+1}, \dots, a_m) \\ &\quad - I_{\mathcal{A}^*}(u; a_1; \dots, a_\ell | a_{\ell+1}, \dots, a_m) \\ J(a_1; \dots; a_\ell | u_1, \dots, u_k, a_{\ell+1}, \dots, a_m) &= I_\mathcal{A}(a_1; \dots; a_\ell | u_1, \dots, u_k, a_{\ell+1}, \dots, a_m) \\ &\quad - I_{\mathcal{A}^*}(a_1; \dots; a_\ell | u, a_{\ell+1}, \dots, a_m) \\ J(u_1; \dots; u_p; a_1; \dots; a_\ell | u_{p+1}, \dots, u_k, a_{\ell+1}, \dots, a_m) &= I_\mathcal{A}(u_1; \dots; u_p; a_1; \dots; a_\ell | u_{p+1}, \dots, u_k, a_{\ell+1}, \dots, a_m). \end{aligned}$$

In light of Equation (5), the values of J are the unique solution to the system of equations

$$\sum_{x \in \delta(V)} J(x) = \begin{cases} H_\mathcal{A}(V) - H_{\mathcal{A}^*}(\{u\} \cup V) & \text{if } V \cap U \neq \varnothing \\ H_\mathcal{A}(V) - H_{\mathcal{A}^*}(V) & \text{if } V \cap U = \varnothing, \end{cases} \tag{A8}$$

as V runs over subsets of A. But Lemma A1 and Equation (A5) imply that the right-hand side of Equation (A8) is zero for each $V \subset A$. Therefore, $J(x) = 0$ for each $x \in \mathfrak{D}_\mathcal{A}$, and Eq. (A7) follows. The claim regarding scale-weighted information then follows from Equations (8), (9) and (A6). □

Theorem A3 shows that all nonzero quantities of shared information and scale-weighted information are preserved when collapsing a set of completely dependent components into a single component. It follows that the complexity profile and MUI are also preserved under this collapsing operation.

Appendix C. Properties of Independence

Here we prove fundamental properties of independent subsystems, which will be used in Appendices D and E to demonstrate the additivity properties of the complexity profile and MUI. Our first target is the *hereditary property of independence* (Theorem A4), which asserts that subsystems of independent subsystems are independent [64]. We then establish in Theorem A5 a simple characterization of information in systems comprised of independent subsystems.

For $i = 1, \dots, k$, let $\mathcal{A}_i = (A_i, H_{\mathcal{A}_i}, \sigma_{\mathcal{A}_i})$ be subsystems of $\mathcal{A} = (A, H_\mathcal{A}, \sigma_\mathcal{A})$, with the subsets $A_i \subset A$ disjoint from each other. We recall the definition of independent subsystems from Section 5.4.

Definition A1. *The subsystems $\mathcal{A}_i$ are* independent *if*

$$H(A_1 \cup \ldots \cup A_k) = H(A_1) + \ldots + H(A_k).$$

We establish the hereditary property of independence first in the case of two subsystems (Lemma A2), using repeated application of the strong subadditivity axiom. We then extend this result in Theorem A4 to arbitrary numbers of subsystems.

Lemma A2. *If $\mathcal{A}_1$ and $\mathcal{A}_2$ are independent subsystems of $\mathcal{A}$, then for every pair of subsets $U_1 \subset A_1$, $U_2 \subset A_2$, $H(U_1 \cup U_2) = H(U_1) + H(U_2)$.*

Proof. The strong subadditivity axiom, applied to the sets A_1 and $U_1 \cup A_2$, yields

$$H(A_1 \cup A_2) \leq H(A_1) + H(U_1 \cup A_2) - H(U_1).$$

Replacing the left-hand side by $H(A_1) + H(A_2)$ and adding $H(U_1) - H(A_1)$ to both sides yields

$$H(U_1) + H(A_2) \leq H(U_1 \cup A_2). \tag{A9}$$

Now applying strong subadditivity to the sets $U_1 \cup U_2$ and A_2 yields

$$H(U_1 \cup A_2) \leq H(U_1 \cup U_2) + H(A_2) - H(U_2).$$

Combining with (A9) via transitivity, we have

$$H(U_1) + H(A_2) \leq H(U_1 \cup U_2) + H(A_2) - H(U_2).$$

Adding $H(U_2) - H(A_2)$ to both sides yields

$$H(U_1) + H(U_2) \leq H(U_1 \cup U_2). \tag{A10}$$

But strong subadditivity applied to U_1 and U_2 yields

$$H(U_1 \cup U_2) \leq H(U_1) + H(U_2) - H(U_1 \cap U_2) \leq H(U_1) + H(U_2). \tag{A11}$$

We conclude from inequalities (A10) and (A11) that

$$H(U_1 \cup U_2) = H(U_1) + H(U_2). \quad \square$$

We now use an induction argument to extend the hereditary property of independence to any number of subsystems.

Theorem A4. *If $\mathcal{A}_1, \ldots, \mathcal{A}_k$ are independent subsystems of $\mathcal{A}$, and $U_i \subset A_i$ for $i = 1, \ldots, k$ then*

$$H(U_1 \cup \ldots \cup U_k) = H(U_1) + \ldots + H(U_k).$$

Proof. This follows by induction on k. The $k = 1$ case is trivial. Suppose inductively that the statement is true for $k = \tilde{k}$, for some integer $\tilde{k} \geq 1$, and consider the case $k = \tilde{k} + 1$. We have

$$H(U_1) + \ldots + H(U_{\tilde{k}}) + H(U_{\tilde{k}+1}) = H(U_1 \cup \ldots \cup U_{\tilde{k}}) + H(U_{\tilde{k}+1})$$

by the inductive hypothesis, and

$$H(U_1 \cup \ldots \cup U_{\tilde{k}}) + H(U_{\tilde{k}+1}) = H(U_1 \cup \ldots \cup U_{\tilde{k}} \cup U_{\tilde{k}+1})$$

by Lemma A2 (since the subsystem of $\mathcal{A}$ with component set $A_1 \cup \ldots \cup A_{\tilde{k}}$ is clearly independent from $\mathcal{A}_{\tilde{k}+1}$). This completes the proof. □

We now examine the information in dependencies for systems comprised of independent subsystems. For convenience, we introduce a new notion: The *power system* of a system $\mathcal{A}$ is a system $2^{\mathcal{A}} = (2^A, H_{2^{\mathcal{A}}})$, where 2^A is the set of all subsets of A (which in set theory is called the *power set* of A). In other words, the components of $2^{\mathcal{A}}$ are the subsets of A. The information function $H_{2^{\mathcal{A}}}$ on $2^{\mathcal{A}}$ is defined by the relation

$$H_{2^{\mathcal{A}}}(U_1, \ldots, U_k) = H_{\mathcal{A}}(U_1 \cup \ldots \cup U_k). \tag{A12}$$

By identifying the singleton subsets of 2^A with the elements of A (that is, identifying each $\{a\} \in 2^A$ with $a \in A$), we can view $\mathcal{A}$ as a subsystem of $2^{\mathcal{A}}$.

This new system allows us to use the following relation: For any integers $k, \ell \geq 0$ and components $a_1, a_2, b_1, \ldots, a_k, c_1, \ldots, c_\ell \in A$,

$$\begin{aligned} I_{\mathcal{A}}(a_1; a_2; b_1; \ldots; b_k | c_1, \ldots, c_\ell) &= I_{\mathcal{A}}(a_1; b_1; \ldots; b_k | c_1, \ldots, c_\ell) \\ &\quad + I_{\mathcal{A}}(a_2; b_1; \ldots; b_k | c_1, \ldots, c_\ell) - I_{2^{\mathcal{A}}}(\{a_1, a_2\}; b_1; \ldots; b_k | c_1, \ldots, c_\ell). \end{aligned} \tag{A13}$$

This relation generalizes the identity $I(a_1; a_2) = H(a_1) + H(a_2) - H(a_1, a_2)$ to conditional mutual information. It follows directly from the mathematical definition of I, Equation (5) of the main text.

We now show that if $\mathcal{B}$ and $\mathcal{C}$ are independent subsystems of $\mathcal{A}$, any conditional mutual information of components $\mathcal{B}$ and components of $\mathcal{C}$ is zero.

Lemma A3. *Let $\mathcal{B} = (B, H_{\mathcal{B}})$ and $\mathcal{C} = (C, H_{\mathcal{C}})$ be independent subsystems of $\mathcal{A}$. For any components $b_1, \ldots, b_m, b'_1, \ldots, b'_{m'} \in B$ and $c_1, \ldots, c_n, c'_1, \ldots, c'_{n'} \in C$, with $m, n \geq 1$, $m', n' \geq 0$,*

$$I(b_1; \ldots; b_m; c_1; \ldots; c_n | b'_1, \ldots, b'_{m'}, c'_1, \ldots, c'_{n'}) = 0. \tag{A14}$$

Proof. We prove this by induction. As a base case, we take $m = n = 1, m' = n' = 0$. In this case, the statement reduces to $I(b; c) = 0$ for every $b \in B$, $c \in C$. Since Lemma A2 guarantees that $H(b, c) = H(b) + H(c)$, this claim follows directly from the identity $I(b; c) = H(b) + H(c) - H(b, c)$.

We now inductively assume that the claim is true for all independent subsystems $\mathcal{B}$ and $\mathcal{C}$ of a system $\mathcal{A}$, and all $m \leq \tilde{m}, n \leq \tilde{n}, m' \leq \tilde{m}'$, and $n' \leq \tilde{n}'$, for some integers $\tilde{m}, \tilde{n} \geq 1$, $\tilde{m}', \tilde{n}' \geq 0$. We show that the truth of the claim is maintained when each of $\tilde{m}, \tilde{n}, \tilde{m}'$, and $\tilde{n}'$ is incremented by one.

We begin by incrementing m to $\tilde{m} + 1$. Applying (A13) yields

$$\begin{aligned} I_{\mathcal{A}}\big(b_{\tilde{m}}; b_{\tilde{m}+1}; b_1; \ldots; b_{\tilde{m}-1}; c_1; \ldots; c_{\tilde{n}} | b'_1, \ldots, b'_{\tilde{m}'}, c'_1, \ldots, c'_{\tilde{n}'}\big) \\ = I_{\mathcal{A}}\big(b_{\tilde{m}}; b_1; \ldots; b_{\tilde{m}-1}; c_1; \ldots; c_{\tilde{n}} | b'_1, \ldots, b'_{\tilde{m}'}, c'_1, \ldots, c'_{\tilde{n}'}\big) \\ + I_{\mathcal{A}}\big(b_{\tilde{m}+1}; b_1; \ldots; b_{\tilde{m}-1}; c_1; \ldots; c_{\tilde{n}} | b'_1, \ldots, b'_{\tilde{m}'}, c'_1, \ldots, c'_{\tilde{n}'}\big) \\ - I_{2^{\mathcal{A}}}\big(\{b_{\tilde{m}}; b_{\tilde{m}+1}\}; b_1; \ldots; b_{\tilde{m}-1}; c_1; \ldots; c_{\tilde{n}} | b'_1, \ldots, b'_{\tilde{m}'}, c'_1, \ldots, c'_{\tilde{n}'}\big). \end{aligned} \tag{A15}$$

The first two terms of the right-hand side of (A15) are zero by the inductive hypothesis. Furthermore, it is clear from the definition of a power system that $2^{\mathcal{B}}$ and $2^{\mathcal{C}}$ are independent subsystems of $2^{\mathcal{A}}$. Thus the final term on the right-hand size of (A15) is also zero by the inductive hypothesis. In sum, the entire right-hand side of (A15) is zero, and the left-hand side must therefore be zero as well. This proves the claim is true for $m = \tilde{m} + 1$.

We now increment m' to $\tilde{m}' + 1$. From Equation (6) of the main text, we have the relation

$$\begin{aligned} &I_{\mathcal{A}}\big(b_1; \ldots; b_{\tilde{m}}; c_1; \ldots; c_{\tilde{n}} | b'_1, \ldots, b'_{\tilde{m}'}, c'_1, \ldots, c'_{\tilde{n}'}\big) \\ &\qquad = I_{\mathcal{A}}\big(b'_{\tilde{m}'+1}; b_1; \ldots; b_{\tilde{m}}; c_1; \ldots; c_{\tilde{n}} | b'_1, \ldots, b'_{\tilde{m}'}, c'_1, \ldots, c'_{\tilde{n}'}\big) \\ &\qquad\qquad + I_{\mathcal{A}}\big(b_1; \ldots; b_{\tilde{m}}; c_1; \ldots; c_{\tilde{n}} | b'_1, \ldots, b'_{\tilde{m}'}, b'_{\tilde{m}'+1}, c'_1, \ldots, c'_{\tilde{n}'}\big). \end{aligned}$$

The left-hand side above is zero by the inductive hypothesis, and the first term on the right-hand side is zero by the case $m = \tilde{m} + 1$ proven above. Thus the second term on the right-hand side is also zero, which proves the claim is true for $m' = \tilde{m}' + 1$.

Finally, the cases $n = \tilde{n} + 1$ and $n' = \tilde{n}' + 1$ follow by interchanging the roles of $\mathcal{B}$ and $\mathcal{C}$. The result now follows by induction. □

We next show that for $\mathcal{B}$ and $\mathcal{C}$ independent subsystems of $\mathcal{A}$, the amounts of information in dependencies of $\mathcal{B}$ are not affected by additionally conditioning on components of $\mathcal{C}$.

Lemma A4. *Let $\mathcal{B} = (B, H_{\mathcal{B}}, \sigma_{\mathcal{B}})$ and $\mathcal{C} = (C, H_{\mathcal{C}}, \sigma_{\mathcal{C}})$ be independent subsystems of $\mathcal{A}$. For integers $m \geq 1$ and $m', n' \geq 0$, let $b_1, \ldots, b_m \in B$, $c_1, \ldots, c_n, c'_1, \ldots, c'_{n'} \in C$. Then*

$$I(b_1; \ldots; b_m | b'_1, \ldots, b'_{m'}, c'_1, \ldots, c'_{n'}) = I(b_1; \ldots; b_m | b'_1, \ldots, b'_{m'}). \tag{A16}$$

Proof. This follows by induction on n'. The claim is trivially true for $n' = 0$. Suppose it is true in the case $n' = \tilde{n}'$, for some $\tilde{n}' \geq 0$. By Equation (6) we have

$$\begin{aligned} &I(b_1; \ldots; b_m | b'_1, \ldots, b'_{m'}, c'_1, \ldots, c'_{\tilde{n}'}) \\ &\qquad = I(b_1; \ldots; b_m; c'_{\tilde{n}'+1} | b'_1, \ldots, b'_{m'}, c'_1, \ldots, c'_{n'}) \\ &\qquad\qquad + I(b_1; \ldots; b_m | b'_1, \ldots, b'_{m'}, c'_1, \ldots, c'_{\tilde{n}'}, c'_{\tilde{n}'+1}). \end{aligned} \tag{A17}$$

The left-hand side is equal to $I(b_1; \ldots; b_m | b'_1, \ldots, b'_{m'})$ by the inductive hypothesis, and the first term on the right-hand side is zero by Lemma A3. This completes the proof. □

Finally, it follows from Lemmas A3 and A4 that if $\mathcal{A}$ separates into independent subsystems, an irreducible dependency of $\mathcal{A}$ has nonzero information only if it includes components from only one of these subsystems. To state this precisely, we introduce a projection mapping from irreducible dependencies of a system $\mathcal{A}$ to those of a subsystem $\mathcal{B}$ of $\mathcal{A}$. This mapping, denoted $\rho^{\mathcal{A}}_{\mathcal{B}} : \mathfrak{D}_{\mathcal{A}} \to \mathfrak{D}_{\mathcal{B}}$, takes an irreducible dependency among the components in A, and "forgets" those components that are not in B, leaving an irreducible dependency among only the components in B. For example, suppose $A = \{a, b, c\}$ and $B = \{b, c\}$. Then

$$\begin{aligned} \rho^{\mathcal{A}}_{\mathcal{B}}(a; b|c) &= b|c \\ \rho^{\mathcal{A}}_{\mathcal{B}}(b; c|a) &= b; c. \end{aligned} \tag{A18}$$

We can now state the following simple characterization of information in systems comprised of independent subsystems:

Theorem A5. *Let $\mathcal{A}_1, \ldots, \mathcal{A}_k$ be independent subsystems of $\mathcal{A}$, with $A = A_1 \cup \ldots \cup A_k$. Then for any irreducible dependency $x \in \mathfrak{D}_{\mathcal{A}}$,*

$$I_{\mathcal{A}}(x) = \begin{cases} I_{\mathcal{A}_i}\big(\rho^{\mathcal{A}}_{\mathcal{A}_i}(x)\big), & \textit{if } x \textit{ includes only components of } \mathcal{A}_i \\ & \textit{for some } i \in \{1, \ldots, k\}, \\ 0 & \textit{otherwise.} \end{cases} \tag{A19}$$

Proof. In the case that x includes only components of $\mathcal{A}_i$ for some i, the statement follows from Lemma A4. In all other cases, the claim follows from Lemma A3. □

Appendix D. Additivity of the Complexity Profile

Here we prove Property 3 of the complexity profile claimed in Section 6: the complexity profile is additive over independent systems.

Theorem A6. *Let $\mathcal{A}_1, \ldots, \mathcal{A}_k$ be independent subsystems of $\mathcal{A}$. Then*

$$C_{\mathcal{A}}(y) = C_{\mathcal{A}_1}(y) + \ldots + C_{\mathcal{A}_k}(y). \tag{A20}$$

Proof. We start with the definition

$$C_{\mathcal{A}}(y) = \sum_{\substack{x \in \mathfrak{D}_{\mathcal{A}} \\ \sigma(x) \geq y}} I_{\mathcal{A}}(x). \tag{A21}$$

Applying Theorem A5 to each term on the right-hand side yields

$$\begin{aligned} C_{\mathcal{A}}(y) &= \sum_{i=1}^{k} \sum_{\substack{x \in \mathfrak{D}_{\mathcal{A}} \\ x \text{ includes only components of } \mathcal{A}_i \\ \sigma(x) \geq y}} I_{\mathcal{A}_i}\left(\rho^{\mathcal{A}}_{\mathcal{A}_i}(x)\right) \\ &= \sum_{i=1}^{k} \sum_{\substack{x \in \mathfrak{D}_{\mathcal{A}_i} \\ \sigma(x) \geq y}} I_{\mathcal{A}_i}(x) = \sum_{i=1}^{k} C_{\mathcal{A}_i}(y). \quad \square \end{aligned}$$

Appendix E. Additivity of Marginal Utility of Information

Here we prove the additivity property of MUI stated in Section 7. We begin by recalling the mathematical context for this result.

The maximal utility of information, $U(y)$, is defined as the maximal value of the quantity

$$u = \sum_{a \in A} \sigma(a) I(d; a), \tag{A22}$$

as the variables in the set $\{I(d; V)\}_{V \subset A}$ vary subject to the following constraints:

(i) $0 \leq I(d; V) \leq H(V)$ for all $V \subset A$.

(ii) For any $W \subset V \subset A$,

$$0 \leq I(d; V) - I(d; W) \leq H(V) - H(W). \tag{A23}$$

(iii) For any $V, W \subset A$,

$$I(d; V) + I(d; W) - I(d; V \cup W) - I(d; V \cap W) \leq H(V) + H(W) - H(V \cup W) - H(V \cap W).$$

(iv) $I(d; A) \leq y$.

The marginal utility of information, $M(y)$ is defined as the derivative of $U(y)$.

We emphasize for clarity that, while we intuitively regard $I(d; V)$ as the information that a descriptor d imparts about utility V, we formally treat the quantities $\{I(d; V)\}_{V \subset A}$ not as functions of two inputs but as variables subject to the above constraints.

Throughout this appendix we consider a system $\mathcal{A} = (A, H_{\mathcal{A}})$ comprising two independent subsystems, $\mathcal{B} = (B, H_{\mathcal{B}})$ and $\mathcal{C} = (C, H_{\mathcal{C}})$. This means that A is the disjoint union of B and C, and $H(A) = H(B) + H(C)$. The additivity property of MUI can be stated as

$$M_{\mathcal{A}}(y) = \min_{\substack{y_1+y_2=y\\ y_1,y_2\geq 0}} \max \left\{ M_{\mathcal{B}}(y_1), M_{\mathcal{C}}(y_2) \right\}. \tag{A24}$$

Alternatively, this property can be stated in terms of the reflection $\tilde{M}_{\mathcal{A}}(x)$ of $M_{\mathcal{A}}(y)$, with the dependent and independent variables interchanged (see Section 7), as

$$\tilde{M}_{\mathcal{A}}(x) = \tilde{M}_{\mathcal{B}}(x) + \tilde{M}_{\mathcal{C}}(x). \tag{A25}$$

The proof of this property is organized as follows. Our first major goal is Theorem A7, which asserts that $I(d;A) = I(d;B) + I(d;C)$ when u is maximized. Lemmas A5 and A6 are technical relations needed to achieve this result. We then apply the decomposition principle of linear programming to prove an additivity property of $U_{\mathcal{A}}$ (Theorem A8). Theorem A9 then deduces the additivity of $M_{\mathcal{A}}$ from the additivity of $\hat{U}_{\mathcal{A}}$. Finally, in Corollary A1, we demonstrate the additivity of the reflected function $\tilde{M}_{\mathcal{A}}$.

Lemma A5. *Suppose the quantities $\{I(d;V)\}_{V\subset A}$ satisfy Constraints (i)–(iv). Then for any subset $V \subset A$,*

$$I(d;V) \geq I(d;V\cap B) + I(d;V\cap C). \tag{A26}$$

Proof. Applying Constraint (iii) to the sets $V\cap B$ and $V\cap C$ we have

$$I(d;V\cap B) + I(d;V\cap C) - I(d;V) \leq H(V\cap B) + H(V\cap C) - H(V). \tag{A27}$$

But by Lemma A2, $H(V) = H(V\cap B) + H(V\cap C)$. Thus the right-hand side above is zero, which proves the claim. □

Lemma A6. *Suppose the quantities $\{I(d;V)\}_{V\subset A}$ satisfy Constraints (i)–(iv). Suppose further that $W \subset V \subset A$ and $I(d;V) = I(d;V\cap B) + I(d;V\cap C)$. Then $I(d;W) = I(d;W\cap B) + I(d;W\cap C)$.*

Proof. Constraint (iii), applied to the sets $V\cap B$ and $W\cup(V\cap C)$, yields

$$\begin{aligned} &I(d;V\cap B) + I\big(d;W\cup(V\cap C)\big) - I(d;V) - I(d;W\cap B) \\ &\qquad \leq H(V\cap B) + H\big(W\cup(V\cap C)\big) - H(V) - H(W\cap B). \end{aligned} \tag{A28}$$

By Lemma A2, we have

$$\begin{aligned} H\big(W\cup(V\cap C)\big) &= H(W\cap B) + H(V\cap C) \\ H(V) &= H(V\cap B) + H(V\cap C). \end{aligned} \tag{A29}$$

With these two relations, the right-hand side of (A28) simplifies to zero. Making this simplification and substituting $I(d;V) = I(d;V\cap B) + I(d;V\cap C)$ (as given), we obtain

$$I\big(d;W\cup(V\cap C)\big) - I(d;W\cap B) - I(d;V\cap C) \leq 0. \tag{A30}$$

We next apply Constraint (iii) to $V\cap C$ and W, yielding

$$\begin{aligned} &I(d;V\cap C) + I(d;W) - I\big(d;W\cup(V\cap C)\big) - I(d;W\cap C) \\ &\qquad \leq H(V\cap C) + H(W) - H\big(W\cup(V\cap C)\big) - H(W\cap C). \end{aligned} \tag{A31}$$

Lemma A2 implies $H(W) = H(W\cap B) + H(W\cap C)$. Combining this relation with (A29), the right-hand side of (A31) simplifies to zero. We then rewrite (A31) as

$$I(d;W) - I(d;W \cap C) \leq I\big(d;W \cup (V \cap C)\big) - I(d;V \cap C). \tag{A32}$$

By (A30), the right-hand side above is less than or equal to $I(d;W \cap B)$. Making this substitution and rearranging, we obtain

$$I(d;W) \leq I(d;W \cap B) + I(d;W \cap C). \tag{A33}$$

Combining now with Lemma A5, it follows that $I(d;W) = I(d;W \cap B) + I(d;W \cap C)$ as desired. □

Theorem A7. *Suppose the quantities* $\{I(\hat{d};V)\}_{V \subset A}$ *maximixe* $u = \sum_{a \in A} \sigma(a) I(d;a)$ *subject to Constraints (i)–(iv) for some* $0 \leq y \leq H(A)$. *Then*

$$I(\hat{d};A) = I(\hat{d};B) + I(\hat{d};C). \tag{A34}$$

Proof. Let $\hat{u} = \sum_{a \in A} \sigma(a) I(\hat{d};a)$ be the maximal value of u. By the duality principle of linear programming, the quantities $\{I(\hat{d};V)\}_{V \subset A}$ minimize the value of $I(d;A)$ as $\{I(d;V)\}_{V \subset A}$ varies subject to Constraints (i)–(iii) along with the additional constraint $u \geq \hat{u}$. (Informally, the descriptor $\hat{d}$ achieves utility $\hat{u}$ using minimal information.)

Assume for the sake of contradiction that $I(\hat{d};A) > I(\hat{d};B) + I(\hat{d};C)$. We will obtain a contradiction by showing that there is another set of quantities $\{I(\tilde{d};V)\}_{V \subset A}$, satisfying (i)–(iii) and $\tilde{u} = \hat{u}$, with $I(\tilde{d};A) < I(\hat{d};A)$. Here, $\tilde{u}$ is the utility associated to $\{I(\tilde{d};V)\}_{V \subset A}$; that is, $\tilde{u} = \sum_{a \in A} \sigma(a) I(\tilde{d};a)$. (Informally, we construct a new descriptor $\tilde{d}$ that achieves the same utility as $\hat{d}$ using less information.)

To obtain such quantities $\{I(\tilde{d};V)\}_{V \subset A}$, we first define $S \subset 2^A$ as the set of all subsets $V \subset A$ that satisfy

$$I(\hat{d};V) > I(\hat{d};V \cap B) + I(\hat{d};V \cap C). \tag{A35}$$

We observe that, by Lemma A5, if $V \notin S$, then $I(\hat{d};V) = I(\hat{d};V \cap B) + I(\hat{d};V \cap C)$. It then follows from Lemma A6 that if $W \subset V \subset A$ and $W \in S$, then $V \in S$ as well.

Next we choose $\epsilon > 0$ sufficiently small that, for each $V \in S$, the following two conditions are satisfied:

(1) $I(\hat{d};V) > I(\hat{d};V \cap B) + I(\hat{d};V \cap C) + \epsilon$,
(2) $I(\hat{d};V) > I(\hat{d};W) + \epsilon$, for all $W \subset V, W \notin S$.

There is no problem arranging for condition (2) to be satisfied for any particular $V \in S$, since it follows readily from Constraint (ii) on $\hat{d}$ that if $W \subset V$ and $W \notin S$, then $I(\hat{d};V) > I(\hat{d};W)$. We also note that since A is finite, there are only a finite number of conditions to be satisfied as V and W vary, so it is possible to choose an $\epsilon > 0$ satisfying all of them.

Having chosen such an ϵ, we define the quantities $\{I(\tilde{d};V)\}_{V \subset A}$ by

$$I(\tilde{d};V) = \begin{cases} I(\hat{d};V) - \epsilon & V \in S \\ I(\hat{d};V) & \text{otherwise.} \end{cases} \tag{A36}$$

In words, we reduce the amount of information that is imparted about the sets in S by an amount ϵ, while leaving fixed the amount that is imparted about sets not in S. Intuitively, one could say that we are exploiting an inefficiency in the amount of information imparted by $\hat{d}$ about sets in S, and that the new descriptor $\tilde{d}$ is more efficient in terms of minimizing the information $I(d;A)$ without sacrificing utility.

We will now show that $\tilde{d}$ satisfies Constraints (i)–(iii) and $\tilde{u} = \hat{u}$. First, since $0 \leq I(\tilde{d};V) \leq I(\hat{d};V) \leq H(V)$ for all $V \subset A$, Constraint (i) is clearly satisfied.

For Constraint (ii), consider any $W \subset V \subset A$. If V and W are either both in S or both not in S then $I(\tilde{d};V) - I(\tilde{d};W) = I(\hat{d};V) - I(\hat{d};W)$, and Constraint (ii) is satisfied for $\tilde{d}$ since it is satisfied for $\hat{d}$. It only remains to consider the case that $V \in S$ and $W \notin S$. In this case, we have

$$I(\tilde{d};V) - I(\tilde{d};W) = I(\hat{d};V) - I(\hat{d};W) - \epsilon > 0, \tag{A37}$$

since V and ϵ satisfy condition (2) above. Furthermore,

$$\begin{aligned} I(\tilde{d};V) - I(\tilde{d};W) &= I(\hat{d};V) - I(\hat{d};W) - \epsilon \\ &\leq H(V) - H(W) - \epsilon \\ &< H(V) - H(W). \end{aligned}$$

Thus Constraint (ii) is satisfied.

To verify Constraint (iii), we must consider a number of cases, only one of which is nontrivial.

- If either
  - none of V, W, $V \cup W$ and $V \cap W$ belong to S,
  - all of V, W, $V \cup W$ and $V \cap W$ belong to S,
  - V and $V \cup W$ belong to S while W and $V \cap W$ do not, or
  - W and $V \cup W$ belong to S while V and $V \cap W$ do not,

  then the difference on the left-hand side of Constraint (iii) has the same value for $d = \hat{d}$ and $d = \tilde{d}$—that is, the changes in each term cancel out in the difference. Thus Constraint (iii) is satisfied for $\tilde{d}$ since it is satisfied for $\hat{d}$.
- If V, W, and $V \cup W$ belong to S while $V \cap W$ does not, then

$$\begin{aligned} &I(\tilde{d};V) + I(\tilde{d};W) - I(\tilde{d};V \cup W) - I(\tilde{d};V \cap W) \\ &\qquad = I(\hat{d};V) + I(\hat{d};W) - I(\hat{d};V \cup W) - I(\hat{d};V \cap W) - \epsilon. \end{aligned}$$

  The left-hand side of Constraint (iii) therefore decreases when moving from $d = \hat{d}$ to $d = \tilde{d}$. So Constraint (iii) is satisfied for $\tilde{d}$ since it is satisfied for $\hat{d}$.
- The nontrivial case is that $V \cup W$ belongs to S while V, W and $V \cap W$ do not. Then

$$\begin{aligned} &I(\tilde{d};V) + I(\tilde{d};W) - I(\tilde{d};V \cup W) - I(\tilde{d};V \cap W) \\ &\qquad = I(\hat{d};V) + I(\hat{d};W) - \Big(I(\hat{d};V \cup W) - \epsilon\Big) - I(\hat{d};V \cap W). \end{aligned} \tag{A38}$$

  By the definition of S and condition (1) on ϵ, we have

$$\begin{aligned} I(\hat{d};V \cup W) - \epsilon &> I\left(\hat{d};(V \cup W) \cap B\right) + I\left(\hat{d};(V \cup W) \cap C\right) \\ I(\hat{d};V) &= I(\hat{d};V \cap B) + I(\hat{d};V \cap C) \\ I(\hat{d};W) &= I(\hat{d};W \cap B) + I(\hat{d};W \cap C) \\ I(\hat{d};V \cap W) &= I\left(\hat{d};(V \cap W) \cap B\right) + I\left(\hat{d};(V \cap W) \cap C\right). \end{aligned}$$

  Substituting into (A38) we have

$$\begin{aligned} &I(\tilde{d};V) + I(\tilde{d};W) - I(\tilde{d};V \cup W) - I(\tilde{d};V \cap W) \\ &\quad < I(\hat{d};V \cap B) + I(\hat{d};W \cap B) \\ &\qquad - I(\hat{d};(V \cup W) \cap B) - I(\hat{d};(V \cap W) \cap B) \\ &\qquad\quad + I(\hat{d};V \cap C) + I(\hat{d};W \cap C) \\ &\qquad\qquad - I(\hat{d};(V \cup W) \cap C) - I(\hat{d};(V \cap W) \cap C). \end{aligned}$$

Applying Constraint (iii) on $\hat{d}$ twice to the right-hand side above, we have

$$\begin{aligned} I(\tilde{d};V) + I(\tilde{d};W) - I(\tilde{d};V\cup W) - I(\tilde{d};V\cap W) \\ < H(V\cap B) + H(W\cap B) - H\big((V\cup W)\cap B\big) - H\big((V\cap W)\cap B\big) \\ + H(V\cap C) + H(W\cap C) - H\big((V\cup W)\cap C\big) - H\big((V\cap W)\cap C\big). \end{aligned}$$

But Lemma A2 implies that $H(Z\cap B) + H(Z\cap C) = H(Z)$ for any subset $Z \subset A$. We apply this to the sets V, W, $V\cup W$ and $V\cap W$ to simplify the right-hand side above, yielding

$$I(\tilde{d};V) + I(\tilde{d};W) - I(\tilde{d};V\cup W) - I(\tilde{d};V\cap W) < H(V) + H(W) - H(V\cup W) - H(V\cap W),$$

as required.

No other cases are possible, since, as discussed above, any superset of a set in S must also be in S.

Finally, it is clear that no singleton subsets of A are in S. Thus $I(\tilde{d};a) = I(\hat{d};a)$ for each $a \in A$, and it follows that $\sum_{a\in A} \sigma(a) I(\tilde{d};a) = \hat{u}$.

We have now verified that $\tilde{d}$ satisfies Constraints (i)–(iii) and $U(\tilde{d}) = \hat{u}$. Furthermore, since $A \in S$ by assumption, we have $I(\tilde{d};A) < I(\hat{d};A)$. This contradicts the assertion that $\hat{d}$ minimizes $I(d;A)$ subject to Constraints (i)–(iii) and $U(d) \geq \hat{u}$. Therefore our assumption that $I(\hat{d};A) > I(\hat{d};B) + I(\hat{d};C)$ was incorrect, and we must instead have $I(\hat{d};A) = I(\hat{d};B) + I(\hat{d};C)$. □

Theorem A8. *The maximal utility function $U(y)$ is additive over independent subsystems in the sense that*

$$U_{\mathcal{A}}(y) = \max_{\substack{y_1+y_2=y \\ y_1,y_2\geq 0}} \left(U_{\mathcal{B}}(y_1) + U_{\mathcal{C}}(y_2)\right). \tag{A39}$$

Proof. For a given $y \geq 0$, let $\{I(\hat{d};V)\}_{V\subset A}$ maximixe $u = \sum_{a\in A} \sigma(a) I(d;a)$ subject to Constraints (i)–(iv). Combining Theorem A7 with Lemma A6, it follows that $I(\hat{d};V) = I(\hat{d};V\cap B) + I(\hat{d};V\cap C)$. We may therefore augment our linear program with the additional constraint,

(v) $I(d;V) = I(d;V\cap B) + I(d;V\cap C)$,

for each $V \subset A$, without altering the optimal solution.

Upon doing so, we can use this new constraint to eliminate the variables $I(d;V)$ for V not a subset of either B or C. We thereby reduce the set of variables from $\big\{I(d;V)\big\}_{V\subset A}$ to

$$\big\{I(d;V)\big\}_{V\subset B} \quad \cup \quad \big\{I(d;W)\big\}_{W\subset C}. \tag{A40}$$

We observe that this reduced linear program has the following structure: The variables in the set $\big\{I(d;V)\big\}_{V\subset B}$ are restricted by Constraints (i)–(iii) as applied to these variables. Separately, variables in the set $\big\{I(d;W)\big\}_{W\subset C}$ are also restricted by Constraints (i)–(iii), as they apply to the variables in this second set. The only constraint that simultaneously involves variables in both sets is Constraint (iv). This constraint can be rewritten as

$$u_{\mathcal{B}} + u_{\mathcal{C}} \leq y, \tag{A41}$$

with

$$u_{\mathcal{B}} = \sum_{b\in B} \sigma(b) I(d;b), \qquad u_{\mathcal{C}} = \sum_{c\in C} \sigma(c) I(d;c). \tag{A42}$$

This structure enables us to apply the decomposition principle for linear programs [112] to decompose the full program into two linear sub-programs, one on the variables $\big\{I(d;V)\big\}_{V\subset B}$ and one on $\big\{I(d;W)\big\}_{W\subset C}$, together with a coordinating program described by Constraint (iv). The desired result then follows from standard theorems of linear program decomposition [112]. □

Theorem A9. *$M_{\mathcal{A}}$ is additive over independent subsystems in the sense that*

$$M_{\mathcal{A}}(y) = \min_{\substack{y_1+y_2=y \\ y_1,y_2\geq 0}} \max\left\{M_{\mathcal{B}}(y_1), M_{\mathcal{C}}(y_2)\right\}. \tag{A43}$$

Proof. We define the function

$$F(y_1;y) = U_{\mathcal{B}}(y_1) + U_{\mathcal{C}}(y-y_1). \tag{A44}$$

The result of Theorem A8 can then be expressed as

$$U(y) = \max_{0\leq y_1 \leq y} F(y_1;y). \tag{A45}$$

We choose and fix an arbitrary y-value $\tilde{y} \geq 0$, and we will prove the desired result for $y = \tilde{y}$.

We observe that $F(y_1;\tilde{y})$ is concave in y_1 since $U_{\mathcal{B}}(y_1)$ and $U_{\mathcal{C}}(\tilde{y}-y_1)$ are. It follows that any local maximum of $F(y_1;\tilde{y})$ in y_1 is also a global maximum. We assume that the maximum of $F(y_1;\tilde{y})$ in y_1 is achieved at a single point $\hat{y}_1$ with $0 < \hat{y}_1 < \tilde{y}$. The remaining cases—that the maximum is achieved at $y_1 = 0$ or $y_1 = \tilde{y}$, or is achieved on a closed interval of y_1-values—are trivial extensions of this case.

Assuming we are in the case described above (and again invoking the concavity of F in y_1), $\hat{y}_1$ must be the unique point at which the derivative $\frac{\partial F}{\partial y_1}(y_1;\tilde{y})$ changes sign from positive to negative. This derivative can be written

$$\frac{\partial F}{\partial y_1}(y_1;\tilde{y}) = M_{\mathcal{B}}(y_1) - M_{\mathcal{C}}(\tilde{y}-y_1). \tag{A46}$$

It follows that $\hat{y}_1$ is the unique real number in $[0,\tilde{y}]$ satisfying

$$\begin{cases} M_{\mathcal{B}}(y_1) > M_{\mathcal{C}}(\tilde{y}-y_1) & y_1 < \hat{y}_1 \\ M_{\mathcal{B}}(y_1) < M_{\mathcal{C}}(\tilde{y}-y_1) & y_1 > \hat{y}_1. \end{cases} \tag{A47}$$

From inequalities (A47), and using the fact that $M_{\mathcal{B}}(y_1)$ and $M_{\mathcal{C}}(y_2)$ are nonincreasing, piecewise-constant functions, we see that either $M_{\mathcal{B}}(y_1)$ decreases at $y_1 = \hat{y}_1$, or $M_{\mathcal{C}}(y_2)$ decreases at $y_2 = \tilde{y} - \hat{y}_1$, or both. We analyze these cases separately.

Case A1. *$M_{\mathcal{B}}(y_1)$ decreases at $y_1 = \hat{y}_1$, while $M_{\mathcal{C}}(y_2)$ is constant in a neighborhood of $y_2 = \tilde{y} - \hat{y}_1$.*

Pick $\epsilon > 0$ sufficiently small so that $M_{\mathcal{C}}(y_2)$ has constant value for $y_2 \in (\tilde{y}-\hat{y}_1-\epsilon, \tilde{y}-\hat{y}_1+\epsilon)$. Then inequalities (A47) remain satisfied with $\tilde{y}$ replaced by any $y \in (\tilde{y}-\epsilon, \tilde{y}+\epsilon)$ and $\hat{y}_1$ fixed. Thus for y in this range, we have

$$U_{\mathcal{A}}(y) = U_{\mathcal{B}}(\hat{y}_1) + U_{\mathcal{C}}(y-\hat{y}_1). \tag{A48}$$

Taking the derivative of both sides in y at $y = \tilde{y}$ yields

$$M_{\mathcal{A}}(\tilde{y}) = M_{\mathcal{C}}(\tilde{y}-\hat{y}_1). \tag{A49}$$

We claim that

$$M_{\mathcal{C}}(\tilde{y}-\hat{y}_1) = \min_{0\leq y_1\leq \tilde{y}} \max\left\{M_{\mathcal{B}}(y_1), M_{\mathcal{C}}(\tilde{y}-y_1)\right\}. \tag{A50}$$

To prove this claim, we first note that, by the inequalities (A47),

$$\max\left\{M_{\mathcal{B}}(y_1), M_{\mathcal{C}}(\tilde{y}-y_1)\right\} = \begin{cases} M_{\mathcal{B}}(y_1) & y_1 < \hat{y}_1 \\ M_{\mathcal{C}}(\tilde{y}-y_1) & y_1 > \hat{y}_1. \end{cases} \tag{A51}$$

Since both $M_{\mathcal{B}}$ and $M_{\mathcal{C}}$ are piecewise-constant and nonincreasing, the minimax in Equation (A50) is achieved for values y_1 near $\hat{y}_1$. We therefore can restrict to the range $y_1 \in (\hat{y}_1 - \epsilon, \hat{y}_1 + \epsilon)$. Combining Equation (A51) with the conditions defining Case A1 and the definition of ϵ, we have

$$\begin{aligned} \max\left\{M_{\mathcal{B}}(y_1), M_{\mathcal{C}}(\tilde{y} - y_1)\right\} &= M_{\mathcal{B}}(y_1) > M_{\mathcal{C}}(\tilde{y} - \hat{y}_1) && \text{for } y_1 \in (\hat{y}_1 - \epsilon, \hat{y}_1) \\ \max\left\{M_{\mathcal{B}}(y_1), M_{\mathcal{C}}(\tilde{y} - y_1)\right\} &= M_{\mathcal{C}}(\tilde{y} - y_1) = M_{\mathcal{C}}(\tilde{y} - \hat{y}_1) && \text{for } y_1 \in (\hat{y}_1, \hat{y}_1 + \epsilon). \end{aligned} \tag{A52}$$

Thus the minimax in Equation (A50) is achieved at a value of $M_{\mathcal{C}}(\tilde{y} - \hat{y}_1)$ when $y_1 \in (\hat{y}_1, \hat{y}_1 + \epsilon)$, verifying Equation (A50). Combining with Equation (A49), we have

$$M_{\mathcal{A}}(y) = \min_{\substack{y_1 + y_2 = y \\ y_1, y_2 \geq 0}} \max\left\{M_{\mathcal{B}}(y_1), M_{\mathcal{C}}(y_2)\right\}, \tag{A53}$$

proving the theorem in this case.

Case A2. *$M_{\mathcal{B}}(y_1)$ is constant in a neighborhood of $y_1 = \hat{y}_1$, while $M_{\mathcal{C}}(y_2)$ decreases at $y_2 = \tilde{y} - \hat{y}_1$.*

In this case, we define $\hat{y}_2 = \tilde{y} - \hat{y}_1$. The proof then follows exactly as in Case 1, with $\mathcal{B}$ and $\mathcal{C}$ interchanged, and the subscripts 1 and 2 interchanged.

Case A3. *$M_{\mathcal{B}}(y_1)$ decreases at $y_1 = \hat{y}_1$ and $M_{\mathcal{C}}(y_2)$ decreases at $y_2 = \tilde{y} - \hat{y}_1$.*

This case only occurs at the y-values for which $U_{\mathcal{A}}(y)$ changes slope and $M_{\mathcal{A}}(y)$ changes value. At these nongeneric points, $M_{\mathcal{A}}(y)$ (defined as the derivative of $U_{\mathcal{A}}(y)$) is undefined. We therefore disregard this case. □

We now define $\tilde{M}_{\mathcal{A}}(x)$ as the reflection of $M_{\mathcal{A}}(y)$ with the dependent and independent variables interchanged. Since $M_{\mathcal{A}}$ is positive and nonincreasing, $\tilde{M}_{\mathcal{A}}$ is a well-defined function given by the formula

$$\tilde{M}_{\mathcal{A}}(x) = \max\{y : M_{\mathcal{A}}(y) \leq x\}. \tag{A54}$$

The following corollary gives a simpler expression of the additivity property of MUI.

Corollary A1. *If $\mathcal{A}$ consists of independent subsystems $\mathcal{B}$ and $\mathcal{C}$, then $\tilde{M}_{\mathcal{A}}(x) = \tilde{M}_{\mathcal{B}}(x) + \tilde{M}_{\mathcal{C}}(x)$ for all $x \geq 0$.*

Proof. Combining the above formula for $\tilde{M}_{\mathcal{A}}(x)$ with the result of Theorem A9, we write

$$\begin{aligned} \tilde{M}_{\mathcal{A}}(x) &= \max\{y : M_{\mathcal{A}}(y) \leq x\} \\ &= \max\left\{ y : \left(\min_{\substack{y_1 + y_2 = y \\ y_1, y_2 \geq 0}} \max\left\{M_{\mathcal{B}}(y_1), M_{\mathcal{C}}(y_2)\right\} \right) \leq x \right\} \\ &= \max\Big\{ y : \Big(\exists y_1, y_2 \geq 0. \left(y_1 + y_2 = y \ \wedge \ \max\left\{M_{\mathcal{B}}(y_1), M_{\mathcal{C}}(y_2)\right\} \leq x\right) \Big) \Big\} \\ &= \max\Big\{ (y_1 + y_2) : \Big(y_1, y_2 \geq 0 \ \wedge \ M_{\mathcal{B}}(y_1) \leq x \ \wedge \ M_{\mathcal{C}}(y_2) \leq x \Big) \Big\} \\ &= \max\{y_1 : M_{\mathcal{B}}(y_1) \leq x\} + \max\{y_2 : M_{\mathcal{B}}(y_2) \leq x\} \\ &= \tilde{M}_{\mathcal{B}}(x) + \tilde{M}_{\mathcal{C}}(x). \quad \square \end{aligned}$$

Appendix F. Marginal Utility of Information for Parity Bit Systems

Here we compute the MUI for a family of systems which exhibit exchange symmetry and have a constraint at the largest scale. Systems in this class have $N \geq 3$ components and information function given by

$$H(V) = H_{|V|} = \begin{cases} |V| & |V| \leq N-1 \\ N-1 & |V| = N. \end{cases} \tag{A55}$$

This includes Example **D** as the case $N = 3$. More generally, this family includes systems of $N-1$ independent random bits together with one parity bit.

Since these systems have exchange symmetry, we can apply the argument of Section 9 to obtain the reduced set of constraints:

(i) $0 \leq I_n \leq H_n$ for all $n \in \{1, \ldots, N\}$,
(ii) $0 \leq I_n - I_{n-1} \leq H_n - H_{n-1}$ for all $n \in \{1, \ldots, N\}$,
(iii) $I_n + I_m - I_{n+m-\ell} - I_\ell \leq H_n + H_m - H_{n+m-\ell} - H_\ell$ for all $n, m, \ell \in \{1, \ldots, N\}$,
(iv) $I_N \leq y$.

Above, I_n denotes information that a descriptor imparts about any set of n components, $0 \leq n \leq N$.

Constraint (ii), in the case $n = N$, yields

$$0 \leq I_N - I_{N-1} \leq H_N - H_{N-1} = 0, \tag{A56}$$

and therefore,

$$I_{N-1} = I_N. \tag{A57}$$

Constraint (iii), in the case $m + n \leq N - 1$, yields

$$I_m + I_n - I_{m+n} \leq H_m + H_n - H_{m+n} = 0, \tag{A58}$$

and therefore

$$I_m + I_n \leq I_{m+n} \qquad \text{for all } m, n \geq 0 \text{ with } m + n \leq N-1. \tag{A59}$$

By iteratively applying Equation (A59) we arrive at the inequality

$$(N-1)I_1 \leq I_{N-1}. \tag{A60}$$

Combining (A57), (A60) and Constraint (iv), we have

$$I_1 \leq \frac{I_{N-1}}{N-1} = \frac{I_N}{N-1} \leq \frac{y}{N-1}. \tag{A61}$$

By definition, the utility of a descriptor is

$$u = NI_1. \tag{A62}$$

Combining with (A61) yields the inequality

$$u \leq \frac{Ny}{N-1}. \tag{A63}$$

Inequality (A63) places a limit on the utility of any descriptor of this system. To complete the argument, we exhibit a descriptor for which equality holds in (A63). This descriptor is defined by

$$I_m = \begin{cases} \frac{m}{N-1} \min\{y, N-1\} & 0 \leq m \leq N-1 \\ \min\{y, N-1\} & m = N. \end{cases} \tag{A64}$$

It is straightforward to verify that Constraints (i)–(iv) are satisfied by this descriptor. Combining Equations (A62) and (A64), the utility of this descriptor is

$$u = \min\left\{\frac{Ny}{N-1}, N\right\}. \tag{A65}$$

By inequality (A63), this descriptor achieves optimal utility. Taking the derivative with respect to y, we obtain the marginal utility of information

$$M(y) = \begin{cases} \frac{N}{N-1} & 0 \leq y \leq N-1 \\ 0 & y > N-1. \end{cases} \tag{A66}$$

Setting $N = 3$, we recover the MUI for example **D** as stated in the main text, Equation (25).

References

1. Bar-Yam, Y. *Dynamics of Complex Systems*; Westview Press: Boulder, CO, USA, 2003.
2. Haken, H. *Information and Self-Organization: A Macroscopic Approach to Complex Systems*; Springer: New York, NY, USA, 2006.
3. Miller, J.H.; Page, S.E. *Complex Adaptive Systems: An Introduction to Computational Models of Social Life*; Princeton University Press: Princeton, NJ, USA, 2007.
4. Boccara, N. *Modeling Complex Systems*; Springer: New York, NY, USA, 2010.
5. Newman, M.E.J. Complex Systems: A Survey. *Am. J. Phys.* **2011**, *79*, 800–810.
6. Kwapień, J.; Drożdż, S. Physical approach to complex systems. *Phys. Rep.* **2012**, *515*, 115–226.
7. Sayama, H. *Introduction to the Modeling and Analysis of Complex Systems*; Open SUNY: Bunghamton, NY, USA, 2015.
8. Sethna, J.P. *Statistical Mechanics: Entropy, Order Parameters, and Complexity*; Oxford University Press: Oxford, UK, 2006.
9. Shannon, C. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423.
10. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*; Wiley: Hoboken, NJ, USA, 1991.
11. Prokopenko, M.; Boschetti, F.; Ryan, A.J. An information-theoretic primer on complexity, self-organization, and emergence. *Complexity* **2009**, *15*, 11–28.
12. Gallagher, R.G. *Information Theory and Reliable Communication*; Wiley: Hoboken, NJ, USA, 1968.
13. Bar-Yam, Y. Multiscale complexity/entropy. *Adv. Complex Syst.* **2004**, *7*, 47–63.
14. Bar-Yam, Y. Multiscale variety in complex systems. *Complexity* **2004**, *9*, 37–45.
15. Bar-Yam, Y.; Harmon, D.; Bar-Yam, Y. Computationally tractable pairwise complexity profile. *Complexity* **2013**, *18*, 20–27..
16. Metzler, R.; Bar-Yam, Y. Multiscale complexity of correlated Gaussians. *Phys. Rev. E* **2005**, *71*, 046114.
17. Gheorghiu-Svirschevski, S.; Bar-Yam, Y. Multiscale analysis of information correlations in an infinite-range, ferromagnetic Ising system. *Phys. Rev. E* **2004**, *70*, 066115.
18. Stacey, B.C.; Allen, B.; Bar-Yam, Y. Multiscale Information Theory for Complex Systems: Theory and Applications. In *Information and Complexity*; Burgin, M., Calude, C.S., Eds.; World Scientific: Singapore, 2017; pp. 176–199.
19. Grassberger, P. Toward a quantitative theory of self-generated complexity. *Int. J. Theor. Phys.* **1986**, *25*, 907–938.
20. Crutchfield, J.P.; Young, K. Inferring statistical complexity. *Phys. Rev. Lett.* **1989**, *63*, 105–108.
21. Crutchfield, J.P. The calculi of emergence: Computation, dynamics and induction. *Phys. D Nonlinear Phenom.* **1994**, *75*, 11–54.
22. Misra, V.; Lagi, M.; Bar-Yam, Y. *Evidence of Market Manipulation in the Financial Crisis*; Technical Report 2011-12-01; NECSI: Cambridge, MA, USA, 2011.
23. Harmon, D.; Lagi, M.; de Aguiar, M.A.; Chinellato, D.D.; Braha, D.; Epstein, I.R.; Bar-Yam, Y. Anticipating Economic Market Crises Using Measures of Collective Panic. *PLoS ONE* **2015**, *10*, e0131871.
24. Green, H. *The Molecular Theory of Fluids*; North–Holland: Amsterdam, The Netherlands, 1952.

25. Nettleton, R.E.; Green, M.S. Expression in terms of molecular distribution functions for the entropy density in an infinite system. *J. Chem. Phys.* **1958**, *29*, 1365–1370.
26. Wolf, D.R. Information and Correlation in Statistical Mechanical Systems. Ph.D. Thesis, University of Texas, Austin, TX, USA, 1996.
27. Kardar, M. *Statistical Physics of Particles*; Cambridge University Press: Cambridge, UK, 2007.
28. Kadanoff, L.P. Scaling laws for Ising models near T_c. *Physics* **1966**, *2*, 263.
29. Wilson, K.G. The renormalization group: Critical phenomena and the Kondo problem. *Rev. Mod. Phys.* **1975**, *47*, 773.
30. McGill, W.J. Multivariate information transmission. *Psychometrika* **1954**, *46*, 26–45.
31. Han, T.S. Multiple mutual information and multiple interactions in frequency data. *Inf. Control* **1980**, *46*, 26–45.
32. Yeung, R.W. A new outlook on Shannon's information measures. *IEEE Trans. Inf. Theory* **1991**, *37*, 466–474.
33. Jakulin, A.; Bratko, I. Quantifying and visualizing attribute interactions. *arXiv* **2003**, arXiv:cs.AI/0308002.
34. Bell, A.J. The co-information lattice. In Proceedings of the Fifth International Workshop on Independent Component Analysis and Blind Signal Separation (ICA), Nara, Japan, 1–4 April 2003; Volume 2003.
35. Bar-Yam, Y. A mathematical theory of strong emergence using multiscale variety. *Complexity* **2004**, *9*, 15–24.
36. Krippendorff, K. Information of interactions in complex systems. *Int. J. Gen. Syst.* **2009**, *38*, 669–680.
37. Leydesdorff, L. Redundancy in systems which entertain a model of themselves: Interaction information and the self-organization of anticipation. *Entropy* **2010**, *12*, 63–79.
38. Kolchinsky, A.; Rocha, L.M. Prediction and modularity in dynamical systems. *arXiv* **2011**, arXiv:1106.3703.
39. James, R.G.; Ellison, C.J.; Crutchfield, J.P. Anatomy of a bit: Information in a time series observation. *Chaos Interdiscip. J. Nonlinear Sci.* **2011**, *21*, 037109.
40. Tononi, G.; Sporns, O.; Edelman, G.M. A measure for brain complexity: Relating functional segregation and integration in the nervous system. *Proc. Natl. Acad. Sci. USA* **1994**, *91*, 5033–5037.
41. Ay, N.; Olbrich, E.; Bertschinger, N.; Jost, J. A unifying framework for complexity measures of finite systems. In Proceedings of the European Complex Systems Society (ECCS06), Oxford, UK, 25 September 2006.
42. Bar-Yam, Y. *Complexity of Military Conflict: Multiscale Complex Systems Analysis of Littoral Warfare*; Technical Report; NECSI: Cambridge, MA, USA, 2003.
43. Granovsky, B.L.; Madras, N. The noisy voter model. *Stoch. Process. Appl.* **1995**, *55*, 23–43.
44. Faddeev, D.K. On the concept of entropy of a finite probabilistic scheme. *Uspekhi Mat. Nauk* **1956**, *11*, 227–231.
45. Khinchin, A.I. *Mathematical Foundations of Information Theory*; Dover: New York, NY, USA, 1957.
46. Lee, P. On the axioms of information theory. *Ann. Math. Stat.* **1964**, *35*, 415–418.
47. Rényi, A. *Probability Theory*; Akadémiai Kiadó: Budapest, Hungary, 1970.
48. Daróczy, Z. Generalized information functions. *Inf. Control* **1970**, *16*, 36–51.
49. Dos Santos, R.J. Generalization of Shannon's theorem for Tsallis entropy. *J. Math. Phys.* **1997**, *38*, 4104–4107.
50. Abe, S. Axioms and uniqueness theorem for Tsallis entropy. *Phys. Lett. A* **2000**, *271*, 74–79.
51. Tsallis, C. Possible generalization of Boltzmann-Gibbs statistics. *J. Stat. Phys.* **1988**, *52*, 479–487.
52. Gell-Mann, M.; Tsallis, C. *Nonextensive Entropy: Interdisciplinary Applications*; Oxford University Press: Oxford, UK, 2004.
53. Furuichi, S. Information theoretical properties of Tsallis entropies. *J. Math. Phys.* **2006**, *47*, 023302.
54. Steudel, B.; Janzing, D.; Schölkopf, B. Causal Markov condition for submodular information measures. *arXiv* **2010**, arXiv:1002.4020.
55. Dougherty, R.; Freiling, C.; Zeger, K. Networks, matroids, and non-Shannon information inequalities. *IEEE Trans. Inf. Theory* **2007**, *53*, 1949–1969.
56. Li, M.; Vitányi, P. *An Introduction to Kolmogorov Complexity and Its Applications*; Springer Science & Business Media: New York, NY, USA, 2009.
57. Chaitin, G.J. A theory of program size formally identical to information theory. *J. ACM* **1975**, *22*, 329–340.
58. May, R.M.; Arinaminpathy, N. Systemic risk: The dynamics of model banking systems. *J. R. Soc. Interface* **2010**, *7*, 823–838.
59. Haldane, A.G.; May, R.M. Systemic risk in banking ecosystems. *Nature* **2011**, *469*, 351–355.
60. Beale, N.; Rand, D.G.; Battey, H.; Croxson, K.; May, R.M.; Nowak, M.A. Individual versus systemic risk and the Regulator's Dilemma. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 12647–12652.

61. Erickson, M.J. *Introduction to Combinatorics*; Wiley: Hoboken, NJ, USA, 1996.
62. Williams, P.L.; Beer, R.D. Nonnegative decomposition of multivariate information. *arXiv* **2010**, arXiv:1004.2515.
63. James, R.G.; Crutchfield, J.P. Multivariate Dependence Beyond Shannon Information. *arXiv* **2016**, arXiv:1609.01233.
64. Perfect, H. Independence theory and matroids. *Math. Gaz.* **1981**, *65*, 103–111.
65. Studenỳ, M.; Vejnarová, J. The multiinformation function as a tool for measuring stochastic dependence. In *Learning in Graphical Models*; Springer: Dodrecht, The Netherlands, 1998; pp. 261–297.
66. Schneidman, E.; Still, S.; Berry, M.J.; Bialek, W. Network information and connected correlations. *Phys. Rev. Lett.* **2003**, *91*, 238701.
67. Polani, D. Foundations and formalizations of self-organization. In *Advances in Applied Self-Organizing Systems*; Springer: New York, NY, USA, 2008; pp. 19–37.
68. Wets, R.J.B. Programming Under Uncertainty: The Equivalent Convex Program. *SIAM J. Appl. Math.* **1966**, *14*, 89–105.
69. James, R.G. Python Package for Information Theory. *Zenodo* **2017**, doi:10.5281/zenodo.235071.
70. Slonim, N.; Tishby, L. Agglomerative information bottleneck. *Adv. Neural Inf. Process. Syst. NIPS* **1999**, *12*, 617–623.
71. Shalizi, C.R.; Crutchfield, J.P. Information bottlenecks, causal states, and statistical relevance bases: How to represent relevant information in memoryless transduction. *Adv. Complex Syst.* **2002**, *5*, 91–95,
72. Tishby, N.; Pereira, F.C.; Bialek, W. The information bottleneck method. *arXiv* **2000**, arXiv:physics/0004057.
73. Ziv, E.; Middendorf, M.; Wiggins, C. An information-theoretic approach to network modularity. *Phys. Rev. E* **2005**, *71*, 046117.
74. Peng, C.K.; Buldyrev, S.V.; Havlin, S.; Simons, M.; Stanley, H.E.; Goldberger, A.L. Mosaic organization of DNA nucleotides. *Phys. Rev. E* **1994**, *49*, 1685–1689.
75. Gell-Mann, M.; Lloyd, S. Information measures, effective complexity, and total information. *Complexity* **1996**, *2*, 44–52.
76. Hu, K.; Ivanov, P.; Chen, Z.; Carpena, P.; Stanley, H.E. Effect of Trends on Detrended Fluctuation Analysis. *Phys. Rev. E* **2002**, *64*, 011114.
77. Vereshchagin, N.; Vitányi, P. Kolmogorov's structure functions and model selection. *IEEE Trans. Inf. Theory* **2004**, *50*, 3265–3290.
78. Grünwald, P.; Vitányi, P. Shannon information and Kolmogorov complexity. *arXiv* **2004**, arXiv:cs.IT/0410002.
79. Vitányi, P. Meaningful information. *IEEE Trans. Inf. Theory* **2006**, *52*, 4617–4626.
80. Moran, P.A.P. Random processes in genetics. *Math. Proc. Camb. Philos. Soc.* **1958**, *54*, 60–71.
81. Harmon, D.; Stacey, B.C.; Bar-Yam, Y. *Networks of Economic Market Independence and Systemic Risk*; Technical Report 2009-03-01 (updated); NECSI: Cambridge, MA, USA, 2010.
82. Stacey, B.C. Multiscale Structure in Eco-Evolutionary Dynamics. Ph.D. Thesis, Brandeis University, Waltham, MA, USA, 2015.
83. Domb, C.; Green, M.S. (Eds.) *Phase Transitions and Critical Phenomena*; Academic Press: New York, NY, USA, 1972.
84. Kardar, M. *Statistical Physics of Fields*; Cambridge University Press: Cambridge, UK, 2007.
85. Jacob, F.; Monod, J. Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.* **1961**, *3*, 318–356.
86. Britten, R.J.; Davidson, E.H. Gene regulation for higher cells: A theory. *Science* **1969**, *165*, 349–357.
87. Carey, M.; Smale, S. *Transcriptional Regulation in Eukaryotes: Concepts, Strategies, and Techniques*; Cold Spring Harbor Laboratory Press: Cold Spring Harbor, NY, USA, 2001.
88. Elowitz, M.B.; Levine, A.J.; Siggia, E.D.; Swain, P.S. Stochastic gene expression in a single cell. *Science* **2002**, *297*, 1183–1186.
89. Lee, T.I.; Rinaldi, N.J.; Robert, F.; Odom, D.T.; Bar-Joseph, Z.; Gerber, G.K.; Hannett, N.M.; Harbison, C.T.; Thompson, C.M.; Simon, I.; et al. Transcriptional Regulatory Networks in *Saccharomyces cerevisiae*. *Science* **2002**, *298*, 799–804.
90. Boyer, L.A.; Lee, T.I.; Cole, M.F.; Johnstone, S.E.; Levine, S.S.; Zucker, J.P.; Guenther, M.G.; Kumar, R.M.; Murray, H.L.; Jenner, R.G.; et al. Core Transcriptional Regulatory Circuitry in Human Embryonic Stem Cells. *Cell* **2005**, *122*, 947–956.
91. Chowdhury, S.; Lloyd-Price, J.; Smolander, O.P.; Baici, W.C.; Hughes, T.R.; Yli-Harja, O.; Chua, G.; Ribeiro, A.S. Information propagation within the Genetic Network of *Saccharomyces cerevisiae*. *BMC Syst. Biol.* **2010**, *4*, 143.

92. Hopfield, J.J. Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci. USA* **1982**, *79*, 2554–2558.
93. Rabinovich, M.I.; Varona, P.; Selverston, A.I.; Abarbanel, H.D.I. Dynamical principles in neuroscience. *Rev. Mod. Phys.* **2006**, *78*, 1213–1265.
94. Schneidman, E.; Berry, M.J.; Segev, R.; Bialek, W. Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature* **2006**, *440*, 1007–1012.
95. Bonabeau, E.; Dorigo, M.; Theraulaz, G. *Swarm Intelligence: From Natural to Artificial Systems*; Oxford University Press: Oxford, UK, 1999.
96. Vicsek, T.; Zafeiris, A. Collective motion. *Phys. Rep.* **2012**, *517*, 71–140.
97. Berdahl, A.; Torney, C.J.; Ioannou, C.C.; Faria, J.J.; Couzin, I.D. Emergent sensing of complex environments by mobile animal groups. *Science* **2013**, *339*, 574–576.
98. Ohtsuki, H.; Hauert, C.; Lieberman, E.; Nowak, M.A. A simple rule for the evolution of cooperation on graphs and social networks. *Nature* **2006**, *441*, 502–505.
99. Allen, B.; Lippner, G.; Chen, Y.T.; Fotouhi, B.; Momeni, N.; Yau, S.T.; Nowak, M.A. Evolutionary dynamics on any population structure. *Nature* **2017**, *544*, 227–230.
100. Mandelbrot, B.; Taylor, H. On the distribution of stock price differences. *Oper. Res.* **1967**, *15*, 1057–1062.
101. Mantegna, R.N. Hierarchical structure in financial markets. *Eur. Phys. J. B Condens. Matter Complex Syst.* **1999**, *11*, 193–197.
102. Sornette, D. *Why Stock Markets Crash: Critical Events in Complex Financial Systems*; Princeton University Press: Princeton, NJ, USA, 2004.
103. May, R.M.; Levin, S.A.; Sugihara, G. Complex systems: Ecology for bankers. *Nature* **2008**, *451*, 893–895.
104. Schweitzer, F.; Fagiolo, G.; Sornette, D.; Vega-Redondo, F.; Vespignani, A.; White, D.R. Economic Networks: The New Challenges. *Science* **2009**, *325*, 422–425.
105. Harmon, D.; De Aguiar, M.; Chinellato, D.; Braha, D.; Epstein, I.; Bar-Yam, Y. Predicting economic market crises using measures of collective panic. *arXiv* **2011**, arXiv:1102.2620.
106. Schrödinger, E. *What Is Life? The Physical Aspect of the Living Cell and Mind*; Cambridge University Press: Cambridge, UK, 1944.
107. Brillouin, L. The negentropy principle of information. *J. Appl. Phys.* **1953**, *24*, 1152–1163.
108. Stacey, B.C. Multiscale Structure of More-than-Binary Variables. *arXiv* **2017**, arXiv:1705.03927.
109. Ashby, W.R. *An Introduction to Cybernetics*; Chapman & Hall: London, UK, 1956.
110. Stacey, B.C.; Bar-Yam, Y. *Principles of Security: Human, Cyber, and Biological*; Technical Report 2008-06-01; NECSI: Cambridge, MA, USA, 2008.
111. Dorogovtsev, S.N. *Lectures on Complex Networks*; Oxford University Press: Oxford, UK, 2010.
112. Dantzig, G.B.; Wolfe, P. Decomposition principle for linear programs. *Oper. Res.* **1960**, 8, 101–111.