entropy

MDPI

Article

A Novel Approach to the Partial Information Decomposition

Artemy Kolchinsky

Santa Fe Institute, Santa Fe, NM 87501, USA; artemyk@gmail.com

Abstract: We consider the "partial information decomposition" (PID) problem, which aims to decompose the information that a set of source random variables provide about a target random variable into separate redundant, synergistic, union, and unique components. In the first part of this paper, we propose a general framework for constructing a multivariate PID. Our framework is defined in terms of a formal analogy with intersection and union from set theory, along with an ordering relation which specifies when one information source is more informative than another. Our definitions are algebraically and axiomatically motivated, and can be generalized to domains beyond Shannon information theory (such as algorithmic information theory and quantum information theory). In the second part of this paper, we use our general framework to define a PID in terms of the well-known Blackwell order, which has a fundamental operational interpretation. We demonstrate our approach on numerous examples and show that it overcomes many drawbacks associated with previous proposals.

Keywords: partial information decomposition; redundancy; synergy

Citation: Kolchinsky, A. A Novel Approach to the Partial Information Decomposition. *Entropy* **2022**, *24*, 403. https://doi.org/10.3390/e24030403

Academic Editor: Eckehard Olbrich

Received: 4 January 2022
Accepted: 23 February 2022
Published: 13 March 2022

1. Introduction

Understanding how information is distributed in multivariate systems is an important problem in many scientific fields. In the context of neuroscience, for example, one may wish to understand how information about an external stimulus is encoded in the activity of different brain regions. In computer science, one might wish to understand how the output of a logic gate reflects the information present in different inputs to that gate. Numerous other examples abound in biology, physics, machine learning, cryptography, and other fields [1–10].

Formally, suppose that we are provided with a random variable Y which we call the "target", as well as a set of n random variables $X_1, \ldots, X_n$ which we call the "sources". The *partial information decomposition* (PID), first proposed by Williams and Beer in 2010 [11], aims to quantify how information about the target is distributed among the different sources. In particular, the PID seeks to decompose the mutual information provided jointly by all sources into a set of nonnegative terms, such as *redundancy* (information present in each individual source), *synergy* (information only provided by the sources jointly, not individually), *union information* (information provided by at least one individual source), and *unique information* (information provided by only one individual source).

As discussed in detail below, the PID is inspired by an analogy between information theory and set theory. In this analogy, the information that the sources provide about the target are imagined as sets, while PID terms such as redundancy, union information, and synergy are imagined as the sizes of intersections, unions, and complements. While the analogy between information-theoretic and set-theoretic quantities is suggestive, it does not specify how to actually define the PID. Moreover, it has also been shown that existing measures from information theory (such as mutual information and conditional mutual information) cannot be used directly to construct the PID, since these measures conflate contributions from different terms like synergy and redundancy [11,12]. In response, many proposals for how to define PID terms have been advanced [5,13–21]. However, existing proposals suffer from various drawbacks, such as behaving counterintuitively on simple examples, being limited to only two sources, or lacking a clear operational interpretation. Today there is no generally agreed-upon way of defining the PID.

In this paper, we propose a new and principled approach to the PID which addresses these drawbacks. Our approach can handle any number of sources and can be justified in algebraic, axiomatic, and operational terms. We present our approach in two parts.

In part I (Section 4), we propose a general framework for defining the PID. Our framework does not prescribe specific definitions, but instead shows how an information-theoretic decomposition can be grounded in a formal analogy with set theory. Specifically, we consider the definitions of "set intersection" and "set union" in set theory: the intersection of sets $S_1, S_2, \ldots$ is the largest set that is contained in all of the S_i, while the union of sets $S_1, S_2, \ldots$ is the smallest set that contains all of the S_i. As we show, these set-theoretic definitions can be mapped into information-theoretic terms by treating "sets" as random variables, "set size" as mutual information between a random variable and the target Y, and "set inclusion" as some externally specified ordering relation $\sqsubset$, which specifies when one random variable is more informative than another. Using this mapping, we define information-theoretic redundancy and union information in the same way that the sizes of intersections and unions are defined in set theory (other PID terms, such as synergy and unique information, can be computed in a straightforward way from redundancy and union information). Moreover, while our approach is motivated by set-theoretic intuitions, as we show in Section 4.2, it can also be derived from an alternative axiomatic foundation. We finish part I by reviewing relevant prior work in information theory and the PID literature. We also discuss how our framework can be generalized beyond the standard setting of the PID and even beyond Shannon information theory, to domains like algorithmic information theory and quantum information theory.

One unusual aspect of our framework is that it provides independent definitions of union information and redundancy. Most prior work on the PID has focused exclusively on the definition of redundancy, because it assumed that union information can be determined from redundancy using the so-called "inclusion-exclusion principle". In Section 4.3, we argue that the inclusion-exclusion principle should not be expected to hold in the context of the PID.

Part I provides a general framework. Concrete definitions of the PID can be derived from this general framework by choosing a specific "more informative" ordering relation $\sqsubset$. In fact, the study of ordering relations between information sources has a long history in statistics and information theory [22–27]. One particularly important relation is the so-called "Blackwell order" [13,28], which has a fundamental operational interpretation in terms of utility maximization in decision theory.

In part II of this paper (Section 5), we combine the general framework developed in part I with the Blackwell order. This gives rise to concrete definitions of redundancy and union information. We show that our measures behave intuitively and have simple operational interpretations in terms of decision theory. Interestingly, while our measure of redundancy is novel, our measure of union information has previously appeared in the literature under a different guise [13,17].

In Section 6, we compare our redundancy measure to previous proposals, and illustrate it with various bivariate and multivariate examples. We finish the paper with a discussion and proposals for future work in Section 7.

We introduce some necessary notation and preliminaries in the next section. In addition, we provide background regarding the PID in Section 3. All proofs, as well as some additional results, are found in the appendix.

2. Notation and Preliminaries

We use uppercase letters ($Y, X, Q, \ldots$) to indicate random variables over some underlying probability space. We use lowercase letters ($y, x, q, \ldots$) to indicate specific outcomes of random variables, and calligraphic letters ($\mathcal{Y}, \mathcal{X}, \mathcal{Q} \ldots$) to indicate sets of outcomes. We often index random variables with a subscript, e.g., the random variable X_i with outcomes $x_i \in \mathcal{X}_i$ (so x_i does not refer to the i^{th} outcome of random variable X, but rather to some generic outcome of random variable X_i). We use notation like $A - B - C$ to indicate that A is conditionally independent of C given B. Except where otherwise noted, we assume that all random variables have a finite number of outcomes.

We use notation like $P_X(x)$ to indicate the probability distribution associated with random variable X, $P_{XY}(x,y)$ to indicate the joint probability distribution associated with random variables X and Y, and $P_{X|Y}(x|y)$ to indicate the conditional probability distribution of X given Y. Given two random variables X and Y with outcome sets $\mathcal{X}$ and $\mathcal{Y}$, we use notation like $\kappa_{X|Y}(x|y)$ to indicate some stochastic *channel* of outputs $x \in \mathcal{X}$ given inputs $y \in \mathcal{Y}$. In general, a channel $\kappa_{X|Y}$ specifies some arbitrary conditional distribution of X given Y, which can be different from $P_{X|Y}$, the actual conditional distribution of X given Y (as determined by the underlying probability space).

As described above, we consider the information that a set of "source" random variables $X_1, \ldots, X_n$ provide a "target" random variable Y. Without loss of generality, we assume that the marginal distributions P_Y and P_{X_i} for all i have full support (if they do not, one can restrict $\mathcal{Y}$ and/or $\mathcal{X}_i$ to outcomes that have strictly positive probability).

Finally, note that despite our use of the terms "source" and "target", we do not assume any causal directionality between the sources and target (see also discussion in [29]). For example, in neuroscience, Y might be an external stimulus which causes the activity of brain regions $X_1, \ldots, X_n$, while in computer science Y might represent the output of a logic gate caused by inputs $X_1, \ldots, X_n$ (so the causal direction is reversed). In yet other contexts, there could be other causal relationships among $X_1, \ldots, X_n$ and Y, or they might not be causally related at all.

3. Background on the Partial Information Decomposition (PID)

Given a set of sources $X_1, \ldots, X_n$ and a target Y, the PID aims to decompose $I(Y; X_1, \ldots, X_n)$, the total mutual information provided by all sources about the target, into a set of nonnegative terms such as [11,12]:

Redundancy $I_\cap(X_1; \ldots; X_n \to Y)$, the information present in each individual source. Redundancy can be considered as the intersection of the information provided by different sources and is sometimes called "intersection information" in the literature [16,18].

Union information $I_\cup(X_1; \ldots; X_n \to Y)$, the information provided by at least one individual source [12,17].

Synergy $S(X_1; \ldots; X_n \to Y)$, the information found in the joint outcome of all sources, but not in any of their individual outcomes. Synergy is defined as [17]

$$S(X_1; \ldots; X_n \to Y) = I(Y; X_1, \ldots, X_n) - I_\cup(X_1; \ldots; X_n \to Y). \tag{1}$$

Unique information in source X_i, $U(X_i \to Y | X_1; \ldots; X_n)$, the non-redundant information in each particular source. Unique information is defined as

$$U(X_i \to Y | X_1; \ldots; X_n) = I(Y; X_i) - I_\cap(X_1; \ldots; X_n \to Y). \tag{2}$$

In addition to the above terms, one can also define *excluded information*,

$$E(X_i \to Y | X_1; \ldots; X_n) = I_\cup(X_1; \ldots; X_n \to Y) - I(Y; X_i), \tag{3}$$

as the information in the union of the sources which is not in a particular source X_i. To our knowledge, excluded information has not been previously considered in the PID literature, although it is the natural "dual" of unique information as defined in Equation (2).

Given the definitions above, once a measure of redundancy $I_\cap$ is chosen, unique information is determined by Equation (2). Similarly, once a measure of union information $I_\cup$ is chosen, synergy and excluded information are determined by Equations (1) and (3). In Figure 1, we illustrate the relationships between these different PID terms for the simple case of two sources, X_1 and X_2. We show two different decompositions of the information provided by the sources jointly, $I(X_1, X_2; Y)$, and individually, $I(X_1; Y)$ and $I(X_2; Y)$. The

diagram on the left shows the decomposition defined in terms of redundancy $I_\cap$, while the diagram on the right shows the decomposition defined in terms of union information $I_\cup$.

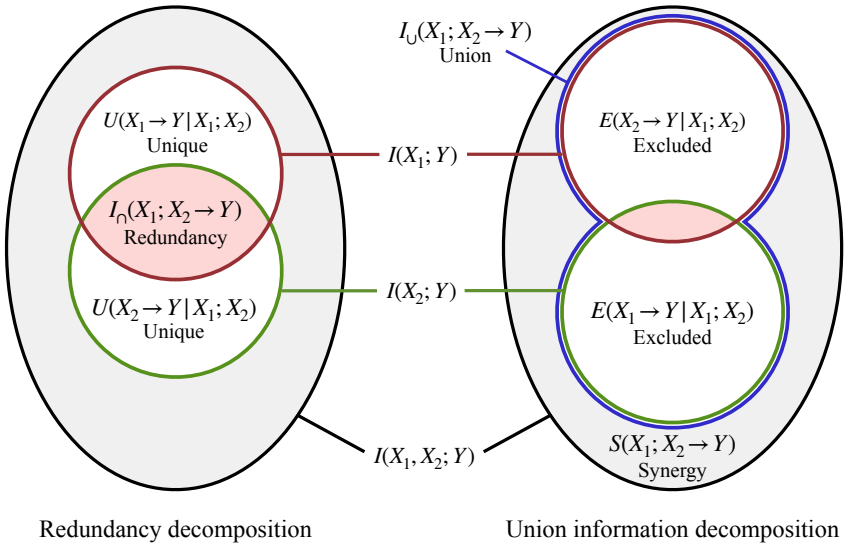

Figure 1. Partial information decomposition of the information provided by two sources about a target. On the left, we show the decomposition induced by redundancy $I_\cap$, which leads to measures of unique information U. On the right, we show the decomposition induced by union information $I_\cup$, which leads to measures of synergy S and excluded information E.

When more than two sources are present, the PID can be used to define additional terms, beyond the ones shown in Figure 1. For example, for three sources, one can define redundancy terms like $I_\cap(X_1, X_2, X_3 \to Y)$ (representing the information found in all individual sources) as well as redundancy terms like $I_\cap((X_1, X_2), (X_1, X_3), (X_2, X_3) \to Y)$ (representing the information found in all pairs of sources), and similarly for union information.

The idea that redundancy and union information lead to two different information decompositions is rarely discussed in the literature. In fact, the very concept of union information is rarely discussed in the literature explicitly (although it often appears in an implicit form via measures of synergy, since synergy is related to union information through Equation (1)). As we discuss below in Section 4.3, the reason for this omission is that most existing work assumes (whether implicitly or explicitly) that redundancy and union information are not independent measures, but are instead related via the so-called "inclusion-exclusion principle". If the inclusion-exclusion principle is assumed to hold, then the distinction between the two decompositions disappears. We discuss this issue in greater detail below, where we also argue that the inclusion-exclusion principle should not be expected to hold in the context of the PID.

We have not yet described how the redundancy and union information measures $I_\cap$ and $I_\cup$ are defined. In fact, this remains an open research question in the field (and one which this paper will address). When they first introduced the idea of the PID, Williams and Beer proposed a set of intuitive axioms that any measure of redundancy should satisfy [11,12], which we summarize in Appendix A. In later work, Griffith and Koch [17] proposed a similar set of axioms that union information should satisfy, which are also summarized in Appendix A. However, these axioms do not uniquely identify a particular measure of redundancy or union information.

Williams and Beer also proposed a particular redundancy measure which satisfies their axioms, which we refer to as $I_\cap^{\mathrm{WB}}$ [11,12]. Unfortunately, $I_\cap^{\mathrm{WB}}$ has been shown to behave counterintuitively in some simple cases [19,20]. For example, consider the so-called "COPY gate", where there are two sources X_1 and X_2 and the target is a copy of their joint outcomes, $Y = (X_1, X_2)$. If X_1 and X_2 are statistically independent, $I(X_1; X_2) = 0$,

then intuition suggests that the two sources provide independent information about Y and therefore that redundancy should be 0. In general, however, $I_\cap^{\mathrm{WB}}(X_1; X_2 \to Y)$ does not vanish in this case. To avoid this issue, Ince [20] proposed that any valid redundancy measure should obey the following property:

$$\text{If } I(X_1; X_2) = 0, \text{ then } I_\cap(X_1; X_2 \to (X_1, X_2)) = 0, \tag{4}$$

which is called the *Independent identity property*.

In recent years, many other redundancy measures have been proposed [13,15,16,18–21]. However, while some of these proposals satisfy the Independent identity property, they suffer various other drawbacks, such as exhibiting other types of counterintuitive behavior, being limited to two sources, and/or lacking a clear operational motivation. We discuss some of these previously proposed measures in Sections 4.4, 5.4 and 6.

Unlike redundancy, to our knowledge only two measures of union information have been advanced. The first one appeared in the original work on the PID [12], and was derived from $I_\cap^{\mathrm{WB}}$ using the inclusion-exclusion principle. The second one appeared more recently [13,17] and is discussed in Section 5.4 below.

4. Part I: Redundancy and Union Information from an Ordering Relation

4.1. Introduction

As mentioned above, PID is motivated by an informal analogy with set theory [12]. In particular, redundancy is interpreted analogously to the size of the intersection of the sources $X_1, \ldots, X_n$, while union information is interpreted analogously to the size of their union.

We propose to define the PID by making this analogy formal, and in particular by going back to the algebraic definitions of intersection and union in set theory. In pursuing this direction, we build on a line of previous work in information theory and PID, which we discuss in Section 4.4.

Recall that in set theory, the intersection of sets $S_1, \ldots, S_n \subseteq U$ (where U is some universal set) is the largest set that is contained in all S_i (Section 7.2, [30]). This means that the size of the intersection can be written as

$$\left|\bigcap_i S_i\right| = \sup_{T \subseteq U} |T| \text{ such that } \forall i \; T \subseteq S_i, \tag{5}$$

Similarly, the union of sets $S_1, \ldots, S_n \subseteq U$ is the smallest set that contains all S_i (Section 7.2, [30]), so the size of the union can be written as

$$\left|\bigcup_i S_i\right| = \inf_{T \subseteq U} |T| \text{ such that } \forall i \; S_i \subseteq T. \tag{6}$$

Equations (5) and (6) are useful because they express the size of the intersection and union via an optimization over simpler terms (the size of individual sets, $|T|$, and the subset inclusion relation, $\subseteq$).

We translate these definitions to the information-theoretic setting of the PID. We take the analogue of a "set" to be some random variable A that provides information about the target Y, and the analogue of "set size" to be the mutual information $I(A; Y)$. In addition, we assume that there is some ordering relation $\sqsubset$ between random variables analogous to set inclusion $\subseteq$. Given such a relation, the expression $A \sqsubset B$ means that random variable B is "more informative" than A, in the sense that the information that A provides about Y is contained within the information that B provides about Y.

At this point, we leave the ordering relation $\sqsubset$ unspecified. In general, we believe that the choice of $\sqsubset$ will not be determined from purely information-theoretic considerations, but may instead depend on the operational setting and scientific domain in which the PID is applied. At the same time, there has been a great deal of research on ordering relations in statistics and information theory. In part II of this paper, Section 5, we will combine our

general framework with a particular ordering relation, the so-called "Blackwell order", which has a fundamental interpretation in terms of decision theory.

We now provide formal definitions of redundancy and union information, relative to the choice of ordering relation $\sqsubset$. In analogy to Equation (5), we define redundancy as

$$I_\cap(X_1;\ldots;X_n \to Y) := \sup_Q I(Q;Y) \;\text{ such that } \forall i\; Q \sqsubset X_i \tag{7}$$

where the maximization is over all random variables with a finite number of outcomes. Thus, redundancy $I_\cap$ is the maximum information about Y in any random variable that is less informative than all of the sources. In analogy with Equation (6), we define union information as

$$I_\cup(X_1;\ldots;X_n \to Y) := \inf_Q I(Q;Y) \;\text{ such that } \forall i\; X_i \sqsubset Q \tag{8}$$

Thus, union information $I_\cup$ is the minimum information about Y in any random variable that is more informative than all of the sources. Given these definitions, other elements of the PID (such as unique information, synergy, and excluded information) can be defined using the expressions found in Section 3. Note that $I_\cap$ and $I_\cup$ depend the choice of ordering relation $\sqsubset$, although for convenience we leave this dependence implicit in our notation.

One of the attractive aspects of our definitions is that they do not simply quantify the amount of redundancy and union information, but also specify the "content" of that redundant and union information. In particular, the random variable Q that achieves the optimum in Equation (7) specifies the content of the redundant information via the joint distribution P_{YQ}. Similarly, the random variable Q which achieves the optimum in Equation (8) specifies the content of the union information via the joint distribution P_{YQ}. Note that these optimizing Q may not be unique, reflecting the fact that there may be different ways to represent the redundancy or union information. (Note also that the supremum or infinitum may not be achieved in Equations (7) and (8), in which case one can consider Q that achieve the optimal values to any desired precision $\epsilon > 0$.)

So far we have not made any assumptions about the ordering relation $\sqsubset$. However, we can derive some useful bounds by introducing three weak assumptions:

I. Monotonicity of mutual information: $A \sqsubset B \implies I(A;Y) \leq I(B;Y)$ (less informative sources have less mutual information).
II. Reflexivity: $A \sqsubset A$ for all A (each source is at least as informative as itself).
III. For all sources X_i, $O \sqsubset X_i \sqsubset (X_1,\ldots,X_n)$, where O indicates a constant random variable with a single outcome and $(X_1,\ldots,X_n)$ indicates all sources considered jointly (each source is more informative than a trivial source and less informative than all sources jointly).

Assumptions I and II imply that the redundancy and union information of a single source are equal to the mutual information in that source:

$$I_\cap(X_1 \to Y) = I_\cup(X_1 \to Y) = I(X_1;Y).$$

Assumptions I and III imply the following bounds on redundancy and union information:

$$0 \leq I_\cap(X_1;\ldots;X_n \to Y) \leq \min_i I(Y;X_i)\,. \tag{9}$$

$$\max_i I(Y;X_i) \leq I_\cup(X_1;\ldots;X_n \to Y) \leq I(Y;X_1,\ldots,X_n). \tag{10}$$

Equation (9) in turn implies that the unique information in each source X_i, as defined in Equation (2), is bounded between 0 and $I(Y;X_i)$. Similarly, Equation (10) implies that the synergy, as defined in Equation (1), obeys

$$0 \leq S(X_1;\ldots;X_n \to Y) \leq \min_i I(Y;X_1,\ldots,X_n|X_i),$$

where we have used the chain rule $I(Y; X_1, \ldots, X_n) = I(Y; X_i) + I(Y; X_1, \ldots, X_n | X_i)$. Equation (10) also implies that excluded information in each source X_i, as defined in Equation (3), is bounded between 0 and $I(Y; X_1, \ldots, X_n | X_i)$.

Note that in general, stronger orders give smaller values of redundancy and larger values of union information. Consider two orders $\sqsubset$ and $\sqsubset'$ where the first one is stronger than the second: $A \sqsubset B \implies A \sqsubset' B$ for all A and B. Then, any Q in the feasible set of Equation (7) under $\sqsubset$ will also be in the feasible set under $\sqsubset'$, and similarly for Equation (8). Therefore, $I_\cap$ defined relative to $\sqsubset$ will have a lower value than $I_\cap$ defined relative to $\sqsubset'$, and vice versa for $I_\cup$.

In the rest of this section, we discuss alternative axiomatic justifications for our general framework, the role of the inclusion-exclusion principle, relation to prior work, and further generalizations. Readers who are more interested in the use of our framework to define concrete measures of redundancy and union information may skip to Section 5.

4.2. Axiomatic Derivation

In Section 4.1, we defined the PID in terms of an algebraic analogy with intersection and union in set theory. This definition can be considered as the primary one in our framework. At the same time, the same definitions can also be derived in an alternative manner from a set of axioms, as commonly sought after in the PID literature. In particular, in Appendix B, we prove the following result regarding redundancy.

Theorem 1. *Any redundancy measure that satisfies the following five axioms is equal to* $I_\cap(X_1; \ldots; X_n \to Y)$ *as defined in Equation* (7).

1. Symmetry*:* $I_\cap(X_1; \ldots; X_n \to Y)$ *is invariant to the permutation of* $X_1, \ldots, X_n$.
2. Self-redundancy*:* $I_\cap(X_1 \to Y) = I(Y; X_1)$.
3. Monotonicity*:* $I_\cap(X_1; \ldots; X_n \to Y) \leq I_\cap(X_1; \ldots; X_{n-1} \to Y)$.
4. Order equality*:* $I_\cap(X_1; \ldots; X_n \to Y) = I_\cap(X_1; \ldots; X_{n-1} \to Y)$ *if* $X_i \sqsubset X_n$ *for some* $i < n$.
5. Existence*: There is some* Q *such that* $I_\cap(X_1; \ldots; X_n \to Y) = I(Y; Q)$ *and* $Q \sqsubset X_i$ *for all* i.

While *Symmetry*, *Self-redundancy*, and *Monotonicity* axioms are standard in the PID literature (see Appendix A), the last two axioms require some explanation. *Order equality* is a generalization of the previously proposed *Deterministic equality* axiom, described in Appendix A, where the condition $X_i = f(X_n)$ (deterministic relationship) is generalized to the "more informative" relation $X_i \sqsubset X_n$. This axiom reflects the idea that if a new source X_n is more informative than an existing source X_i, then redundancy shouldn't decrease when X_n is added.

Existence is the most novel of our proposed axioms. It says that for any set of sources $X_1, \ldots, X_n$, there exists some random variable which captures the redundant information. It is similar to the statement in axiomatic set theory that the intersection of a collection of sets is itself a set (note that in Zermelo-Fraenkel set theory, this statement is derived from the Axiom of Separation).

We can derive a similar result for union information (proof in Appendix B).

Theorem 2. *Any union information measure that satisfies the following five axioms is equal to* $I_\cup(X_1; \ldots; X_n \to Y)$ *as defined in Equation* (8).

1. Symmetry*:* $I_\cup(X_1; \ldots; X_n \to Y)$ *is invariant to the permutation of* $X_1, \ldots, X_n$.
2. Self-union*:* $I_\cup(X_1 \to Y) = I(Y; X_1)$.
3. Monotonicity*:* $I_\cup(X_1; \ldots; X_n \to Y) \geq I_\cup(X_1; \ldots; X_{n-1} \to Y)$.
4. Order equality*:* $I_\cup(X_1; \ldots; X_n \to Y) = I_\cup(X_1; \ldots; X_{n-1} \to Y)$ *if* $X_n \sqsubset X_i$ *for some* $i < n$.
5. Existence*: There is some* Q *such that* $I_\cup(X_1; \ldots; X_n \to Y) = I(Y; Q)$ *and* $X_i \sqsubset Q$ *for all* i.

These axioms are dual to the redundancy axioms outlined above. Compared to previously proposed axioms for union information, as described in Appendix A, the most unusual of our axioms is *Existence*. It says that given a set of sources $X_1, \ldots, X_n$, there exists some random variable which captures the union information. It is similar in spirit to the "Axiom of Union" in axiomatic set theory [31].

Finally, note that for some choices of $\sqsubset$, there may not exist measures of redundancy and/or union information that satisfy the axioms in Theorem 1 and Theorem 2, in which case these theorems still hold but are trivial. However, even in such "pathological" cases, $I_\cap$ and $I_\cup$ can still be defined via Equations (7) and (8), as long as $\sqsubset$ has a "least informative" and a "most informative" element (e.g., as provided by Assumption III above), so that the feasible sets are not empty. In this sense, the definitions in Equations (7) and (8) are more general than the axiomatic derivations provided by Theorems 1 and 2.

4.3. Inclusion-Exclusion Principle

One unusual aspect of our approach is that, unlike most previous work, we propose separate measures of redundancy and union information.

Recall that in set theory, the size of the intersection and the union are not independent of each other, but are instead related by the *inclusion-exclusion principle* (IEP). For example, given any two sets S and T, the IEP states that the size of the union of S and T is given by the sum of their individual sizes minus the intersection,

$$|S \cup T| = |S| + |T| - |S \cap T|. \tag{11}$$

More generally, the IEP relates the sizes of intersection and unions for any number of sets, via the following inclusion-exclusion formulas:

$$\left| \bigcup_{i=1}^{n} S_i \right| = \sum_{\varnothing \neq J \subseteq \{1,\ldots,n\}} (-1)^{|J|-1} \left| \bigcap_{i \in J} S_i \right|. \tag{12}$$

$$\left| \bigcap_{i=1}^{n} S_i \right| = \sum_{\varnothing \neq J \subseteq \{1,\ldots,n\}} (-1)^{|J|-1} \left| \bigcup_{i \in J} S_i \right|. \tag{13}$$

Historically, the IEP has played an important role in analogies between set theory and information theory, which began to be explored in 1950s and 1960s [32–36]. Recall that the entropy $H(X)$ quantifies the amount of information gained by learning the outcome of random variable X. It has been observed that, for a set of random variables $X_1, \ldots, X_n$, the joint entropy $H(X_1, \ldots, X_n)$ behaves somewhat like the size of the union of the information in the individual variables. For instance, like the size of the union, joint entropy is sub-additive ($H(X_1) + H(X_2) \geq H(X_1, X_2)$) and increases with additional random variables ($H(X_1, X_2) \geq H(X_1)$). Moreover, for two random variables X_1 and X_2, the mutual information $I(X_1; X_2) = H(X_1) + H(X_2) - H(X_1, X_2)$ acts like the size of the intersection of the information provided by X_1 and X_2, once intersection is defined analogously to the IEP expression in Equation (11) [35,36]. Given the general IEP formula in Equation (13), this can be used to define the size of the intersection between any number of random variables. For instance, the size of a three-way intersection is

$$\begin{aligned} I(X_1; X_2; X_3) = {} & H(X_1) + H(X_2) + H(X_3) \\ & - H(X_1, X_2) - H(X_1, X_3) - H(X_2, X_3) + H(X_1, X_2, X_3), \end{aligned}$$

a quantity called *co-information* or *interaction information* in the literature [32,33,35–37].

Unfortunately, interaction information, as well as other higher-order interaction terms defined via the IEP, can take negative values [32,35,37]. This conflicts with the intuition that information measures should always be non-negative, in the same way that set size is always non-negative.

One of the primary motivations for the PID, as originally proposed by Williams and Beer [11,12], was to solve the problem of negativity encountered by interaction information. To develop a non-negative information decomposition, Williams and Beer took two steps. First, they considered the information that a set of sources $X_1, \ldots, X_n$ provide about some target random variable Y. Second, they developed a non-negative measure of redundancy ($I_\cap^{\mathrm{WB}}$) which leads to a non-negative union information once an IEP formula like

Equation (12) is applied (Theorem 4.7, [12]). For example, in the original proposal, union information and redundancy are related via

$$I_\cup(X_1; X_2 \to Y) \stackrel{?}{=} I(Y; X_1) + I(Y; X_2) - I_\cap(X_1; X_2 \to Y), \tag{14}$$

which is the analogue of Equation (11). This can be plugged into expressions like Equation (1), so as to express synergy in terms of redundancy as

$$S(X_1; \ldots; X_n \to Y) \stackrel{?}{=} I(Y; X_1, \ldots, X_n) - I(Y; X_1) - I(Y; X_2) + I_\cap(X_1; X_2 \to Y). \tag{15}$$

The meaning of IEP-based identities such as Equations (14) and (15) can be illustrated using the Venn diagrams in Figure 1. In particular, they imply that the pink region in the right diagram is equal in size to the pink region in the left diagram, and that the grey region in the left diagram is equal in size to the grey region in the right diagram. More generally, IEP implies an equivalence between the information decomposition based on redundancy and the one based on union information.

As mentioned in Section 3, due to shortcomings in the original redundancy measure $I_\cap^{\mathrm{WB}}$, numerous other proposals for the PID have been advanced. Most of these proposals introduce new measures of redundancy, while keeping the general structure of the PID as introduced by Williams and Beer. In particular, most of these proposals assume that the IEP holds, so that union information can be derived from a measure of redundancy. While the assumption of the IEP is sometimes stated explicitly, more frequently it is implicit in the definitions used. For example, many proposals assume that synergy is related to redundancy via an expression like Equation (15), although (as shown above) this implicitly assumes that the IEP holds. In general, the IEP has been largely an unchallenged and unexamined assumption in the PID field. It is easy to see the appeal of the IEP: it builds on deep-seated intuitions about intersection/union from set theory and Venn diagrams, it has a long history in the information-theoretic literature, and it simplifies the problem of defining the PID since it only requires a measure of redundancy to be defined — rather than a measure of redundancy and a measure of union information. (Note that one can also start from union information and then derive redundancy via the IEP formula in Equation (13), as in Appendix B of Ref. [17], although this is much less common in the literature.)

However, there is a different way to define a non-negative PID, which is still grounded in a formal analogy with set theory but does not assume the IEP. Here, one defines measures of redundancy and union information based on the underlying algebra of intersection and union: the intersection of $X_1, \ldots, X_n$ is the largest element that is less than each X_i, while the union is the smallest element that is greater than each X_i. Given these definitions, intersections and unions are not necessarily related to each numerically, as in the IEP, but are instead related by an algebraic duality.

This latter approach is the one we pursue in our definitions (it has also appeared in some prior work, which we review in the next subsection). In general, the IEP will not hold for redundancy and union information as defined in Equations (7) and (8). (To emphasize this point, we put a question mark in Equations (14) and (15), and made the sizes of the pink and grey regions visibly different in Figure 1). However, given the algebraic and axiomatic justifications for $I_\cap$ and $I_\cup$, we do not see the violation of the IEP as a fatal issue. In fact, there are many domains where generalizations of intersections and unions do not obey the IEP. For example, it is well-known that the IEP is violated in the domain of vector spaces, once the size of a vector space is measured in terms of its dimension [38]. The PID is simply another domain where the IEP should not be expected to hold.

We believe that many problems encountered in previous work on the PID — such as the failure of certain redundancy measures to generalize to more than two sources, or the appearance of uninterpretable negative synergy values — are artifacts of the IEP assumption. In fact, the following result shows that any measures of redundancy and union information which satisfy several reasonable assumptions must violate the IEP as

soon as 3 or more sources are present (the proof, in Appendix I, is based on a construction from [39,40]).

Lemma 1. *Let $I_\cap$ be any nonnegative redundancy measure which obeys* Symmetry, Self-redundancy, Monotonicity, *and* Independent identity. *Let $I_\cup$ be any union information measure which obeys $I_\cup(X_1;\ldots;X_n \to Y) \leq I(Y;X_1,\ldots,X_n)$. Then, $I_\cap$ and $I_\cup$ cannot be related by the inclusion-exclusion principle for 3 or more sources.*

The idea that different information decompositions may arise from redundancy versus synergy (and therefore union information) has recently appeared in the PID literature [15,40–43]. In particular, Chicharro and Panzeri proposed a PID that involves two decomposition: an "information gain" decomposition based on redundancy and an "information loss" decomposition based on synergy [41]. These decompositions correspond to the two Venn diagrams shown in Figure 1.

4.4. Relation to Prior Work

Here we discuss prior work which is relevant to our algebraic approach to the PID.

First, note that our definitions of redundancy and union information in Equations (7) and (8) are closely related to notions of "meet" and "join" in a field of algebra called order theory, which generalize intersections and unions to domains beyond set theory [44]. Given a set of objects S and an order $\sqsubset$, the *meet* of $a, b \in S$ is the unique largest $c \in S$ that is smaller than both a and b: $c \sqsubset a, c \sqsubset b$ and $d \sqsubset c$ for any d that obeys $d \sqsubset a, d \sqsubset b$. Similarly, the *join* of $a, b \in S$ is the unique smallest c that is larger than both a and b: $a \sqsubset c, a \sqsubset c$ and $c \sqsubset d$ for any d that obeys $a \sqsubset d, b \sqsubset d$. Note that meets and joins are only defined when $\sqsubset$ is a special type of partial order called a *lattice.* This is a strict requirement, and many important ordering relations in information theory are not lattices (this includes the "Blackwell order", which we will consider in part II of this paper [45]).

In our approach, we do not require the ordering relation $\sqsubset$ to be a lattice, or even a partial order. We do not require these properties because we do not aim to find the unique union random variable or the unique redundancy random variable. Instead, we aim to quantify the *size of the intersection* and *the size of the union*, which we do by optimizing mutual information subject to constraints, as Equations (7) and (8). These definitions are well-defined even when $\sqsubset$ is not a lattice, which allows us to consider a much broader set of ordering relations.

We mention three important precursors of our approach that have been proposed in the PID literature. First, Griffith et al. [16] considered the following order between random variables:

$$A \lhd B \text{ iff } A = f(B) \text{ for some deterministic function } f. \tag{16}$$

This ordering relation $\lhd$ was first considered in a 1953 paper by Shannon [22], who showed that it defines a lattice over random variables. That paper was the first to introduce the algebraic idea of meets and joins into information theory, leading to an important line of subsequent research [46–50]. Using this order, Ref. [16] defined redundancy as the maximum mutual information in any random variable that is a deterministic function of all of the sources,

$$I_\cap^\lhd(X_1;\ldots;X_n \to Y) := \max_Q I(Q;Y) \quad \text{such that} \quad \forall i \; Q \lhd X_i, \tag{17}$$

which is clearly a special case of Equation (7). Unfortunately, in practice, $I_\cap^\lhd$ is not a useful redundancy measure, as it tends to give very small values and is highly discontinuous. For example, $I_\cap^\lhd(X_1;\ldots;X_n \to Y) = 0$ whenever the joint distribution $P_{X_1\ldots X_n Y}$ has full support, meaning that it vanishes on almost all joint distributions [16,18,47]. The reason for this counterintuitive behavior is that the order $\lhd$ formalizes an extremely strict notion of "more informative", which is not robust to noise.

Given the deficiencies of $I_\cap^\triangleleft$, Griffith and Ho [18] proposed another measure of redundancy (also discussed as $I_\cap^2$ in Ref. [49]),

$$I_\cap^{\mathrm{GH}}(X_1;\ldots;X_n \to Y) := \max_Q I(Q;Y) \quad \text{such that} \quad \forall i\ Q - X_i - Y. \tag{18}$$

This measure is also a special case of Equation (7), where the more informative relation $A \sqsubset B$ is formalized via the conditional independence condition $A - B - Y$. This measure is similar to the redundancy measure we propose in part II of this paper, and we discuss it in more detail in Section 5.4. (Note that there are some incorrect claims about $I_\cap^{\mathrm{GH}}$ in the literature: Lemmas 6 and 7 of Ref. [49] incorrectly state that $I_\cap^{\mathrm{GH}}(X_1; X_2 \to Y) = 0$ whenever X_1 and X_2 are independent — see the AND gate counterexample in Section 6 — while Ref. [18] incorrectly states that $I_\cap^{\mathrm{GH}}$ obeys a property called *Target Monotonicity*.)

Finally, we mention the so-called "minimum mutual information" redundancy $I_\cap^{\mathrm{MMI}}$ [51]. This is perhaps the simplest redundancy measure, being equal to the minimal mutual information in any source: $I_\cap^{\mathrm{MMI}}(X_1;\ldots;X_n \to Y) := \min_i I(X_i;Y)$. It can be written in the form of Equation (7) as

$$I_\cap^{\mathrm{MMI}}(X_1;\ldots;X_n \to Y) := \max_Q I(Q;Y) \quad \text{such that} \quad \forall i\ I(Q;Y) \leq I(X_i;Y). \tag{19}$$

This redundancy measure has been criticized for depending only on the amount of information provided by the different sources, being completely insensitive to the content of that information. Nonetheless, $I_\cap^{\mathrm{MMI}}$ can be useful in some settings, and it plays an important role in the context of Gaussian random variables [51].

Interestingly, unlike $I_\cap^{\mathrm{MMI}}$, the original redundancy measure proposed by Williams and Beer [11], $I_\cap^{\mathrm{WB}}$, does not appear to be a special case of Equation (7) (at least not under the natural definition of the ordering relation $\sqsubset$). We demonstrate this using a counter-example in Appendix H.

As mentioned in Section 4.1, stronger ordering relations give smaller values of redundancy. For the orders considered above, it is easy to show that

$$A \lhd B \implies A - B - Y \implies I(A;Y) \leq I(B;Y). \tag{20}$$

This implies that $I_\cap^\triangleleft \leq I_\cap^{\mathrm{GH}} \leq I_\cap^{\mathrm{MMI}}$. In fact, $I_\cap^{\mathrm{MMI}}$ is the largest measure that is compatible with the monotonicity of mutual information (Assumption I in Section 4.1).

4.5. Further Generalizations

We finish part I of this paper by noting that one can further generalize our approach, by considering other analogues of "set", "set size", and "set inclusion" beyond the ones considered in Section 4.1. Such generalizations allow one to analyze notions of information intersection and union in a wide variety of domains, including setups different from the standard one considered in the PID, and domains not based on Shannon information theory.

At a general level, consider a set of object Ω that represents possible "sources", which may be random variables, as in Section 4.1, or otherwise. Assume there is some function $\phi : \Omega \to \mathbb{R}$ that quantifies the "amount of information" in a given source Ω, and some relation $\sqsubset$ on Ω that indicates which sources are more informative than others. Then, in analogy to Equations (5) and (6), for any set of sources $\{b_1,\ldots,b_n\} \subseteq \Omega$, one can define redundancy and union information as

$$I_\cap(b_1;\ldots;b_n) := \sup_{a\in\Omega} \phi(a) \quad \text{such that} \quad \forall i\ a \sqsubset b_i \tag{21}$$

$$I_\cup(b_1;\ldots;b_n) := \inf_{a\in\Omega} \phi(a) \quad \text{such that} \quad \forall i\ b_i \sqsubset a. \tag{22}$$

Synergy, unique, and excluded information can then be defined via Equations (1) to (3).

There are many possible examples of such generalizations, of which we mention a few as illustrations.

Shannon information theory (beyond mutual information). In Section 4.1, ϕ was the mutual information between each random variable and some target Y. This can be generalized by choosing a different "amount of information" function ϕ, so that redundancy and union information are quantified in terms of other measures of statistical dependence. Among many other options, possible choices of ϕ include Pearson's correlation (for continuous random variables) and measures of statistical dependency based f-divergences [52], Bregman divergences [53], and Fisher information [54].

Shannon information theory (without a fixed target). The PID can also be defined for a different setup than the typical one considered in the literature. For example, consider a situation where the sources are channels $\kappa_{X_1|Y}, \ldots, \kappa_{X_n|Y}$, while the marginal distribution over the target Y is left unspecified. Here one may take Ω as the set of channels, ϕ as the channel capacity $\phi(\kappa_{A|Y}) := \max_{P_Y} I_{P_Y \kappa_{A|Y}}(A;Y)$, and $\sqsubset$ as some ordering relation on channels [24]

Algorithmic information theory. The PID can be defined for other notions of information, such as the ones used in Algorithmic Information Theory (AIT) [55]. In AIT, "information" is not defined in terms of statistical uncertainty, but rather in terms of the program length necessary to generate strings. For example, one may take Ω as the set of finite strings, $\sqsubset$ as algorithmic conditional independence ($a \sqsubset b$ iff $K(y|b) - K(y|b,a) \leq$ const, where $K(\cdot|\cdot)$ is conditional Kolmogorov complexity), and $\phi(a) := K(y) - K(y|a)$ as the "algorithmic mutual information" with some target string y. (This setup is closely related to the notion of algorithmic "common information" [47].)

Quantum information theory. As a final example, the PID can be defined in the context of quantum information theory. For example, one may take Ω as the set of quantum channels, $\sqsubset$ as quantum Blackwell order [56–58], and $\phi(\Phi) = \mathcal{I}(\rho, \Phi)$, where $\mathcal{I}$ is the Ohya mutual information for some target density matrix ρ under channel $\Phi \in \Omega$ [59].

5. Part II: Blackwell Redundancy and Union Information

In the first part of this paper, we proposed a general framework for defining PID terms. In this section, which forms part II of this paper, we develop a concrete definition of redundancy and union information by combining our general framework with a particular ordering relation $\sqsubset$. This ordering relation is called the "Blackwell order", and it plays a fundamental role in statistics and decision theory [28,45,60]. We first introduce the Blackwell order, then use it to define measures of redundancy and union information, and finally discuss various properties of our measures.

5.1. The Blackwell Order

We begin by introducing the ordering relation that we use to define our PID. Given three random variables B, C and Y, the ordering relation $B \prec_Y C$ is defined as follows:

$$B \prec_Y C \text{ iff } P_{B|Y}(b|y) = \sum_c \kappa_{B|C}(b|c) P_{C|Y}(c|y) \text{ for some channel } \kappa_{B|C} \text{ and all } b, y. \quad (23)$$

We refer to the relation $\prec_Y$ as the *Blackwell order* relative to random variable Y. (Note that the Blackwell order and Blackwell's Theorem are usually formulated in terms of channels — that is, conditional distributions like $\kappa_{B|Y}$ and $\kappa_{C|Y}$ — rather than of random variables as done here. However, these two formulations are equivalent, as shown in [45].)

In words, Equation (23) means the conditional distribution by $P_{B|Y}$ can be generated by first sampling from the conditional distribution $P_{C|Y}$, and then applying some channel $\kappa_{B|C}$ to the outcome. The relation $B \prec_Y C$ implies that $P_{B|Y}$ is more noisy than $P_{C|Y}$ and, by the "data processing inequality" [61], B must have less mutual information about Y than C:

$$B \prec_Y C \implies I(B;Y) \leq I(C;Y). \quad (24)$$

Intuition suggests that when $B \prec_Y C$, the information that B provides about Y is contained in the information that C provides about Y. This intuition is formalized within a

decision-theoretic framework using the so-called Blackwell's Theorem [28,45,60]. To introduce this theorem, imagine a scenario in which Y represents the state of the environment. Imagine also that there is an agent who acquires information about the environment via the conditional distribution $P_{B|Y}(b|y)$, and then uses outcome $B = b$ to select actions $a \in \mathcal{A}$ according to some "decision rule" given by the channel $\kappa_{A|B}$. Finally, the agent gains utility according to some utility function $u(a, y)$, which depends on the agent's action a and the environment's state y. The maximum expected utility achievable by any decision rule is given by

$$V_Y^{\max}(B, u) := \max_{\kappa_{A|B}} \sum_{y,b,a} P_Y(y) P_{B|Y}(b|y) \kappa_{A|B}(a|b) u(a, y). \tag{25}$$

From an operational perspective, it is natural to say that B is less informative than C about Y if there is no utility function such that an agent with access to B can achieve higher expected utility than an agent with access to C. Blackwell's Theorem states that this is precisely the case if and only if $B \prec_Y C$ [28,45]:

$$B \prec_Y C \quad \text{iff} \quad V_Y^{\max}(B, u) \leq V_Y^{\max}(C, u) \quad \text{for all } u. \tag{26}$$

In some sense, this operational description of the relation $\prec_Y$ is deeper than the data processing inequality, Equation (24), which says that $B \prec_Y C$ is sufficient (but not necessary) for $I(B;Y) \leq I(C;Y)$. In fact, it can happen that $I(B;Y) \leq I(C;Y)$ even though $B \not\prec_Y C$ [26,60,62].

A connection between PID and Blackwell's theorem was first proposed in [13], which argued that the PID should be defined in an operational manner (see Section 5.3 for further discussion of [13]).

5.2. Blackwell Redundancy

We now define a measure of redundancy based on the Blackwell order. Specifically, we use our general definition of redundancy, Equation (7), while using the Blackwell order relative to Y as the "more informative" relation $\sqsubset$:

$$I_\cap^{\prec}(X_1; \ldots; X_n \to Y) := \sup_Q \; I(Q;Y) \quad \text{such that} \quad \forall i \; Q \prec_Y X_i. \tag{27}$$

We refer to this measure as *Blackwell redundancy*.

Given Blackwell's Theorem, $I_\cap^{\prec}$ has a simple operational interpretation. Imagine two agents, Alice and Bob, who can acquire information about Y via different random variables, and then use this information to maximize their expected utility. Suppose that Alice has access to one of the sources X_i. Then, the Blackwell redundancy $I_\cap^{\prec}$ is the maximum information that Bob can have about Y without being able to do better than Alice on any utility function, regardless of which source Alice has access to.

Blackwell redundancy can also be used to define a measure of Blackwell unique information, $U^{\prec}(X_i \to Y | X_1; \ldots; X_n) := I(Y; X_i) - I_\cap^{\prec}(X_1; \ldots; X_n \to Y)$, via Equation (2). As we show in Appendix I, $U^{\prec}$ satisfies the following property, which we term the *Multivariate Blackwell property*.

Theorem 3. *$U^{\prec}(X_i \to Y | X_1; \ldots; X_n) = 0$ if and only if $X_i \prec_Y X_j$ for all $j \neq i$.*

Operationally, Theorem 3 means that source X_i has non-zero unique information iff there exists a utility function such that an agent with access to source X_i can achieve higher utility than an agent with access to any other source X_j.

Computing $I_\cap^{\prec}$ involves maximizing a convex function subject to a set of linear constraints. These constraints define a feasible set which is a convex polytope, and the maximum must lie on one of the vertices of this polytope [63]. In Appendix C, we show how to solve this optimization problem. In particular, we use a computational geometry package to enumerate the vertices of the feasible set, and then choose the best vertex (code is available at [64]). In that appendix, we also prove that an optimal solution to Equation (27) can

always be achieved by Q with cardinality $|\mathcal{Q}| = (\sum_i |\mathcal{X}_i|) - n + 1$. Note that the supremum in Equation (27) is always achieved. Note also that $I_\cap^\prec$ satisfies the redundancy axioms in Section 4.2.

As discussed above, solving the optimization problem in Equation (27) gives a (possibly non-unique) optimal random variable Q which specifies the content of the redundant information. As shown in Appendix C, solving Equation (27) also provides a set of channels $\kappa_{Q|X_i}$ for each source X_i, which identify the redundant information in each source.

Note that the Blackwell order satisfies assumptions I-III in Section 4.1, thus Blackwell redundancy satisfies the bounds derived in that section. Finally, note that like many other redundancy measures, Blackwell redundancy becomes equivalent to the measure $I_\cap^{\mathrm{MMI}}$ (as defined in Equation (19)) when applied to Gaussian random variables (for details, see Appendix E).

5.3. Blackwell Union Information

We now define a measure of union information using our general definition in Equation (8), while using the Blackwell order relative to Y as the "more informative" relation:

$$I_\cup^\prec(X_1; \ldots; X_n \to Y) := \inf_Q \; I(Q;Y) \quad \text{such that} \quad \forall i \; X_i \prec_Y Q. \tag{28}$$

We refer to this measure as *Blackwell union information*.

As for Blackwell redundancy, Blackwell union information can be understood in operational terms. Consider two agents, Alice and Bob, whose use information about Y to maximize their expected utility. Suppose that Alice has access to one of the sources X_i. Then, the Blackwell union information $I_\cup^\prec$ is the minimum information that Bob must have about Y in order to do better than Alice on any utility function, regardless of which source Alice has access to.

Blackwell union information can be used to define measures of synergy and excluded information via Equations (1) and (3). The resulting measure of excluded information $E^\prec(X_i \to Y | X_1; \ldots; X_n) := I_\cup^\prec(X_1; \ldots; X_n \to Y) - I(Y; X_i)$ satisfies the following property, which is the "dual" of the *Multivariate Blackwell property* considered in Theorem 3. (See Appendix I for the proof.)

Theorem 4. *$E^\prec(X_i \to Y | X_1; \ldots; X_n) = 0$ if and only if $X_j \prec_Y X_i$ for all $j \neq i$.*

Operationally, Theorem 4 means that there is excluded information for source X_i iff there exists a utility function such that an agent with access to one of the other sources X_j can achieve higher expected utility than an agent with access to X_i.

We discuss the problem of numerically solving the optimization problem in Equation (28) in the next subsection.

5.4. Relation to Prior Work

Our measure of Blackwell redundancy $I_\cap^\prec$ is new to the PID literature. The most similar existing redundancy measure is $I_\cap^{\mathrm{GH}}$ [18], which is discussed above in Section 4.4. $I_\cap^{\mathrm{GH}}$ is a special case of Equation (7), once the "more informative" relation $B \sqsubset C$ is defined in terms of conditional independence $B - C - Y$. Note that conditional independence is stronger than the Blackwell order: given the definition of $\prec_Y$ in Equation (23), it is clear that $B - C - Y$ implies $B \prec_Y C$ (the channel $\kappa_{B|C}$ can be taken to be $P_{B|C}$), but not vice versa. As discussed in Section 4.1, stronger ordering relations give smaller values of redundancy, so in general $I_\cap^{\mathrm{GH}} \leq I_\cap^\prec$. Note also that $B \prec_Y C$ depends only on the pairwise marginals P_{BY} and P_{CY}, while conditional independence $B - C - Y$ depends on the joint distribution P_{BCY}. As we discuss in Appendix F, the conditional independence order can be interpreted in decision-theoretic terms, which suggests an operational interpretation for $I_\cap^{\mathrm{GH}}$.

Interestingly, Blackwell union information $I_\cup^\prec$ is equivalent to two measures that have been previously proposed in the PID literature, although they were formulated in a different way. Bertschinger et al. [13] considered the following measure of bivariate redundancy:

$$I_\cap^{\text{BROJA}}(X_1; X_2 \to Y) := I(Y; X_1) + I(Y; X_2) - I_\cup^{\text{BROJA}}(X_1; X_2 \to Y), \tag{29}$$

where $I_\cup^{\text{BROJA}}$ is defined via the optimization problem

$$I_\cup^{\text{BROJA}}(X_1; X_2 \to Y) = \min_{\tilde{X}_1, \tilde{X}_2} I(Y; \tilde{X}_1, \tilde{X}_2) \quad \text{such that} \quad P_{\tilde{X}_1 Y} = P_{X_1 Y}, P_{\tilde{X}_2 Y} = P_{X_2 Y}, \tag{30}$$

and reflects the minimal mutual information that two random variables can have about Y, given that their pairwise marginals with Y are fixed to be $P_{X_1 Y}$ and $P_{X_2 Y}$. Note that Ref. [13] did not refer to $I_\cup^{\text{BROJA}}$ as a measure of union information (we use our notation in writing it as $I_\cup^{\text{BROJA}}$). Instead, these measures were derived from an operational motivation, with the goal of deriving a unique information measure that obeys the so-called *Blackwell property*: $I(Y; X_1) - I_\cap^{\text{BROJA}}(X_1; X_2 \to Y) = 0$ if $X_1 \prec_Y X_2$ (see Theorems 3 and 4 above).

Starting from a different motivation, Griffith and Koch [17] proposed a multivariate version of $I_\cup^{\text{BROJA}}$,

$$I_\cup^{\text{BROJA}}(X_1; \ldots; X_n \to Y) = \min_{\tilde{X}_i, \ldots, \tilde{X}_n} I(Y; \tilde{X}_1, \ldots, \tilde{X}_n) \quad \text{such that} \quad \forall i \; P_{\tilde{X}_i Y} = P_{X_i Y}. \tag{31}$$

The goal of Ref. [17] was to derive a measure of multivariate synergy from a measure of union information, as in Equation (1). In that paper, $I_\cup^{\text{BROJA}}$ was explicitly defined as a measure of union information. To our knowledge, Ref. [17] was the first (and perhaps only) paper to propose a measure of union information that was not derived from redundancy via the inclusion-exclusion principle.

While $I_\cup^{\text{BROJA}}(X_1; \ldots; X_n \to Y)$ and $I_\cup^\prec(X_1; \ldots; X_n \to Y)$ are stated as different optimization problems, we prove in Appendix G that these optimization problems are equivalent, in that they will always achieve the same optimum value. Interestingly, since $I_\cup^{\text{BROJA}}$ and $I_\cup^\prec$ are equivalent, our measure of Blackwell redundancy $I_\cap^\prec$ appears as the natural dual to $I_\cup^{\text{BROJA}}$. Another implication of this equivalence is that Blackwell union information $I_\cup^\prec$ can be quantified by solving the optimization problem in Equation (31), rather than Equation (28). This is advantageous, because Equation (31) involves the minimization of a convex function over a convex polytope, which can be solved using standard convex optimization techniques [65].

In Ref. [13], the redundancy measure $I_\cap^{\text{BROJA}}$ in Equation (29) was only defined for the bivariate case. Since then, it has been unclear how to extend this redundancy measure to more than two sources. However, by comparing Equation (29) and Equation (14), we see the root of the problem: $I_\cap^{\text{BROJA}}$ is derived by applying the inclusion-exclusion principle to a measure of union information, $I_\cup^{\text{BROJA}}$. It cannot be extended to more than two sources because the inclusion-exclusion principle generally leads to counterintuitive results for more than 2 sources, as shown in Lemma 1. Note also that what Ref. [13] called the unique information in X_1, $I_\cup^{\text{BROJA}}(X_1; X_2 \to Y) - I(Y; X_2)$, in our framework would be considered a measure of the excluded information for X_2.

At the same time, the union information measure $I_\cup^{\text{BROJA}}$, and the corresponding synergy from Equation (1), does not use the inclusion-exclusion principle. Therefore, it can be easily extended to any number of sources [17].

5.5. Continuity of Blackwell Redundancy and Union Information

It is often desired that information-theoretic measures are continuous, meaning that small changes in underlying probability distributions lead to small changes in the resulting measures. In this section, we consider the continuity of our proposed measures, $I_\cap^\prec$ and $I_\cup^\prec$.

We first consider Blackwell redundancy $I_\cap^\prec$. It turns out that this measure is not always continuous in the joint probability $P_{X_1 \ldots X_n Y}$ (a discontinuous example is provided in Section 5.6). However, the discontinuity of $I_\cap^\prec$ is not necessarily pathological, and we can derive an interpretable geometric condition that guarantees that $I_\cap^\prec$ is continuous.

Consider the conditional distribution of the target Y given some source X_i, $P_{Y|X_i}$. Let rank $P_{Y|X_i}$ indicate its *rank*, meaning the dimension of the space spanned by the vectors $\{P_{Y|X_i=x_i}\}_{x_i \in \mathcal{X}_i}$. The rank of $P_{Y|X_i}$ quantifies the number of independent directions that the target distribution P_Y can be moved by manipulating the source distribution P_{X_i}, and it cannot be larger than $|\mathcal{Y}|$. The next theorem shows that $I_\cap^\prec$ is locally continuous, as long as $n-1$ or more of the source conditional distributions have this maximal rank.

Theorem 5. *As a function of the joint distribution $P_{X_1,\ldots,X_n,Y}$, $I_\cap^\prec$ is locally continuous whenever $n-1$ or more of the conditional distributions $P_{Y|X_i}$ have* rank $P_{Y|X_i} = |\mathcal{Y}|$.

In proving this result, we also show that $I_\cap^\prec$ is continuous almost everywhere (see proof in Appendix D). Finally, in that appendix we also use Theorem 5 to show that $I_\cap^\prec$ is continuous everywhere if Y is a binary random variable.

We illustrate the meaning of Theorem 5 visually in Figure 2. We show two situations, both of which involve two sources X_1 and X_2 and a target Y with cardinality $|\mathcal{Y}| = 3$. In one situation, both pairwise conditional distributions have rank equal to $|\mathcal{Y}|$, so $I_\cap^\prec$ is locally continuous. In the other situation, both pairwise conditional distributions are rank deficient (e.g., this might happen because X_1 and X_2 have cardinality $|\mathcal{X}_1| = |\mathcal{X}_2| = 2$), so $I_\cap^\prec$ is not guaranteed to be continuous. From the figure it is easy to see how the discontinuity may arise. Given the definition of the Blackwell order and $I_\cap^\prec$, for any random variable Q in the feasible set of Equation (27), the conditional distributions $P_{Y|Q=q}$ must fall within the intersection of the distributions spanned by $P_{Y|X_1}$ and $P_{Y|X_2}$ (the intersection of the red and green shaded regions in Figure 2). On the right, the size of this intersection can discontinuously jump from a line (when $P_{Y|X_1}$ and $P_{Y|X_2}$ are perfectly aligned) to a point (when $P_{Y|X_1}$ and $P_{Y|X_2}$ are not perfectly aligned). Thus, the discontinuity of $I_\cap^\prec$ arises from a geometric phenomenon, which is related to the discontinuity of the intersection of low-dimensional vector subspaces.

We briefly comment on the continuity of $I_\cup^\prec$. As we described above, this measure turns out to be equivalent to $I_\cup^{\text{BROJA}}$. The continuity of $I_\cup^{\text{BROJA}}$ in the bivariate case was proven in Theorem 35 of Ref. [66]. We believe that the continuity of $I_\cup^{\text{BROJA}}$ for an arbitrary number of sources can be shown using similar methods, although we leave this for future work.

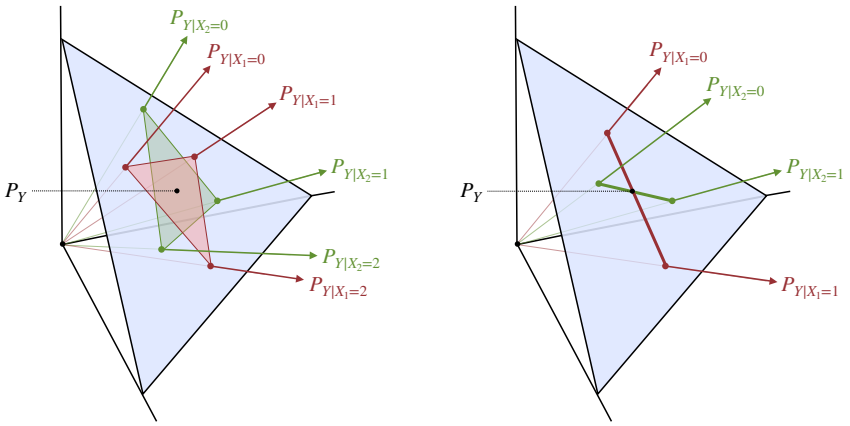

Figure 2. Illustration of Theorem 5, which provides a sufficient condition for the local continuity of $I_\cap^\prec$. Consider two scenarios, both of which involves two sources X_1 and X_2 and a target Y with cardinality $|\mathcal{Y}| = 3$. The blue areas indicate the simplex of probability distributions over $\mathcal{Y}$, with the marginal P_Y and the pairwise conditionals $P_{Y|X_i=x_i}$ marked. On the left, both sources have rank $P_{Y|X_i} = 3 = |\mathcal{Y}|$, so $I_\cap^\prec$ is locally continuous. On the right, both sources have rank $P_{Y|X_i} = 2 < |\mathcal{Y}|$, so $I_\cap^\prec$ is not necessarily locally continuous. Note that $I_\cap^\prec$ is also continuous if only source has rank $P_{Y|X_i} = 3$.

5.6. Behavior on the COPY Gate

As mentioned in Section 3, the "COPY gate" example is often used to test the behavior of different redundancy measures. The COPY gate has two sources, X_1 and X_2, and a target $Y = (X_1, X_2)$ which is a copy of the joint outcome. It is expected that redundancy should vanish if X_1 and X_2 are statistically independent, as formalized by the *Independent identity* property in Equation (4).

Blackwell redundancy $I_\cap^\prec$ satisfies the *Independent identity*. In fact, we prove a more general result, which shows that $I_\cap^\prec(X_1, X_2 \to (X_1, X_2))$ is equal to an information-theoretic measure called *Gács-Körner common information* $C(X \wedge Y)$ [16,47,67]. $C(X \wedge Y)$ quantifies the amount of information that can be deterministically extracted from both random variables X or Y, and it is closely related to the "deterministic function" order $\lhd$ defined in Equation (16). Formally, it can be written as

$$C(X \wedge Y) = \sup_Q H(Q) \qquad \text{such that} \qquad Q \lhd X, Q \lhd Y, \tag{32}$$

where H is Shannon entropy. In Appendix I, we prove the following result.

Theorem 6. $I_\cap^\prec(X_1, X_2 \to (X_1, X_2)) = C(X_1 \wedge X_2)$.

Note that $0 \leq C(X_1 \wedge X_2) \leq I(X_1; X_2)$ [47], so $I_\cap^\prec$ satisfies the *Independent identity* property. At the same time, $C(X_1 \wedge X_2)$ can be strictly less than $I(X_1; X_2)$. For example, if $P_{X_1X_2}$ has full support, then $I(X_1; X_2)$ can be arbitrarily large while $C(X_1 \wedge X_2) = 0$ (see proof of Theorem 6). This means that $I_\cap^\prec$ violates a previously proposed property, sometimes called the *Identity property*, that suggests that redundancy should satisfy $I_\cap(X_1; X_2 \to (X_1, X_2)) = I(X_1; X_2)$. However, the validity of the *Identity property* is not clear, and several papers have argued against it [15,39].

The value of $C(X_1 \wedge X_2)$ depends on the precise pattern of zeros in the joint distribution $P_{X_1X_2}$ and is therefore not continuous. For instance, for the bivariate COPY gate, redundancy can change discontinuously as one goes from the situation where $X_1 = X_2$ (so that all information is redundant, $I_\cap^\prec = I(X_1; X_2)$) to one where X_1 and X_2 are almost, but not entirely, identical. This discontinuity can be understood in terms of Theorem 5 and Figure 2: in the COPY gate, the cardinality of the target variable $|\mathcal{Y}| = |\mathcal{X}_1| \times |\mathcal{X}_2|$ is larger than the cardinality of the individual sources. In other words, when the sources X_1 and X_2 are not perfectly correlated, they provide information about different "subspaces" of the target (X_1, X_2), and so it is possible that very little (or none) of their information is redundant.

At the same time, the Blackwell property, Theorem 3, implies that

$$I_\cap^\prec(X_1, X_2 \to X_1) = I(X_1; X_2) = I_\cap^\prec(X_1, X_2 \to X_2) \tag{33}$$

In other words, the redundancy in X_1 and X_2, where either one of the individual sources is taken as the target, is given by the mutual information $I(X_1; X_2)$. This holds even though the redundancy in the COPY gate can be much lower than $I(X_1; X_2)$.

It is also interesting to consider how Blackwell union information, $I_\cup^\prec$, behaves on the COPY gate. Using techniques from [13], it can be shown that the union information is simply the joint entropy,

$$I_\cup^\prec(X_1; X_2 \to (X_1, X_2)) = H(X_1, X_2). \tag{34}$$

Since $H(X_1, X_2) = I(X_1, X_2; X_1, X_2)$, Equation (34) and Equation (1) together imply that the COPY gate has no synergy.

Note that we can use Theorem 6 and Equation (34) to illustrate that $I_\cap^\prec$ and $I_\cup^\prec$ violate the inclusion-exclusion principle, Equation (14). Using Equation (34) and a bit of rearranging, Equation (14) becomes equivalent to $I_\cap^\prec(X_1; X_2 \to (X_1, X_2)) \stackrel{?}{=} I(X_1; X_2)$, which is the *Identity property* mentioned above. $I_\cap^\prec$ violates this property, since redundancy for the COPY gate can be smaller than $I(X_1; X_2)$.

6. Examples and Comparisons to Previous Measures

In this section, we compare our proposed measure of Blackwell redundancy $I_\cap^\prec$ to existing redundancy measures. We focus on redundancy, rather than union information, because redundancy has seen much more development in the literature, and because Blackwell union information $I_\cup^\prec$ is equivalent to an existing measure (see Section 5.4).

6.1. Qualitative Comparison

In Table 1, we compare $I_\cap^\prec$ to six existing measures of multivariate redundancy:

- $I_\cap^{\mathrm{WB}}$, the redundancy measure first proposed by Williams and Beer [11].
- $I_\cap^{\mathrm{MMI}}$, the "minimum mutual information" [51], Equation (19) in Section 4.4.
- $I_\cap^\lhd$, proposed by Griffith et al. [16], Equation (17) in Section 4.4.
- $I_\cap^{\mathrm{GH}}$, proposed by Griffith and Ho [18], Equation (18) in Section 4.4.
- $I_\cap^{\mathrm{Ince}}$, proposed by Ince [20].
- $I_\cap^{\mathrm{FL}}$, proposed by Finn and Lizier [21].

We also compare $I_\cap^\prec$ to three existing measures of bivariate redundancy (i.e., for 2 sources):

- $I_\cap^{\mathrm{BROJA}}$, proposed by Bertschinger et al. [13], defined in Equation (29).
- $I_\cap^{\mathrm{Harder}}$, proposed by Harder et al. [19].
- $I_\cap^{\mathrm{dep}}$, proposed by James et al. [15].

For $I_\cap^\prec$ as well as the 9 existing measures, we consider the following properties, which are chosen to highlight differences between our approach and previous proposals:

1. Has it been defined for more than 2 sources
2. Does it obey the *Monotonicity* axiom from Section 4.2
3. Is it compatible with the inclusion-exclusion principle (IEP) for the bivariate case, such that union information as defined in Equation (14) obeys $I_\cup(X_1; X_2 \to Y) \leq I(X_1, X_2; Y)$
4. Does it obey the *Independent identity* property, Equation (4)
5. Does it obey the *Blackwell property* (possibly in its multivariate form, Theorem 3)

We also consider two additional properties, which require a bit of introduction.

The first property was suggested by Ref. [13], who argued that redundancy should only depend on the pairwise marginal distributions of each source with the target,

$$\text{If } p_{X_iY} = p_{\tilde{X}_i\tilde{Y}} \text{ for all } i, \text{ then } I_\cap(X_1; \ldots; X_n \to Y) = I_\cap(\tilde{X}_1; \ldots; \tilde{X}_n \to \tilde{Y}). \tag{35}$$

In Table 1, we term this property *Pairwise marginals*. We believe that the validity of Equation (35) is not universal, but may depend on the particular setting in which the PID is being used. However, redundancy redundancy measures that satisfy this property have one important advantage: they are well-defined not only when the sources are random variables $X_1, \ldots, X_n$, but also in the more general case when the sources are channels $\kappa_{X_1|Y}, \ldots, \kappa_{X_n|Y}$.

Table 1. Comparison of different redundancy measures. ? indicate properties that we could not easily establish.

	$I_\cap^\prec$	$I_\cap^{\mathrm{WB}}$	$I_\cap^{\mathrm{MMI}}$	$I_\cap^\lhd$	$I_\cap^{\mathrm{GH}}$	$I_\cap^{\mathrm{Ince}}$	$I_\cap^{\mathrm{FL}}$	$I_\cap^{\mathrm{BROJA}}$	$I_\cap^{\mathrm{Harder}}$	$I_\cap^{\mathrm{dep}}$
More than 2 sources	✓	✓	✓	✓	✓	✓	✓			
Monotonicity	✓	✓	✓	✓	✓			✓	✓	✓
IEP for bivariate case		✓	✓			?	?	✓	✓	✓
Independent identity	✓			✓	✓	✓		✓	✓	✓
Blackwell property	✓							✓	✓	
Pairwise marginals	✓	✓	✓	✓				✓	✓	
Target equality	✓	✓	✓		✓			✓	✓	

The second property has not been previously considered in the literature, although it appears to be highly intuitive. Observe that the target random variable Y contains all possible information about itself. Thus, it may be expected that adding the target to the set of sources should not decrease the redundancy:

$$I_\cap(X_1;\ldots;X_n;Y\to Y) = I_\cap(X_1;\ldots;X_n\to Y). \tag{36}$$

In Table 1, we term this property *Target equality*. Note that for redundancy measures which can be put in the form of Equation (7), *Target Equality* is satisfied if the order $\sqsubset$ obeys $X_i \sqsubset Y$ for all sources X_i. (Note also that *Target Equality* is unrelated to the previously proposed *Strong Symmetry* property; for instance, it is easy to show that the redundancy measures $I_\cap^{\mathrm{WB}}$ and $I_\cap^{\mathrm{MMI}}$ satisfy *Target Equality*, even though they violate *Strong Symmetry* [68].)

6.2. Quantitative Comparison

We now illustrate our proposed measure of redundancy $I_\cap^\prec$ on some simple examples, and compare its behavior to existing redundancy measures.

The values of $I_\cap^\prec$ were computed with our code, provided at [64]. The values of all other redundancy measures except $I_\cap^{\mathrm{GH}}$ were computed using the `dit` Python package [69]. To our knowledge, there have been no previous proposals for how to compute $I_\cap^{\mathrm{GH}}$. In fact, this measure involves maximizing a convex function subject to linear constraints, and can be computed using similar methods as $I_\cap^\prec$. We provide code for computing $I_\cap^{\mathrm{GH}}$ at [64].

We begin by considering some simple bivariate examples. In all cases, the sources X_1 and X_2 are binary and uniformly distributed. The results are shown in Table 2.

Table 2. Behavior of $I_\cap^\prec$ and other redundancy measures on bivariate examples.

Target	$I_\cap^\prec$	$I_\cap^{\mathrm{WB}}$	$I_\cap^{\mathrm{MMI}}$	$I_\cap^\wedge$	$I_\cap^{\mathrm{GH}}$	$I_\cap^{\mathrm{Ince}}$	$I_\cap^{\mathrm{FL}}$	$I_\cap^{\mathrm{BROJA}}$ $I_\cap^{\mathrm{Harder}}$	$I_\cap^{\mathrm{dep}}$
$Y = X_1$ AND X_2	0.311	0.311	0.311	0	0.123	0.104	0.561	0.311	0.082
$Y = X_1 + X_2$	0.5	0.5	0.5	0	0	0	0.5	0.5	0.189
$Y = X_1$	$I(X_1;X_2)$	$I(X_1;X_2)$	$I(X_1;X_2)$	$C(X_1\wedge X_2)$	$I(X_1;X_2)$	*	1	$I(X_1;X_2)$	$I(X_1;X_2)$
$Y = (X_1, X_2)$	$C(X_1\wedge X_2)$	1	1	$C(X_1\wedge X_2)$	$C(X_1\wedge X_2)$	*	1	$I(X_1;X_2)$	$I(X_1;X_2)$

1. The AND gate, $Y = X_1$ AND X_2, with X_1 and X_2 independent. (It is incorrectly stated in Refs. [18,49] that $I_\cap^{\mathrm{GH}}$ vanishes here; actually $I_\cap^{\mathrm{GH}}(X_1; X_2 \to X_1 \text{ AND } X_2) \approx 0.123$, which corresponds to the maximum achieved in Equation (18) by $Q = X_1$ OR X_2.)
2. The SUM gate: $Y = X_1 + X_2$, with X_1 and X_2 independent.
3. The UNQ gate: $Y = X_1$. Here $I_\cap^{\mathrm{Ince}}$ (marked with $*$) gave values that increased with the amount of correlation between X_1 and X_2 but were typically larger than $I(X_1;X_2)$.
4. The COPY gate: $Y = (X_1, X_2)$. Here, our redundancy measure is equal to the Gács-Körner common information between X and Y, as discussed in Section 5.6. The same holds for the redundancy measures $I_\cap^{\mathrm{GH}}$ and $I_\cap^\lhd$, which can be shown using a slight modification of the proof of Theorem 6. For this gate, $I_\cap^{\mathrm{Ince}}$ (marked with $*$) gave the same values as for the UNQ gate, which increased with the amount of correlation between X_1 and X_2 but were typically larger than $I(X_1;X_2)$.

We also analyze several examples with three sources, with the results shown in Table 3. We considered those previously proposed measures which can be applied to more than two sources (we do not show $I_\cap^{\mathrm{GH}}$, as our implementation was too slow for these examples).

1. Three-way AND gate: $Y = X_1$ AND X_2 AND X_3, where the sources are binary and uniformly and independently distributed.
2. Three-way SUM gate: $Y = X_1 + X_2 + X_3$, where the sources are binary and uniformly and independently distributed.
3. "Overlap" gate: we defined four independent uniformly distributed binary random variables, A, B, C, D. These were grouped into three sources X_1, X_2, X_3 as $X_1 = (A, B)$,

$X_2 = (A, C)$, $X_3 = (A, D)$. The target was the joint outcome of all three sources, $Y = (X_1, X_2, X_3) = ((A, B), (A, C), (A, D))$. Note that the three sources overlap on a single random variable A, which suggests that the redundancy should be 1 bit.

Table 3. Behavior of $I_\cap^\prec$ and other redundancy measures on three sources.

Target	$I_\cap^\prec$	$I_\cap^{\mathbf{WB}}$	$I_\cap^{\mathbf{MMI}}$	$I_\cap^\wedge$	$I_\cap^{\mathbf{Ince}}$	$I_\cap^{\mathrm{FL}}$
$Y = X_1$ AND X_2 AND X_3	0.138	0.138	0.138	0	0.024	0.294
$Y = X_1 + X_2 + X_3$	0.311	0.311	0.311	0	0	0.561
$Y = ((A, B), (A, C), (A, D))$	1	2	2	1	1	2

7. Discussion and Future Work

In this paper, we proposed a new general framework for defining the partial information decomposition (PID). Our framework was motivated in several ways, including a formal analogy with intersections and unions in set theory as well as an axiomatic derivation.

We also used our general framework to propose concrete measures of redundancy and union information, which have clear operational interpretations based on Blackwell's theorem. Other PID measures, such as synergy and unique information, can be computed from our measures of redundancy and union information via simple expressions.

One unusual aspect of our framework is that it provides separate measures of redundancy and union information. As we discuss above, most prior work on the PID assumed that redundancy and union information are related to each other via the so-called "inclusion-exclusion" principle. We argue that the inclusion-exclusion principle should not be expected to hold in the context of the PID, and in fact that it leads to counterintuitive behavior once 3 or more sources are present. This suggests that different information decompositions should be derived for redundancy vs. union information. This idea is related to a recent proposal in the literature, which argues that two different PIDs are needed, one based on redundancy and one based on synergy [41]. An interesting direction for future work is to relate our framework with the dual decompositions proposed in [41].

From a practical standpoint, an important direction for future work is to develop better schemes for computing our redundancy measure. This measure is defined in terms of a convex maximization problem, which in principle can be NP-hard (a similar convex maximization problem was proven to be NP-hard in [70]). Our current implementation, which enumerates the vertices of the feasible set, works well for relatively small state spaces, but we do not expect it to scale to situations with many sources, or where the sources have large cardinalities. However, the problem of convex maximization with linear constraints is a very active area of optimization research, with many proposed algorithms [63,71,72]. Investigating these algorithms, as well as various approximation schemes such as relaxations and variational bounds, is of interest.

Finally, we showed how our framework can be used to define measures of redundancy and union information in situations that go beyond the standard setting of the PID (e.g., when the probability distribution of the target is not specified). Our framework can even be applied in domains beyond Shannon information theory, such as algorithmic information theory and quantum information theory. Future work may exploit this flexibility to explore various new applications of the PID.

Funding: This research received no external funding.

Data Availability Statement: Not applicable.

Acknowledgments: We thank Paul Williams, Alexander Gates, Nihat Ay, Bernat Corominas-Murtra, Pradeep Banerjee, and especially Johannes Rauh for helpful discussions and suggestions. We also thank the Santa Fe Institute for helping to support this research.

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

Appendix A. PID Axioms

In developing the PID framework, Williams and Beer [11,12] proposed that any measure of redundancy should obey a set of axioms. In slightly modified form, these axioms can be written as follows:

- *Symmetry*: $I_\cap(X_1;\ldots;X_n \to Y)$ is invariant to the permutation of $X_1,\ldots,X_n$.
- *Self-redundancy*: $I_\cap(X_1 \to Y) = I(Y;X_1)$.
- *Monotonicity*: $I_\cap(X_1;\ldots;X_n \to Y) \leq I_\cap(X_1;\ldots;X_{n-1} \to Y)$.
- *Deterministic equality*: $I_\cap(X_1;\ldots;X_n \to Y) = I_\cap(X_1;\ldots;X_{n-1} \to Y)$ if $X_i = f(X_n)$ for some $i < n$ and deterministic function f.

These axioms are based on intuitions regarding the behavior of intersection in set theory [12]. The *Symmetry* axiom is self-explanatory. *Self-redundancy* states that if only a single-source is present, all of its information is redundant. *Monotonicity* states that redundancy should not increase when an additional source is considered (consider that the size of set intersection can only decrease as more sets are considered). *Deterministic equality* states that redundancy should remain the same when an additional source X_n is added that contains all (or more) of the same information that is already contained in an existing source X_i (which is formalized as the condition $X_i = f(X_n)$).

Union information was considered the original PID proposal [12,73], as well as a more recent paper [17]. Ref. [17] proposed that any measure of union information should satisfy the following set of natural axioms, stated here in slightly modified form:

- *Symmetry*: $I_\cup(X_1;\ldots;X_n \to Y)$ is invariant to the permutation of $X_1,\ldots,X_n$.
- *Self-union*: $I_\cup(X_1 \to Y) = I(Y;X_1)$.
- *Monotonicity*: $I_\cup(X_1;\ldots;X_n \to Y) \geq I_\cup(X_1;\ldots;X_{n-1} \to Y)$.
- *Deterministic equality*: $I_\cup(X_1;\ldots;X_n \to Y) = I_\cup(X_1;\ldots;X_{n-1} \to Y)$ if $X_n = f(X_i)$ for some $i < n$ and deterministic function f.

These axioms are based on intuitions concerning the behavior of the union operator in set theory, and are the natural "duals" of the redundancy axioms mentioned above.

Appendix B. Uniqueness Proofs

Proof of Theorem 1. Assume there is a redundancy measure $I'_\cap$ that obeys the five axioms stated in the theorem. We will show that $I'_\cap = I_\cap$, as defined in Equation (7).

Given Equation (7) and the definition of the supremum, for any $\epsilon > 0$ there exists a random variable Q such that $Q \sqsubset X_i$ for $i \in \{1,\ldots,n\}$ and

$$I(Q;Y) \geq I_\cap(X_1;\ldots;X_n \to Y) - \epsilon, \tag{A1}$$

By *Order equality*, $I'_\cap(Q;X_1;\ldots;X_k \to Y) = I'_\cap(Q;X_1;\ldots;X_{k-1} \to Y)$. Induction gives

$$I'_\cap(Q;X_1;\ldots;X_n \to Y) = I'_\cap(Q \to Y) = I(Q;Y) \geq I_\cap(X_1;\ldots;X_n \to Y) - \epsilon$$

where we used *Self-redundancy* and Equation (A1). We also have $I'_\cap(Q;X_1;\ldots;X_n \to Y) \leq I'_\cap(X_1;\ldots;X_n \to Y)$ by *Symmetry* and *Monotonicity*. Combining gives

$$I_\cap(X_1;\ldots;X_n \to Y) - \epsilon \leq I'_\cap(X_1;\ldots;X_n \to Y).$$

We now show that $I_\cap$ is the largest measure that satisfies *Existence*. Let Q be a random variable that obeys $Q \sqsubset X_i$ for all $i \in \{1,\ldots,n\}$ and $I'_\cap(X_1;\ldots;X_n \to Y) = I(Y;Q)$. Since Q falls within the feasible set of the optimization problem in Equation (7),

$$I'_\cap(X_1;\ldots;X_n \to Y) = I(Q;Y) \leq I_\cap(X_1;\ldots;X_n \to Y).$$

Combining gives

$$I'_\cap(X_1;\ldots;X_n \to Y) - \epsilon \leq I_\cap(X_1;\ldots;X_n \to Y) - \epsilon \leq I'_\cap(X_1;\ldots;X_n \to Y).$$

Since this holds for all $\epsilon > 0$, taking the limit $\epsilon \to 0$ gives $I'_\cap = I_\cap$. □

Proof of Theorem 2. Assume there is a union information measure $I'_\cup$ that obeys the five axioms stated in the theorem. We will show that $I'_\cup = I_\cup$, as defined in Equation (8).

Given Equation (8) and the definition of the infinitum, for any $\epsilon > 0$ there exists a random variable Q such that $Q \sqsubset X_i$ for $i \in \{1, \ldots, n\}$ and

$$I(Q;Y) \leq I_\cup(X_1; \ldots; X_n \to Y) + \epsilon, \tag{A2}$$

By *Order equality*, $I'_\cup(Q; X_1; \ldots; X_k \to Y) = I'_\cup(Q; X_1; \ldots; X_{k-1} \to Y)$. Induction gives

$$I'_\cup(Q; X_1; \ldots; X_n \to Y) = I'_\cup(Q \to Y) = I(Q;Y) \leq I_\cup(X_1; \ldots; X_n \to Y) + \epsilon$$

where we used *Self-union* and Equation (A2). We also have $I'_\cup(Q; X_1; \ldots; X_n \to Y) \geq I'_\cup(X_1; \ldots; X_n \to Y)$ by *Symmetry* and *Monotonicity*. Combining gives

$$I_\cup(X_1; \ldots; X_n \to Y) + \epsilon \geq I'_\cup(X_1; \ldots; X_n \to Y).$$

We now show that $I_\cap$ is the smallest measure that satisfies *Existence*. Let Q be a random variable that obeys $X_i \sqsubset Q$ for all $i \in \{1, \ldots, n\}$ and $I'_\cup(X_1; \ldots; X_n \to Y) = I(Y;Q)$. Since Q falls within the feasible set of the optimization problem in Equation (8),

$$I'_\cup(X_1; \ldots; X_n \to Y) = I(Q;Y) \geq I_\cup(X_1; \ldots; X_n \to Y).$$

Combining gives

$$I'_\cup(X_1; \ldots; X_n \to Y) \leq I_\cup(X_1; \ldots; X_n \to Y) + \epsilon \leq I'_\cap(X_1; \ldots; X_n \to Y) + \epsilon.$$

Since this holds for all $\epsilon > 0$, taking the limit $\epsilon \to 0$ gives $I'_\cup = I_\cup$. □

Appendix C. Computing $I_\cap^\prec$

Here we consider the optimization problem that defines our proposed measure of redundancy, Equation (27). We first prove a bound on the required cardinality of Q.

Theorem A1. *For optimizing Equation* (27)*, it suffices to consider Q with cardinality $|\mathcal{Q}| = (\sum_i |\mathcal{X}_i|) - n + 1$.*

Proof. Consider any random variable Q with outcome set $\mathcal{Q}$ which satisfies $Q \prec_Y X_i$ for all i. We show that whenever Q has full support on $|\mathcal{Q}| > (\sum_i |\mathcal{X}_i|) - n + 1$ outcomes, there is another random variable $\tilde{Q}$ which achieves $I(\tilde{Q};Y) \geq I(Q;Y)$, while satisfying $\tilde{Q} \prec_Y X_i$ for all i and having support on at most $(\sum_i |\mathcal{X}_i|) - n + 1$ outcomes.

To begin, let Ω indicate the set of random variables over outcomes $\mathcal{Q}$, such that all $\tilde{Q} \in \Omega$ satisfy:

$$P_{Y|\tilde{Q}}(y|q) = P_{Y|Q}(y|q) \qquad \text{for all } y, q \in \{q \in \mathcal{Q} : P_{\tilde{Q}}(q) > 0\} \tag{A3}$$

$$\sum_q P_{\tilde{Q}}(q) P_{X_i|Q}(x_i|q) = P_{X_i}(x_i) \qquad \text{for all } i, x_i. \tag{A4}$$

Since $Q \prec_Y X_i$, by Equation (23) there exist channels $\kappa_{Q|X_i}(q|x_i)$ that satisfy $P_{Q|Y}(q|y) = \sum_{x_i} \kappa_{Q|X_i}(q|x_i) P_{X_i|Y}(x_i|y)$. Now write the conditional distribution over $\tilde{Q}$ and Y as

$$\begin{aligned} P_{\tilde{Q}Y}(q,y) = \frac{P_{\tilde{Q}}(q)}{P_Q(q)} P_{QY}(q,y) &= \sum_{x_i} \frac{P_{\tilde{Q}}(q)}{P_Q(q)} \kappa_{Q|X_i}(q|x_i) P_{X_i|Y}(x_i|y) P_Y(y) \\ &= \sum_{x_i} \kappa'_{\tilde{Q}|X_i}(q|x_i) P_{X_i|Y}(x_i|y) P_Y(y), \end{aligned} \tag{A5}$$

where we used Equation (A3) and defined the channel $\kappa'_{\tilde{Q}|X_i}$ as

$$\kappa'_{\tilde{Q}|X_i} = \frac{P_{\tilde{Q}}(q)}{P_{X_i}(x_i)}\left[\frac{P_{X_i}(x_i)}{P_Q(q)}\kappa_{Q|X_i}(q|x_i)\right],$$

(Note this is a kind of double Bayesian inverse, given Equation (A4)). Equation (A5) implies that $\tilde{Q} \prec_Y X_i$ for all i.

We now show that there is $\tilde{Q} \in \Omega$ that achieves $I(Q;Y) \leq I(\tilde{Q};Y)$ and has support on at most $(\sum_i |\mathcal{X}_i|) - n + 1$ outcomes in $\mathcal{Q}$. Write the mutual information between any $\tilde{Q} \in \Omega$ and Y as

$$I(\tilde{Q};Y) = \sum_q P_{\tilde{Q}}(q) D_{\mathrm{KL}}(P_{Y|\tilde{Q}=q} \| P_Y) = \sum_q P_{\tilde{Q}}(q) D_{\mathrm{KL}}(P_{Y|Q=q} \| P_Y), \tag{A6}$$

where D_{KL} is the Kullback-Leibler divergence. We consider the maximum of this mutual information across Ω, $I^* = \max_{\tilde{Q}\in\Omega} I(\tilde{Q};Y)$. Using Equations (A4) and (A6), this maximum can be written as

$$I^* = \max_{\omega\in\Delta} \sum_q \omega(q) D(P_{Y|Q=q} \| P_Y) \quad \text{such that} \quad \forall i, x_i : \sum_q \omega(q) P_{X_i|Q}(x_i|q) = P_{X_i}(x_i),$$

where Δ is the set of all distributions over $\mathcal{Q}$. By conservation of probability, $\sum_{x_i} P_{X_i}(x_i) = 1$, so we can eliminate a constraint for one of the outcomes x_i of each source i. Thus, I^* is the maximum of a linear function over Δ, subject to $\sum_i(|\mathcal{X}_i| - 1) = (\sum_i |\mathcal{X}_i|) - n$ hyperplane constraints.

The feasible set is compact, and the maximum will be achieved at one of the extreme points of the feasible set. By Dubin's Theorem [74], any extreme point of this feasible set can be expressed as a convex combination of at most $(\sum_i |\mathcal{X}_i|) - n + 1$ extreme points of Δ. In other words, the maximum in Equation (27) is achieved by a random variable $\tilde{Q}$ with support on at most $(\sum_i |\mathcal{X}_i|) - n + 1$ values of $\mathcal{Q}$. This random variable satisfies

$$I(\tilde{Q};Y) = I^* \geq I(Q;Y),$$

where the last inequality comes from the fact that Q is an element of Ω. □

We now return to the optimization problem in Equation (27). Given Theorem A1 and the definition of the Blackwell order in Equation (23), it can be rewritten as

$$\begin{aligned} I^{\prec}_{\cap}(X_1;\ldots;X_n \to Y) = \max_{\kappa_{Q|Y}, \kappa_{Q|X_1}, \ldots, \kappa_{Q|X_n}} & I_\kappa(Q,Y) \\ \text{such that } \forall i, y, x_i : & \sum_{x_i} \kappa_{Q|X_i}(q|x_i) P_{X_i|Y}(x_i|y) = \kappa_{Q|Y}(q|y). \end{aligned} \tag{A7}$$

where the optimization is over channels with $\mathcal{Q}$ of cardinality $(\sum_i |\mathcal{X}_i|) - n + 1$. The notation $I_\kappa(Q;Y)$ indicates the mutual information that arises from the marginal distribution P_Y and the conditional distribution $\kappa_{Q|Y}$,

$$I_\kappa(Q;Y) = \sum_y P_Y(y) \kappa_{Q|Y}(q|y) \ln \frac{\kappa_{Q|Y}(q|y)}{\sum_{y'} \kappa_{Q|Y}(q|y') P_Y(y')}$$

Equation (A7) involves maximizing a convex function over the convex polytope defined by the following system of linear (in)equalities:

$$\Lambda = \Big\{(\kappa_{Q|Y}, \kappa_{Q|X_1,\ldots,}\kappa_{Q|X_n}) :$$

$$\forall i, x_i, q \quad \kappa_{Q|X_i}(q|x_i) \geq 0, \tag{A8}$$

$$\forall q, y \quad \kappa_{Q|Y}(q|y) \geq 0, \tag{A9}$$

$$\forall y \quad \sum_q \kappa_{Q|Y}(q|y) = 1, \tag{A10}$$

$$\forall i, x_i \quad \sum_q \kappa_{Q|X_i}(q|x_i) = 1, \tag{A11}$$

$$\forall i, y, q \in \mathcal{Q} \setminus \{0\} \quad \left[\sum_{x_i} \kappa_{Q|X_i}(q|x_i) P_{X_iY}(x_i, y)\right] - \kappa_{Q|Y}(q|y) P_Y(y) = 0\Big\}, \tag{A12}$$

We do not place a constraint on $q = 0$ in Equation (A12) because that would be redundant with the constraints Equations (A10) and (A11). Also, note that we replaced the sup in Equation (27) with max in Equation (A7), which is justified since we are optimizing over a finite dimensional, closed, and bounded region (thus the supremum is always achieved).

The maximum of a convex function over a convex polytope is found at one of the vertices of the polytope. To find the solution to Equation (A7), we use a computational geometry package to enumerate the vertices of Λ. We evaluate $I_\kappa(Y;Q)$ at each vertex, and pick the maximum value. This procedure also finds optimal conditional distributions $\kappa_{Q|Y}, \kappa_{Q|X_1,\ldots,}\kappa_{Q|X_n}$. Code is available at [64].

Appendix D. Continuity of $I_\cap^\prec$

To prove the continuity of $I_\cap^\prec$, we begin by considering the feasible set of the optimization problem in Equation (A7), as specified by the system (in)equalities in Equations (A8) to (A12). For convenience, write this system of (in)equalities in matrix notation,

$$\Lambda = \Big\{\vec{\kappa} \in \mathbb{R}^{|\mathcal{Q}||\mathcal{Y}| + \sum_i |\mathcal{Q}||\mathcal{X}_i|} : \kappa \geq 0, A\kappa = a\Big\}, \tag{A13}$$

where $\vec{\kappa}$ is a vector representation of $(\kappa_{Q|Y}, \kappa_{Q|X_1,\ldots,}\kappa_{Q|X_n})$, the matrix A encodes the left-hand side of Equations (A10) to (A12), and the vector a is filled with 1s and 0s, as appropriate.

We first prove the following lemma.

Lemma A1. *The matrix A defined in Equation (A13) is full rank if $n-1$ or more of the pairwise conditional distributions have* rank $P_{Y|X_i} = |\mathcal{Y}|$.

Proof. Without loss of generality, assume that P_Y has full support (otherwise none of the pairwise marginals P_{X_iY} can achieve rank $|\mathcal{Y}|$). Write A in block matrix form as $A = \begin{bmatrix} B \\ C \end{bmatrix}$, where the matrix B has $|\mathcal{Y}| + \sum_i |\mathcal{X}_i|$ rows and encodes the constraints of Equations (A10) and (A11), and the matrix C has $n|\mathcal{Y}|(|\mathcal{Q}| - 1)$ rows and encodes the constraints of Equation (A12).

Each row in B has a 1 in some column which is zero in every other row of B and every row of C. This column corresponds either to $\kappa_{Q|Y}(0|y)$ for a particular y (for constraints like Equation (A10)), or to $\kappa_{Q|X_i}(0|x_i)$ for a particular i and x_i (for constraints like Equation (A11)). These columns are 0 in C because $q = 0$ is omitted Equation (A12). This means that no row of B is a linear combination of other rows in B or C, and that no row in C is a linear combination of any set of other rows that includes a row in B. Therefore, if the rows of A are linearly dependent, it must be that the rows of C are linearly dependent.

Next, let $\vec{c}^{i,y,q}$ indicate the row of C that represents the constraints in Equation (A12) for some source i and outcomes $y, q \neq 0$. Any such row has a column for each $x_i \in \mathcal{X}_i$

with value $P_{X_iY}(x_i, y)$ (at the same index as the row in $\vec{\kappa}$ that represents $\kappa_{Q|X_i}(q|x_i)$). Since $P_{X_iY}(x_i, y) > 0$ for at least one $x_i \in \mathcal{X}_i$, one of these columns must be non-zero. At the same time, these columns are zero in every row $\vec{c}^{j,y,q'}$ where $j \neq i$ or $q' \neq q$. This means that row $\vec{c}^{i,y,q}$ can only be a linear combination of other rows in C if, for all x_i, $P_{X_iY}(x_i, y)$ is a linear combination of $P_{X_iY}(x_i, y')$ for $y' \neq y$. In linear algebra terms, this can be stated as rank $P_{Y|X_i} < |\mathcal{Y}|$.

The previous argument shows that if A is linearly dependent, there must be at least one source i with rank $P_{Y|X_i} < |\mathcal{Y}|$ and some row $\vec{c}^{i,y,q}$ which is a linear combination of other rows from C. Observe that this row $\vec{c}^{i,y,q}$ has a column with value $P_Y(y) > 0$ (at the same index as the row in $\vec{\kappa}$ that represents $\kappa_{Q|Y}(q|y)$). This column is zero in every other row $\vec{c}^{i,y',q'}$ for $y' \neq y$ or $q' \neq q$. This means that $\vec{c}^{i,y,q}$ is a linear combination of a set of other rows in C that include some row $\vec{c}^{j,y,q}$ for $j \neq i$. This implies that $\vec{c}^{j,y,q}$ is also a linear combination of other rows in C, which means that rank $P_{Y|X_j} < |\mathcal{Y}|$.

We have shown that if A is linearly dependent, there must be at least two pairwise conditionals with rank $P_{Y|X_i} < |\mathcal{Y}|$. □

We are now ready to prove Theorem 5.

Proof of Theorem 5. For the case of a single source ($n = 1$), $I_\cap^\prec$ reduces to the mutual information $I_\cap^\prec = I(Y; X_1)$, which is continuous (Section 2.3, [75]). Thus, without loss of generality, we assume that $n \geq 2$.

Next, we define some notation. Note that the optimum value ($I_\cap^\prec$) and the feasible set of the optimization problem in Equation (A7) is a function of the pairwise marginal distributions $P_{X_1Y}, \ldots, P_{X_nY}$. We write Ω for the set of all pairwise marginal distributions which have the same marginal over Y:

$$\Omega = \left\{ (q_{X_1Y}, \ldots, q_{X_nY}) : \sum_{x_i} q_{X_iY}(x_i, y) = \sum_{x_j} q_{X_jY}(x_j, y) \quad \forall i, j \right\}.$$

For any $r \in \Omega$, let $I_\cap^\prec(r)$ indicate the corresponding optimum value in Equation (A7), given the marginals in r, and let $\Lambda(r)$ indicate the feasible set of the optimization problem, as defined in Equation (A13).

Note that the matrix A in Equation (A13) depends on the choice of r, which we indicate by writing it as the matrix-valued function $A(r)$. Given any $r = (q_{X_1Y}, \ldots, q_{X_nY}) \in \Omega$ and feasible solution $\kappa = (\kappa_{Q|Y}, \kappa_{Q|X_1,\ldots,}\kappa_{Q|X_n}) \in \Lambda(r)$, let $I(r, \kappa)$ indicate the corresponding mutual information $I(Q; Y)$, where the marginal distribution over Y is specified by r and the conditional distribution of Q given Y is specified by $\kappa_{Q|Y}$. Using this notation, $I_\cap^\prec(r) = \max_{\kappa \in \Lambda(r)} I(r, \kappa)$.

Below, we show that $I_\cap^\prec(r)$ is continuous if r is *rank regular* [76], which means that there is a neighborhood $U \subseteq \Omega$ of r such that rank $A(r') =$ rank $A(r)$ for all $r' \in U$. Then, to prove the theorem, we assume that $A(r)$ is full rank. Given Lemma A1, this is true as long as $n - 1$ or more of the pairwise conditionals $P_{Y|X_i}$ have rank $P_{Y|X_i} = |\mathcal{Y}|$. Note that a matrix M is full rank iff the singular values $\sigma(M)$ are all strictly positive. Since $A(r)$ is full rank, and $A(r)$ and $\sigma(M)$ are continuous, there is a neighborhood U of r such that the singular values $\sigma(A(r'))$ are all strictly positive for all $r' \in U$, therefore all $A(r'))$ have full rank. This shows that r is rank regular and so $I_\cap^\prec$ is continuous at r.

We now prove that $I_\cap^\prec(r)$ is continuous if $A(r)$ is rank regular. To do so, we will use Hoffman's Theorem [77,78]. In our case, it states that for any pair of marginals $r, r' \in \Omega$ and a feasible solution $\kappa' \in \Lambda(r')$, there exists a feasible solution $\kappa \in \Lambda(r)$ such that

$$\left\| \kappa - \kappa' \right\| \leq \alpha \left\| A(r) - A(r') \right\|, \tag{A14}$$

where α is a constant that does not depend on r' or κ'. (In the notation of [78], we take $G = G'$, $g = g'$ and $d' = d$, and use that the norm of $s = \kappa'$ is bounded, given that it is finite dimensional and has entries in $[0, 1]$). We will also use Daniel's theorem

(Theorem 4.2, [78]), which states that for any $r, r' \in \Omega$ such that rank $A(r) =$ rank $A(r')$, and any feasible solution $\kappa \in \Lambda(r)$, there exists $\kappa' \in \Lambda(r')$ such that

$$\|\kappa - \kappa'\| \leq \beta \|A(r) - A(r')\|, \tag{A15}$$

where β is a constant that doesn't depend on r' (in the notation of [78], $\varepsilon' = \|A(r) - A(r'_i)\|$ and again use that κ have a bounded norm).

Now consider also any sequence $r'_1, r'_2, \cdots \in \Omega$ that converges to a marginal $r \in \Omega$. Let $\kappa'_i \in \Lambda(r'_i)$ indicate an optimal solution of Equation (A7) for r_i, so that $I(r'_i, \kappa'_i) = I^{\prec}_{\cap}(r'_i)$. Given Equation (A14), there is a corresponding sequence $\kappa_1, \kappa_2, \cdots \in \Lambda(r)$ such that

$$\|\kappa_i - \kappa'_i\| \leq \alpha \|A(r) - A(r'_i)\|.$$

Since $A(\cdot)$ is continuous and r'_i converges to r, we have $\lim_{i\to\infty} A(r'_i) = A(r)$ and therefore $\lim_{i\to\infty} \|\kappa_i - \kappa'_i\| = 0$. This implies

$$0 = \lim_{i\to\infty} I(r'_i, \kappa'_i) - I(r, \kappa_i) \geq \lim_{i\to\infty} I^{\prec}_{\cap}(r'_i) - I^{\prec}_{\cap}(r) \tag{A16}$$

where we first used continuity of mutual information, $I(r'_i, \kappa'_i) = I^{\prec}_{\cap}(r'_i)$ and $I(r, \kappa_i) \leq I^{\prec}_{\cap}(r)$.

Now assume that r is rank regular. Since r_i converges to r, rank $A(r'_i) =$ rank $A(r)$ for all sufficiently large i. Let $\kappa \in \Lambda(r)$ be an optimal solution of Equation (A7) for r, so that $I(r, \kappa) = I^{\prec}_{\cap}(r)$. Given Equation (A15), for all sufficiently large i there exists $\kappa'_i \in \Lambda(r'_i)$ such that

$$\|\kappa - \kappa'_i\| \leq \beta \|A(r) - A(r'_i)\|.$$

As before, we have $\lim_{i\to\infty} A(r'_i) = A(r)$ and $\lim_{i\to\infty} \|\kappa - \kappa'_i\| = 0$, which implies

$$0 = \lim_{i\to\infty} I(r'_i, \kappa'_i) - I(r, \kappa) \leq \lim_{i\to\infty} I^{\prec}_{\cap}(r'_i) - I^{\prec}_{\cap}(r) \tag{A17}$$

where we first used continuity of mutual information, $I(r'_i, \kappa'_i) \leq I^{\prec}_{\cap}(r'_i)$, and $I(r, \kappa) \leq I^{\prec}_{\cap}(r)$.

Combining Equation (A16) and Equation (A17) proves continuity, $\lim_{i\to\infty} I^{\prec}_{\cap}(r_i) = I^{\prec}_{\cap}(r)$, under the assumption that $A(r)$ is rank regular.

Finally, note that $A(r)$ is a real analytic function of r. This means that almost all r rank regular, because those r which are not rank regular form a proper analytic subset of Ω (which has measure zero) [76]. Thus, $I^{\prec}_{\cap}(r)$ is continuous almost everywhere. □

We finish our analysis of the continuity of $I^{\prec}_{\cap}$ by showing global continuity when the target is a binary random variable.

Corollary A1. *$I^{\prec}_{\cap}(X_1; \ldots; X_n \to Y)$ is continuous everywhere when Y is a binary random variable.*

Proof. In an overloading of notation, let $I^{\prec}_{\cap}(r)$ and $I_r(X_i; Y)$ indicate $I^{\prec}_{\cap}(X_1; \ldots; X_n \to Y)$ and the mutual information $I(X_i; Y)$, respectively, for the joint distribution $r_{X_1 \ldots X_n Y}$. By Theorem 5, $I^{\prec}_{\cap}$ can only be discontinuous at the joint distribution $P_{X_1 \ldots X_n Y}$ if there is a source X_i with rank $P_{Y|X_i} = 1 < |\mathcal{Y}|$. However, if source X_i has rank 1, then the conditional distributions $P_{Y|X_i = x_i}$ are the same for all x_i, so $I_P(X_i; Y) = 0$ and $I^{\prec}_{\cap}(P) = 0$ (since $0 \leq I^{\prec}_{\cap}(P) \leq I_P(X_i; Y)$). Finally, consider any sequence of joint distributions $s^{(n)}_{X_1 \ldots X_n Y}$ for $n = 1, 2, \ldots$ that converges to $P_{X_1 \ldots X_n Y}$. We have

$$0 \leq \lim_{n\to\infty} I^{\prec}_{\cap}(s^{(n)}) \leq \lim_{n\to\infty} I_{s^{(n)}}(X_i; Y) = I_P(X_i; Y) = 0,$$

where we used the continuity of mutual information. This shows that $\lim_{n\to\infty} I^{\prec}_{\cap}(s^{(n)}) = 0 = I^{\prec}_{\cap}(P)$, proving continuity. □

Appendix E. Behavior of $I_\cap^\prec$ on Gaussian Random Variables

Although in this paper we focused on random variables with finite sets of outcomes, we can briefly comment on the behavior of Blackwell redundancy on Gaussian random variables. Suppose that all sources $X_1, \ldots, X_n$ and the target Y are continuous-valued, and that the pairwise marginals $P_{X_i Y}$ are multivariate Gaussians. In addition, suppose that Y is one-dimensional (the sources X_i can be multi-dimensional). Given these assumptions, Barrett [51] analyzed the $I_\cup^{\text{BROJA}}$ measure and showed that the corresponding excluded information obeys $E(X_j \to Y | X_i; X_j) = 0$ whenever $I(X_i; Y) \leq I(X_j; Y)$. Recall that $I_\cup^{\text{BROJA}}$ is equivalent to Blackwell union information $I_\cup^\prec$. Then, given the Blackwell property, Theorem 4, and the data processing inequality, Equation (24), the result in Ref. [51] implies that $X_i \prec_Y X_j$ if and only if $I(X_i; Y) \leq I(X_j; Y)$. Thus, for Gaussian random variables, Blackwell redundancy $I_\cap^\prec$ is equivalent to $I_\cap^{\text{MMI}}$ redundancy, as defined in Equation (19). This parallels the case for most other redundancy measures [51].

Appendix F. Operational Interpretation of the $I_\cap^{\text{GH}}$

As mentioned in the main text, the redundancy measure $I_\cap^{\text{GH}}$ is a special case of Equation (7), where the "more informative" order $B \sqsubset C$ is defined in terms of conditional independence $B - C - Y$. Here we show that this ordering relation can be given an operational interpretation, which is similar but distinct from the operational interpretation of the Blackwell order $\prec_Y$ discussed in Section 5.1.

To introduce this operational interpretation, let the random variable Y represent the state of the environment, and assume there are two random variables B and C which have some information about Y. Suppose that an agent tries to maximize expected utility $u(a, y)$ by using a strategy that depends either on the outcomes of B or C. Blackwell's theorem tells us that $B \prec_Y C$ iff an agent with access to C can always achieve higher expected utility than an agent with access to B. It is possible, however, the agent with access to C may do worse than the agent with access to B, *conditional on the event that random variable C has some particular outcome c*. In the following theorem, we show $B - C - Y$ iff the agent cannot do better with B than C, even when conditioned on any particular outcome $C = c$. (We thank Johannes Rauh for suggesting this simplified proof).

Theorem A2. *Given random variables B, C, and Y, $B - C - Y$ if and only if*

$$\max_{\kappa_{A|B}} \sum_{y,a,b} P_{YB|C}(y, b|c) \kappa_{A|B}(a|b) u(a, y) \leq \max_{\kappa_{A|C}} \sum_{y,a} P_{Y|C}(y|c) \kappa_{A|C}(a|c) u(a, y). \tag{A18}$$

for all utility functions $u(a, y)$ and all $c \in \mathcal{C}$ with $P_C(c) > 0$.

Proof. Consider any $c \in \mathcal{C}$ with $P_C(c) > 0$. By multiplying both sides of Equation (A18) by $P_C(c)$ and rearranging, this inequality can be rewritten as

$$\begin{aligned} \max_{\kappa_{A|B}} \sum_{y,a,b} P_Y(y) P_{C|Y}(c|y) P_{B|Y,C}(b|y, c) \kappa_{A|B}(a|b) u(a, y) \\ \leq \max_{\kappa_{A|C}} \sum_{y,a} P_Y(y) P_{C|Y}(c|y) \kappa_{A|C}(a|c) u(a, y). \end{aligned} \tag{A19}$$

Note that if Equation (A18) holds for a given c and all utility functions, then it must also hold for the utility function $u'(a, y) := P_{C|Y}(c|y) u(a, y)$. Plugging into Equation (A19) gives

$$\max_{\kappa_{A|B}} \sum_{y,a,b} P_Y(y) P_{B|Y,C}(b|y, c) \kappa_{A|B}(a|b) u'(a, y) \leq \max_{\kappa_{A|C}} \sum_{y,a} P_Y(y) \kappa_{A|C}(a|c) u'(a, y). \tag{A20}$$

Now define two random variables: a constant random variable $\hat{C}_c$ with a single outcome c and $\hat{B}_c$ with the same outcomes as B but having the conditional distribution $P_{\hat{B}_c|Y} = P_{B|Y,C=c}$. Then, Equation (A20) can be written in terms of these random variables as

$$\max_{\kappa_{A|B}} \sum_{y,a,b} P_Y(y) P_{\hat{B}_c|Y}(b|y) \kappa_{A|B}(a|b) u'(a,y) \leq \max_{\kappa_{A|C}} \sum_{y,a} P_Y(y) P_{\hat{C}_c|Y}(c|y) \kappa_{A|C}(a|c) u'(a,y). \quad \text{(A21)}$$

Given Equations (25) and (26), Equation (A21) holds for all u' iff $\hat{B}_c \prec_Y \hat{C}_c$. Since $\hat{C}_c$ has a single outcome, it is independent of Y. That means $\hat{B}_c$ must be also independent of Y and so $P_{\hat{B}_c|Y=y} = P_{B|Y=y,C=c}$ is the same for all y, implying that $P_{B|Y=y,C=c} = P_{B|C=c}$. Since this holds for all $c \in \mathcal{C}$, $P_{B|YC} = P_{B|C}$ and therefore $B - C - Y$. □

Given Theorem A2, $I_\cap^{\mathrm{GH}}$ can be given the following operational interpretation. Imagine two agents, Alice and Bob, who can acquire information about Y via different random variables, and then use this information to maximize their expected utility. Suppose that Alice has access to one of the sources X_i. Then, $I_\cap^{\mathrm{GH}}$ is the maximum information that Bob can have about Y without being able to do better than Alice on any utility function, regardless of which source X_i Alice has access to, and even when conditioned on X_i having any particular outcome x_i.

Appendix G. Equivalence of $I_\cup^\prec$ and $I_\cup^{\mathbf{BROJA}}$

The following proves that $I_\cup^\prec$ and $I_\cup^{\mathrm{BROJA}}$, as defined via the optimization problems in Equations (28) and (31), are equivalent.

Theorem A3. $I_\cup^\prec(X_1;\ldots;X_n \to Y) = I_\cup^{\mathrm{BROJA}}(X_1;\ldots;X_n \to Y)$.

Proof. Let $\tilde{X}_1,\ldots,\tilde{X}_n$ be a set of random variables that achieve $I(Y;\tilde{X}_1,\ldots,\tilde{X}_n) = I_\cup^{\mathrm{BROJA}}(X_1;\ldots;X_n \to Y)$. Define the random variable $Q := (\tilde{X}_1,\ldots,\tilde{X}_n)$, and note that $\tilde{X}_i \prec_Y Q$ for all i. Since $P_{\tilde{X}_i Y} = P_{X_i Y}$ for all i, it must be that $X_i \prec_Y Q$ for all i. Thus Q satisfies the constraints of the optimization problem in Equation (28), so

$$I_\cup^\prec(X_1;\ldots;X_n \to Y) \leq I(Y;Q) = I_\cup^{\mathrm{BROJA}}(X_1;\ldots;X_n \to Y). \quad \text{(A22)}$$

Next, consider the optimization in Equation (28). For any $\epsilon > 0$, let Q be a random variable that satisfies $X_i \prec_Y Q$ and achieves

$$I(Y;Q) \leq I_\cup^\prec(X_1;\ldots;X_n \to Y) + \epsilon. \quad \text{(A23)}$$

For each i, let $\kappa_{X_i|Q}$ be a channel that obeys $P_{X_i|Y}(x_i|y) = \sum_q \kappa_{X_i|Q}(x_i|q) P_{Q|Y}(q|y)$ (such a channel must exist since $X_i \prec_Y Q$). Define the random variables $\tilde{X}_1,\ldots,\tilde{X}_n$ with the joint distribution

$$P_{YQ\tilde{X}_1\ldots\tilde{X}_n}(y,q,x_1,\ldots,x_n) = P_Y(y) P_{Q|Y}(q|y) \prod_i \kappa_{X_i|Q}(x_i|q). \quad \text{(A24)}$$

Note that the pairwise marginals obey $P_{\tilde{X}_i Y} = P_{X_i Y}$. Thus, all of the $\tilde{X}_i$ satisfy the marginal constraints in the right hand side of Equation (31), so

$$I_\cup^{\mathrm{BROJA}}(X_1;\ldots;X_n \to Y) \leq I(Y;\tilde{X}_1,\ldots,\tilde{X}_n). \quad \text{(A25)}$$

By elementary properties of mutual information, we have

$$I(Y;\tilde{X}_1,\ldots,\tilde{X}_n) \leq I(Y;Q,\tilde{X}_1,\ldots,\tilde{X}_n) \quad \text{(A26)}$$

Given Equation (A24), the Markov condition $Y - Q - \tilde{X}_1, \ldots, \tilde{X}_n$ holds, so

$$I(Y; \tilde{X}_1, \ldots, \tilde{X}_n) \leq I(Y; Q) \tag{A27}$$

by the data processing inequality. Combining Equations (A23) and (A25) to (A27) implies

$$I_{\cup}^{\text{BROJA}}(X_1; \ldots; X_n \to Y) \leq I(Y; Q) \leq I_{\cup}^{\prec}(X_1; \ldots; X_n \to Y) + \epsilon.$$

Since this holds for all ϵ, we can take the limit $\epsilon \to 0$ to give $I_{\cup}^{\text{BROJA}}(X_1; \ldots; X_n \to Y) \leq I_{\cup}^{\prec}(X_1; \ldots; X_n \to Y)$. The result follows by combining with Equation (A22). □

Appendix H. Relation between $I_{\cap}^{\text{WB}}$ and Our General Framework

Here we consider whether the redundancy measure $I_{\cap}^{\text{WB}}$ proposed by Williams and Beer [11] can be put in the general form Equation (7). This measure is defined as

$$I_{\cap}^{\text{WB}}(X_1; \ldots; X_n \to Y) := \sum_y P_Y(y) \min_i I(X_i; Y = y), \tag{A28}$$

where $I(X_i; Y = y)$ is called the "specific information" between X_i and target outcome y,

$$I(X_i; Y = y) := D_{\text{KL}}(P_{X_i|Y=y} \| P_{X_i}) = \sum_{x_i} P_{X_i|Y}(x_i|y) \log \frac{P_{X_i|Y}(x_i|y)}{P_{X_i}(x_i)},$$

and D_{KL} is Kullback-Leibler (KL) divergence.

Specific information obeys $I(X; Y) = \sum_y P_Y(y) I(X; Y = y)$. Thus, Equation (A28) looks similar to a mutual information expression, where each specific information term is given by the smallest specific information that y carries about any of the sources. Motivated by this interpretation, one might ask whether there exists a random variable Q whose specific information terms are equal to $I(Q; Y = y) = \min_i I(X_i; Y = y)$ for each y. If such a random variable existed, then $I_{\cap}^{\text{WB}}$ could be written as

$$I_{\cap}^{\text{WB}}(X_1; \ldots; X_n \to Y) \stackrel{?}{=} \max_Q I(Q; Y) \text{ such that } \forall i, y : I(Q; Y = y) \leq I(X_i; Y = y), \tag{A29}$$

which has the form of Equation (7), with the $\sqsubset$ order defined as

$$A \sqsubset B \text{ iff } I(A; Y = y) \leq I(B; Y = y) \text{ for all } y \in \mathcal{Y}. \tag{A30}$$

Here we provide a counterexample to demonstrate that such a variable does not exist in general, and so therefore Equation (A29) is not generally valid. Suppose Y has three outcomes $\mathcal{Y} = \{0, 1, 2\}$ with a uniform distribution, and consider two binary sources X_1, X_2 with the following conditional distributions,

$$P_{X_1|Y}(x_1|y) = \begin{cases} \delta(x_1, y) & \text{if } y \in \{0, 1\} \\ \frac{1}{2}\delta(x_1, 0) + \frac{1}{2}\delta(x_1, 1) & \text{if } y = 2 \end{cases}$$

$$P_{X_2|Y}(x_1|y) = \begin{cases} \delta(x_1, y) & \text{if } y \in \{0, 2\} \\ \frac{1}{2}\delta(x_1, 0) + \frac{1}{2}\delta(x_1, 2) & \text{if } y = 1 \end{cases}$$

In this case, a simple calculation shows that the specific information obeys (in bits)

$$\begin{aligned} I(X_1; Y = 0) &= 1 & I(X_2; Y = 0) &= 1 \\ I(X_1; Y = 1) &= 1 & I(X_2; Y = 1) &= 0 \\ I(X_1; Y = 2) &= 0 & I(X_2; Y = 2) &= 1 \end{aligned}$$

Plugging into Equation (A28) gives $I_{\cap}^{\text{WB}}(X_1; X_2 \to Y) = 1/3$.

Now consider the optimization problem in Equation (A29). Since $I(X_1; Y = 2) = I(X_2; Y = 1) = 0$, any allowed Q must satisfy $I(Q; Y = 1) = I(Q; Y = 2) = 0$ and therefore $P_{Q|Y=1} = P_Q = P_{Q|Y=2}$. Combined with the marginalization identity $P_Y(0)P_{Q|Y=0} + P_Y(1)P_{Q|Y=1} + P_Y(2)P_{Q|Y=2} = P_Q$, this implies that $P_{Q|Y=0} = P_Q$ and therefore that $I(Q; Y = 0) = 0$. Thus, any allowed Q obeys $I(Q; Y) = 0 \neq I_\cap^{\mathrm{WB}}$. This means that $I_\cap^{\mathrm{WB}}$ cannot be expressed in the form of Equation (7) when $\sqsubset$ is defined as Equation (A30).

Appendix I. Miscellaneous Derivations

Proof of Lemma 1. We use a modified version of the example in [39,68]. Consider a set of $n \geq 3$ sources. The inclusion-exclusion principle states that

$$I_\cup(X_1; \ldots; X_n \to Y) = \sum_{J \subseteq \{1,\ldots,n\} \setminus \{\varnothing\}} (-1)^{|J|-1} I_\cap(X_{J_1}; \ldots; X_{J_{|J|}} \to Y). \tag{A31}$$

Now, let $X_1, \ldots, X_{n-1}$ be uniformly distributed and statistically independent binary random variables, and take $X_n = X_1$ XOR X_2 and $Y = (X_1, X_2, X_n)$. Note that $I(Y; X_i) = 1$ bit for $i \in \{1, 2, n\}$ and $I(Y; X_i) = 0$ for $i \in \{3, \ldots, n-1\}$, and that $I(Y; X_1, \ldots, X_n) = 2$ bit. Thus, $I_\cap(X_i; X_j \to Y) = 0$ whenever $i \in \{3, \ldots, n-1\}$ or $j \in \{3, \ldots, n-1\}$, as follows from *Symmetry*, *Self-redundancy*, and *Monotonicity*. Note also that the outcomes of Y are simply a relabelling of (X_1, X_2), and similarly for (X_1, X_n) and (X_2, X_n). Then, since by *Independent identity* property, $I_\cap(X_i; X_j \to Y) = 0$ for $i \neq j$ where $i, j \in \{1, 2, n\}$. Thus, $I_\cap(X_i; X_j \to Y) = 0$ for all pairs $i \neq j$. By *Monotonicity*, redundancy is 0 for any set of 2 or more sources.

Plugging this into Equation (A31) gives

$$I_\cup(X_1; \ldots; X_n \to Y) = \sum_i I_\cap(X_i \to Y) = \sum_i I(X_i \to Y) = 3 \text{ bit}$$

Note that this exceeds the total amount of information about the target provided jointly by all sources, which is only 2 bits, so $I_\cup(X_1; \ldots; X_n \to Y) \not\leq I(Y; X_1, \ldots, X_n)$. □

Proof of Theorem 3. Without loss of generality, let $i = 1$. We will use that $U^\prec(X_1 \to Y | X_1; \ldots; X_n) = 0$ is equivalent to

$$I_\cap^\prec(X_1; \ldots; X_n \to Y) = I(X_1; Y). \tag{A32}$$

We will use that by monotonicity of mutual information with respect to $\prec_Y$ (see Section 4.1),

$$I(X_1; Y) \geq I_\cap^\prec(X_1; \ldots; X_n \to Y). \tag{A33}$$

We first prove the "if" direction. Since $Q = X_1$ is in the feasible set of Equation (27), $I_\cap^\prec(X_1; \ldots; X_n \to Y) \geq I(X_1; Y)$. Combining with Equation (A33) gives Equation (A32).

We now prove the "only if" direction. As described in Appendix C, $I_\cap^\prec$ can be expressed as an optimization over a finite dimensional, closed, and bounded region, so the supremum in Equation (27) is achieved. Thus, there is some Q such that $Q \prec_Y X_i$ for all i and

$$I(Y; Q) = I_\cap^\prec(X_1; \ldots; X_n \to Y). \tag{A34}$$

Since $Q \prec_Y X_1$, there is a conditional probability distribution $\kappa_{Q|\mathcal{X}_1}$ such that $P_{Q|Y}(q|y) = \sum_{x_1} \kappa_{Q|X_1}(q|x_1) P_{X_1|Y}(x_1|y)$. Define a random variable $\tilde{Q}$ with the joint distribution

$$P_{YX_1\tilde{Q}}(y, x_1, q) = \kappa_{Q|X_1}(q|x_1) P_{YX_1}(y, x_1).$$

We will use that $P_{QY} = P_{\tilde{Q}Y}$. Then, the chain rule for mutual information gives

$$I(Y; X_1, \tilde{Q}) = I(Y; \tilde{Q}) + I(Y; X_1|\tilde{Q}) = I(Y; X_1) + I(Y; \tilde{Q}|X_1) = I(Y; X_1),$$

where we used the Markov condition $Y - X_1 - \tilde{Q}$. Combining and rearranging gives

$$I(Y;\tilde{Q}) = I(Y;X_1) - I(Y;X_1|\tilde{Q}). \tag{A35}$$

Now assume that Equation (A32) holds. Combining with Equation (A34) and $P_{QY} = P_{\tilde{Q}Y}$ gives $I(Y;X_1) = I(Y;Q) = I(Y;\tilde{Q})$. Combining with Equation (A35) gives $I(Y;X_1|\tilde{Q}) = 0$, meaning that the Markov condition $Y - \tilde{Q} - X_1$ holds and therefore $X_1 \prec_Y \tilde{Q}$. Since $Q \prec_Y X_i$ for all i and $P_{QY} = P_{\tilde{Q}Y}$, it also the case that $\tilde{Q} \prec_Y X_i$ for all i. Finally, since $\prec_Y$ is transitive, $X_1 \prec_Y X_i$ for all i, which is the desired result. □

Proof of Theorem 4. Without loss of generality, let $i = 1$. We will use that $E^{\prec}(X_1 \to Y|X_1;\ldots;X_n) = 0$ is equivalent to

$$I_{\cup}^{\prec}(X_1;\ldots;X_n \to Y) = I(X_1;Y). \tag{A36}$$

We will use that by monotonicity of mutual information with respect to $\prec_Y$ (see Section 4.1),

$$I(X_1;Y) \leq I_{\cup}^{\prec}(X_1;\ldots;X_n \to Y). \tag{A37}$$

We first prove the "if" direction. Since $Q = X_1$ is in the feasible set of Equation (28), $I_{\cup}^{\prec}(X_1;\ldots;X_n \to Y) \leq I(X_1;Y)$. Combining with Equation (A37) gives Equation (A36).

We now prove the "only if" direction. As we show in Appendix G, $I_{\cup}^{\prec}$ is equivalent to $I_{\cup}^{\text{BROJA}}$, which is defined as an optimization over a finite dimensional, closed, and bounded region. Thus the infimum in Equation (28) is always achieved, so there is some Q such that $X_i \prec_Y Q$ for all i and

$$I(Y;Q) = I_{\cup}^{\prec}(X_1;\ldots;X_n \to Y). \tag{A38}$$

Moreover, since $X_1 \prec_Y Q$, there is a conditional probability distribution $\kappa_{X_1|Q}$ such that $P_{X_1|Y}(x_1|y) = \sum_q \kappa_{X_1|Q}(x_1|q)P_{Q|Y}(q|y)$. Define a random variable $\tilde{X}_1$ with the joint distribution

$$P_{Y\tilde{X}_1Q}(y,x_1,q) = \kappa_{X_1|Q}(x_1|q)P_{QY}(q,y).$$

We will use that $P_{X_1Y} = P_{\tilde{X}_1Y}$. Then, using the chain rule for mutual information,

$$I(Y;\tilde{X}_1,Q) = I(Y;\tilde{X}_1) + I(Y;Q|\tilde{X}_1) = I(Y;Q) + I(Y;\tilde{X}_1|Q) = I(Y;Q),$$

where we used the Markov condition $Y - Q - \tilde{X}_1$. Combining and rearranging gives

$$I(Y;\tilde{X}_1) = I(Y;Q) - I(Y;Q|\tilde{X}_1). \tag{A39}$$

Now assume that Equation (A36) holds. Combining with Equation (A38) and $P_{X_1Y} = P_{\tilde{X}_1Y}$ gives $I(Y;X_1) = I(Y;\tilde{X}_1) = I(Y;Q)$. Combining with Equation (A39) gives $I(Y;Q|\tilde{X}_1) = 0$, meaning that the Markov condition $Y - \tilde{X}_1 - Q$ holds and therefore $Q \prec_Y \tilde{X}_1$. Since $P_{X_1Y} = P_{\tilde{X}_1Y}$, it is also the case that $Q \prec_Y X_1$. Finally, since $X_i \prec_Y Q$ for all i and $\prec_Y$ is transitive, $X_i \prec_Y X_1$ for all i, which is the desired result. □

Proof of Theorem 6. Consider any random variable Q which achieves the maximum in Equation (27). This implies there are channels $\kappa_{Q|X_1}$ and $\kappa_{Q|X_2}$ such that for any $q \in \mathcal{Q}$ and $(x_1,x_2) \in \mathcal{X}_1 \times \mathcal{X}_2$ with $p_{X_1X_2}(x_1,x_2) > 0$,

$$P_{Q|X_1X_2}(q|x_1,x_2) = \sum_{x_1'} \kappa_{Q|X_1}(q|x_1')P_{X_1|X_1X_2}(x_1'|x_1,x_2)$$
$$P_{Q|X_1X_2}(q|x_1,x_2) = \sum_{x_2'} \kappa_{Q|X_2}(q|x_2')P_{X_2|X_1X_2}(x_2'|x_1,x_2).$$

We now equate the above two expressions, while using that $P_{X_1|X_1X_2}(x_1'|x_1,x_2) = \delta(x_1',x_1)$ and $P_{X_2|X_1X_2}(x_2'|x_1,x_2) = \delta(x_2',x_2)$ (where $\delta(\cdot,\cdot)$ is the Kronecker delta). This gives

$$\kappa_{Q|X_1}(q|x_1) = \kappa_{Q|X_2}(q|x_2) \tag{A40}$$

for all q and any (x_1,x_2) where $p_{X_1X_2}(x_1,x_2) > 0$.

Now consider a bipartite graph with vertex set $\mathcal{X}_1 \cup \mathcal{X}_2$ and an edge between vertex x_1 and vertex x_2 if $P_{X_1X_2}(x_1,x_2) > 0$. Define Π to be the set of connected components of this bipartite graph, and let $f_1 : \mathcal{X}_1 \to \Pi$ be a function that maps each x_1 to its corresponding connected component (for any x_1 with $P_{X_1}(x_1) = 0$, $f_1(x_1)$ can be any value). Equation (A40) implies that if x_1 and x_1' both belong to the same connected component, then the constraint Equation (A40) will "propagate" from x_1 to x_1', so that $\kappa_{Q|X_1}(q|x_1) = \kappa_{Q|X_1}(q|x_1')$. Said differently, this means that $\kappa_{Q|X_1}(q|x_1) = \kappa_{Q|X_1}(q|f_1(x_1))$ and that the Markov condition $(X_1,X_2) - X_1 - f_1(X_1) - Q$ holds. This gives

$$I(X_1,X_2;Q) \le I(X_1,X_2;f_1(X_1)) = H(f_1(X_1)), \tag{A41}$$

where the first inequality uses the data processing inequality, and the second equality uses that $f_1(X_1)$ is a deterministic function of (X_1,X_2). The upper bound in Equation (A41) is achieved when $Q = f_1(X_1)$, thus $Q \lhd X_1$. A similar argument shows that $Q \lhd X_2$.

We have shown that the constraints in Equation (27) can be replaced by $Q \lhd X_1, Q \lhd X_2$. It is also clear that any Q which is a deterministic function of either X_1 or X_2 must also be a deterministic function of the target $Y = (X_1,X_2)$, hence $I(Y;Q) = H(Q)$. Combining these results shows that Equation (27) is equivalent to Equation (32) for the COPY gate. □

References

1. Schneidman, E.; Bialek, W.; Berry, M.J. Synergy, Redundancy, and Independence in Population Codes. *J. Neurosci.* **2003**, *23*, 11539–11553. [CrossRef] [PubMed]
2. Daniels, B.C.; Ellison, C.J.; Krakauer, D.C.; Flack, J.C. Quantifying collectivity. *Curr. Opin. Neurobiol.* **2016**, *37*, 106–113. [CrossRef] [PubMed]
3. Tax, T.; Mediano, P.; Shanahan, M. The partial information decomposition of generative neural network models. *Entropy* **2017**, *19*, 474. [CrossRef]
4. Amjad, R.A.; Liu, K.; Geiger, B.C. Understanding individual neuron importance using information theory. *arXiv* **2018**, arXiv:1804.06679.
5. Lizier, J.; Bertschinger, N.; Jost, J.; Wibral, M. Information decomposition of target effects from multi-source interactions: Perspectives on previous, current and future work. *Entropy* **2018**, *20*, 307. [CrossRef]
6. Wibral, M.; Priesemann, V.; Kay, J.W.; Lizier, J.T.; Phillips, W.A. Partial information decomposition as a unified approach to the specification of neural goal functions. *Brain Cogn.* **2017**, *112*, 25–38. [CrossRef]
7. Timme, N.; Alford, W.; Flecker, B.; Beggs, J.M. Synergy, redundancy, and multivariate information measures: An experimentalist's perspective. *J. Comput. Neurosci.* **2014**, *36*, 119–140. [CrossRef]
8. Chan, C.; Al-Bashabsheh, A.; Ebrahimi, J.B.; Kaced, T.; Liu, T. Multivariate Mutual Information Inspired by Secret-Key Agreement. *Proc. IEEE* **2015**, *103*, 1883–1913. [CrossRef]
9. Rosas, F.E.; Mediano, P.A.; Jensen, H.J.; Seth, A.K.; Barrett, A.B.; Carhart-Harris, R.L.; Bor, D. Reconciling emergences: An information-theoretic approach to identify causal emergence in multivariate data. *PLoS Comput. Biol.* **2020**, *16*, e1008289. [CrossRef]
10. Cang, Z.; Nie, Q. Inferring spatial and signaling relationships between cells from single cell transcriptomic data. *Nat. Commun.* **2020**, *11*, 1–13. [CrossRef]
11. Williams, P.L.; Beer, R.D. Nonnegative decomposition of multivariate information. *arXiv* **2010**, arXiv:1004.2515.
12. Williams, P.L. Information dynamics: Its theory and application to embodied cognitive systems. Ph.D. Thesis, Indiana University, Bloomington, IN, USA, 2011.
13. Bertschinger, N.; Rauh, J.; Olbrich, E.; Jost, J.; Ay, N. Quantifying unique information. *Entropy* **2014**, *16*, 2161–2183. [CrossRef]
14. Quax, R.; Har-Shemesh, O.; Sloot, P. Quantifying synergistic information using intermediate stochastic variables. *Entropy* **2017**, *19*, 85. [CrossRef]
15. James, R.G.; Emenheiser, J.; Crutchfield, J.P. Unique information via dependency constraints. *J. Phys. Math. Theor.* **2018**, *52*, 014002. [CrossRef]
16. Griffith, V.; Chong, E.K.; James, R.G.; Ellison, C.J.; Crutchfield, J.P. Intersection information based on common randomness. *Entropy* **2014**, *16*, 1985–2000. [CrossRef]

17. Griffith, V.; Koch, C. Quantifying synergistic mutual information. In *Guided Self-Organization: Inception*; Springer: Berlin/Heidelberg, Germany, 2014; pp. 159–190.
18. Griffith, V.; Ho, T. Quantifying redundant information in predicting a target random variable. *Entropy* **2015**, *17*, 4644–4653. [CrossRef]
19. Harder, M.; Salge, C.; Polani, D. Bivariate measure of redundant information. *Phys. Rev.* **2013**, *87*, 012130. [CrossRef]
20. Ince, R. Measuring Multivariate Redundant Information with Pointwise Common Change in Surprisal. *Entropy* **2017**, *19*, 318. [CrossRef]
21. Finn, C.; Lizier, J. Pointwise Partial Information Decomposition Using the Specificity and Ambiguity Lattices. *Entropy* **2018**, *20*, 297. [CrossRef]
22. Shannon, C. The lattice theory of information. *Trans. Ire Prof. Group Inf. Theory* **1953**, *1*, 105–107. [CrossRef]
23. Shannon, C.E. A note on a partial ordering for communication channels. *Inf. Control.* **1958**, *1*, 390–397. [CrossRef]
24. Cohen, J.; Kempermann, J.H.; Zbaganu, G. *Comparisons of Stochastic Matrices with Applications in Information Theory, Statistics, Economics and Population*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 1998.
25. Le Cam, L. Sufficiency and approximate sufficiency. *Ann. Math. Stat.* **1964**, *35*, 1419–1455. [CrossRef]
26. Korner, J.; Marton, K. Comparison of two noisy channels. *Top. Inf. Theory* **1977**, *16*, 411–423.
27. Torgersen, E. *Comparison of Statistical Experiments*; Cambridge University Press: Cambridge, UK, 1991; Volume 36.
28. Blackwell, D. Equivalent comparisons of experiments. *Ann. Math. Stat.* **1953**, *24*, 265–272. [CrossRef]
29. James, R.; Emenheiser, J.; Crutchfield, J. Unique information and secret key agreement. *Entropy* **2019**, *21*, 12. [CrossRef]
30. Whitelaw, T.A. *Introduction to Abstract Algebra*, 2nd ed.; OCLC: 17440604; Blackie & Son: London, UK, 1988.
31. Halmos, P.R. *Naive Set Theory*; Courier Dover Publications: Mineola, NY, USA; 2017.
32. McGill, W. Multivariate information transmission. *Trans. Ire Prof. Group Inf. Theory* **1954**, *4*, 93–111. [CrossRef]
33. Fano, R.M. *The Transmission of Information: A Statistical Theory of Communications*; Massachusetts Institute of Technology: Cambridge, MA, USA, 1961.
34. Reza, F.M. *An Introduction to Information Theory*; Dover Publications, Inc.: Mineola, NY, USA, 1961.
35. Ting, H.K. On the amount of information. *Theory Probab. Its Appl.* **1962**, *7*, 439–447. [CrossRef]
36. Yeung, R.W. A new outlook on Shannon's information measures. *IEEE Trans. Inf. Theory* **1991**, *37*, 466–474. [CrossRef]
37. 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.
38. Tilman. Examples of Common False Beliefs in Mathematics (Dimensions of Vector Spaces). MathOverflow, 2010. Available online: https://mathoverflow.net/q/23501 (accessed on 4 January 2022).
39. Rauh, J.; Bertschinger, N.; Olbrich, E.; Jost, J. Reconsidering unique information: Towards a multivariate information decomposition. In Proceedings of the 2014 IEEE International Symposium on Information Theory, Honolulu, HI, USA, 29 June–4 July 2014; pp. 2232–2236.
40. Rauh, J. Secret Sharing and Shared Information. *Entropy* **2017**, *19*, 601. [CrossRef]
41. Chicharro, D.; Panzeri, S. Synergy and Redundancy in Dual Decompositions of Mutual Information Gain and Information Loss. *Entropy* **2017**, *19*, 71. [CrossRef]
42. Ay, N.; Polani, D.; Virgo, N. Information decomposition based on cooperative game theory. *arXiv* **2019**, arXiv:1910.05979.
43. Rosas, F.E.; Mediano, P.A.; Rassouli, B.; Barrett, A.B. An operational information decomposition via synergistic disclosure. *J. Phys. A Math. Theor.* **2020**, *53*, 485001. [CrossRef]
44. Davey, B.A.; Priestley, H.A. *Introduction to Lattices and Order*; Cambridge University Press: Cambridge, UK, 2002.
45. Bertschinger, N.; Rauh, J. The Blackwell relation defines no lattice. In Proceedings of the 2014 IEEE International Symposium on Information Theory, Honolulu, HI, USA, 29 June–4 July 2014; IEEE: Piscataway, NJ, USA, 2014; pp. 2479–2483.
46. Li, H.; Chong, E.K. On a connection between information and group lattices. *Entropy* **2011**, *13*, 683–708. [CrossRef]
47. Gács, P.; Körner, J. Common information is far less than mutual information. *Probl. Control Inf. Theory* **1973**, *2*, 149–162.
48. Aumann, R.J. Agreeing to disagree. *Ann. Stat.* **1976**, *4*, 1236–1239. [CrossRef]
49. Banerjee, P.K.; Griffith, V. Synergy, Redundancy and Common Information. *arXiv* **2015**, arXiv:1509.03706v1.
50. Hexner, G.; Ho, Y. Information structure: Common and private (Corresp.). *IEEE Trans. Inf. Theory* **1977**, *23*, 390–393. [CrossRef]
51. Barrett, A.B. Exploration of synergistic and redundant information sharing in static and dynamical Gaussian systems. *Phys. Rev. E* **2015**, *91*, 052802. arXiv:1411.2832.
52. Pluim, J.P.; Maintz, J.A.; Viergever, M.A. F-information measures in medical image registration. *IEEE Trans. Med. Imaging* **2004**, *23*, 1508–1516. [CrossRef]
53. Banerjee, A.; Merugu, S.; Dhillon, I.S.; Ghosh, J.; Lafferty, J. Clustering with Bregman divergences. *J. Mach. Learn. Res.* **2005**, *6*, 1705–1749.
54. Brunel, N.; Nadal, J.P. Mutual information, Fisher information, and population coding. *Neural Comput.* **1998**, *10*, 1731–1757. [CrossRef] [PubMed]
55. Li, M.; Vitányi, P.. *An Introduction to Kolmogorov Complexity and Its Applications*; Springer: Berlin/Heidelberg, Germany, 2008; Volume 3.
56. Shmaya, E. Comparison of information structures and completely positive maps. *J. Phys. A Math. Gen.* **2005**, *38*, 9717. [CrossRef]
57. Chefles, A. The quantum Blackwell theorem and minimum error state discrimination. *arXiv* **2009**, arXiv:0907.0866.

58. Buscemi, F. Comparison of quantum statistical models: Equivalent conditions for sufficiency. *Commun. Math. Phys.* **2012**, *310*, 625–647. [CrossRef]
59. Ohya, M.; Watanabe, N. Quantum entropy and its applications to quantum communication and statistical physics. *Entropy* **2010**, *12*, 1194–1245. [CrossRef]
60. Rauh, J.; Banerjee, P.K.; Olbrich, E.; Jost, J.; Bertschinger, N.; Wolpert, D. Coarse-Graining and the Blackwell Order. *Entropy* **2017**, *19*, 527. [CrossRef]
61. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*; John Wiley & Sons: Hoboken, NJ, USA, 2006.
62. Makur, A.; Polyanskiy, Y. Comparison of channels: Criteria for domination by a symmetric channel. *IEEE Trans. Inf. Theory* **2018**, *64*, 5704–5725. [CrossRef]
63. Benson, H.P. Concave minimization: Theory, applications and algorithms. In *Handbook of Global Optimization*; Springer: Berlin/Heidelberg, Germany, 1995; pp. 43–148.
64. Kolchinsky, A. Code for Computing $I_{\cap}^{\prec}$. 2022. Available online: https://github.com/artemyk/redundancy (accessed on 3 January 2022).
65. Banerjee, P.K.; Rauh, J.; Montúfar, G. Computing the unique information. In Proceedings of the 2018 IEEE International Symposium on Information Theory (ISIT), Vail, CO, USA, 17–22 June 2018 IEEE: Piscataway, NJ, USA, 2018; pp. 141–145.
66. Banerjee, P.K.; Olbrich, E.; Jost, J.; Rauh, J. Unique informations and deficiencies. In Proceedings of the 2018 56th Annual Allerton Conference on Communication, Control, and Computing (Allerton), Monticello, IL, USA, 2–5 October 2018 IEEE: Piscataway, NJ, USA, 2018; pp. 32–38.
67. Wolf, S.; Wultschleger, J. Zero-error information and applications in cryptography. In Proceedings of the Information Theory Workshop, San Antonio, TX, USA, 24–29 October 2004; IEEE: Piscataway, NJ, USA, 2004; pp. 1–6.
68. Bertschinger, N.; Rauh, J.; Olbrich, E.; Jost, J. Shared information - new insights and problems in decomposing information in complex systems. In *Proceedings of the European Conference on Complex Systems 2012*; Springer: Berlin/Heidelberg, Germany, 2013; pp. 251–269.
69. James, R.G.; Ellison, C.J.; Crutchfield, J.P. dit: A Python package for discrete information theory. *J. Open Source Softw.* **2018**, *3*, 738. [CrossRef]
70. Kovačević, M.; Stanojević, I.; Šenk, V. On the entropy of couplings. *Inf. Comput.* **2015**, *242*, 369–382. [CrossRef]
71. Horst, R. On the global minimization of concave functions. *Oper.-Res.-Spektrum* **1984**, *6*, 195–205. [CrossRef]
72. Pardalos, P.M.; Rosen, J.B. Methods for global concave minimization: A bibliographic survey. *Siam Rev.* **1986**, *28*, 367–379. [CrossRef]
73. Williams, P.L.; Beer, R.D. Generalized measures of information transfer. *arXiv* **2011**, arXiv:1102.1507.
74. Dubins, L.E. On extreme points of convex sets. *J. Math. Anal. Appl.* **1962**, *5*, 237–244. [CrossRef]
75. Yeung, R.W. *A First Course in Information Theory*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2012.
76. Lewis, A.D. Semicontinuity of Rank and Nullity and Some Consequences. 2009. Available online: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.709.7290&rep=rep1&type=pdf (accessed on 3 January 2022).
77. Hoffman, A.J. On Approximate Solutions of Systems of Linear Inequalities. *J. Res. Natl. Bur. Stand.* **1952**, *49*, 174–176. [CrossRef]
78. Daniel, J.W. On Perturbations in Systems of Linear Inequalities. *SIAM J. Numer. Anal.* **1973**, *10*, 299–307. [CrossRef]