OPEN ACCESS

IOP Publishing

Journal of Physics A: Mathematical and Theoretical

J. Phys. A: Math. Theor. **53** (2020) 485001 (27pp)

https://doi.org/10.1088/1751-8121/abb723

An operational information decomposition via synergistic disclosure

Fernando E Rosas[1,2,3,7,*], Pedro A M Mediano[4,7,*], Borzoo Rassouli[5] and Adam B Barrett[6]

[1] Data Science Institute, Imperial College London, London SW7 2AZ
[2] Center for Psychedelic Research, Department of Medicine, Imperial College London, London SW7 2DD
[3] Center for Complexity Science, Imperial College London, London SW7 2AZ
[4] Department of Psychology, University of Cambridge, Cambridge CB2 3EB, United States of America
[5] School of Computer Science and Electronic Engineering, University of Essex, Colchester CO4 3SQ, United States of America
[6] Sackler Center for Consciousness Science, Department of Informatics, University of Sussex, Brighton BN1 9RH, United States of America

E-mail: f.rosas@imperial.ac.uk and pam83@cam.ac.uk

Received 23 April 2020, revised 22 August 2020
Accepted for publication 10 September 2020
Published 5 November 2020

Abstract

Multivariate information decompositions hold promise to yield insight into complex systems, and stand out for their ability to identify synergistic phenomena. However, the adoption of these approaches has been hindered by there being multiple possible decompositions, and no precise guidance for preferring one over the others. At the heart of this disagreement lies the absence of a clear operational interpretation of what synergistic information is. Here we fill this gap by proposing a new information decomposition based on a novel operationalisation of informational synergy, which leverages recent developments in the literature of data privacy. Our decomposition is defined for any number of information sources, and its atoms can be calculated using elementary optimisation techniques. The decomposition provides a natural coarse-graining that scales gracefully with the system's size, and is applicable in a wide range of scenarios of practical interest.

[7] FR and PM contributed equally to this work.
[*] Authors to whom any correspondence should be addressed.

Keywords: complex systems, synergy, information theory, high-order statistics, partial information decomposition

(Some figures may appear in colour only in the online journal)

1. Introduction

The familiarity with which we relate to the notion of 'information'—due to its central role in our modern worldview—is at odds with the mysteries still surrounding some of its fundamental properties. One such mystery is the nature and role of *synergistic information*, which is present in systems that exhibit global interdependencies that are not traceable from any of their subsystems. Synergistic relationships have shown to be instrumental in a wide range of systems, including the nervous system [1, 2], artificial neural networks [3], cellular automata [4], and music scores [5]. Furthermore, the concept of synergy traces a particularly promising road to formalise the notion of 'the whole being greater than the sum of the parts', one of the long-standing aims of complexity science [6].

Informational synergy has been studied following various approaches, including redundancy-synergy balances [5, 7–9], information geometry [10, 11], and others. Within this literature, one of the most elegant and powerful proposals is the *partial information decomposition* (PID) framework [12], which divides information into *redundant* (contained in every part of the system), *unique* (contained in only one part), and *synergistic* (contained in the whole, but not in any part) components. One peculiarity of the PID framework is the absence of precise prescriptions about how synergy should be quantified [13]; and despite numerous efforts, an agreed-upon measure of synergy remains elusive [14–17]. Most approaches to quantify synergy proceed by postulating axioms encoding some 'intuitive' desiderata, which should ideally lead towards a unique measure—following the well-known axiomatic derivation of Shannon's entropy [18]. Unfortunately, a number of critical incompatibilities between some of these axioms have been reported [19, 20], which reveals the limitations of our intuition as a guide within the counterintuitive realms of high-order statistics.

Building on these remarks, we argue that measures of synergy with little concrete, operational meaning provide a limited advance from mere qualitative criteria. Moreover, as argued by Kolchinsky [20], there might exist not a single but multiple reasonable definitions of synergy, and hence it is crucial to clarify what each proposed measure is capturing [21]. There have been a few attempts to formulate operational measures of redundant [14, 16] and unique information [15, 22, 23], but these efforts are still in progress, and apply only indirectly to synergy [24]. Providing a clear operational meaning for synergy is, to the best of our knowledge, an important unresolved challenge.

In this paper we put forward a *synergy-centered information decomposition*, rooted on the notion of *synergistic data disclosure* from the literature of data privacy [25, 26] and synergistic variables introduced in reference [27]. In this decomposition, synergy corresponds to the information that can be disclosed about a system without revealing the state of any of its parts. This measure is computable via elementary optimisation techniques and is, to the best of our knowledge, the first to provide a direct operational interpretation for synergistic information. Moreover, our proposed decomposition is applicable to any number of source variables, and its operational meaning provides natural coarse-grainings that enable useful tools for practical analysis.

The paper is structured as follows. First, section 2 introduces our operational definition of synergy, and section 3 uses it to build our proposed decomposition. The decomposition's

coarse-graining is discussed in section 4, and the special case of *self-synergy* in section 5. Finally, the relationship with other decompositions is studied in section 6.

2. Synergy and data disclosure

Our goal is to develop a method to decompose the information that a multivariate system $\boldsymbol{X} := (X_1, \ldots, X_n)$ provides about a target variable Y, as quantified by Shannon's mutual information $I(\boldsymbol{X}; Y)$. Our approach consists of three steps:

(a) Introduce *synergistic channels*, which convey information about $\boldsymbol{X}$ but not about any of its parts (section 2.1);
(b) Define *synergistic disclosure* as the maximum amount of information about Y that can be obtained through a synergistic channel on $\boldsymbol{X}$ (section 2.2); and
(c) Build an information decomposition by computing the synergistic disclosure for every node of a lattice, and using the Möbius inversion formula (section 3).

The rest of this section provides technical details about synergistic disclosure, building upon the work recently reported in references [25, 26].

2.1. Synergistic channels

Consider a system described by n variables, $\boldsymbol{X} := (X_1, \ldots, X_n)$, where each X_k takes values on a discrete alphabet $\mathcal{X}_k$ of cardinality $|\mathcal{X}_k|$. Consider also a channel that is applied on $\boldsymbol{X}$ to generate a scalar observable V, which is characterised by a conditional distribution $p_{V|\boldsymbol{X}}$. We are interested in a particular class of observables, which carry information about $\boldsymbol{X}$ while revealing no information about specific subsystems.

Subsystems of $\boldsymbol{X}$ can be represented by sets of indices of the form $\alpha = \{n_1, \ldots, n_k\} \subset [n]$, with $[n] := \{1, \ldots, n\}$ being a shorthand notation, and the corresponding subsystem being denoted by $\boldsymbol{X}^\alpha = (X_{n_1}, \ldots, X_{n_k})$. We consider collections of subsystems, which are represented by *source-sets* of the form $\boldsymbol{\alpha} = \{\alpha_1, \ldots, \alpha_L\}$, where $\alpha_j \subset [n]$ for all $i = 1, \ldots, L$. For example, possible source-sets for $n = 2$ are $\{\varnothing\}, \{\{1\}\}, \{\{1\}, \{1, 2\}\}$, etc. With the notion of source-set in hand, we can formally define synergistic channels as follows:

Definition 1. A channel $p_{V|\boldsymbol{X}}$ is $\boldsymbol{\alpha}$-synergistic for $\boldsymbol{\alpha} = \{\alpha_1, \ldots, \alpha_L\}$, if $V \perp\!\!\!\perp \boldsymbol{X}^{\alpha_i}$, $\forall i = 1, \ldots, L$. The set of all $\boldsymbol{\alpha}$-synergistic channels is denoted by

$$\mathcal{C}(\boldsymbol{X}; \boldsymbol{\alpha}) = \left\{ p_{V|\boldsymbol{X}} \;\middle|\; V \perp\!\!\!\perp \boldsymbol{X}^{\alpha_i}, \forall i \in [L] \right\}. \tag{1}$$

A variable V generated via an $\boldsymbol{\alpha}$-synergistic channel is said to be an $\boldsymbol{\alpha}$-synergistic observable.

Due to the independence constraints (denoted by $\perp\!\!\!\perp$), an $\boldsymbol{\alpha}$-synergistic observable V satisfies $I(\boldsymbol{X}^{\alpha_i}; V) = 0$ for all $i = 1, \ldots, L$. Thus the name synergistic: by construction, an $\boldsymbol{\alpha}$-synergistic observable V might convey information about the whole, $\boldsymbol{X}$, while disclosing no information about the corresponding parts $\boldsymbol{X}^{\alpha_1}, \ldots, \boldsymbol{X}^{\alpha_L}$.

Example 1. If $\boldsymbol{X} = (X_1, X_2)$ are two independent fair coins, then the observable $V = X_1 \mathtt{xor} X_2$ given by

$$X_1 \mathtt{xor} X_2 := \begin{cases} 0 & \text{if } X_1 = X_2 \\ 1 & \text{if } X_1 \neq X_2. \end{cases} \tag{2}$$

is $\boldsymbol{\alpha}$-synergistic for $\boldsymbol{\alpha} = \{\{1\}, \{2\}\}$.

Note that $p_{V|\boldsymbol{X}}$ can be depicted as a rectangular matrix. Elegant algebraic methods for characterising synergistic channels based on this matrix representation are available, and are discussed in appendix A.

2.2. Synergistic disclosure: definition and operational interpretation

Now that synergistic channels have been defined, let us formulate our measure of synergistic disclosure. To do this, consider a target variable Y, potentially having some dependence on $\boldsymbol{X}$ according to a given joint distribution $p_{\boldsymbol{X},Y}$. We are interested in quantifying to what extent the collective properties of $\boldsymbol{X}$ can predict Y without revealing any information about the subsystems $\boldsymbol{X}^{\alpha_1}, \ldots, \boldsymbol{X}^{\alpha_L}$. This intuition can be naturally operationalised by the mutual information between Y and the $\boldsymbol{\alpha}$-synergistic observables of $\boldsymbol{X}$, as described in the next definition.

Definition 2. The $\boldsymbol{\alpha}$-synergy between sources $\boldsymbol{X}$ and target Y is defined as

$$S^{\boldsymbol{\alpha}}(\boldsymbol{X} \to Y) := \sup_{p_{V|\boldsymbol{X}} \in \mathcal{C}(\boldsymbol{X};\boldsymbol{\alpha}) : V - \boldsymbol{X} - Y} I(V;Y). \tag{3}$$

Above, the notation $V - \boldsymbol{X} - Y$ means that V and Y are conditionally independent given $\boldsymbol{X}$ (i.e. they form a Markov chain). Consequently, the supremum in equation (3) is calculated over all the $\boldsymbol{\alpha}$-synergistic channels $p_{V|\boldsymbol{X}}$, so that the joint distribution of (V, Y) over which $I(V;Y)$ is calculated is of the form

$$\mathbb{P}\{V = v, Y = y\} = \sum_{\boldsymbol{x} \in \prod_{i=1}^{n} \mathcal{X}_i} p_{V|\boldsymbol{X}}(v|\boldsymbol{x}) p_{\boldsymbol{X},Y}(\boldsymbol{x}, y).$$

Additionally, it can be verified that the supremum in (3) is attained, and hence, it is a maximum [28]. Finally, this definition can used to extend the notion of synergy over any f-information, as discussed in appendix B.

Definition 2 has a straightforward operational interpretation in the context of data transmission over noisy channels, following Shannon's channel coding theorem [29]. To make this explicit, let us consider a point-to-point communication system where a sender wishes to reliably communicate a message M at a rate R bits per channel usage to a receiver over a noisy communication channel, specified as $p_{V|Y}(v|y) = \sum_{\boldsymbol{x}} p_{V|\boldsymbol{X}}(v|\boldsymbol{x}) p_{\boldsymbol{X}|Y}(\boldsymbol{x}|y)$, where $p_{V|\boldsymbol{X}} \in \mathcal{C}(\boldsymbol{X};\boldsymbol{\alpha})$ can be arbitrarily chosen. To this end, the sender encodes the message into a codeword $(Y_1, \ldots, Y_t)$ and transmits it over the channel in t successive channel uses. Upon receiving the noisy sequence $(V_1, \ldots, V_t)$, the receiver decodes it to obtain the estimate $\hat{M}$ of the message. In this scenario, $S^{\boldsymbol{\alpha}}(\boldsymbol{X} \to Y)$ denotes the highest rate R, over the choice of $p_{V|\boldsymbol{X}}$, such that the probability of decoding error can be made to decay asymptotically to zero as the code block length t grows [[30], chapter 7].

Additionally, definition 2 has a second operational interpretation related to data privacy, as discussed in references [25, 26]. For this, consider a given database $\boldsymbol{X}$ of n random variables that in this case are considered to be 'data samples'. A disclosure mechanism $p_{V|\boldsymbol{X}}$ is said to be ϵ-Bayesian differentially private [31, 32] if, for any index i and subset of indices α, we have $I(X_i; V|\boldsymbol{X}^{\alpha}) \leqslant \epsilon$. Interestingly, when considering adversaries with no background knowledge about the database, then all ϵ-Bayesian differentially private mechanisms satisfy $I(X_i; V) \leqslant \epsilon$—i.e. the conditioning in the mutual information (which accounts for the adversary's background knowledge) is dropped. Now, let us consider the utility-privacy trade-off in which one aims to reveal information about a variable of interest Y contained in the database $\boldsymbol{X}$ while observing privacy constraints [33–36]. Then, $S^{\boldsymbol{\alpha}}(\boldsymbol{X} \to Y)$ corresponds to

the maximum utility that can be attained if an ϵ-Bayesian differentially private mechanism is harnessed against an adversary with no background knowledge, for $\epsilon = 0$. While this simple construction applies to $\boldsymbol{\alpha} = \{\{1\}, \ldots, \{n\}\}$, it can be easily extended to other sets of indices.

These results strongly contrast with some previous approaches to information decomposition, which have typically proceeded by writing down an axiomatic base and then formulating a measure consistent with those desiderata. Is worth noting, however, some notable exceptions that have attempted to operationalise PID via decision theory [15, 20], game theory [14], probability mass exclusions [16], and information theoretic secrecy [22, 23]. However, most of these efforts are not entirely satisfactory, as they do not provide concrete implications to the exact value that an information atom attains [37]. Also, is worth mentioning that quantities with known operational interpretation such as Wiener's common information have been considered as redundancy metric; unfortunately, they fail to provide some of the basic properties that are desirable for a PID [38].

Finally, it is important to remark that all the existent operationalisations of PID apply on either measures of redundancy or unique information, and their results apply to synergy only indirectly. To the best of our knowledge, ours is the first operationalisation that applies directly to synergy itself.

2.3. Fundamental properties

Let us explore some basic properties of our measure of synergy, $S^{\boldsymbol{\alpha}}$. A first fortunate feature is that this quantity is computable via elementary optimisation techniques, which is a direct extension of reference [[25], theorem 1].

Theorem 1. *The supremum in equation* (3) *is always attained, and the corresponding synergistic channel can be obtained as the solution to a standard linear-programming problem.*

While the proof of theorem 1 is omitted, interested readers can find the corresponding details in reference [[26], section 3]. Additionally, software alternatives to compute S^{α} are discussed in section 7. It is worth noting that our algorithm is efficient for systems with moderate state spaces (e.g. binary systems with up to 6 variables), but scales poorly with system size. While some methods for providing lower bounds on $S^{\boldsymbol{\alpha}}$ have been outlined in reference [[26], section 5], finding efficient algorithms for large systems is an interesting avenue for future research.

Despite the guarantees provided by theorem 1, it is useful to have simple bounds. Note that, due to the data processing inequality, $S^{\boldsymbol{\alpha}}$ satisfies $S^{\boldsymbol{\alpha}}(\boldsymbol{X} \to Y) \leqslant I(\boldsymbol{X}; Y)$ for all $\boldsymbol{\alpha}$. The following result introduces a less trivial upper bound.

Proposition 1. *The following upper bound holds for* $S^{\boldsymbol{\alpha}}$*:*

$$S^{\boldsymbol{\alpha}}(\boldsymbol{X} \to Y) \leqslant \min_{j \in \{1,\ldots,L\}} I(Y; \boldsymbol{X}^{-\alpha_j} | \boldsymbol{X}^{\alpha_j}), \tag{4}$$

where $\boldsymbol{X}^{-\alpha_j} \triangleq \{X_1, \ldots, X_n\} \backslash \{X_{n_1}, \ldots, X_{n_k}\}$ *with* $\alpha_j = \{n_1, \ldots, n_k\}$.

Proof. See appendix C. □

The above property sometimes provide a shortcut to calculate $S^{\boldsymbol{\alpha}}$, as if one finds a particular synergistic observable that attains this upper bound then it is clear that it is maximal. One immediate consequence of this proposition, noting that $I(\boldsymbol{X}^{-\alpha_j}; Y | \boldsymbol{X}^{\alpha_j}) = I(\boldsymbol{X}; Y) - I(\boldsymbol{X}^{\alpha_j}; Y)$,

is that

$$I(\boldsymbol{X};Y) - S^{\boldsymbol{\alpha}}(\boldsymbol{X} \to Y) \geqslant \max_{j\in\{1,\ldots,n\}} I(\boldsymbol{X}^{\alpha_j};Y). \tag{5}$$

In other words, the amount of non-synergistic information is lower-bounded by the amount of information carried by the most strongly correlated subgroup.

Further details on $S^{\boldsymbol{\alpha}}$, including properties of its bounds, algebraic properties, and a data processing inequality, are presented in section 6.1 and appendix D.

3. Information decomposition

This section uses the functional definition of $\boldsymbol{\alpha}$-synergy to formulate our proposed information decomposition. For this, we focus on the study of the sets of constraints of the form $\boldsymbol{\alpha} = \{\alpha_1, \ldots, \alpha_L\}$, which are the argument in the synergy $S^{\boldsymbol{\alpha}}(\boldsymbol{X} \to Y)$. For such sets, we say $|\boldsymbol{\alpha}| := L$ is the cardinality of the set.

3.1. The extended constraint lattice

Let us start by observing that not all source-sets yield unique synergistic channels. As a simple example, if $\boldsymbol{\alpha} = \{\{1,2\}\}$ and $\boldsymbol{\beta} = \{\{1\},\{1,2\}\}$ one has that $\mathcal{C}(\boldsymbol{X};\boldsymbol{\alpha}) = \mathcal{C}(\boldsymbol{X};\boldsymbol{\beta})$, as all the additional constraints in $\boldsymbol{\beta}$ are subsumed by the constraints in $\boldsymbol{\alpha}$. More formally, we say that two source-sets are equivalent, denoted by $\boldsymbol{\alpha} \equiv_I \boldsymbol{\beta}$, if $\mathcal{C}(\boldsymbol{X};\boldsymbol{\beta}) = \mathcal{C}(\boldsymbol{X};\boldsymbol{\alpha})$. Our next result shows that the set of anti-chains

$$\mathcal{A}^* = \{\boldsymbol{\alpha} = \{\alpha_1, \ldots, \alpha_L\} : \alpha_i \subset [n], \alpha_i \not\subset \alpha_j \quad \forall\, i \neq j\} \tag{6}$$

contains exactly one member of each equivalence class, and this member is the simplest such source-set.

Lemma 1. *For any $\boldsymbol{\beta} = \{\beta_1, \ldots, \beta_M\}$ with $\beta_i \subset [n]$, there exists one and only one $\boldsymbol{\alpha} \in \mathcal{A}^*$ such that $\boldsymbol{\beta} \equiv_I \boldsymbol{\alpha}$. Moreover, if $\boldsymbol{\beta} \equiv_I \boldsymbol{\alpha}$ and $\boldsymbol{\alpha} \in \mathcal{A}^*$, then $|\boldsymbol{\beta}| \geqslant |\boldsymbol{\alpha}|$.*

Proof. See appendix E. □

In other words, considering collections of indices that are not anti-chains would not provide new classes of channels, as broader subunits subsume smaller ones. This property brings strong reminiscences of Williams and Beer's redundancy lattice [12]—which we will discuss in detail in section 6 [39].

In addition to the set of nodes, to build a lattice on which one can formulate a decomposition one needs a partial order relationship. Considering our setup, a natural candidate is the order introduced by James *et al* in their proposed *constraint lattice* [40], defined by

$$\boldsymbol{\alpha} \preceq_{\mathrm{c}} \boldsymbol{\beta} \iff \forall\, \alpha \in \boldsymbol{\alpha}, \quad \exists \beta \in \boldsymbol{\beta} : \alpha \subseteq \beta \tag{7}$$

for $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathcal{A}^*$. Intuitively, $\boldsymbol{\alpha} \preceq_{\mathrm{c}} \boldsymbol{\beta}$ means that all the constraints imposed by $\boldsymbol{\alpha}$ are included within those imposed by $\boldsymbol{\beta}$, and therefore $\mathcal{C}(\boldsymbol{X};\boldsymbol{\beta}) \subseteq \mathcal{C}(\boldsymbol{X};\boldsymbol{\alpha})$.

Putting these structures together generates the *extended constraint lattice* $\mathcal{L}^* := (\mathcal{A}^*, \preceq_{\mathrm{c}})$, which extends the lattice introduced by James *et al* [40], and has been recently used by Ay *et al* [41]. The cases $n = 2$ and $n = 3$ are depicted in figure 1. Importantly, in contrast with James' proposal, $\mathcal{L}^*$ includes nodes that do not cover all the sources. The resulting lattice is isomorphic in shape to Williams and Beer's redundancy lattice, but with different relationships

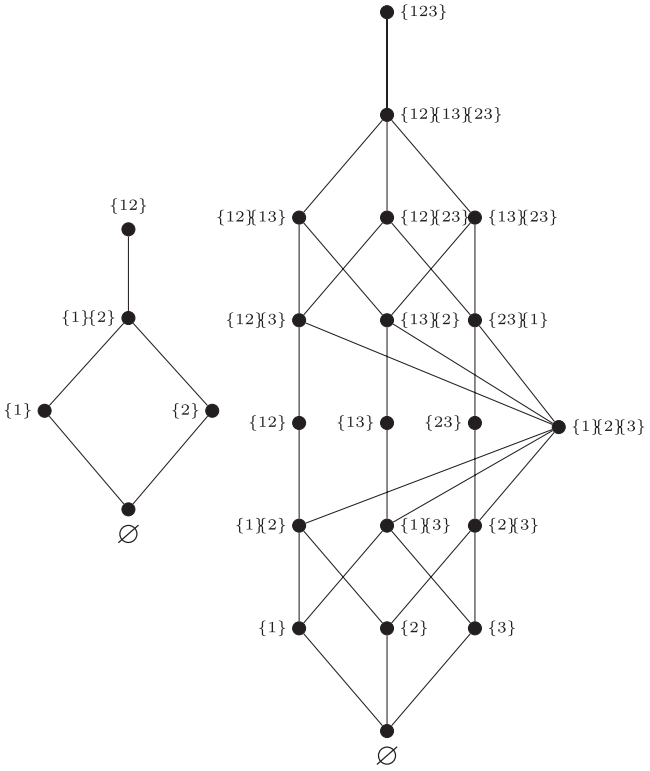

Figure 1. Extended constraint lattice for systems of $n = 2$ (*left*) and $n = 3$ (*right*) sources.

between the nodes. Despite this similarity, however, comparisons between these two lattices are not straightforward (cf section 6).

The lattice $\mathcal{L}^*$ possesses some interesting properties, most prominently:

Lemma 2. *If* $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathcal{L}^*$ *and* $\boldsymbol{\alpha} \preceq_{\mathrm{c}} \boldsymbol{\beta}$*, then*

$$S^{\boldsymbol{\alpha}}(\boldsymbol{X} \to Y) \geqslant S^{\boldsymbol{\beta}}(\boldsymbol{X} \to Y) \ . \tag{8}$$

Proof. See appendix E. □

This result shows that $S^{\boldsymbol{\alpha}}(\boldsymbol{X} \to Y)$ is a non-increasing function of $\boldsymbol{\alpha} \in \mathcal{L}^*$ for any given variables $\boldsymbol{X}, Y$. With this, one can propose the following decomposition based on the Möbius inversion formula [42]:

Definition 3. For a given $p_{\boldsymbol{X},Y}$, the atoms $S^{\boldsymbol{\alpha}}_{\partial}(\boldsymbol{X} \to Y)$ correspond to the terms given by the *Möbius inverse* of $S^{\boldsymbol{\alpha}}(\boldsymbol{X} \to Y)$; i.e. the unique set of values that satisfy

$$S^{\boldsymbol{\alpha}}_{\partial}(\boldsymbol{X} \to Y) := S^{\boldsymbol{\alpha}}(\boldsymbol{X} \to Y) - \sum_{\boldsymbol{\beta} \in \mathcal{A}^* \boldsymbol{\beta} \succ \boldsymbol{\alpha}} S^{\boldsymbol{\beta}}_{\partial}(\boldsymbol{X} \to Y) \tag{9}$$

for all $\boldsymbol{\alpha} \in \mathcal{A}^*$.

Intuitively, the Möbius inversion can be understood as a discrete derivative over a lattice. In effect, an equivalent representation of the Möbius relationship is given by

$$S^{\alpha}(\boldsymbol{X} \to Y) = \sum_{\beta \in \mathcal{A}^* \beta \succeq \alpha} S^{\beta}_{\partial}(\boldsymbol{X} \to Y) \ , \tag{10}$$

which is analogous to the fundamental theorem of calculus. The Möbius inversion yields *synergy atoms* of the form S^{α}_{∂}, which quantify how much information about the target is contained in the collective effects of variables $\boldsymbol{\alpha}$. For example, $S^{[n]}(\boldsymbol{X} \to Y) = S^{[n]}_{\partial}(\boldsymbol{X} \to Y) = 0$, and $S^{\varnothing}(\boldsymbol{X} \to Y) = I(\boldsymbol{X}; Y)$ for any $p_{\boldsymbol{X},Y}$ [43]. This last identity, combined with equation (10), gives the following important result:

Proposition 2. (Information decomposition). *The mutual information between* $\boldsymbol{X}$ *and* Y *can be decomposed as*

$$I(\boldsymbol{X}; Y) = \sum_{\alpha \in \mathcal{A}^*} S^{\alpha}_{\partial}(\boldsymbol{X} \to Y) \ . \tag{11}$$

Proof. Follows directly from noting that $S^{\varnothing}(\boldsymbol{X} \to Y) = I(\boldsymbol{X}; Y)$, and combining this with equation (10). □

Please note that the synergy atoms are not guaranteed to be non-negative. However, section 4 presents a coarse-grained decomposition with provably non-negative atoms (cf proposition 3).

3.2. The case $n = 2$

After having formally presented the decomposition for n variables, let us focus on the bivariate ($n = 2$) case, and develop some intuitions about the resulting synergy atoms. For two predictors $\boldsymbol{X} = (X_1, X_2)$, equation (11) yields

$$I(\boldsymbol{X}; Y) = \ S^{\{1\}\{2\}}_{\partial}(\boldsymbol{X} \to Y) + S^{\{1\}}_{\partial}(\boldsymbol{X} \to Y) + S^{\{2\}}_{\partial}(\boldsymbol{X} \to Y) + S^{\varnothing}_{\partial}(\boldsymbol{X} \to Y).$$

Above, $S^{\{1\}\{2\}}_{\partial}(\boldsymbol{X} \to Y)$ can be understood as the information about Y that is related to collective properties of $\boldsymbol{X}$ that can be disclosed without compromising either X_1 or X_2 [44]. Similarly, $S^{\{1\}}_{\partial}(\boldsymbol{X} \to Y)$ is the information about Y that can be disclosed without revealing parts of X_1 but compromising X_2 (otherwise it would have been included in $S^{\{1\}\{2\}}_{\partial}(\boldsymbol{X} \to Y)$). Finally, $S^{\varnothing}_{\partial}(\boldsymbol{X} \to Y)$ is information about Y that compromises both variables; put differently, information that is neither in $S^{\{1\}}(\boldsymbol{X} \to Y)$ or $S^{\{2\}}(\boldsymbol{X} \to Y)$. Loosely speaking, $S^{\varnothing}_{\partial}$ can be associated with the standard PID redundancy, $S^{\{i\}}_{\partial}$ with the unique information, and $S^{\{1\}\{2\}}_{\partial}$ with the synergy. A detailed comparison of these and the standard PID atoms is presented in section 6.

For the particular case where X_1 and X_2 are binary variables, then the optimal synergistic channel only depends on their joint distribution—and not on the target variable, as shown in reference [25]. Interestingly, if X_1 and X_2 are independent fair coin flips, then [45]

$$S^{\{1\}\{2\}}(\boldsymbol{X} \to Y) = I(X_1 \texttt{xor} X_2; Y). \tag{12}$$

This result shows that our definition of synergy effectively captures high-order statistical effects, which are most purely exhibited by XOR logic gates [46]. Analytical results for the more general case where X_1 and X_2 are binary, though not necessarily independent, are presented in appendix F.

Table 1. Common distributions and their S^{α} decomposition.

	XOR	COPY	Unq.1	AND	TBC
$S_{\partial}^{\{1\}\{2\}}$	1	0	0	0.3113	1
$S_{\partial}^{\{2\}}$	0	0	1	0	0
$S_{\partial}^{\{1\}}$	0	0	0	0	0
$S_{\partial}^{\varnothing}$	0	1	0	0.5	1

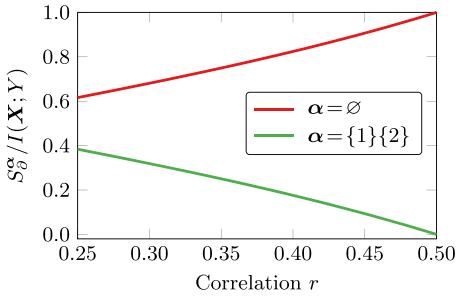

Figure 2. Normalised atoms of the disclosure decomposition for the AND gate with correlated inputs, with $\mathbb{P}\{X_1 = 1\} = \mathbb{P}\{X_2 = 1\} = 0.5$ and correlation $\langle x_1 x_2 \rangle = r$.

With these results, it is straightforward to compute the decomposition in equation (11) for a few illustrative examples; results are presented in table 1. First, we notice that the paradigmatic distributions copy and xor have the expected 1 bit of redundancy and synergy, respectively, in agreement with our intuition for these cases. Similarly, the unq.1 distribution shows only one non-zero atom, $S_{\partial}^{\{2\}}$, which corresponds to unique information. The index of the atom, however, might seem counterintuitive; the confusion is explained by the fact that the superscript $\{2\}$ refers to a constraint (the impossibility to disclose what is in X_2), and hence $S_{\partial}^{\{2\}}$ is more related with the contents of X_1. This shows a general theme: that S^{α}, while operationally meaningful and intuitive, needs to be interpreted differently from other PIDs (cf section 6).

As a further example, we compute the disclosure decomposition S_{∂}^{α} for the result of an AND gate with correlated inputs (figure 2). As the inputs become more correlated, there is less information that can be disclosed without compromising either of them, and therefore the fraction of the total information that corresponds to $S_{\partial}^{\varnothing}$ grows as correlation increases.

4. The backbone decomposition

As the extended constraint lattice $\mathcal{L}^{*}$ grows extremely rapidly with system size, it is unfeasible to examine every element of our proposed decomposition in all but very small systems. Luckily, the nature of S^{α} allows us to formulate a reduced collection of source-sets that form the 'backbone' of the constraint lattice, which provides a natural summary of the system's high-order interactions.

In the sequel, subsection 4.1 introduces the backbone lattice, then subsection 4.2 discusses the backbone decomposition, and finally subsection 4.3 illustrates these ideas with some examples.

4.1. The backbone constraint lattice

We introduce the *backbone constraint lattice*, denoted by $\mathcal{B} \subset \mathcal{L}^*$, as the sublattice composed by the elements of $\mathcal{A}^*$ of the form $\gamma_m = \{\alpha \subset [n] : |\alpha| = m\}$ for $m = 0, \ldots, n$ (the dependency on n is left implicit). Importantly, $\preceq_c$ restricted to $\mathcal{B}$ provides a total order:

$$\gamma_0 \preceq_c \gamma_1 \cdots \preceq_c \gamma_n. \tag{13}$$

For example, for the case of $n = 3$ then $\mathcal{B}$ is composed by $\gamma_0 = \{\varnothing\}$, $\gamma_1 = \{\{1\}, \{2\}, \{3\}\}$, $\gamma_2 = \{\{1,2\}, \{2,3\}, \{1,3\}\}$, and $\gamma_3 = \{\{1,2,3\}\}$. Hence, the constraint γ_m corresponds to the synergistic channel that discloses no information about any of the mth-order marginals.

For the synergy terms associated with $\mathcal{B}$, we use the shorthand notation $B^m(\boldsymbol{X} \to Y) := S^{\gamma_m}(\boldsymbol{X} \to Y)$. In simple words, $B^m(\boldsymbol{X} \to Y)$ accounts for the information about Y that can be disclosed without compromising any group of m variables. Furthermore, as $\gamma_{m-1} \preceq_c \gamma_m$, the following chain of inequalities is guaranteed:

$$0 = B^n(\boldsymbol{X} \to Y) \leqslant \cdots \leqslant B^0(\boldsymbol{X} \to Y) = I(\boldsymbol{X}; Y). \tag{14}$$

4.2. Backbone atoms

A new application of the Möbius inversion formula allows us to define *backbone atoms*, $B^m_\partial(\boldsymbol{X} \to Y)$, which we define as

$$\begin{aligned} B^m_\partial(\boldsymbol{X} \to Y) &:= B^{m-1}(\boldsymbol{X} \to Y) - \sum_{k=m+1}^{n} B^k_\partial(\boldsymbol{X} \to Y) \\ &= B^{m-1}(\boldsymbol{X} \to Y) - B^m(\boldsymbol{X} \to Y). \end{aligned} \tag{15}$$

Equivalently, the backbone atoms are the values $B^k_\partial(\boldsymbol{X} \to Y)$ that satisfy, for all $m \in [n]$,

$$B^{m-1}(\boldsymbol{X} \to Y) = \sum_{k=m}^{n} B^k_\partial(\boldsymbol{X} \to Y), \tag{16}$$

Intuitively, B^{m-1} corresponds to the amount of information about Y that $\boldsymbol{X}$ can reveal without compromising any group of $m-1$ variables; or, equivalently, information revealed by compromising only groups of m or more variables. Consequently, B^m_∂ quantifies the marginal gain of information that can be disclosed by relaxing the constraints from groups of m variables to groups of $m-1$. For example, for $m = 1$ then $B^1(\boldsymbol{X} \to Y)$ measures how much information can be disclosed while keeping each X_j confidential, while $B^1_\partial(\boldsymbol{X} \to Y)$ corresponds to how much is gained when these constraints are relaxed. Additionally, note that these backbone atoms can be directly related to the synergy atoms in equation (9), as

$$B^m_\partial(\boldsymbol{X} \to Y) = \sum_{\gamma_{m-1} \preceq_c \alpha \preceq_c \gamma_m} S^\alpha_\partial(\boldsymbol{X} \to Y). \tag{17}$$

Putting all these results together one finds a reduced decomposition, which is formalised by the following result.

Proposition 3 (Backbone decomposition). *The following decomposition always holds:*

$$I(\boldsymbol{X} \to Y) = \sum_{m=1}^{n} B^m_\partial(\boldsymbol{X} \to Y). \tag{18}$$

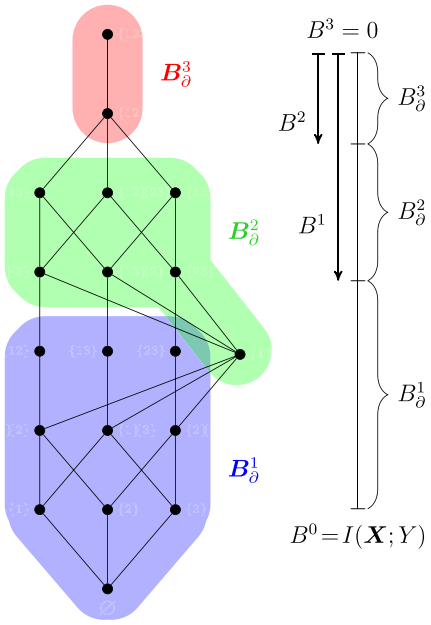

Figure 3. Schematic representation of the backbone lattice. (*left*) Correspondence between backbone atoms and S^α_∂ for the $n = 3$ lattice. (*right*) Representation of the backbone lattice as a totally ordered set.

Moreover, $B^m_\partial(\boldsymbol{X} \to Y) \geqslant 0$ *for all* $m = 1, \ldots, n$.

Proof. One can obtain equation (18) by evaluating equation (16) for $m = 1$. The non-negativity of the atoms is a consequence of equations (14) and (15).

These backbone atoms provide a coarse-graining of the full decomposition in equation (11). A basic schematic of this backbone decomposition, as well as its relationship with the S^α_∂ atoms in the extended constraint lattice are shown in figure 3. Importantly, note that the cardinality of the backbone lattice grows linearly with system size, and hence the number of atoms in equation (18) remains tractable for large systems.

4.3. Examples

As an illustrative example of the potential of the backbone decomposition, let us apply it to scenarios where the relationship between $\boldsymbol{X}$ and Y can be expressed as a Gibbs distribution. In particular, we consider systems of $n + 1$ spins (i.e. $\mathcal{X}_i = \{-1, 1\}$ for $i = 1, \ldots, n + 1$) whose joint probability distributions can be expressed in the form

$$p_{\boldsymbol{X}^{n+1}}(\boldsymbol{x}^{n+1}) = \frac{\mathrm{e}^{-\beta \mathcal{H}_k(\boldsymbol{x}^{n+1})}}{Z}, \tag{19}$$

where β is the inverse temperature, Z a normalisation constant, and $\mathcal{H}_k(\boldsymbol{x}^n)$ a Hamiltonian function of the form

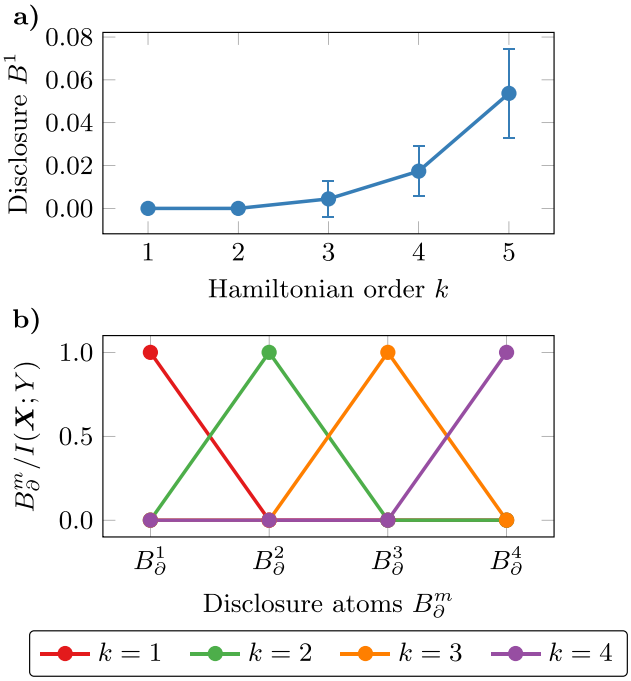

Figure 4. Synergistic disclosure in Ising models (**a**) with terms up to order k and (**b**) with terms only of order k.

$$\mathcal{H}_k(\boldsymbol{x}^{n+1}) = -\sum_{i=1}^{n+1} J_i x_i - \sum_{i=1}^{n}\sum_{j=i+1}^{n+1} J_{i,j} x_i x_j \\ \ldots - \sum_{|\boldsymbol{I}|=k} J_\gamma \prod_{i\in\boldsymbol{I}} x_i, \tag{20}$$

with the last sum running over all collections of indices $\boldsymbol{I} \subseteq [n+1]$ of cardinality $|\gamma| = k$. Hamiltonians of the form of equation (20) correspond to maximum entropy distributions with constraints on the ^{k}th-order marginals [10, 11], which arise naturally in scenarios where only ^{k}th-order properties are observed [47, 48]. To calculate all quantities in this section we consider $Y = X_{n+1}$ as target variable. Full simulation details are reported in appendix G.

As a first test case, we consider Hamiltonians with interactions up to order k, as in equation (20) above. For these systems, we calculated the backbone term $B^1(\boldsymbol{X} \to Y)$, which measures the strength of the high-order statistical effects beyond pairwise interactions (figure 4(a)). As expected, our results show that if the Hamiltonian only possesses first or second order interactions (i.e. $k = 1$ or 2) then $B^1(\boldsymbol{X} \to Y)$ is negligible; and for $k \geqslant 3$, $B^1(\boldsymbol{X} \to Y)$ grows monotonically with k.

As a second test case, we studied Hamiltonians with source-target interactions only of order k, and compute their full backbone decomposition. Figure 4(b) shows all the backbone atoms B^m_∂ on the X-axis, normalised by $I(\boldsymbol{X};Y)$. Interestingly, for each Hamiltonian order k there is only one non-zero backbone atom, which suggests that $I(\boldsymbol{X};Y) \approx B^k_\partial(\boldsymbol{X} \to Y)$. Note that this relationship between Hamiltonian interaction order and backbone atom is highly non-trivial, and finding analytical methods to make this connection more explicit is an open question.

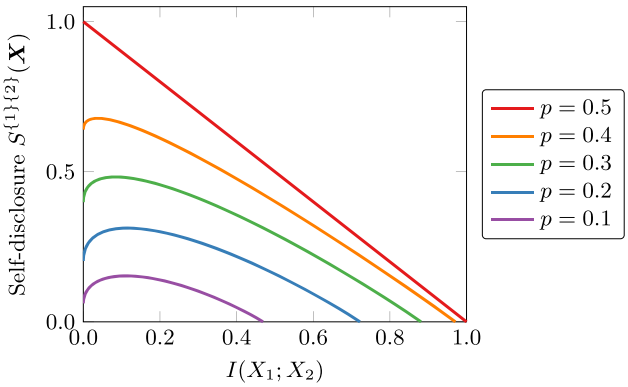

Figure 5. Self-disclosure capacity of two correlated bits.

These findings suggest that the backbone decomposition may provide an analogue to the measure of *connected information* introduced in references [10, 11], which captures the effects of Hamiltonian high-order terms over their corresponding Gibbs distributions [49]. The main difference between the connected information and the backbone decomposition is that in the former all variables play an equivalent role, while in the latter they are divided between sources and target.

As a final remark, we note that in statistical physics the structure of Gibbs models are typically explored via two-point correlation functions. However, pairwise statistics do not generally capture high-order statistical phenomena—see references [[5], section 3.1] and [[50], section 5] for related discussions.

5. Synergistic capacity and private self-disclosure

So far, we have investigated the usual information decomposition scenario, in which a group of source variables $\boldsymbol{X}$ hold information about *another*, target variable Y. Using the tools developed so far, we can ask a new question: how much information can $\boldsymbol{X}$ disclose *about itself* under specific constraints? Answering this question will provide further intuitions on the nature of synergistic disclosure, while revealing some unexpected properties.

We start by presenting the definition of the *self-disclosure* of a system, which is a particular case of the formalism presented above.

Definition 4. The α-self-synergy of $\boldsymbol{X}$ is given by $S^{\alpha}(\boldsymbol{X} \to \boldsymbol{X})$, and denoted simply by $S^{\alpha}(\boldsymbol{X})$.

This definition makes it straightforward to extend the concepts above to define self-synergy atoms $S^{\alpha}_{\partial}(\boldsymbol{X})$, as well as backbone self-synergy terms and atoms, denoted by $B^{m}(\boldsymbol{X})$ and $B^{m}_{\partial}(\boldsymbol{X})$, respectively.

Let us begin with an example, by computing the self-disclosure of binary bivariate distributions. Consider two binary variables $\boldsymbol{X} = (X_1, X_2)$, with $\mathbb{P}\{X_1 = 1\} = \mathbb{P}\{X_2 = 1\} = p$ and $\mathbb{P}\{X_1 = 1, X_2 = 1\} = r$ (figure 5). Perhaps surprisingly, a direct calculation shows that maximal synergy is achieved for X_1, X_2 independent and $p = 1/2$—which is equivalent to the much-debated two-bit-copy (TBC) gate commonly discussed in the PID literature [16, 51, 52]. To make sense of this result, consider the following bounds on the self-disclosure:

Lemma 3. *For any $\boldsymbol{X}, Y$ the following bound holds:*

$$H(\boldsymbol{X}) - \max_{\alpha_j \in \boldsymbol{\alpha}} H(\boldsymbol{X}^{\alpha_j}) \geqslant S^{\boldsymbol{\alpha}}(\boldsymbol{X}) \geqslant S^{\boldsymbol{\alpha}}(\boldsymbol{X} \to Y). \tag{21}$$

Proof. The upper bound is proven by an application of proposition 1 with $Y = \boldsymbol{X}$, and the lower bound by an application of lemma 6. □

This lower bound is particularly insightful, as it suggests that the synergistic self-disclosure of $\boldsymbol{X}$ is the *tightest upper bound on the synergistic information that $\boldsymbol{X}$ could hold about any other target*. Therefore, this (admittedly heterodox) perspective of synergy provides a clear explanation of why the TBC could have non-zero synergy, since it accounts for the 'synergistic capacity' of its inputs.

Additionally, the upper bound in lemma 3 provides a quick way to estimate how much synergy can be found with respect to a given set of sources $\boldsymbol{X}$. For example, if (X_1, X_2) are two i.i.d. fair coins, lemma 3 states that their synergy cannot be larger than 1 bit, which is attained by the optimal self-synergistic channel $V^* = X_1 \text{xor} X_2$ [53].

Another natural conjecture, in the light of the findings reported in section 4.3, would be to argue a relationship between self-synergy and connected information, as both measures treat symmetrically all the corresponding variables. However, numerical evaluations show there is no relationship between them. As a matter of fact, systems with low degrees of interdependency have high levels of self-synergy, while having low levels of connected information.

A final lesson that can be learnt from studying self-synergy is that high-order synergies are not rare corner cases, but are in fact prevalent in the space of probability distributions. More formally, our next result shows that $B^m(\boldsymbol{X})$ takes most of the information contained in $\boldsymbol{X}$ as the system size grows.

Proposition 4. *Consider a sequence of random variables $\boldsymbol{X} := (X_1, \ldots, X_n)$ for which there exists $K \in \mathbb{N}$ such that $|\mathcal{X}_k| \leqslant K$ for all $k \in \mathbb{N}$. If $\lim_{n\to\infty} H(\boldsymbol{X})/n$ exists and is not zero, then for any fixed $m \in \mathbb{N}$*

$$\lim_{n\to\infty} \frac{B^m(\boldsymbol{X})}{H(\boldsymbol{X})} = 1. \tag{22}$$

Proof. See appendix H. □

Let us work an example to gain intuition on this seemingly counterintuitive result.

Example 2. Consider a system $\boldsymbol{X}$ where the components X_k are independent fair coins. The mapping $V = (X_1 \text{xor} X_2, \ldots, X_{n-1} \text{xor} X_n) : \{0,1\}^n \to \{0,1\}^{n-1}$ belongs to $C(\boldsymbol{X}, \{\{1\}, \ldots, \{n\}\})$, and $I(V; \boldsymbol{X}) = n - 1$, attaining the upper bound provided in lemma 3. This implies that $B^1(\boldsymbol{X}) = n - 1$. Similarly, one can notice that $V_m = (X_1 \text{xor} \ldots \text{xor} X_m, \ldots, X_{n-m+1} \text{xor} \ldots \text{xor} X_n) : \{0,1\}^n \to \{0,1\}^{n-m+1}$ also attains the bound for the class $C(\boldsymbol{X}, \{\{1, \ldots, m\}, \ldots, \{n-m+1, \ldots, n\}\})$, and hence $B^m(\boldsymbol{X}) = n - m$. Therefore, $B^m_\partial(\boldsymbol{X}) = 1$ for all $m = 1, \ldots, n-1$, and

$$\lim_{n\to\infty} \frac{B^m(\boldsymbol{X})}{H(\boldsymbol{X})} = \lim_{n\to\infty} \frac{n-m}{n} = 1. \tag{23}$$

The theoretical and practical consequences of the prevalence of synergy will be discussed in a separate publication.

6. Relationship with other information decompositions

This section explores the relationship of our proposed framework with other information decompositions. For this, subsection 6 1 explores various properties of our definition of synergy under the light of various axioms typically used in the PID literature, then subsection 6 2 explores relationships of our decomposition with other PID, and finally subsection 6.3 carries out numerical comparisons between our metrics and other well-known decompositions.

6.1. Axioms

In previous literature, PID is usually discussed in terms of axioms, which encode various desirable properties that measures might—or might not—satisfy. These axioms are often formulated for redundancy measures, which, given that the basic constituent of our decomposition is a synergy measure, makes assessing our framework in these terms non-trivial. Nevertheless, this subsection explores some of the common axioms from the point of view of S^{α}, using as guideline the set of axioms discussed in reference [38].

The following axioms are satisfied by our measure:

- **(GP)** Global positivity: $S^{\alpha}(\boldsymbol{X} \to Y) \geqslant 0$ for all $\boldsymbol{X}, Y$ and $\boldsymbol{\alpha} \in \mathcal{A}^{*}$.
- **(Eq)** Equivalence-class invariance: $S^{\alpha}(\boldsymbol{X} \to Y)$ is invariant under substitution of X_i or Y by an informationally equivalent random variable (i.e. re-labeling).
- **(wS)** Weak symmetry: $S^{\alpha}(\boldsymbol{X} \to Y)$ is invariant under reordering of $X_1, \ldots, X_n$.
- **(wM)** Weak monotonicity: $S^{\alpha}(\boldsymbol{X} \to Y) \leqslant S^{\alpha}((\boldsymbol{X}, Z) \to Y)$ (see appendix E). Note that this does not hold for the backbone terms, as the $\boldsymbol{\alpha}$'s are not equal.
- **(CCx)** Channel convexity: $S^{\alpha}(\mathbf{X} \to Y)$ is a convex function of $p_{Y|\mathbf{X}}$ for a given $p_{\mathbf{X}}$ (proof in appendix D).
- **(T-DPI)** Target data processing inequality: if $\boldsymbol{X} - Y - Z$ is a Markov chain, then $S^{\alpha}(\boldsymbol{X} \to Y) \geqslant S^{\alpha}(\boldsymbol{X} \to Z)$ (proof in appendix D).

The proposed measure does not satisfy strong symmetry (**sS**), as it might be the case that $S^{\alpha}((X_1, X_2) \to Y) \neq S^{\alpha}((X_1, Y) \to X_2)$ [54].

We can prove by counterexample that S^{α}_{∂} does not satisfy strong local positivity (**LP**), i.e. that there exist $S^{\alpha}_{\partial}(\boldsymbol{X} \to Y) < 0$ for some $\boldsymbol{\alpha} \in \mathcal{A}^{*}$ [55]. On the other hand, note that the backbone atoms $B^{m}_{\partial}(\boldsymbol{X} \to Y)$ do satisfy (**LP**), as shown in section 4.

6.2. General relationship with PID

In this section we focus on the relationship between our decomposition for the case of $n = 2$ (cf section 3.2), and the standard PID. When considering $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathcal{A} := \mathcal{A}^{*}/\{\varnothing\}$, the classic work of Williams and Beer [12] introduces the following partial ordering:

$$\boldsymbol{\alpha} \preceq_{\mathrm{wb}} \boldsymbol{\beta} \iff \forall \beta \in \boldsymbol{\beta} \quad \exists \alpha \in \boldsymbol{\alpha}, \alpha \subseteq \beta. \tag{24}$$

While the difference between $\preceq_{\mathrm{wb}}$ and $\preceq_{\mathrm{c}}$ might seem subtle, they induce drastically different lattice structures. For example, if $\boldsymbol{\alpha} = \{\{1\}\}$ and $\boldsymbol{\beta} = \{\{1\}\{2\}\}$, then $\boldsymbol{\beta} \preceq_{\mathrm{wb}} \boldsymbol{\alpha}$ while $\boldsymbol{\alpha} \preceq_{\mathrm{c}} \boldsymbol{\beta}$. The lattices for $n = 2$ for both orderings are shown in figure 6.

Traditional PID-type decompositions for two sources are based on the following conditions:

$$I(X_i; Y) = \mathtt{Red}(X_1, X_2 \to Y) + \mathtt{Un}(X_i; Y|X_j)$$
$$I(X_i; Y|X_j) = \mathtt{Un}(X_i; Y|X_j) + \mathtt{Syn}(X_1, X_2 \to Y),$$

Table 2. Correspondence between PID atoms and S^{α}_{∂}.

	Disclosure decomposition	PID
Synergy	$S^{\{1\}\{2\}}(\boldsymbol{X} \to Y)$	$\mathtt{Syn}(X_1, X_2 \to Y)$
Unique	$S^{\{i\}}_{\partial}(\boldsymbol{X} \to Y)$	$\mathtt{Un}(X_j; Y \vert X_i)$
Redundancy	$S^{\varnothing}_{\partial}(\boldsymbol{X} \to Y)$	$\mathtt{Red}(X_1, X_2 \to Y)$

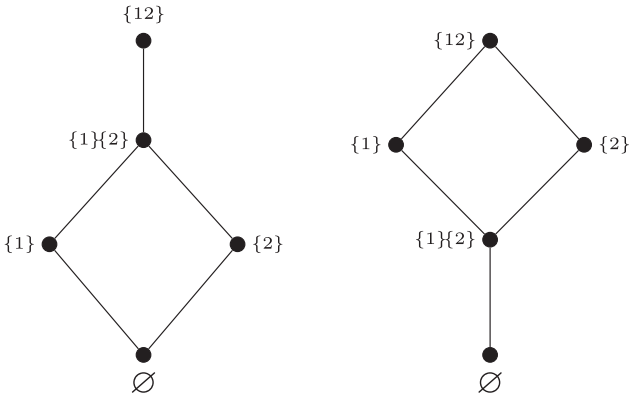

Figure 6. Extended constraint (*left*) and redundancy (*right*) lattices for $n = 2$.

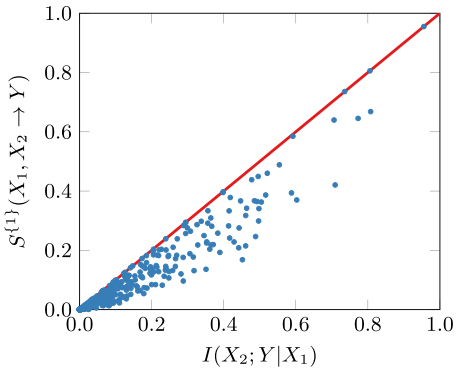

Figure 7. Conditional information is an upper bound on discloseable information.

which are valid for $i, j \in \{1, 2\}$ with $i \neq j$. A direct parallel between these terms and our framework can be made, and is shown in table 2.

A key relationship between any PID and our decomposition comes from noticing that, considering proposition 1 for $\boldsymbol{X} = (X_1, X_2)$ and $\boldsymbol{\alpha} = \{\{1\}\}$, one finds that

$$S^{\{1\}}(\boldsymbol{X} \to Y) \leqslant I(X_2; Y|X_1). \tag{25}$$

Moreover, numerical evaluations show that this bound is often not attained, as illustrated by figure 7 (see appendix G for more details).

As a consequence of this, one has that

$$\begin{aligned} I(X_i; Y|X_j) &= \mathtt{Un}(X_i; Y|X_j) + \mathtt{Syn}(X_1, X_2 \to Y) \\ &\geqslant S^{\{j\}}(X_1, X_2 \to Y) \\ &= S_\partial^{\{j\}}(X_1, X_2 \to Y) + S_\partial^{\{1\}\{2\}}(X_1, X_2 \to Y). \end{aligned}$$

Conversely, an opposite relationship holds for the marginal mutual information:

$$\begin{aligned} I(X_i; Y) &= \mathtt{Red}(X_1, X_2 \to Y) + \mathtt{Un}(X_i; Y|X_j). \\ &\leqslant I(X_1, X_2; Y) - S^{\{i\}}(X_1, X_2 \to Y) \\ &= S_\partial^{\varnothing}(X_1, X_2 \to Y) + S_\partial^{\{i\}}(X_1, X_2 \to Y). \end{aligned}$$

By combining these two results, one can compare the co-information with a corresponding co-information obtained from our decomposition, as follows:

$$\begin{aligned} I(X_1; X_2; Y) &= \mathtt{Red}(X_1, X_2 \to Y) - \mathtt{Syn}(X_1, X_2 \to Y) \\ &= I(X_i; Y) - I(X_i; Y|X_j) \\ &\geqslant S_\partial^{\varnothing}(\boldsymbol{X} \to Y) - S_\partial^{\{1\}\{2\}}(\boldsymbol{X} \to Y). \end{aligned}$$

This result implies that, when assessing the balance between redundancy and synergy, our decomposition always tends towards redundancy over synergy with respect to any PID decomposition. In this sense, one can say that—at least for $n = 2$—our decomposition is conservative when attributing dominance of synergies. The next section provides further evidence to support this claim.

6.3. Numerical comparisons with other PIDs

Let us now study how our proposed measure of synergy relates to the ones corresponding to other well-known decompositions. Our analysis includes the I_{BROJA} decomposition by Bertschinger *et al* [15], *common change in surprisal* (I_{ccs}) by Ince [14], $I_{\min}$ by Williams and Beer [12], I_{dep} by James *et al* [40], and the pointwise decomposition by Finn and Lizier ($I_\pm$) [16]; all computed using the `dit` package [56]. To do this comparison, we draw random distributions from the probability simplex following an NSB prior (see appendix G for details), and then compute their synergy values with all measures.

A first, somewhat striking result is the overwhelming correlation found between most proposed measures—BROJA, CCS, $I_{\min}$, and I_{dep} are all related with each other with correlations greater than 0.94 for every pair (figure 8(a)). The two oddballs in this plot are our proposed measure S^α and $I_\pm$, which are less well correlated with the rest *and* with each other (correlations range around 0.70 for S^{γ_1} and around 0.50 for $I_\pm$).

To examine this discrepancy, we computed the inverse cumulative function of the resulting values of the synergy for the various measures (figure 8(b)). This curve shows the fraction of all sampled distributions that have a synergy greater than a given threshold, to gauge how prevalent synergy is judged to be according to each measure. Consistent with figures 8(a) and (b) shows that the measures BROJA, CCS, $I_{\min}$, and I_{dep} all follow similar profiles. Interestingly, S^{γ_1} falls much faster than the rest, while $I_\pm$ does it much more slowly. Therefore, our measure S^{γ_1} can be said to be more 'restrictive', in the sense that it tends to assign lower values of

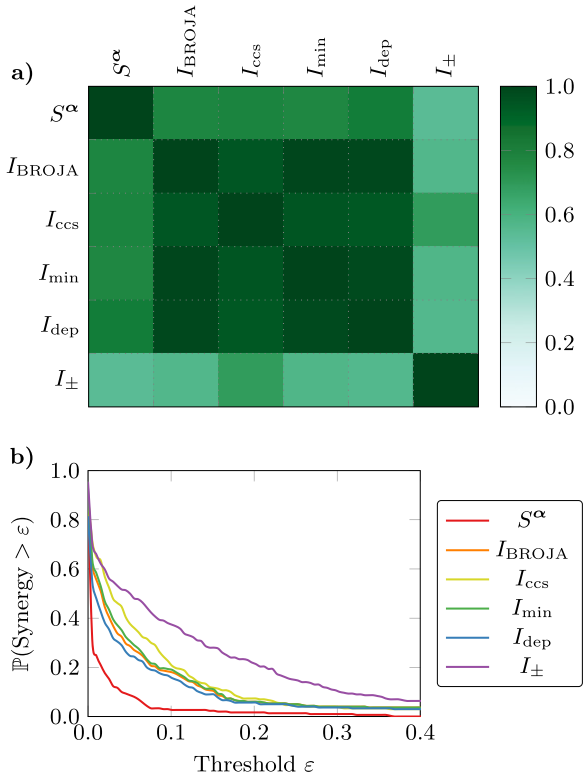

Figure 8. Numerical comparison between synergy values according to different proposed decompositions. (**a**) Correlation matrix of synergy values of random distributions. (**b**) Fraction of distributions with synergy greater than a given threshold.

synergy, while $I_\pm$ is more lenient. We hypothesise this 'overestimation' of synergy by $I_\pm$ happens because of its tendency to assign negative values to the redundant or unique information [57].

Finally, it is worth illustrating a simple case where S^α disagrees with previous decompositions. As a straightforward example, consider the TBC gate (cf section 5). In this distribution, S^α yields one bit of redundancy and one of synergy, while I_{BROJA} yields two bits of unique information. The reason for this discrepancy is simple, and easily understood within the operational interpretation of S^α: when dealing with two independent bits one can always disclose their XOR, which, for the TBC problem, excludes half of the possible outcomes and therefore attains $I(V;Y) = 1$ bit.

7. Conclusion

This paper puts forward an operational definition of informational synergy, and uses it as a foundation to build a multivariate information decomposition. Compared to previous approaches to information decomposition, our framework possesses two key features:

(a) It is a 'synergy-first' decomposition, which begins by positing a measure of synergy and builds a decomposition after it, as opposed to previous approaches that are based on redundancy or unique information, and have synergy as a by-product.
(b) It is based on a quantity that is the optimal solution of a well-defined problem in the data privacy literature, which makes reasoning about the measure more transparent while bringing the decomposition closer to standard information-theoretic formulations.

We illustrated the capabilities of the proposed decomposition on various examples, and showed that it gives a complementary perspective compared to other information decompositions. In particular, our results show that our measure of synergy is in general more conservative than other approaches, as it tends to attribute smaller values of synergy. We also showed how its operational interpretation provides clear explanations to open questions in the field of information decomposition, such as the well-known TBC problem [14, 16, 20].

Moreover, our measure has an associated 'backbone' decomposition, which provides a natural coarse-graining of the information atoms. Our results show that in some scenarios the backbone atoms provide a directed version of the well-known connected information, which captures the effect of high-order interaction terms within Gibbs distributions. The number of backbone atoms grows linearly with system size, which makes this decomposition practical for studying a wide range of systems of interest.

The operational approach taken in this work represents a step towards establishing a solid foundation in the field of information decomposition. Additionally, we provide an open-source software package [58] implementing the key quantities in this paper, opening the door for a wide range of applications in data analysis, neuroscience, and information dynamics.

Acknowledgments

The authors thank Michael Gastpar and Shamil Chandaria for insightful discussions, and Yike Guo for supporting this research. FR is supported by the Ad Astra Chandaria foundation. PM is funded by the Wellcome Trust (Grant Nos. 210920/Z/18/Z). A.B.B. was supported by the Dr. Mortimer and Theresa Sackler Foundation through affiliation with the Sackler Centre for Consciousness Science.

Appendix A. Characterising synergistic channels

Here we provide a characterisation of $\mathcal{C}(\boldsymbol{X};\boldsymbol{\alpha})$ in terms of matricial properties of it constituents. To do this, let us introduce the matrix $\mathbf{P}_{\boldsymbol{\alpha}}$ defined as

$$\mathbf{P}_{\boldsymbol{\alpha}} \triangleq \begin{bmatrix} \mathbf{P}_{\boldsymbol{X}^{\alpha_1}|\boldsymbol{X}} \\ \vdots \\ \mathbf{P}_{\boldsymbol{X}^{\alpha_L}|\boldsymbol{X}} \end{bmatrix}_{G\times|\hat{\mathcal{X}}|}, \tag{A1}$$

where $G \coloneqq \sum_{k=1}^{L} \prod_{i\in\alpha_k} |\mathcal{X}_i|$, and $\hat{\mathcal{X}}$ is the set of tuples $\boldsymbol{x} \in \prod_{k=1}^{n} \mathcal{X}_k$ such that $p_{\boldsymbol{X}}(\boldsymbol{x}) > 0$. This matrix is designed such that the matrix product $\mathbf{P}_{\boldsymbol{\alpha}}\mathbf{p}_{\boldsymbol{X}}$ (with $\mathbf{p}_{\boldsymbol{X}}$ being the probability vector of $\boldsymbol{X}$) yields the marginals within $p_{\boldsymbol{X}}$ that need to be 'masked' by the synergistic channel—so that $p_{\boldsymbol{X}^{\alpha}|V=v}$ is a uniform distribution for all $\alpha \in \boldsymbol{\alpha}$. Note that $\mathbf{P}_{\boldsymbol{\alpha}}$ is a binary matrix, since the $\boldsymbol{X}^{\boldsymbol{\alpha}}$'s are deterministic functions of $\boldsymbol{X}$. As an example, if $|\mathcal{X}_i| = 2, \quad \forall\, i \in [n]$ and $\boldsymbol{\alpha} =$

$\{\{1\},\ldots,\{n\}\}$, then $\mathbf{P}_{\boldsymbol{\alpha}}$ is a $2n \times 2^n$ matrix that can be built recursively according to

$$\mathbf{P}_{k+1} = \begin{bmatrix} 1\ldots1 & 0\ldots0 \\ 0\ldots0 & 1\ldots1 \\ \mathbf{P}_k & \mathbf{P}_k \end{bmatrix},$$

with $\mathbf{P}_{\boldsymbol{\alpha}} = \mathbf{P}_n$ and $\mathbf{P}_1 = \begin{bmatrix} 10 \\ 01 \end{bmatrix}$. With this matrix, one can characterise the channels in $\mathcal{C}(\boldsymbol{X};\boldsymbol{\alpha})$ as shown in the next lemma, this being a straightforward extension of reference [[25], lemma 1].

Lemma 4. *$p_{V|\boldsymbol{X}} \in \mathcal{C}(\boldsymbol{X};\boldsymbol{\alpha})$ if and only if $(\mathbf{p}_{\boldsymbol{X}} - \mathbf{p}_{\boldsymbol{X}|v}) \in Null(\mathbf{P}_{\boldsymbol{\alpha}})$, $\forall\ v \in \mathcal{V}$.*

Proof. Let X, Y and Z be discrete r.v.'s that form a Markov chain as $X - Y - Z$. Having $X \perp\!\!\!\perp Z$ is equivalent to $p_X(\cdot) = p_{X|Z}(\cdot|z)$, i.e., $\mathbf{p}_X = \mathbf{p}_{X|z}$, $\forall\ z \in \mathcal{Z}$. Furthermore, due to the Markov chain assumption, we have $\mathbf{p}_{X|z} = \mathbf{P}_{X|Y}\mathbf{p}_{Y|z}$, $\forall\ z \in \mathcal{Z}$, and in particular, $\mathbf{p}_X = \mathbf{P}_{X|Y}\mathbf{p}_Y$. Therefore, having $\mathbf{p}_X = \mathbf{p}_{X|z}$, $\forall\ z \in \mathcal{Z}$ results in

$$\mathbf{P}_{X|Y}\left(\mathbf{p}_Y - \mathbf{p}_{Y|z}\right) = \mathbf{0}, \quad \forall\ z \in \mathcal{Z},$$

or equivalently, $\left(\mathbf{p}_Y - \mathbf{p}_{Y|z}\right) \in \mathrm{Null}(\mathbf{P}_{X|Y})$, $\forall\ z \in \mathcal{Z}$.

The proof is complete by noting that (i) $\boldsymbol{X}^{\alpha_i} - \boldsymbol{X} - Y$ form a Markov chain for each index $i \in [L]$, and (ii) $\mathrm{Null}(\mathbf{P}_{\boldsymbol{\alpha}}) = \bigcap_{i=1}^{n} \mathrm{Null}(\mathbf{P}_{\boldsymbol{X}^{\alpha_i}|\boldsymbol{X}})$. □

In summary, the matrix form of the (reverse) synergistic channel is related to $p_{\boldsymbol{X}}$ and the null space of $\mathbf{P}_{\boldsymbol{\alpha}}$. They key take-away from this lemma is that one can compute the conditional distributions $p_{\boldsymbol{X}|v}$ of the synergistic channel by algebraic manipulation of $\mathbf{P}_{\boldsymbol{\alpha}}$ and $\mathbf{p}_{\boldsymbol{X}}$. Furthermore, this lemma has one important implication: the synergistic channels needed to compute the synergistic components of $I(\boldsymbol{X};Y)$ with respect to a target variable Y depend only on $p_{\boldsymbol{X}}$, not on $p_{Y|\boldsymbol{X}}$.

Appendix B. Synergy based on *f*-information

Let p, q be two probability mass functions on $\mathcal{X}$, such that $q(x) > 0, \quad \forall\ x \in \mathcal{X}$. For a convex function f such that $f(1) = 0$, the f-divergence of p from q is defined as

$$D_f(p\|q) := \sum_{x\in\mathcal{X}} f\left(\frac{p(x)}{q(x)}\right) q(x). \tag{B1}$$

Many well-known divergences are special cases of the f-divergence, including the Kullbac–Leibler divergence (for $f(p) = p\log p$) and the total variation distance (for $f(p) = |p - 1|/2$).

Using D_f, one can define the *f-information* of a pair of discrete random variables (X, Y) as

$$I_f(X;Y) := D_f(p_{X,Y}\|p_X \cdot p_Y), \tag{B2}$$

which assesses the difference in terms of the f-divergence between their joint pmf (i.e. $p_{X,Y}$) and the product of the marginals (i.e., $p_X \cdot p_Y$). As a special case, one can obtain Shannon's mutual information when $f(p) = p\log p$.

Since the main tools used in this paper (such as convexity and data processing inequality) are also valid for the f-information, the main results of this paper continue to hold for the more

general *f-synergy*, defined as

$$S_f^{\alpha}(\boldsymbol{X} \to Y) := \sup_{\substack{p_{V|\boldsymbol{X}} \in \mathcal{C}(\boldsymbol{X};\boldsymbol{\alpha}): \\ V - \boldsymbol{X} - Y}} I_f(V;Y). \tag{B3}$$

That being said, please note that the *f*-information in general does not satisfy a chain rule, unlike mutual information which stems from the logarithmic nature of the KL-divergence. Hence, results that rely on the chain rule (such as proposition 1) fail to hold in this more general setup.

Appendix C. Proofs of section 2

The following proof is an extension of results presented in reference [25].

Proof of proposition 1. Let $j \in [n]$ be an arbitrary index. First note that,

$$\begin{aligned} I(Y;\boldsymbol{X}|V) &= I(Y;\boldsymbol{X}^{\alpha_j}|V) + I(Y;\boldsymbol{X}^{-\alpha_j}|V,\boldsymbol{X}^{\alpha_j}) && \text{(C1)}\\ &\geqslant I(Y;\boldsymbol{X}^{\alpha_j}|V) && \text{(C2)}\\ &= I(Y,V;\boldsymbol{X}^{\alpha_j}). && \text{(C3)} \end{aligned}$$

where the second equality is due to the independence between V and $\boldsymbol{X}^{\alpha_j}$. Then, we find that

$$\begin{aligned} I(Y;V) =&\ I(Y;\boldsymbol{X}) - I(Y;\boldsymbol{X}|V) && \text{(C4)}\\ \leqslant&\ I(Y;\boldsymbol{X}^{-\alpha_j}|\boldsymbol{X}^{\alpha_j}) + I(Y;\boldsymbol{X}^{\alpha_j}) - I(Y,V;\boldsymbol{X}^{\alpha_j})\\ =&\ I(Y;\boldsymbol{X}^{-\alpha_j}|\boldsymbol{X}^{\alpha_j}) - I(V;\boldsymbol{X}^{\alpha_j}|Y)\\ \leqslant&\ I(Y;\boldsymbol{X}^{-\alpha_j}|\boldsymbol{X}^{\alpha_j}), && \text{(C5)} \end{aligned}$$

where (C4) follows from the Markov chain $Y - \boldsymbol{X} - V$. Since j is chosen arbitrarily, (C5) holds for all $j \in [n]$, resulting in (4).

The inequalities in the above derivation turn into an equality when

$$I(Y;\boldsymbol{X}^{-\alpha_j}|V,\boldsymbol{X}^{\alpha_j}) = I(V;\boldsymbol{X}^{\alpha_j}|Y) = 0. \tag{C6}$$

□

Appendix D. Further properties of synergistic disclosure

We first consider the convexity of the synergy over channels, i.e. conditional probabilities relating sources $\boldsymbol{X}$ and target Y.

Lemma 5 (Convexity of S^{α} over target channels). *$S^{\alpha}(\boldsymbol{X} \to Y)$ is a convex function of $p_{Y|\mathbf{X}}$ for a given $p_{\mathbf{X}}$.*

Proof. Let us denote the maximiser of $S^{\alpha}(\mathbf{X} \to Y)$ by $p^{\alpha*}_{V|\mathbf{X}}$ (cf the corresponding discussion below (3)), and consider $p_{Y|\mathbf{X}} = \theta p^1_{Y|\mathbf{X}} + (1-\theta)p^2_{Y|\mathbf{X}}$ for $\theta \in [0,1]$. From the convexity of mutual information, we have

$$S^{\alpha}(\mathbf{X} \to Y) \leqslant \theta I_1(V;Y) + (1-\theta) I_2(V;Y),$$

where I_1, and I_2 are evaluated over

$$p_1(v,y) = \sum_{\mathbf{X}} p^{\alpha*}_{V|\mathbf{X}} \cdot p_{\mathbf{X}} \cdot p^1_{Y|\mathbf{X}}, \text{ and} \tag{D1}$$

$$p_2(v, y) = \sum_{\mathbf{x}} p^{\alpha*}_{V|\mathbf{X}} \cdot p_{\mathbf{X}} \cdot p^2_{Y|\mathbf{X}}, \tag{D2}$$

respectively. It is also readily verified that

$$p_1(v|\mathbf{x}) = p_2(v|\mathbf{x}) = p^{\boldsymbol{\alpha}*}(v|\mathbf{x}) \in \mathcal{C}(\boldsymbol{X}; \boldsymbol{\alpha}).$$

As a result, one finds that

$$S^{\alpha}(\mathbf{X} \to Y) \leqslant \theta S^{\alpha}_1(\mathbf{X} \to Y) + (1 - \theta) S^{\alpha}_2(\mathbf{X} \to Y). \qquad \square$$

We now considered an extension of the data processing inequality of the mutual information to the case of $\boldsymbol{\alpha}$ synergies.

Lemma 6 (Data processing inequality for α-synergy). *If $\boldsymbol{X} - Y - Z$ is a Markov chain, then*

$$S^{\alpha}(\boldsymbol{X} \to Y) \geqslant S^{\alpha}(\boldsymbol{X} \to Z). \tag{D3}$$

Proof. A direct calculation shows that

$$S^{\alpha}(\boldsymbol{X} \to Y) = \sup_{\substack{p_{V|X^n} \in \mathcal{C}(\boldsymbol{X};\boldsymbol{\alpha}): \\ V - \boldsymbol{X} - Y}} I(V; Y) \tag{D4}$$

$$\geqslant \sup_{\substack{p_{V|X^n} \in \mathcal{C}(\boldsymbol{X};\boldsymbol{\alpha}): \\ V - \boldsymbol{X} - Z}} I(V; Z) \tag{D5}$$

$$= S^{\beta}(\boldsymbol{X} \to Z). \tag{D6}$$

Above, (D5) uses the fact that $\mathcal{C}(\boldsymbol{X}; \boldsymbol{\alpha})$ depends only on $\boldsymbol{X}$ and not on the target variable, and the traditional data processing inequality over the Markov Chain $V - \boldsymbol{X} - Y - Z$. $\square$

Finally, the last proposition explored here characterises conditions under which when there is no $\boldsymbol{\alpha}$-synergy. The proof of this result is omitted, as it is a direct extension of reference [[25], proposition 1]

Lemma 7. $S^{\alpha}(\boldsymbol{X} \to Y) = 0$ *if and only if* $\text{Null}(\mathbf{P}_{\boldsymbol{\alpha}}) \not\subset \text{Null}(\mathbf{P}_{Y|\boldsymbol{X}})$.

Appendix E. Proofs of section 3

Proof of lemma 1. As per lemma 4, the synergistic channel of interest depends only on the null space of $\mathbf{P}_{\boldsymbol{\alpha}}$. Recall that adding a new source α' to an existing source-set $\boldsymbol{\alpha} = \{\alpha_1, \ldots, \alpha_L\}$ corresponds to appending rows to $\mathbf{P}_{\boldsymbol{\alpha}}$ (cf equation (A1)). If $\alpha' \subset \alpha_i$, the new rows added to $\mathbf{P}_{\boldsymbol{\alpha}}$ corresponding to α' are linearly dependent on the existing rows, and therefore the null space of the matrix (and thus the synergistic channel) remains unchanged.

From this same line of reasoning, the smallest such source-set is that in which no source is contained in another one—i.e. an anti-chain. $\square$

Proof of lemma 2. Consider $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathcal{A}^*$ with $\boldsymbol{\alpha} \preceq_{\text{c}} \boldsymbol{\beta}$. Then, it is direct to check that $\mathcal{C}(\boldsymbol{X}; \boldsymbol{\beta}) \subseteq \mathcal{C}(\boldsymbol{X}; \boldsymbol{\alpha})$, and therefore

$$S^{\alpha}(\boldsymbol{X} \to Y) = \sup_{\substack{p_{V|X^n} \in \mathcal{C}(\boldsymbol{X};\boldsymbol{\alpha}): \\ V - \boldsymbol{X} - Y}} I(V; Y) \tag{E1}$$

$$\geqslant \sup_{\substack{p_{V|X^n} \in \mathcal{C}(\boldsymbol{X};\boldsymbol{\beta}):\\ V-\boldsymbol{X}-Y}} I(V;Y) \tag{E2}$$

$$= S^{\boldsymbol{\alpha}}(\boldsymbol{X} \to Y). \tag{E3}$$

Above, the inequality is because the supremum is taken over a smaller set of parameters. □

Note that the proof of the *weak monotonicity* property (cf section 6.1) follows exactly the same pattern, but leveraging the fact that $\mathcal{C}(\boldsymbol{X}, \boldsymbol{\alpha}) \subseteq \mathcal{C}((\boldsymbol{X}, Z), \boldsymbol{\alpha})$. The details are left to the interested reader.

Appendix F. Characterisation of synergistic channels in binary bivariate systems

Without loss of generality, let us consider the joint distribution of binary variables (X_1, X_2) described by

$$p_{X_1,X_2} = \left[r, a-r, b-r, 1-a-b+r\right], \tag{F1}$$

where $\mathbb{P}\{X_1 = 1\} = a$ and $\mathbb{P}\{X_2 = 1\} = b$ with $a \geqslant b$ determine the marginal distributions, and $\mathbb{P}\{X_1 = 1, X_2 = 1\} = r \in [0, R]$ with $R = \min\{a, b\}$ gives the interdependency (note that X_1 and X_2 are independent when $r = ab$).

The optimal $\boldsymbol{\alpha}$-synergistic channel for $\boldsymbol{\alpha} = \{\{1\}, \{2\}\}$ for this system has been shown to be (see reference [25])

$$P_{V^*|\boldsymbol{X}} = \begin{bmatrix} \frac{r(a-R)}{R(a-r)} & 1 & \frac{r(1-a-b+R)}{R(1-a-b+r)} & \frac{r(b-R)}{R(b-r)} \\ \frac{a(R-r)}{R(a-r)} & 0 & \frac{r(1-a-b)(R-r)}{R(1-a-b-r)} & \frac{b(R-r)}{R(b-r)} \end{bmatrix}. \tag{F2}$$

Interestingly, $P_{V^*|\boldsymbol{X}}$ depends on the distribution of $\boldsymbol{X}$ but not on Y. For the particular case of $a = b = 1/2$ and $r = ab = 1/4$, $P_{V^*|\boldsymbol{X}}$ reduces to an XOR.

Appendix G. Simulation details

This Appendix provides simulation details for the numerical results in the paper.

Simulations in section 4.3: for figure 4 we use Gibbs distributions as specified in equation (19) with the Hamiltonian given by equation (20). In contrast, for figure 4(b) we also considered Gibbs distributions but this time with Hamiltonians that only have terms of order k, i.e.

$$\mathcal{H}_k(\boldsymbol{x}^n) = -x_{n+1} \sum_{|\gamma|=k} J_\gamma \prod_{i \in \gamma} x_i.$$

In all simulations, all interaction coefficients J in the Hamiltonians are drawn i.i.d. from a normal distribution with zero mean and standard deviation 0.1. In both simulations, 25 Hamiltonians are sampled at random for each k, and the mean and standard deviation of the resulting quantities (B^m or B^m_∂) are reported in both panes of figure 4.

Simulations in sections 6.2 and 6.3: for these simulations, each distribution $p_{\boldsymbol{X}}$ is sampled from a symmetric Dirichlet distribution with concentration parameter α. Let us define $\mathrm{Dir}(K, \alpha)$

as the Dirichlet distribution over the $(K-1)$ simplex with all parameters $\alpha_1 = \cdots = \alpha_K = \alpha$. Sampling from this Dirichlet with a fixed α, however, has the undesirable effect of generating distributions with a very narrow distribution of entropy $H(\boldsymbol{X})$ [59]. To generate distributions with a near-uniform of entropy, we sample α from a Nemenman–Shafee–Bialek (NSB) prior [60]

$$p(\alpha) \propto K\psi(K\alpha + 1) + \psi(\alpha + 1),$$

for a distribution over an alphabet of size K. For simulations of n binary variables, we set $K = 2^n$, sample α using the equation above, and then sample $p_{\boldsymbol{X}}$ from a symmetric Dirichlet using standard algorithms.

Appendix H. Asymptotic limits of self-disclosure

Proof of proposition 4. From Corollary 1.2 of reference [26] one can see that

$$\min_{p_{V|\boldsymbol{X}} \in \mathcal{C}(\boldsymbol{X};\gamma_m)} H(\boldsymbol{X}|V) \leqslant \log\left(\operatorname{rank}(\mathbf{P}_{\gamma_m})\right). \tag{H1}$$

As the rank of a matrix cannot be larger than its number of rows, for a given $m \in \{1, \ldots, n-1\}$ is clear that

$$\operatorname{rank}(\mathbf{P}_{\gamma_m}) \leqslant m \sum_{k=1}^{n} |\mathcal{X}_k| \leqslant mnK, \tag{H2}$$

and therefore $\min_{p_{V|\boldsymbol{X}} \in \mathcal{C}(\boldsymbol{X};\gamma_m)} H(\boldsymbol{X}|V) \leqslant \log(mnK)$. By definition of $B^m(\boldsymbol{X})$, this implies that

$$H(\boldsymbol{X}) - \log(mnK) \leqslant B^m(\boldsymbol{X}) \leqslant H(\boldsymbol{X}). \tag{H3}$$

Taking the limit of $n \to \infty$ for m fixed gives that

$$\lim_{n\to\infty} \frac{1}{n} B^m(\boldsymbol{X}) = \lim_{n\to\infty} \frac{1}{n} H(\boldsymbol{X}), \tag{H4}$$

which is equivalent to what we want to prove (note that the cardinality of $\boldsymbol{X}$ grows with n as well). □

ORCID iDs

Fernando E Rosas https://orcid.org/0000-0001-7790-6183
Pedro A M Mediano https://orcid.org/0000-0003-1789-5894
Borzoo Rassouli https://orcid.org/0000-0002-1171-8507

References

[1] Ganmor E, Segev R and Schneidman E 2011 Sparse low-order interaction network underlies a highly correlated and learnable neural population code *Proc. Natl Acad. Sci.* **108** 9679–84
[2] Wibral M, Finn C, Wollstadt P, Lizier J T and Priesemann V 2017 Quantifying information modification in developing neural networks via partial information decomposition *Entropy* **19** 494
[3] Tax T M, Mediano A M and Shanahan M 2017 The partial information decomposition of generative neural network models *Entropy* **19** 474

[4] Rosas F, Mediano A, Ugarte M and Jensen H J 2018 An information-theoretic approach to self-organisation: emergence of complex interdependencies in coupled dynamical systems *Entropy* **20** 793
[5] Rosas F E, Mediano A M, Gastpar M and Jensen H J Sep 2019 Quantifying high-order interdependencies via multivariate extensions of the mutual information *Phys. Rev.* E **100** 032305
[6] Waldrop M M 1993 *Complexity: The Emerging Science at the Edge of Order and Chaos* (Simon and Schuster)
[7] Chechik G, Globerson A, Anderson M J, Young E D, Nelken I and Tishby N 2002 Group redundancy measures reveal redundancy reduction in the auditory pathway *Advances in Neural Information Processing Systems* pp 173–80
[8] Varadan V, Miller D M III and Anastassiou D 2006 Computational inference of the molecular logic for synaptic connectivity in C Elegans *Bioinformatics* **22** e497–506
[9] Barrett A B 2015 Exploration of synergistic and redundant information sharing in static and dynamical Gaussian systems *Phys. Rev.* E **91** 052802
[10] Amari S-I 2001 Information geometry on hierarchy of probability distributions *IEEE Trans. Inf. Theory* **47** 1701–11
[11] Schneidman E, Still S, Berry M J and Bialek W 2003 Network information and connected correlations *Phys. Rev. Lett.* **91** 238701
[12] Williams P L and Beer R D 2010 Nonnegative decomposition of multivariate information (arXiv:1004.2515)
[13] In effect, PID merely states formal relationships between its atoms and various Shannon's mutual information terms.
[14] Ince R A 2017 Measuring multivariate redundant information with pointwise common change in surprisal *Entropy* **19** 318
[15] Bertschinger N, Rauh J, Olbrich E, Jost J and Ay N 2014 Quantifying unique information *Entropy* **16** 2161–83
[16] Finn C and Lizier J T 2018 Pointwise partial information decomposition using the specificity and ambiguity lattices *Entropy* **20** 297
[17] James R, Emenheiser J and Crutchfield J 2019 Unique information and secret key agreement *Entropy* **21** 12
[18] Thurner S, Hanel R and Klimek P 2018 *Introduction to the Theory of Complex Systems* (Oxford: Oxford University Press)
[19] Rauh J, Bertschinger N, Olbrich E and Jost J 2014 Reconsidering unique information: towards a multivariate information decomposition *IEEE Int. Symp. on Information Theory 2014 (IEEE)* 2232–6
[20] Kolchinsky A 2019 A novel approach to multivariate redundancy and synergy (arXiv:1908.08642)
[21] Feldman D P and Crutchfield J P 1998 Measures of statistical complexity: why? *Phys. Lett.* A **238** 244–52
[22] Banerjee P K, Olbrich E, Jost J and Rauh J 2018 Unique informations and deficiencies *2018 56th Annual Allerton Conf. on Communication, Control, and Computing (Allerton) (IEEE)* 32–8
[23] Rauh J, Banerjee K, Olbrich E and Jost J 2019 Unique information and secret key decompositions *IEEE Int. Symp. on Information Theory (ISIT) IEEE, 2019* 3042–6
[24] The direct implications of these approaches are about the unique information, and apply to the synergy only via additional equalities with mutual information terms.
[25] Rassouli B, Rosas F and Gündüz D 2018 Latent feature disclosure under perfect sample privacy *IEEE Int. Workshop on Information Forensics and Security (WIFS) (IEEE) 2018* 1–7
[26] Rassouli B, Rosas F E and Gündüz D 2019 Data disclosure under perfect sample privacy *IEEE Trans. Inf. Forensics Secur.* **15** 2012–25
[27] Quax R, Har-Shemesh O and Sloot P 2017 Quantifying synergistic information using intermediate stochastic variables *Entropy* **19** 85
[28] To verify this, first one shows that it suffices to have a random variable V with a finite alphabet by means of cardinality bounding techniques. Then, using the fact that any finite probability simplex is a compact set, the supremum in (3) has to be attained due to the continuity of the mutual information.
[29] Shannon C E 1948 A mathematical theory of communication *Bell Syst. Tech. J.* **27** 379–423
[30] Cover T M and Thomas J A 2012 *Elements of Information Theory* (New York: Wiley)
[31] Yang B, Sato I and Nakagawa H 2015 Bayesian differential privacy on correlated data *Proc. of the 2015 ACM SIGMOD Int. Conf. on Management of Data* 747–62

[32] Cuff P and Yu L 2016 Differential privacy as a mutual information constraint *Proc. of the 2016 ACM SIGSAC Conf. on Computer and Communications Security*
[33] Li T and Li N 2009 On the tradeoff between privacy and utility in data publishing *Proc. of the 15th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining* 517–26
[34] Wang Y, Basciftci Y O and Ishwar P 2017 Privacy-utility tradeoffs under constrained data release mechanisms (arXiv:1710.09295)
[35] Huang C, Kairouz and Sankar L 2018 Generative adversarial privacy: a data-driven approach to information-theoretic privacy *52nd Asilomar Conf. on Signals, Systems, and Computers* 2162–6
[36] Rassouli B and Gündüz D 2019 Optimal utility-privacy trade-off with total variation distance as a privacy measure *IEEE Trans. Inf. Forensics Secur.* **15** 594–603
[37] Paradigmatic examples of proper operational interpretations are Shannon's source and channel coding theorems [30], which show that the entropy and mutual information correspond to the solution of precise engineering problems As a matter of fact, most problems in information theory are operational in nature; Shannon himself employed the known formula for entropy because it had a strict operational meaning in various communication scenarios, not because it was the result of the four Shannon–Khinchin axioms.
[38] Griffith V, Chong E, James R, Ellison C and Crutchfield J 2014 Intersection information based on common randomness *Entropy* **16** 1985–2000
[39] In fact, we employ the notation $\mathcal{A}^*$ to differentiate this set from with their definition of antichains set $\mathcal{A}$ that doesn't include the empty set—which plays an important role in our framework.
[40] James R, Emenheiser J and Crutchfield J 2018 Unique information via dependency constraints *J. Phys. A: Math. Theor.* **52** 014002
[41] Ay N, Polani D and Virgo N 2019 Information decomposition based on cooperative game theory (arXiv:1910.05979)
[42] Charalambides C A 2018 *Enumerative Combinatorics* (London: Chapman and Hall)
[43] Strictly speaking, since $\boldsymbol{\alpha}$ is a set of sets, these atoms should be denoted by $S_\partial^{\{\varnothing\}}$, $S_\partial^{\{\{1\}\{2\}\}}$, etc. For clarity, and for consistency with prior literature, we omit the outer bracket and denote these symbols by shortened expressions (e.g. $S_\partial^{\varnothing}$, $S_\partial^{\{1\}\{2\}}$, etc).
[44] Note that since $S^{\{12\}} = 0$, we have $S_\partial^{\{1\}\{2\}} = S^{\{1\}\{2\}}$
[45] Proofs of this result can be found in references [25, 27].
[46] For more information on this relationship, please see references [[5], lemma 3] and [[4], section 6].
[47] Jaynes E T 2003 *Probability Theory: the Logic of Science* (Cambridge: Cambridge University Press)
[48] Sakellariou J, Tria F, Loreto V and Pachet F 2017 Maximum entropy models capture melodic styles *Sci. Rep.* **7** 1–9
[49] Note that the presented findings consider a particular ensemble of distributions. Therefore, more work is needed in order to explore their generality.
[50] Rosas F, Ntranos V, Ellison C J, Pollin S and Verhelst M 2016 Understanding interdependency through complex information sharing *Entropy* **18** 38
[51] Griffith V and Koch C 2014 Quantifying synergistic mutual information *Guided Self-Organization: Inception* (Berlin: Springer) pp 159–90
[52] Harder M, Salge C and Polani D 2013 Bivariate measure of redundant information *Phys. Rev.* E **87** 012130
[53] Since our framework provides a algorithm to build the optimal self-synergistic channel for arbitrary sources X, it would be natural to conjecture that this channel could also be optimal for other target variables—i.e., that it could serve as sufficient statistic under the corresponding constraints. Unfortunately, numerical explorations show that, while this works for two binary variables (appendix F), it is in general not the case.
[54] As an example of this, if (X_1, X_2) are two independent fair coins and $Y = (X_1, X_2)$, then a direct calculation shows that, if $\boldsymbol{\alpha} = \{\{1\}, \{2\}\}$, then $S^{\boldsymbol{\alpha}}((X_1, X_2) \to Y) = 1$ and $S^{\boldsymbol{\alpha}}((X_1, Y) \to X_2) = 0$
[55] Consider a 'double-XOR' distribution, with 3 independent bits as inputs, and $Y = (X_1 \text{xor} X_2, X_2 \text{xor} X_3)$ as output. For this distribution, all atoms $S_\partial^{\{ij\}\{k\}}(\boldsymbol{X} \to Y) = -1$, violating (LP). To see why, note that $S^{\{12\}\{13\}}(\boldsymbol{X} \to Y) = S^{\{12\}\{3\}}(\boldsymbol{X} \to Y) = 1$, since in both cases $X_2 \text{xor} X_3$ can be disclosed, yielding the second bit of Y; and $S^{\{12\}\{23\}}(\boldsymbol{X} \to \text{Y}) = 1$, since $X_1 \text{xor} X_3$ can be disclosed, yielding the parity of Y. Hence, applying the Möbius inversion, we have $S_\partial^{\{12\}\{3\}}(\boldsymbol{X} \to Y) = -1$
[56] James R G, Ellison C J and Crutchfield J P 2018 dit: a Python package for discrete information theory *J. Open Source Softw.* **3** 738

[57] We do not take a stance here with respect to the non-negativity of information atoms; but since the atoms have to sum to the same mutual information, negative values in the lower atoms necessarily entail inflated synergy values.
[58] Pedro A M Mediano and Fernando E Rosas 2020 SYNDISC: SYNergistic information via data DISClosure A Python implementation of synergistic disclosure and the corresponding decomposition can be found online at https://github.com/pmediano/syndisc
[59] Nemenman I, Shafee F and Bialek W 2002 Entropy and inference, revisited *Advances in Neural Information Processing Systems* pp 471–8
[60] Archer E, Park I and Pillow J 2013 Bayesian and quasi-Bayesian estimators for mutual information from discrete data *Entropy* **15** 1738–55