

Article

Not peer-reviewed version

---

# A Measure of Synergy Based on Union Information

---

[André F. C. Gomes](#)<sup>\*</sup> and [Mário A. T. Figueiredo](#)

Posted Date: 22 January 2024

doi: 10.20944/preprints202401.1531.v1

Keywords: information theory; partial information decomposition; union information; synergy; communication channels.; mutual information



Preprints.org is a free multidiscipline platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Article

# A Measure of Synergy Based on Union Information

André F. C. Gomes <sup>1,†</sup>  and Mário A. T. Figueiredo <sup>2,†</sup> 

Instituto de Telecomunicações and LUMILIS (Lisbon ELLIS Unit), Instituto Superior Técnico, Universidade de Lisboa, Portugal; mario.figueiredo@tecnico.ulisboa.pt

\* Correspondence: andrefcgomes@tecnico.ulisboa.pt

**Abstract:** The *partial information decomposition* (PID) framework is concerned with decomposing the information that a set of (two or more) random variables (the sources) has about another variable (the target) into three types of information: unique, redundant, and synergistic. Classical information theory alone does not provide a unique way to decompose information in this manner and additional assumptions have to be made. One often overlooked way to do this decomposition is using a so-called measure of union information – which quantifies the information that is present in at least one of the sources – from which a synergy measure stems. In this paper, we introduce a new measure of union information based on adopting a communication channel perspective, compare it with existing measures, and study some of its properties. We also include a comprehensive critical review of characterizations of union information and synergy measures that have been proposed in the literature.

**Keywords:** information theory; partial information decomposition; union information; synergy; communication channels

## 1. Introduction

Williams and Beer [1] introduced the *partial information decomposition* (PID) framework as a way to characterize, or analyze, the information that a set of random variables (often called *sources*) has about another variable (referred to as the *target*). PID is a useful tool for gathering insights and analyzing the way information is stored, modified, and transmitted within complex systems [2], [3]. It has been applied in several areas such as cryptography [4] and neuroscience [5,6], with many other potential use cases, such as in studying information flows in gene regulatory networks [7], neural coding [8], financial markets [9], and network design [10].

Consider the simplest case: a three-variable joint distribution  $p(y_1, y_2, t)$  describing three random variables: two so-called sources,  $Y_1$  and  $Y_2$ , and a target  $T$ . Notice that, despite what the names *sources* and *target* might suggest, there is no directionality (causal or otherwise) assumption. The goal of PID is to *decompose* the information that the sources  $Y = (Y_1, Y_2)$  have about  $T$  into the sum of 4 non-negative quantities: the information that is present in both  $Y_1$  and  $Y_2$ , known as *redundant* information,  $R$ ; the information that only  $Y_1$  (respectively  $Y_2$ ) has about  $T$ , known as *unique* information,  $U_1$  (respectively  $U_2$ ); the *synergistic* information,  $S$ , that is present in the pair  $(Y_1, Y_2)$  but not in  $Y_1$  or  $Y_2$  alone. In this case with two variables, the goal is thus to write

$$I(T; Y) = R + U_1 + U_2 + S, \quad (1)$$

where  $I(T; Y)$  is the mutual information between  $T$  and  $Y$  [11]. The redundant information  $R$ , because it is present in both  $Y_1$  and  $Y_2$ , is also referred to as *intersection* information and denoted as  $I_{\cap}$ . Finally,  $I_{\cup}$  refers to *union* information, *i.e.*, the amount of information provided by at least one of the sources; in the case of two sources,  $I_{\cup} = U_1 + U_2 + R$ , thus  $S = I(T; Y) - I_{\cup}$ .

Because unique information and redundancy satisfy the relationship  $U_i = I(T; Y_i) - R$  (for  $i \in \{1, 2\}$ ), it turns out that defining how to compute one of these quantities ( $R$ ,  $U_i$ , or  $S$ ) is enough to fully determine the others [1]. Williams and Beer [1] suggested a set of axioms that a measure of redundancy should satisfy, and proposed a measure of their own. Those axioms became well known

as the Williams-Ber axioms, although the measure they proposed has subsequently been criticized for not capturing informational content, but only information *size* [12]. It is worth noting that, as the number of variables grows, the number of terms appearing in the PID of  $I(T; Y)$  grows super exponentially [13].

Stimulated by that initial work, other measures of information and other sets of axioms for information decomposition have been introduced; see, for example, the work by Bertschinger *et al.* [14], Griffith and Koch [15], and James *et al.* [16], for different measures of redundant, unique, and synergistic information. There is no consensus about what axioms any measure should satisfy or whether a given measure *captures the information* that it should capture, except for the Williams-Ber axioms. Today, there is still debate about what axioms different measures of information should satisfy, and there is no general agreement on what is an appropriate PID [16–20].

Most PID measures that have been suggested thus far are either measures of redundant information [1,12,20–25] or measures of unique information [14,16]. Alternatively, it is possible to define *union information* of a set of sources as the amount of information provided by at least one of those sources. Synergy is then defined as the difference between the total information and union information [21].

In this paper, we introduce a new measure of union information based on the information channel perspective that we already pursued in earlier work [25] and study some of its properties. The resulting measure leads to a novel information decomposition that is particularly suited for analyzing how information is distributed in channels.

The rest of the paper is organized as follows. A final subsection of this section introduces the notation used throughout the paper. In Section 2, we recall some properties of PID and take a look at how the degradation measure for redundant information introduced by Kolchinsky [21] decomposes information in bivariate systems, while also pointing out some drawbacks of that measure. Section 3 presents the motivation for our proposed measure, its operational interpretation, its multivariate definition, as well as some of its drawbacks. In Section 4, we propose an extension of the Williams-Ber axioms for measures of union information and show that our proposed measure satisfies those axioms. We review all properties that have been proposed both for measures of union information and synergy, and either accept or reject them. We also compare different measures of synergy and relate them, whenever possible. Finally, Section 5 presents concluding remarks and suggestions for future work.

### 1.1. Notation

For two discrete random variables  $X \in \mathcal{X}$  and  $Z \in \mathcal{Z}$ , their Shannon mutual information  $I(X; Z)$  is given by  $I(X; Z) = I(Z; X) = H(X) - H(X|Z) = H(Z) - H(Z|X)$ , where  $H(X) = -\sum_{x \in \mathcal{X}} p(x) \log p(x)$  and  $H(X|Z) = -\sum_{x \in \mathcal{X}} \sum_{z \in \mathcal{Z}} p(x, z) \log p(x|z)$  are the entropy and conditional entropy, respectively [11]. The conditional distribution  $p(z|x)$  corresponds, in an information-theoretical perspective, to a discrete memoryless channel with a channel matrix  $K$ , *i.e.*, such that  $K[x, z] = p(z|x)$  [11]. This matrix is row-stochastic:  $K[x, z] \geq 0$ , for any  $x \in \mathcal{X}$  and  $z \in \mathcal{Z}$ , and  $\sum_{z \in \mathcal{Z}} K[x, z] = 1$ , for any  $x$ .

Given a set of  $n$  discrete random variables (sources),  $Y_1 \in \mathcal{Y}_1, \dots, Y_n \in \mathcal{Y}_n$ , and a discrete random variable  $T \in \mathcal{T}$  (target) with joint distribution (probability mass function)  $p(y_1, \dots, y_n, t)$ , we consider the channels  $K^{(i)}$  between  $T$  and each  $Y_i$ , that is, each  $K^{(i)}$  is a  $|\mathcal{T}| \times |\mathcal{Y}_i|$  row-stochastic matrix with the conditional distribution  $p(y_i|t)$ .

We say that three random variables, say  $X, Y, Z$ , form a Markov chain (which we denote by  $X - Y - Z$  or by  $X \perp Z | Y$ ) if  $X$  and  $Z$  are conditionally independent, given  $Y$ .

## 2. Background

### 2.1. PID Based of Channel Orders

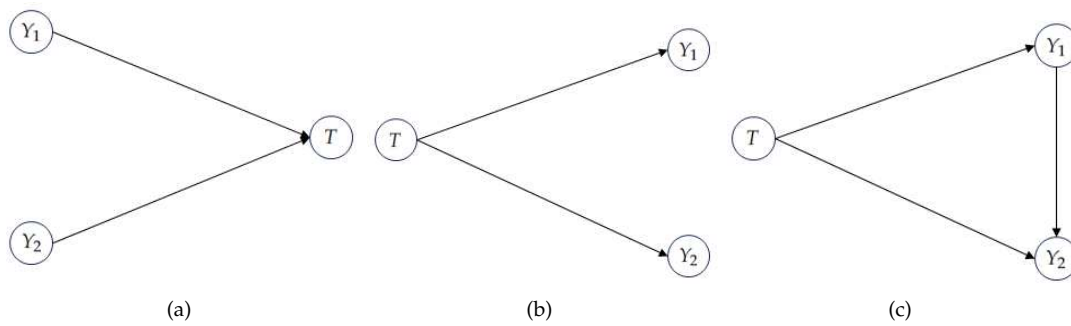
In its current versions, PID is agnostic to causality in the sense that, like mutual information, it is an undirected measure, *i.e.*,  $I(T; Y) = I(Y; T)$ . Some measures indirectly presuppose some kind of directionality to perform PID. Take for instance the redundancy measure introduced by Kolchinsky [21], based on the so-called degradation order  $\preceq_d$  between communication channels (see recent work by Kolchinsky [21] and Gomes and Figueiredo [25] for definitions):

$$I_{\cap}^d(Y_1, Y_2, \dots, Y_m \rightarrow T) := \sup_{K^Q: K^Q \preceq_d K^{(i)}, i \in \{1, \dots, m\}} I(Q; T). \quad (2)$$

When computing the information shared by the  $m$  sources,  $I_{\cap}^d(Y_1, Y_2, \dots, Y_m \rightarrow T)$ , the perspective is that there is a channel with a single input  $T$  and  $m$  outputs  $Y_1, \dots, Y_m$ . This definition of  $I_{\cap}^d$  corresponds to the mutual information of the most informative channel, under the constraint that this channel is dominated (in the degradation order  $\preceq_d$  sense) by all channels  $K^{(1)}, \dots, K^{(m)}$ . Since mutual information was originally introduced to formalize the capacity of communication channels, it is not surprising that measures that presuppose channel directionality are found useful in this context.

Although it is not guaranteed that the structure of the joint distribution  $p(y_1, \dots, y_n, t)$  is compatible with the causal model of single input, and multiple output channels (which implies that the sources are conditionally independent, given  $T$ ), one may always compute such measures, which have interesting and relevant operational interpretations. In the context of PID, where the goal is to study how information is decomposed, such measures provide an excellent starting point. Although it is not guaranteed that there is actually a channel (or a direction) from  $T$  to  $Y_i$ , we can characterize how information about  $T$  is spread through the sources. In the case of the degradation order,  $I_{\cap}^d$  provides insight about which variable  $Y_i$  to observe in order to obtain the most information about  $T$ , under the constraint that only one  $Y_i$  can be observed.

Arguably, the most common scenario in PID is finding out something about the structure of the information the variables  $Y_1, \dots, Y_n$  have about  $T$ . In a particular system of variables characterized by its joint distribution, we do not make causal assumptions, so we can adopt the perspective that the variables  $Y_i$  are functions of  $T$ , hence obtaining the channel structure. Although it might not be faithful to the conditional independence properties implied by  $p(y_1, \dots, y_n, t)$ , this approach allows decomposing  $I(Y; T)$  and drawing conclusions about the inner structure of the information that  $Y$  has about  $T$ . Some distributions, however, *cannot* have this causal structure. Take for instance the distribution generated by  $T = Y_1 \text{ xor } Y_2$ , where  $Y_1$  and  $Y_2$  are two equiprobable and independent binary random variables. We will call this distribution XOR. For this well-known distribution, we have  $Y_1 \perp Y_2$  and  $Y_1 \not\perp Y_2|T$ , whereas the implied channel distribution that  $I_{\cap}^d$  assumes yields the exact opposite dependencies, that is,  $Y_1 \not\perp Y_2$  and  $Y_1 \perp Y_2|T$ . See Figure 1 for more insight.



**Figure 1.** (a) Assuming faithfulness [26], this is the only three-variable directed acyclic graph (DAG) that satisfies  $Y_1 \perp Y_2$  and  $Y_1 \not\perp Y_2|T$ , in general [26]. (b) The DAG that is "implied" by the perspective of  $I_{\cap}^d$ . (c) A DAG that can generate the XOR distribution, but doesn't satisfy the dependencies implied by  $T = Y_1 \text{ xor } Y_2$ . In fact, any DAG that is in the same Markov equivalence class as (c) can generate the XOR distribution (or any other joint distribution), but none satisfy the earlier dependencies, assuming faithfulness.

Consider the computation of  $I_{\cap}^d(Y_1, Y_2 \rightarrow T)$  for the XOR distribution. This measure argues that, since

$$K^{(1)} = K^{(2)} = \begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{bmatrix},$$

then a solution to  $I_{\cap}^d(Y_1, Y_2 \rightarrow T)$  is given by the channel  $K^Q = K^{(1)}$  and redundancy is computed as  $I(Q; T)$ , yielding 0 bits of redundancy, and consequently 1 bit of synergy (as computed from (1)). Under this channel perspective (as in Figure 1 (b)),  $I_{\cap}^d$  is not concerned with, for example  $p(Y_1|T, Y_2)$  or  $p(Y_1, Y_2)$ . If all that is needed to compute redundancy is  $p(T)$ ,  $K^{(1)}$  and  $K^{(2)}$ , this would lead to the wrong conclusion that the outcome  $(Y_1, Y_2, T) = (0, 0, 1)$  has non-null probability, which it does not. With this, we don't mean that  $I_{\cap}^d$  is an incomplete or incorrect measure to do PID, we are using its insights to point us in a different direction.

## 2.2. PID Based on Redundancy Measures

At this point, the most often used approaches to PID are based on redundancy measures. Usually, these are in one of the two following classes:

- Measures that are not concerned with information content, only *information size*, which makes them easy to compute even for distributions with many variables, but at the cost that the resulting decompositions may not give realistic insights into the system, precisely because they are not sensitive to informational content. Examples are  $I_{\cap}^{\text{WB}}$  [1] or  $I_{\cap}^{\text{MMI}}$  [22], and applications of these can be found in [27–30].
- Measures that satisfy the Blackwell property - which arguably do measure information content - but are insensitive to changes in the sources' distribution  $p(y_1, \dots, y_n) = \sum_t p(y_1, \dots, y_n, t)$  (as long as  $p(T), K^{(1)}, \dots, K^{(n)}$  remain the same). Examples are  $I_{\cap}^d$  [21] (see Equation (2)) or  $I_{\cap}^{\text{BROJA}}$  [14]. It should be noted that  $I_{\cap}^{\text{BROJA}}$  is only defined for the bivariate case, that is, for distributions with at most two sources, described by  $p(y_1, y_2, t)$ . Applications of these can be found in [31–33].

Particularly  $I_{\cap}^d$  and  $I_{\cap}^{\text{BROJA}}$  satisfy the so-called (\*) assumption [14], which argues that redundant and unique information should only depend on the marginal distribution of the target  $p(T)$  and on the conditional distributions of the sources given the target, that is, on the stochastic matrices  $K^{(i)}$ . James *et al.* [16, Section 4] and Ince [20, Section 5], provide great arguments as to why the (\*) assumption should not hold in general, and we agree with them.

Towards motivating a new PID, let us look at how  $I_{\cap}^d$  decomposes information in the bivariate case. Any measure that is based on a preorder between channels and which satisfies Kolchinsky's

axioms yields similar decompositions [25], thus there is no loss of generality in focusing on  $I_{\cap}^d$ . We next analyze three different cases.

- Case 1: there is an ordering between the channels, that is, w.l.o.g.,  $K^{(2)} \preceq_d K^{(1)}$ . This means that  $I(Y_2; T) \leq I(Y_1; T)$  and the decomposition (as in (1)) is given by  $R = I(Y_2; T)$ ,  $U_2 = 0$ ,  $U_1 = I(Y_1; T) - I(Y_2; T)$  and  $S = I(Y; T) - I(Y_1; T)$ . Moreover, if  $K^{(1)} \not\preceq_d K^{(2)}$ , then  $S = 0$ .

As an example, consider the leftmost distribution in Table 1, which satisfies  $T = Y_1$ . In this case,

$$K^{(1)} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \succeq_d K^{(2)} = \begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{bmatrix},$$

yielding  $R = 0$ ,  $U_1 = 1$ ,  $U_2 = 0$  and  $S = 0$ , as expected, because  $T = Y_1$ .

**Table 1.** Three joint distributions  $p(y_1, y_2, y_2)$  used to exemplify the three cases. Left: joint distribution satisfying  $T = Y_1$ . Middle: distribution satisfying  $T = (Y_1, Y_2)$ , known as the COPY distribution. Right: the so-called BOOM distribution (see text).

$t$	$y_1$	$y_2$	$p(t, y_1, y_2)$
0	0	0	0.25
0	0	1	0.25
1	1	0	0.25
1	1	1	0.25
$t$	$y_1$	$y_2$	$p(t, y_1, y_2)$
(0,0)	0	0	0.25
(0,1)	0	1	0.25
(1,0)	1	0	0.25
(1,1)	1	1	0.25
$t$	$y_1$	$y_2$	$p(t, y_1, y_2)$
0	0	2	1/6
1	0	0	1/6
1	1	2	1/6
2	0	0	1/6
2	2	0	1/6
2	2	1	1/6

- Case 2: there is no ordering between the channels and the solution of  $I_{\cap}^d(Y_1, Y_2 \rightarrow T)$  is a trivial channel, in the sense that it has no information about  $T$ . The decomposition is given by  $R = 0$ ,  $U_2 = I(Y_2; T)$ ,  $U_1 = I(Y_1; T)$  and  $S = I(Y; T) - I(Y_1; T) - I(Y_2; T)$ , which may lead to a negative value of synergy.

As an example, consider the COPY distribution with  $Y_1$  and  $Y_2$  i.i.d. Bernoulli variables with parameter 0.5, shown in the center of Table 1. In this case, channels  $K^{(1)}$  and  $K^{(2)}$  have the form

$$K^{(1)} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}, \quad K^{(2)} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix},$$

with no degradation order between them. This yields  $R = 0$ ,  $U_1 = U_2 = 1$  and  $S = 0$ .

- Case 3: there is no ordering between the channels and  $I_{\cap}^d(Y_1, Y_2 \rightarrow T)$  is achieved by a nontrivial channel  $K^Q$ . The decomposition is given by  $R = I(Q; T)$ ,  $U_2 = I(Y_2; T) - I(Q; T)$ ,  $U_1 = I(Y_1; T) - I(Q; T)$  and  $S = I(Y; T) + I(Q; T) - I(Y_1; T) - I(Y_2; T)$ .

As an example, consider the BOOM distribution [16], shown on the right side of Table 1. In this case, channels  $K^{(1)}$  and  $K^{(2)}$  are

$$K^{(1)} = \begin{bmatrix} 1 & 0 & 0 \\ 1/2 & 1/2 & 0 \\ 1/3 & 0 & 2/3 \end{bmatrix}, \quad K^{(2)} = \begin{bmatrix} 0 & 0 & 1 \\ 1/2 & 0 & 1/2 \\ 2/3 & 1/3 & 0 \end{bmatrix},$$

and there is no degradation order between them. However, there is a nontrivial channel  $K^Q$  that is dominated by both  $K^{(1)}$  and  $K^{(2)}$  that maximizes  $I(Q; T)$ . One of its versions is

$$K^Q = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 3/4 & 1/4 \\ 1/3 & 1/3 & 1/3 \end{bmatrix},$$

yielding  $R \approx 0.322$ ,  $U_1 = U_2 \approx 0.345$  and  $S \approx 0.114$ .

This class of approaches has some limitations, as all PID measures do. In the bivariate case, the definition of synergy  $S$  from a measure of redundant information is the completing term such that  $I(Y; T) = S + R + U_1 + U_2$  holds. The definition of  $I_{\cap}^d$  supports the argument that if  $K^{(2)} \preceq_d K^{(1)}$  and  $K^{(1)} \not\preceq_d K^{(2)}$ , then there is no synergy. This makes intuitive sense because, in this case,  $T - Y_1 - Y_2$  is a Markov chain (see Section 1.1 for the definition), consequently,  $I(Y; T) = I(Y_1; T)$ , that is,  $Y_1$  has the same information about  $T$  as the pair  $Y = (Y_1, Y_2)$ .

If there is no  $\preceq_d$  ordering between the channels, as in the COPY distribution (Table 1, middle), the situation is more complicated. We saw that the decomposition for this distribution yields  $R = 0$ ,  $U_1 = U_2 = 1$  and  $S = 0$ . However, suppose we change the distribution such that  $T = (1, 1)$  has probability 0 and the other outcomes have probability  $1/3$ . For example, consider the distribution in Table 2. For this distribution, we have  $I(Y; T) \approx 1.585$ . Intuitively, we would expect that  $I(Y; T)$  would be decomposed as  $R = 0$ ,  $U_1 = U_2 = I(Y; T)/2$  and  $S = 0$ , just as before, so that the proportions  $U_i/I(Y; T)$ , for  $i \in \{1, 2\}$ , in both distributions remain the same, whereas redundancy and synergy would remain zero. That is, we don't expect that removing one of the outcomes while maintaining the remaining outcomes equiprobable would change the types of information in the system. However, if we do so and compute the decomposition yielded by  $I_{\cap}^d$ , we obtain  $R = 0$ ,  $U_1 = U_2 = 0.918 \neq I(Y; T)/2$  and  $S = -0.251$ , *i.e.*, a negative synergy, arguably meaningless.

**Table 2.** Tweaked COPY distribution, now without the outcome  $(T, Y_1, Y_2) = ((1, 1), 1, 1)$ .

T	$Y_1$	$Y_2$	$p(t, y_1, y_2)$
(0,0)	0	0	1/3
(0,1)	0	1	1/3
(1,0)	1	0	1/3

There are still many open questions in PID. One of those questions is: should measures of redundant information be used to measure synergy, given that they compute it as the completing term in equation (1). We agree that using a measure of redundant information to compute the synergy in this way may not be appropriate, especially because the *inclusion-exclusion principle* (IEP) should not necessarily hold in the context of PID. See [21] for comments on the IEP.

With these motivations, we propose a measure of union information for PID that shares with  $I_{\cap}^d$  the implicit view of channels. However, unlike  $I_{\cap}^d$  and  $I_{\cap}^{\text{BROJA}}$  – which satisfy the (\*) assumption and thus are not concerned with the conditional dependencies in  $p(y_i|t, y_{-i})$  – our measure defines synergy as the information that cannot be computed from  $p(y_i|t)$ , but can be computed from  $p(y_i|t, y_{-i})$ . That is,

we propose that synergy be computed as the information that is not captured by assuming conditional independence of the sources, given the target.

### 3. A New Measure of Union Information

#### 3.1. Motivation and Bivariate Definition

Consider a distribution  $p(y_1, y_2, t)$  and suppose there are two agents, agent 1 and agent 2, whose goal is to reduce their uncertainty about  $T$  by observing  $Y_1$  and  $Y_2$ , respectively. Suppose also that the agents know  $p(t)$ , and that agent  $i$  has access to its channel distribution  $p(y_i|t)$ . Many PID measures make this same assumption, including  $I_{\cap}^d$ . When agent  $i$  works alone to reduce the uncertainty about  $T$ , since it has access to  $p(t)$  and  $p(y_i|t)$ , it also knows  $p(y_i)$  and  $p(y_i, t)$ , which allows it to compute  $I(Y_i; T)$ : the amount of uncertainty reduction about  $T$  achieved by observing  $Y_i$ .

Now, if the agents can work together, that is, if they have access to  $Y = (Y_1, Y_2)$ , then they can compute  $I(Y; T)$ , because they have access to  $p(y_1, y_2|t)$  and  $p(t)$ . On the other hand, if the agents are not able to work together (in the sense that they are not able to observe  $Y$  together, but only  $Y_1$  and  $Y_2$ , separately) yet can communicate, then they can construct a different distribution  $q$  given by  $q(y_1, y_2, t) := p(t)p(y_1|t)p(y_2|t)$ , i.e., a distribution under which  $Y_1$  and  $Y_2$  are conditionally independent given  $T$ , but have the same marginal  $p(t)$  and the same individual conditionals  $p(y_1|t)$  and  $p(y_2|t)$ .

The form of  $q$  in the previous paragraph should be contrasted with the following factorization of  $p$  which entails no conditional independence assumption:  $p(y_1, y_2, t) = p(t)p(y_1|t)p(y_2|t, y_1)$ . In this sense, we would propose to define union information, for the bivariate case, as follows

$$\begin{aligned} I_{\cup}(Y_1 \rightarrow T) &= I_q(Y_1; T) = I_p(Y_1; T), \\ I_{\cup}(Y_2 \rightarrow T) &= I_q(Y_2; T) = I_p(Y_2; T), \\ I_{\cup}(Y_1, Y_2 \rightarrow T) &= I_q(Y; T), \\ I_{\cup}((Y_1, Y_2) \rightarrow T) &= I_p(Y; T), \end{aligned} \quad (3)$$

where the subscript refers to the distribution under which the mutual information is computed. From this point forward, the absence of a subscript means that the computation is done under the true distribution  $p$ . As we will see, this is not yet the final definition, for reasons to be addressed below.

Using the definition of synergy derived from a measure of union information [21], for the bivariate case we have

$$S(Y_1, Y_2 \rightarrow T) := I(Y; T) - I_{\cup}(Y_1, Y_2 \rightarrow T). \quad (4)$$

Synergy is often posited as *the difference between the whole and the union of the parts*. For our measure of union information, the 'union of the parts' corresponds to the reduction of uncertainty about  $T$  - under  $q$  - that agents 1 and 2 can obtain by sharing their conditional distributions. Interestingly, there are cases where the union of the parts is better than the whole, in the sense that  $I_{\cup}(Y_1, Y_2 \rightarrow T) > I(Y; T)$ . An example of this is given by the *Adapted ReducedOR* distribution, originally introduced by Ince [20] and adapted by James *et al.* [16], which is shown in the left side of Table 3, where  $r \in [0, 1]$ . This distribution is such that  $I_q(Y; T)$  does not depend on  $r$  ( $I_q(Y; T) \approx 0.549$ ), since neither  $p(t)$  nor  $p(y_1|t)$  and  $p(y_2|t)$  depend on  $r$ ; consequently,  $q(t, y_1, y_2)$  also does not depend on  $r$ , as show in the right side of Table 3.

**Table 3.** Left: the *Adapted ReducedOR* distribution, where  $r \in [0, 1]$ . Right: the corresponding distribution  $q(y, y_1, y_2) = p(t)p(y_1|t)p(y_2|t)$ .

t	$y_1$	$y_2$	$p(t, y_1, y_2)$
0	0	0	0.5
1	0	0	$r/4$
1	1	0	$(1-r)/4$
1	0	1	$(1-r)/4$
1	1	1	$r/4$
t	$y_1$	$y_2$	$q(t, y_1, y_2)$
0	0	0	0.5
1	0	0	1/8
1	1	0	1/8
1	0	1	1/8
1	1	1	1/8

It can be easily shown that if  $r > 0.5$ , then  $I_q(Y; T) > I(Y; T)$ , which implies that synergy, if defined as in (4), could be negative. How do we interpret the fact that there exist distributions such that  $I_q(Y; T) > I(Y; T)$ ? This means that under distribution  $q$ , which assumes  $Y_1$  and  $Y_2$  are conditionally independent given  $T$ ,  $Y_1$  and  $Y_2$  reduce the uncertainty about  $T$  more than in the original distribution. Arguably, the parts working independently and achieving better results than the whole should mean there is no synergy, as opposed to negative synergy.

The observations in the previous paragraphs motivate our definition of a new measure of union information as

$$I_{\cup}^{\text{CI}}(Y_1, Y_2 \rightarrow T) := \min\{I(Y; T), I_q(Y; T)\}, \quad (5)$$

with the superscript CI standing for *conditional independence*, yielding a non-negative synergy:

$$S^{\text{CI}}(Y_1, Y_2 \rightarrow T) = I(Y; T) - I_{\cup}^{\text{CI}}(Y_1, Y_2 \rightarrow T) = \max\{0, I(Y; T) - I_q(Y; T)\}. \quad (6)$$

Note that, for the bivariate case, we have 0 synergy if  $p(t, y_2, y_2)$  is such that  $Y_1 \perp_p Y_2|T$ , that is, if the outputs are indeed conditionally independent given  $T$ . Moreover,  $I_{\cup}^{\text{CI}}$  satisfies the monotonicity axiom from the extension of the Williams-Beer axioms to measures of union information (to be mentioned in Section 4.1), which further supports this definition.

### 3.2. Operational Interpretation

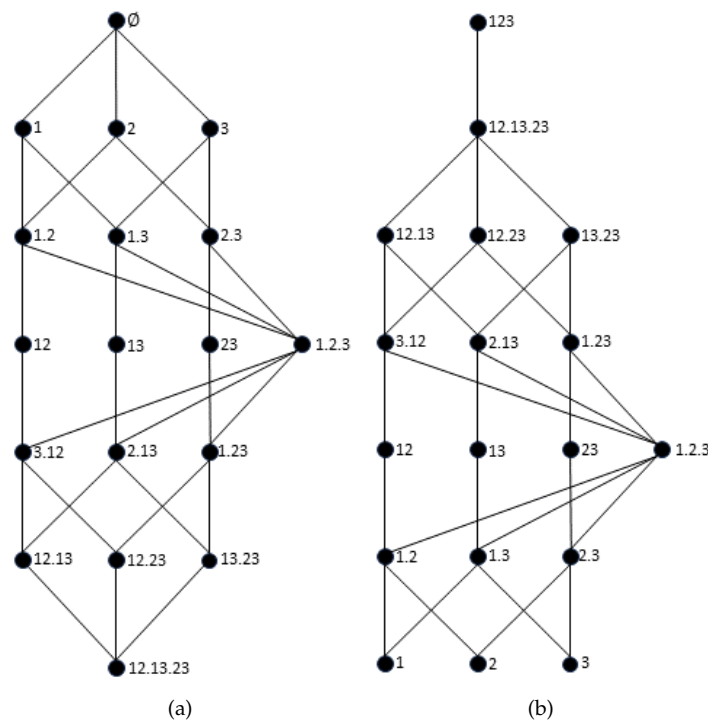
For the bivariate case, if  $Y_1$  and  $Y_2$  are conditionally independent given  $T$  (Figure 1 (b)), then  $p(y_1|t)$  and  $p(y_2|t)$  (and  $p(t)$ ) suffice to reconstruct the original joint distribution  $p(y_1, y_2, t)$ , which means the union of the parts is enough to reconstruct the whole, *i.e.*, there is no synergy between  $Y_1$  and  $Y_2$ . Conversely, a distribution generated by the DAG in Figure 1 (c) does not satisfy conditional independence (given  $T$ ), hence we expect positive synergy, as is the case for the XOR distribution, and indeed our measure yields 1 bit of synergy for this distribution. These two cases motivate the operational interpretation of our measure of synergy: it is the amount of information that is not captured by assuming conditional independence of the sources (given the target).

Recall, however, that some distributions are such that  $I_q(Y; T) > I_p(Y; T)$ , *i.e.*, such that the union of the parts ‘outperforms’ the whole. What does this mean? It means that under  $q$ ,  $Y_1$  and  $Y_2$  have more information about  $T$  than under  $p$ : the constructed distribution  $q$ , which drops the conditional dependence of  $Y_1$  and  $Y_2$  given  $T$ , reduces the uncertainty that  $Y$  has about  $T$  more than the original distribution  $p$ . In some cases, this may happen because the support of  $q$  is larger than that of  $p$ , which may lead to a reduction of uncertainty under  $q$  that cannot be achieved under  $p$ . In these cases, since we are decomposing  $I_p(Y; T)$ , we revert to saying that the union information that a set of variables has

about  $T$  is equal to  $I_p(Y; T)$ , so that our measure satisfies the monotonicity axiom (later introduced in Definition 2). We will comment on this compromise between satisfying the monotonicity axiom and ignoring dependencies later.

### 3.3. General (Multivariate) Definition

To extend the proposed measure to an arbitrary number  $n \geq 2$  of sources, we briefly recall the synergy lattice [17,34] and the union information semi-lattice [34]. For  $n = 3$ , these two lattices are shown in Figure 2. For the sake of brevity, we will not address the construction of the lattices or the different orders between sources. We refer the reader to the work of Gutknecht *et al.* [34], for an excellent overview of the different lattices, the orders between sources, and the construction of different PID measures.



**Figure 2.** Trivariate distribution lattices and their respective ordering of sources. Left: synergy lattice [17]. Right: union information semi-lattice [34].

In the following, we use the term *source* to mean a subset of the variables  $\{Y_1, \dots, Y_n\}$ , or a set of such subsets, we drop the curly brackets for clarity, and refer to the different variables by their indices, as is common in most works on PID. The decomposition resulting from a measure of union information is not as direct to obtain as one obtained from a measure of redundant information, as the solution for the information atoms is not a Möbius inversion [13]. One must first construct the measure of synergy for source  $\alpha$  by writing

$$S^{\text{CI}}(\alpha \rightarrow T) = I(Y; T) - I_{\cup}^{\text{CI}}(\alpha \rightarrow T), \quad (7)$$

which is the generalization of (6) for an arbitrary *source*  $\alpha$ . In the remainder of this paper, we will often omit " $\rightarrow T$ " from the notation (unless it is explicitly needed), with the understanding that the target variable is always referred to as  $T$ . Also for simplicity, in the following, we identify the different agents that have access to different distributions as the distributions they have access to.

It is fairly simple to extend the proposed measure to an arbitrary number of sources, as illustrated in the following two examples.

**Example 1:** to compute  $I_{\cup}^{\text{CI}}((Y_1, Y_2), Y_3)$ , agent  $(Y_1, Y_2)$  knows  $p(y_1, y_2|t)$ , thus it can also compute, by marginalization,  $p(y_1|t)$  and  $p(y_2|t)$ . On the other hand, agent  $Y_3$  only knows  $p(y_3|t)$ . Recall that both agents also have access to  $p(t)$ . By sharing their conditionals, the agents can compute  $q_1(y_1, y_2, y_3, t) := p(t)p(y_1, y_2|t)p(y_3|t)$ , and also  $q_2(y_1, y_2, y_3, t) := p(t)p(y_1|t)p(y_2|t)p(y_3|t)$ . After this, they may choose whichever distribution has the highest information about  $T$ , while still holding the view that any information gain larger than  $I(Y; T)$  must be disregarded. Consequently, we write

$$I_{\cup}^{\text{CI}}((Y_1, Y_2), Y_3) = \min \left\{ I(Y; T), \max \{ I_{q_1}(Y; T), I_{q_2}(Y; T) \} \right\}.$$

**Example 2:** slightly more complicated is the computation of  $I_{\cup}^{\text{CI}}((Y_1, Y_2), (Y_1, Y_3), (Y_2, Y_3))$ . In this case, the three agents may compute four different distributions, two of which are the same  $q_1$  and  $q_2$  defined in the previous paragraph, and the other two are  $q_3(y_1, y_2, y_3, t) := p(t)p(y_1, y_3|t)p(y_2|t)$ , and  $q_4(y_1, y_2, y_3, t) := p(t)p(y_1|t)p(y_2, y_3|t)$ .

Given these insights, we propose the following measure of union information.

**Definition 1.** Let  $A_1, \dots, A_m$  be an arbitrary collection of  $m \geq 1$  sources (recall sources may be subsets of variables). Without loss of generality, assume that no source is a subset of another source and no source is a deterministic function of other sources. We define

$$I_{\cup}^{\text{CI}}(A_1, \dots, A_m \rightarrow T) = \min \left\{ I(A; T), \max_{q \in \mathcal{Q}} I_q(A; T) \right\},$$

where  $A = \bigcup_{i=1}^m A_i$  and  $\mathcal{Q}$  is the set of all different distributions that the  $m$  agents can construct by combining their conditional distributions and marginalizations thereof.

For instance, in Examples 1 above,  $A = \{Y_1, Y_2\} \cup \{Y_3\} = \{Y_1, Y_2, Y_3\}$ ; in Example 2,  $A = \{Y_1, Y_2\} \cup \{Y_1, Y_3\} \cup \{Y_2, Y_3\} = \{Y_1, Y_2, Y_3\}$ . In Example 1,  $\mathcal{Q} = \{q_1, q_2\}$ , whereas in Example 2,  $\mathcal{Q} = \{q_1, q_2, q_3, q_4\}$ .

We now justify the conditions in Definition 1 and the fact that they do not entail any loss of generality.

- The condition that no source is a subset of another source (which also excludes the case where two sources are the same) implies no loss of generality: if one source is a subset of another, say  $A_i \subseteq A_j$ , then  $A_i$  may be removed without affecting either  $A$  or  $\mathcal{Q}$ , thus yielding the same value for  $I_{\cup}^{\text{CI}}$ . The removal of source  $A_i$  is also done for measures of intersection information, but under the opposite condition: whenever  $A_j \subseteq A_i$ .
- The condition that no source is a deterministic function of other sources is slightly more nuanced. In our perspective, an intuitive and desired property of measures of both union and synergistic information is that their value should not change whenever one adds a source that is a deterministic function of sources that are already considered. We provide arguments in favor of this property in Section 4.2.1. This property may not be satisfied by computing  $I_{\cup}^{\text{CI}}$  without previously excluding such sources. For instance, consider  $p(t, y_1, y_2, y_3)$ , where  $Y_1$  and  $Y_2$  are two i.i.d. random variables following a Bernoulli distribution with parameter 0.5,  $Y_3 = Y_2$  (that is,  $Y_3$  is deterministic function of  $Y_2$ ), and  $T = Y_1 \text{ AND } Y_2$ . Computing  $I_{\cup}^{\text{CI}}(Y_1, Y_2, Y_3)$  without excluding  $Y_3$  (or  $Y_2$ ) yields  $I_{\cup}^{\text{CI}}(Y_1, Y_2, Y_3) = I_q(Y_1, Y_2, Y_3; T) \approx 0.6810$  and  $I_{\cup}^{\text{CI}}(Y_1, Y_2) = I_q(Y_1, Y_2; T) \approx 0.5409$ . This issue is resolved by removing deterministic sources before computing  $I_{\cup}^{\text{CI}}$ .

We conclude this section by commenting on the monotonicity of our measure. Suppose we wish to compute the union information of sources  $\{(Y_1, Y_2), Y_3\}$  and  $\{Y_1, Y_2, Y_3\}$ . PID theory demands

that  $I_{\cup}^{\text{CI}}((Y_1, Y_2), Y_3) \geq I_{\cup}^{\text{CI}}(Y_1, Y_2, Y_3)$  (monotonicity of union information). Recall our motivation for  $I_{\cup}^{\text{CI}}((Y_1, Y_2), Y_3)$ : there are two agents, the first has access to  $p(y_1, y_2|t)$  and the second to  $p(y_3|t)$ . The two agents assume conditional independence of their variables and construct  $q'(y_1, y_2, y_3, t) = p(t)p(y_1, y_2|t)p(y_3|t)$ . The story is similar for the computation of  $I_{\cup}^{\text{CI}}(Y_1, Y_2, Y_3)$ , in which case we have three agents that construct  $q''(y_1, y_2, y_3, t) = p(t)p(y_1|t)p(y_2|t)p(y_3|t)$ . Now, it may be the case that  $I_{q'}(Y; T) < I_{q''}(Y; T)$ ; considering only these two distributions would yield  $I_{\cup}^{\text{CI}}((Y_1, Y_2), Y_3) < I_{\cup}^{\text{CI}}(Y_1, Y_2, Y_3)$ , contradicting monotonicity for measures of union information. To overcome this issue, for the computation of  $I_{\cup}^{\text{CI}}((Y_1, Y_2), Y_3)$  - and other sources in general - the agent that has access to  $p(y_1, y_2|t)$  must be allowed to disregard the conditional dependence of  $Y_1$  and  $Y_2$  on  $T$ , even if it holds in the original distribution  $p$ .

#### 4. Properties of Measures of Union Information and Synergy

##### 4.1. Extension of the Williams-Beer Axioms for Measures of Union Information

As Gutknecht *et al.* [34] rightfully notice, the so-called Williams-Beer axioms [1] can actually be derived from parthood distribution functions and the consistency equation [34]. Consequently, they are not really axioms, but consequences of the PID framework. As far as we know, there has been no proposal in the literature for the equivalent of the Williams-Beer axioms (which refer to measures of redundant information) for measures of union information. In the following, we extend the Williams-Beer axioms to measures of union information and show that the proposed  $I_{\cup}^{\text{CI}}$  satisfies these axioms. Although we just argued against calling them axioms, we keep the designation *Williams-Beer axioms* because of its popularity.

**Definition 2.** Let  $A_1, \dots, A_m$  be an arbitrary number  $m \geq 2$  of sources. A measure of union information  $I_{\cup}$  is said to satisfy the Williams-Beer axioms for union information measures if it satisfies:

1. *Symmetry:*  $I_{\cup}$  is symmetric in the  $A_i$ 's.
2. *Self-redundancy:*  $I_{\cup}(A_i) = I(A_i; T)$ .
3. *Monotonicity:*  $I_{\cup}(A_1, \dots, A_{m-1}, A_m) \geq I_{\cup}(A_1, \dots, A_{m-1})$ .
4. *Equality for monotonicity:*  $A_m \subseteq A_{m-1} \Rightarrow I_{\cup}(A_1, \dots, A_{m-1}, A_m) = I_{\cup}(A_1, \dots, A_{m-1})$ .

**Theorem 1.**  $I_{\cup}^{\text{CI}}$  satisfies the Williams-Beer axioms for measures of union information given in Definition 2.

**Proof.** We address each of the axioms in turn.

1. Symmetry follows from the symmetry of mutual information, which in turn is a consequence of the well-known symmetry of joint entropy.
2. Self-redundancy follows from the fact that agent  $i$  has access to  $p(A_i|T)$  and  $p(T)$ , which means that  $p(A_i, T)$  is one of the distributions in the set  $\mathcal{Q}$ , which implies that  $I_{\cup}(A_i) = I(A_i; T)$ .
3. To show that monotonicity holds, begin by noting that

$$I\left(\bigcup_{i=1}^m A_i; T\right) \geq I\left(\bigcup_{i=1}^{m-1} A_i; T\right),$$

due to the monotonicity of mutual information. Let  $\mathcal{Q}_m$  be the set of distributions that the sources  $A_1, \dots, A_m$  can construct and  $\mathcal{Q}_{m-1}$  that which the sources  $A_1, \dots, A_{m-1}$  can construct. Since  $\mathcal{Q}_{m-1} \subseteq \mathcal{Q}_m$ , it is clear that

$$\max_{q \in \mathcal{Q}_m} I_q\left(\bigcup_{i=1}^m A_i; T\right) \geq \max_{q \in \mathcal{Q}_{m-1}} I_q\left(\bigcup_{i=1}^{m-1} A_i; T\right).$$

Consequently,

$$\min \left\{ I \left( \bigcup_{i=1}^m A_i; T \right), \max_{q \in \mathcal{Q}_m} I_q \left( \bigcup_{i=1}^m A_i; T \right) \right\} \geq \min \left\{ I \left( \bigcup_{i=1}^{m-1} A_i; T \right), \max_{q \in \mathcal{Q}_{m-1}} I_q \left( \bigcup_{i=1}^{m-1} A_i; T \right) \right\},$$

which means monotonicity holds.

4. Finally, the proof that equality for monotonicity is the same that was used above to show that the assumption that no source is a subset of another source entails no loss of generality. If  $A_m \subseteq A_{m-1}$ , then the presence of  $A_m$  is irrelevant:  $A = \bigcup_{i=1}^m A_i = \bigcup_{i=1}^{m-1} A_i$  and  $\mathcal{Q}_m = \mathcal{Q}_{m-1}$ , which implies that  $I_{\cup}(A_1, \dots, A_{m-1}, A_m) = I_{\cup}(A_1, \dots, A_{m-1})$ .

□

#### 4.2. Review of Suggested Properties: Griffith and Koch [15]

We now review properties of measures of union information and synergy that have been suggested in the literature, doing so in chronological order. The first set of properties was suggested by Griffith and Koch [15], with the first two being the following.

- *Duplicating a predictor does not change synergistic information*; formally,

$$S(A_1, \dots, A_m \rightarrow T) = S(A_1, \dots, A_m, A_{m+1} \rightarrow T),$$

where  $A_{m+1} = A_i$ , for some  $i = 1, \dots, m$ . Griffith and Koch [15] show that this property holds if the equality for monotonicity property holds for the "corresponding" measure of union information ("corresponding" in the sense of equation (7)). As shown in the previous subsection,  $I_{\cup}^{\text{CI}}$  satisfies this property, and so does the corresponding synergy  $S^{\text{CI}}$ .

- *Adding a new predictor can decrease synergy*, which is a weak statement. We suggest a stronger property: *adding a new predictor cannot increase synergy*, which is formally written as

$$S(A_1, \dots, A_m \rightarrow T) \geq S(A_1, \dots, A_m, A_{m+1} \rightarrow T).$$

This property simply follows from monotonicity for the corresponding measure of union information, which we proved above holds for  $I_{\cup}^{\text{CI}}$ .

The next properties for any measure of union information were also suggested by Griffith and Koch [15]:

1. Global positivity:  $I_{\cup}(A_1, \dots, A_m) \geq 0$ .
2. Self-redundancy:  $I_{\cup}(A_i) = I(A_i; T)$ .
3. Symmetry:  $I_{\cup}(A_1, \dots, A_m)$  is invariant under permutations of  $A_1, \dots, A_m$ .
4. Stronger monotonicity:  $I_{\cup}(A_1, \dots, A_m) \leq I_{\cup}(A_1, \dots, A_m, A_{m+1})$ , with equality if there is some  $A_i$  such that  $H(A_{m+1}|A_i) = 0$ .
5. Target monotonicity: for any (discrete) random variables  $T$  and  $Z$ ,  $I_{\cup}(A_1, \dots, A_m \rightarrow T) \leq I_{\cup}(A_1, \dots, A_m \rightarrow (T, Z))$ .
6. Weak local positivity: for  $n = 2$  the derived partial informations are nonnegative. This is equivalent to

$$\max \{ I(Y_1; T), I(Y_2; T) \} \leq I_{\cup}(Y_1, Y_2) \leq I(Y; T).$$

7. Strong identity:  $I_{\cup}(T \rightarrow T) = H(T)$ .

We argued before that global positivity, self-redundancy, and symmetry are all properties that follow trivially from a well-defined measure of union information [34]. In the following, we discuss in more detail properties 4 and 5.

#### 4.2.1. Stronger Monotonicity

Property 4 in the above list was originally called monotonicity by Griffith and Koch [15]; we changed its name because we had already defined monotonicity in Definition 2, a weaker condition than stronger monotonicity. The proposed inequality clearly follows from the monotonicity of union information (the third Williams-Beer axiom). Now, if there is some  $A_i$  such that  $H(A_{m+1}|A_i) = 0$  (equivalently, if  $A_{m+1}$  is a deterministic function of  $A_i$ ), Griffith and Koch [15] suggest that we must have equality. Recall Axiom 4 (equality for monotonicity) in the extension of the WB axioms (Definition 2). It states that equality must hold if  $A_m \subseteq A_{m-1}$ . In this context,  $A_m$  and  $A_{m-1}$  are sets of random variables, for example  $A_m = \{Y_1, Y_2\}$  and  $A_{m-1} = \{Y_1, Y_2, Y_3\}$ . There is a different point of view we may take. The only way that  $A_m$  is a subset of  $A_{m-1}$  is if  $A_m$ , when viewed as a random vector (in this case, write  $A_m = (Y_1, Y_2)$  and  $A_{m-1} = (Y_1, Y_2, Y_3)$ ), is a subvector of  $A_{m-1}$ . A subvector of a random vector is a deterministic function, and no information gain can come from applying a deterministic function to a random vector. As such, there is no information gain when one considers  $A_m$ , a function of  $A_{m-1}$ , if one already has access to  $A_{m-1}$ . Griffith and Koch [15] argue similarly: there is no information gain by considering  $A_{m+1}$  – a function of  $A_i$  – in addition to  $A_i$ . In conclusion, considering the ‘equality for monotonicity’ strictly through a set inclusion perspective, stronger monotonicity does not follow. On the other hand, extending the idea of set inclusion to the more general context of functions of random variables, then stronger monotonicity follows, because  $\{A_{m+1}\} = \{f(A_i)\}$  is a subset of  $\{A_i\}$ , hence there is no information gain by considering  $A_{m+1} = f(A_i)$  in addition to  $A_i$ . As such, we obtain  $I_{\cup}(A_1, \dots, A_m) = I_{\cup}(A_1, \dots, A_m, A_{m+1})$ . Consequently, it is clear that stronger monotonicity must hold for any measure of union information.

#### 4.2.2. Target Monotonicity

Let us move on to target monotonicity, which we argue should not hold. This precise same property was suggested, but for a measure of redundant information, by Bertschinger *et al.* [18]; they argue that a measure of redundant information should satisfy

$$I_{\cap}(A_1, \dots, A_m \rightarrow T) \leq I_{\cap}(A_1, \dots, A_m \rightarrow (T, Z)),$$

for any discrete random variable  $Z$ , as they argue that this property ‘captures the intuition that if  $A_1, \dots, A_m$  share some information about  $T$ , then at least the same amount of information is available to reduce the uncertainty about the joint outcome of  $(T, Z)$ ’. Since most PID approaches have been built upon measures of redundant information, it is simpler to refute this property. Consider  $I_{\cap}^d$ , one of the most well-motivated and accepted measures of redundant information (as defined in (2)) and the following distribution, which satisfies  $T = Y_1 \text{ AND } Y_2$  and  $Z = (Y_1, Y_2)$ .

From a game theory perspective, since neither agent ( $Y_1$  or  $Y_2$ ) has an advantage when predicting  $T$  (because the channels that each agent has access to have the same conditional distributions), neither agent has any unique information. Moreover, redundancy – as computed by  $I_{\cap}^d(Y_1, Y_2 \rightarrow T)$  – evaluates to approximately 0.311. However, when considering the pair  $(T, Z)$ , the structure that was present in  $T$  is now destroyed, in the sense that now there is no degradation order between the channels that each agent has access to. Note that  $p((t, z), y_1, y_2)$  is a relabelling of the COPY distribution. As such,  $I_{\cap}^d(Y_1, Y_2 \rightarrow (T, Z)) = 0 < I_{\cap}^d(Y_1, Y_2 \rightarrow T)$ , contradicting the property proposed by Bertschinger *et al.* [18].

For a similar reason, we believe that this property should not hold (in general) for measures of union information, even if they satisfy the extension of the Williams-Beer axioms, as our proposed measure does. For instance, the following distribution

**Table 4.** Counter-example distribution for target monotonicity.

T	Z	$Y_1$	$Y_2$	$p(t, z, y_1, y_2)$
0	0	1	0	0.419
1	1	2	1	0.203
2	1	3	0	0.007
0	0	3	1	0.346
2	2	4	4	0.025

satisfies  $I_{\cup}^{\text{Cl}}(Y_1, Y_2 \rightarrow T) \approx 0.91 > 0.90 \approx I_{\cup}^{\text{Cl}}(Y_1, Y_2 \rightarrow (T, Z))$ , meaning target monotonicity does not hold. This happens because, although  $I_p(Y; T) \leq I_p(Y; T, Z)$ , it is not necessarily true that  $I_q(Y; T) \leq I_q(Y; T, Z)$ . The union information measure derived from the degradation order between channels, defined as the ‘dual’ of (2), also agrees with our conclusion [21]. For the distribution in Table 5 we have  $I_{\cup}^{\text{d}}(Y_1, Y_2 \rightarrow T) \approx 0.331 > 0 = I_{\cup}^{\text{d}}(Y_1, Y_2 \rightarrow (T, Z))$ , for the same reason as above: considering  $(T, Z)$  as the target variable destroys the structure present in  $T$ . We agree with the remaining properties suggested by Griffith and Koch [15] and we will address those later.

**Table 5.** Counter-example distribution for target monotonicity.

T	Z	$Y_1$	$Y_2$	$p(t, z, y_1, y_2)$
0	(0,0)	0	0	0.25
0	(0,1)	0	1	0.25
0	(1,0)	1	0	0.25
1	(1,1)	1	1	0.25

#### 4.3. Review of Suggested Properties: Quax et al. [35]

Moving on to additional properties, Quax *et al.* [35] suggest the following properties for a measure of synergy:

1. Nonnegativity:  $S(A_1, \dots, A_m \rightarrow T) \geq 0$ .
2. Upper-Bounded by Mutual Information:  $S(Y \rightarrow T) \leq I(Y; T)$ .
3. Weak Symmetry:  $S(A_1, \dots, A_m \rightarrow T)$  is invariant under any reordering of  $A_1, \dots, A_m$ .
4. Zero synergy about a single variable:  $S(Y_i \rightarrow T) = 0$  for any  $i \in \{1, \dots, n\}$ .
5. Zero synergy in a single variable:  $S(Y \rightarrow Y_i) = 0$  for any  $i \in \{1, \dots, n\}$ .

Let us comment on the proposed ‘zero synergy’ properties (4 and 5). Property 4 seems to have been proposed with the rationale that synergy can only exist for at least two sources, which intuitively makes sense, as synergy is often defined as ‘the information that is present in the pair, but that is not retrievable from any individual variable’. However, because of the way a synergy-based PID is constructed — or weak-synergy, as Gutknecht *et al.* [34] call it — synergy must be defined as in (7), so that, for example in the bivariate case,  $S(Y_1 \rightarrow T) := I(Y; T) - I_{\cup}(Y_1 \rightarrow T) = I(Y_2; T|Y_1)$ , because of self-redundancy of union information and the chain rule of mutual information [11], and since  $I(Y_2; T|Y_1)$  is in general larger than 0, we reject the property ‘Zero synergy about a single variable’.

Property 5, ‘Zero synergy in a single variable’, on the other hand, must hold because of self-redundancy. That is because, for any  $i \in \{1, \dots, n\}$ ,  $S(Y \rightarrow Y_i) := I(Y; Y_i) - I_{\cup}(Y \rightarrow Y_i) = I(Y_i; Y_i) - I(Y; Y_i) = H(Y_i) - H(Y_i) = 0$ .

#### 4.4. Relationship with the Extended Williams-Beer axioms

We now prove which of the introduced properties are implied by the extension of the Williams-Beer axioms for measures of union information. In what follows, assume that the goal is to decompose the information present in the distribution  $p(y, t) = p(y_1, \dots, y_n, t)$ .

**Theorem 2.** Let  $I_{\cup}$  be a measure of UI that satisfies the extension of the Williams-Beer axioms (symmetry, self-redundancy, monotonicity, and equality for monotonicity) for measures of UI as in Definition (2). Then,  $I_{\cup}$  also satisfies the following properties of Griffith and Koch [15]: global positivity, weak local positivity, and strong identity.

**Proof.** Global positivity is a direct consequence of monotonicity and the non-negativity of mutual information:  $I_{\cup}(A_1, \dots, A_m) \geq I_{\cup}(A_1) = I(A_1; T) \geq 0$

Weak local positivity holds because monotonicity and self-redundancy imply that  $I_{\cup}(Y_1, Y_2) \geq I_{\cup}(Y_1) = I(Y_1; T)$ , as well as  $I_{\cup}(Y_1, Y_2) \geq I_{\cup}(Y_2) = I(Y_2; T)$ , hence  $\max\{I(Y_1; T), I(Y_2; T)\} \leq I_{\cup}(Y_1, Y_2)$ . Moreover,  $I_{\cup}(Y_1, Y_2) \leq I_{\cup}(Y_1, Y_2, Y_{12}) = I_{\cup}(Y_{12}) = I(Y; T)$ .

Strong identity follows trivially from self-redundancy, since  $I_{\cup}(T \rightarrow T) = I(T; T) = H(T)$ .  $\square$

**Theorem 3.** Consider a measure of UI that satisfies the conditions of Theorem 2. If synergy is defined as in Equation (7), it satisfies the following properties of [35]: nonnegativity, upper-bounded by mutual information, weak symmetry, and zero synergy in a single variable.

**Proof.** Nonnegativity of synergy and upper-bounded by mutual information follow from the definition of synergy and from the fact that for whichever source  $(A_1, \dots, A_m)$ , with  $m \geq 1$ , we have that  $I_{\cup}(A_1, \dots, A_m \rightarrow T) \leq I(Y; T)$ .

Weak symmetry follows trivially from the fact that both  $I(Y; T)$  and  $I_{\cup}(A_1, \dots, A_m \rightarrow T)$  are symmetric in the relevant arguments.

Finally, zero synergy in a single variable follows from self-redundancy together with the definition of synergy, as shown above.  $\square$

#### 4.5. Review of Suggested Properties: Rosas et al. [36]

Based on the proposals of Griffith et al. [23], Rosas et al. [36] suggested the following properties for a measure of synergy:

- Target data processing inequality: if  $Y - T_1 - T_2$  is a Markov chain, then  $S(Y \rightarrow T_1) \geq S(Y \rightarrow T_2)$ .
- Channel convexity:  $S(Y \rightarrow T)$  is a convex function of  $P(T|Y)$  for a given  $P(Y)$ .

We disagree with both suggested properties, for the reasons presented next. Consider the following distribution, in which  $T_1$  is a relabeling of the COPY distribution and  $T_2 = Y_1 \text{ xor } Y_2$ .

**Table 6.**  $T_1 = \text{COPY}$ ,  $T_2 = \text{XOR}$ .

$T_2$	$T_1$	$Y_1$	$Y_2$	$p(t_2, t_1, y_1, y_2)$
0	0	0	0	0.25
1	1	0	1	0.25
1	2	1	0	0.25
0	3	1	1	0.25

Start by noting that since  $T_2$  is a deterministic function of  $T_1$ , then  $Y - T_1 - T_2$  is a Markov chain. Since  $Y_1 \perp Y_2 | T_1$ , our measure  $S^{\text{CI}}(Y_1, Y_2 \rightarrow T_1) = I(Y; T_1) - I_{\cup}^{\text{CI}}(Y_1, Y_2 \rightarrow T_1) = 0$  leads to zero synergy. On the other hand,  $S^{\text{CI}}(Y_1, Y_2 \rightarrow T_2) = 1$ , contradicting the first property suggested by Rosas et al. [36]. This happens because  $Y_1 \not\perp_p Y_2 | T_2$ , so synergy is positive. The loss of conditional independence of the inputs (given the target), when one goes from considering the target  $T_1$  to  $T_2$  is the reason why synergy increases.

The second suggested property argues that synergy should be a convex function of  $P(T|Y)$ , for fixed  $P(Y)$ . Our measure of synergy does not satisfy this property, even though it is derived from a measure of union information that satisfies the extension of the WB axioms. For instance, consider

the XOR distribution with one extra outcome. We introduce it below and parameterize it using  $r = p(T = 0|Y = (0,0)) \in [0, 1]$ . Notice that this modification does not affect  $P(Y)$

Synergy, as measured by  $S^{\text{CI}}(Y_1, Y_2 \rightarrow T)$ , is maximized when  $r$  equals 1 (the distribution becomes the standard XOR) and minimized when  $r$  equals 0. We don't see an immediate reason as to why a general synergy function should be convex in  $p(t|y)$ , or why it should have a unique minimizer as a function of  $r$ . Recall that a function  $S$  is convex if  $\forall t \in [0, 1], \forall x_1, x_2 \in D$ , we have

$$S(tx_1 + (1-t)x_2) \leq tS(x_1) + (1-t)S(x_2).$$

In the following, we slightly abuse the notation of the input variables of a synergy function. Our synergy measure  $S^{\text{CI}}$ , when considered as a function of  $r$ , does not satisfy this inequality. For the adapted XOR distribution, take  $t = 0.5$ ,  $x_1 = 0$  and  $x_2 = 0.5$ . We have

$$S^{\text{CI}}(0.5 \times 0 + 0.5 \times 0.5) = S^{\text{CI}}(0.25) \approx 0.552$$

and

$$0.5 \times S^{\text{CI}}(0) + 0.5 \times S^{\text{CI}}(0.5) \approx 0.5 \times 0.270 + 0.5 \times 0.610 \approx 0.440,$$

contradicting the property of Channel convexity.

$S^{\text{d}}$ , the synergy measure derived from Kolchinsky's proposed union information measure  $I_{\cup}^{\text{d}}$ , agrees with this [21]. We slightly change  $p(y)$  in the above distribution to obtain a new distribution, which we present in Table 7.

**Table 7.** Adapted XOR distribution v2.

$T$	$Y_1$	$Y_2$	$p(t, y_1, y_2)$
0	0	0	$0.1r$
1	0	0	$0.1(1-r)$
1	1	0	0.4
1	0	1	0.4
0	1	1	0.1

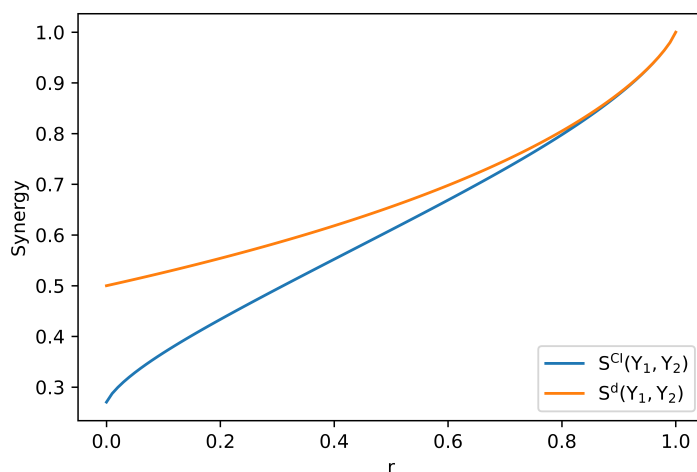
This distribution does not satisfy the convexity inequality, since

$$S^{\text{d}}(0.5 \times 0 + 0.5 \times 0.5) \approx 0.338 > 0.3095 \approx 0.5S^{\text{d}}(0) + 0.5S^{\text{d}}(0.5).$$

This can be easily seen since  $K^{(1)} = K^{(2)}$  for any  $r \in [0, 1]$ , hence we may choose  $K^{\text{Q}} = K^{(1)}$  to compute  $S^{\text{d}} = I(Y; T) - I(Q; T)$ , which is not convex for this particular distribution. To conclude this section, we present a plot of  $S^{\text{CI}}(Y_1, Y_2)$  and  $S^{\text{d}}(Y_1, Y_2)$  as a function of  $r$ , for the distribution presented in Table 8.

**Table 8.** Adapted XOR distribution.

$T$	$Y_1$	$Y_2$	$p(t, y_1, y_2)$
0	0	0	$0.25r$
1	0	0	$0.25(1-r)$
1	1	0	0.25
1	0	1	0.25
0	1	1	0.25



**Figure 3.** Computation of  $S^{CI}$  and  $S^d$  as functions of  $r = p(T = 0|Y = (0,0))$ . As we showed for this distribution,  $S^{CI}$  is not a convex function of  $r$ .

## 5. Previous Measures of Union Information and Synergy

We now review other measures of union information and synergy proposed in the literature. For the sake of brevity, we will not recall all their definitions, only some important conclusions. We suggest the interested reader consult the bibliography for more information.

### 5.1. Qualitative Comparison

Griffith and Koch [15] review three previous measures of synergy:

- $S^{WB}$ , derived from  $I_{\cap}^{WB}$ , the original redundancy measure proposed by Williams and Beer [1], using the IEP;
- the *whole-minus-sum* (WMS) synergy,  $S^{WMS}$ ;
- the *correlational importance* synergy,  $S^{\Delta I}$ .

These synergies can be interpreted as resulting directly from measures of union information; that is, they are explicitly written as  $S(\alpha \rightarrow T) = I(Y; T) - I_{\cup}(\alpha \rightarrow T)$ , where  $I_{\cup}$  may not necessarily satisfy our intuitions of a measure of union information, as in Definition 2, except for  $S^{\Delta I}$ , which has the form of a Kullback-Leibler divergence.

Griffith and Koch [15] argue that  $S^{WB}$  *overestimates* synergy, which is not a surprise, as many authors criticized  $I_{\cap}^{WB}$  for not measuring informational content, only informational values [12]. The WMS synergy, on the other hand, which can be written as a difference of total correlations, can be shown to be equal to the difference between synergy and redundancy for  $n = 2$ , which is not what is desired in a measure of synergy. For  $n > 2$ , the authors show that the problem becomes even more exacerbated:  $S^{WMS}$  equals synergy minus the redundancy *counted multiple times*, which is why the authors argue that  $S^{WMS}$  *underestimates* synergy. Correlational importance,  $S^{\Delta I}$ , is known to be larger than  $I(Y; T)$  for some distributions, excluding it from being an appropriately interpretable measure of synergy.

Faced with these limitations, Griffith and Koch [15] introduce their measure of union information, which they define as

$$I_{\cup}^{VK}(A_1, \dots, A_m \rightarrow T) := \min_{p^*} I_{p^*} \left( \bigcup_{i=1}^m A_i; T \right) \quad (8)$$

$$\text{s.t. } p^*(A_i, T) = p(A_i, T), i = 1, \dots, m,$$

where the minimization is over joint distributions of  $A_1, \dots, A_m, T$ , alongside the derived measure of synergy  $S^{\text{VK}}(\alpha \rightarrow T) = I(Y; T) - I_{\cup}^{\text{VK}}(\alpha \rightarrow T)$ . This measure quantifies union information as the least amount of information that source  $\alpha$  has about  $T$  when the source-target marginals (as determined by  $\alpha$ ) are fixed by  $p$ . Griffith and Koch [15] also establish the following inequalities for the synergistic measures they reviewed:

$$\max \left\{ 0, S^{\text{WMS}}(\alpha \rightarrow T) \right\} \leq S^{\text{VK}}(\alpha \rightarrow T) \leq S^{\text{WB}}(\alpha \rightarrow T) \leq I \left( \bigcup_{i=1}^m A_i; T \right), \quad (9)$$

where  $\alpha = (A_1, \dots, A_m)$ . At the time, Griffith and Koch [15] did not provide a way to analytically compute their measure. Later, Kolchinsky [21] showed that the measure of union information derived from the degradation order,  $I_{\cup}^{\text{d}}$ , is equivalent to  $I_{\cup}^{\text{VK}}$ , and provided a way to compute it. For this reason, we will only consider  $I_{\cup}^{\text{d}}$ .

After the work of Griffith and Koch [15] in 2014, we are aware of only three other suggested measures of synergy:

- $S^{\text{MSRV}}$ , proposed by Quax *et al.* [35], where MSRV stands for *maximally synergistic random variable*;
- *synergistic disclosure*,  $S^{\text{SD}}$ , proposed by Rosas *et al.* [36];
- $S^{\text{d}}$ , proposed by Kolchinsky [21].

The first two proposals do not define synergy via a measure of union information. They define synergy through an auxiliary random variable,  $Z$ , which has positive information about the whole – that is,  $I(Z; Y) > 0$  – but no information about any of the parts – that is,  $I(Z; Y_i) = 0, i = 1, \dots, n$ . While this property has an appealing operational interpretation, we believe that it is too restrictive; that is, we believe that information *can* be synergistic, even if it provides some positive information about some part of  $Y$ .

The authors of  $S^{\text{MSRV}}$  show that their proposed measure is incompatible with PID and that it cannot be computed for all distributions, as it requires the ability to compute orthogonal random variables, which is not always possible [35]. A counter-intuitive example for the value of this measure can be seen for the AND distribution, defined by  $T = Y_1 \text{ AND } Y_2$ , with  $Y_1$  and  $Y_2$  i.i.d. taking values in  $\{0, 1\}$  with equal probability. In this case,  $S^{\text{MSRV}} = 0.5$ , a value that we argue is too large, because whenever  $Y_1$  (respectively  $Y_2$ ) is 0, then  $T$  does not depend on  $Y_2$  (respectively  $Y_1$ ) (which happens with probability 0.75). Consequently,  $S^{\text{MSRV}}/I(Y; T) \approx 0.5/0.811 \approx 0.617$  may be too large of a synergy ratio for this distribution. As the authors note, the only other measure that agrees with  $S^{\text{MSRV}}$  for the AND distribution is  $S^{\text{WB}}$ , which Griffith and Koch [15] argued also overestimates synergy.

Concerning  $S^{\text{SD}}$ , we do not have any criticism, except for the one already pointed out by Gutknecht *et al.* [34]: they note that the resulting decomposition from  $S^{\text{SD}}$  is not a standard PID, in sense that it does not satisfy a consistency equation (see [34] for more details), which implies that ‘(...) the atoms cannot be interpreted in terms of parthood relations with respect to mutual information terms (...). For example, we do not obtain any atoms interpretable as unique or redundant information in the case of two sources’ [34]. Gutknecht *et al.* [34] suggest a very simple modification to the measure so that it satisfies the consistency equation.

For the AND distribution,  $S^{\text{SD}}$  evaluates to approximately 0.311, as does  $S^{\text{d}}$ , whereas our measure yields  $S^{\text{CI}} \approx 0.270$ , as the information that the parts cannot obtain when they combine their marginals, under distribution  $q$ . This shows that these four measures are not equivalent.

## 5.2. Quantitative Comparison

Griffith and Koch [15] applied the synergy measures they reviewed to other distributions. We show their results below and compare them with the synergy resulting from our measure of union information,  $S^{\text{CI}}$ , with the measure of Rosas *et al.* [36],  $S^{\text{SD}}$ , and that of Kolchinsky [21],  $S^{\text{d}}$ . Since the code for the computation of  $S^{\text{MSRV}}$  is no longer available online, we do not present it.

We already saw the definition of the AND, COPY and XOR distributions. The XORDUPLICATE and ANDDUPLICATE are built from the XOR and the AND distributions by inserting a duplicate source variable  $Y_3 = Y_2$ . The goal is to test if the presence of a duplicate predictor impacts the different synergy measures. The definitions of the remaining distributions are presented in the appendix. Some of these are trivariate, and for those we compute synergy as

$$S(Y_1, Y_2, Y_3 \rightarrow T) = I(Y_1, Y_2, Y_3; T) - I_{\cup}(Y_1, Y_2, Y_3 \rightarrow T), \quad (10)$$

unless the synergy measure is directly defined (as opposed to being defined via a union information measure). We now comment on the results. It should be noted that Kolchinsky [21] suggested that unique information  $U_1$  and  $U_2$  should be computed from measures of redundant information, and excluded information  $E_1$  and  $E_2$  should be computed from measures of union information, as in our case. However, since we will only present the decompositions for the bivariate case and in this case  $E_1 = U_2$  and  $E_2 = U_1$ , we present the results considering unique information, as is mostly done in the literature.

- XOR yields  $I(Y; T) = 1$ . The XOR distribution is the hallmark of synergy. Indeed, the only solution of (1) is  $(S, R, U_1, U_2) = (1, 0, 0, 0)$ , and all of the above measures yield 1 bit of synergy.
- AND yields  $I(Y; T) \approx 0.811$ . Unlike XOR, there are multiple solutions for (1), and none is universally agreed upon, since different information measures capture different concepts of information.
- COPY yields  $I(Y; T) = 2$ . Most PID measures argue one of two different possibilities for this distribution. They suggest that the solution is either  $(S, R, U_1, U_2) = (1, 1, 0, 0)$  or  $(0, 0, 1, 1)$ . Our measure suggests that all information flows uniquely from each source.
- RDNXOR yields  $I(Y; T) = 2$ . In words, this distribution is the concatenation of two XOR 'blocks', each of which with its own symbols, and not allowing the two blocks to mix. That is, both  $Y_1$  and  $Y_2$  can determine in which XOR block the resulting value  $T$  will be - which intuitively means that they both have this information, meaning it is redundant - but neither  $Y_1$  nor  $Y_2$  have information about the outcome of the XOR operation - as is expected in the XOR distribution - which intuitively means that such information must be synergistic. All measures except  $S^{\text{WMS}}$  agree with this.
- RDNUNQXOR yields  $I(Y; T) = 4$ . According to Griffith and Koch [15], it was constructed to carry 1 bit of each information type. Although the solution is not unique, it must satisfy  $U_1 = U_2$ . Indeed our measure yields the solution  $(S, R, U_1, U_2) = (1, 1, 1, 1)$ , like most measures except  $S^{\text{WB}}$  and  $S^{\text{WMS}}$ . This confirms the intuition by Griffith and Koch [15] that  $S^{\text{WB}}$  and  $S^{\text{WMS}}$  overestimate and underestimate synergy, respectively. In fact, in the decomposition resulting from  $S^{\text{WB}}$ , there are 2 bits of synergy and 2 bits of redundancy, which we argue cannot be the case, as this would imply that  $U_1 = U_2 = 0$ , and given the construction of this distribution, it is clear that there is some unique information since, unlike in RDNXOR, the XOR blocks are allowed to mix, thus  $(T, Y_1, Y_2) = (1, 0, 1)$  is a possible outcome, but so is  $(T, Y_1, Y_2) = (2, 0, 2)$ . That is not the case with RDNXOR. On the other hand,  $S^{\text{WMS}}$  yields zero synergy and redundancy, with  $U_1$  and  $U_2$  each evaluating to 2 bits. Since this distribution is a mix of blocks satisfying a relation of the form  $T = Y_1 \text{ xor } Y_2$ , we argue that there must be some non-null amount of synergy, which is why we claim that  $S^{\text{WMS}}$  is not valid.
- XORDUPLICATE yields  $I(Y; T) = 1$ . All measures correctly identify that the duplication of a source shouldn't change synergy, at least for this particular distribution.
- ANDDUPLICATE yields  $I(Y; T) \approx 0.811$ . Unlike in the previous example, both  $S^{\text{WMS}}$  and  $S^{\Delta I}$  yield a change in their synergy value. This is a shortcoming since duplicating a source should not increase either synergy or union information. The other measures are not affected by the duplication of a source.

- XORLOSES yields  $I(Y; T) = 1$ . Its distribution is the same as XOR but with a new source  $Y_3$  satisfying  $T = Y_3$ . As such, since  $Y_3$  uniquely determines  $T$ , we expect no synergy. All measures agree with this.
- XORMULTICOAL yields  $I(Y; T) = 1$ . Its distribution is such that any pair  $(Y_i, Y_j), i, j = 1, 2, 3, i \neq j$  is able to determine  $T$  with no uncertainty. All measures agree that the information present in this distribution is purely synergistic.

From these results, we agree with Griffith and Koch [15] that  $S^{WB}$ ,  $S^{WMS}$ , and  $S^{\Delta I}$  are not good measures of synergy: they do not satisfy many of our intuitions and overestimate synergy, not being invariant to duplicate sources or taking negative values. For these reasons and those presented in Section 5, we reject those measures of synergy. In the next Section, we comment on the remaining measures  $S^d$ ,  $S^{SD}$  and  $S^{CI}$ .

### 5.3. Relation to Other PID Measures

Kolchinsky [21] introduced  $I_{\cup}^d$  and showed that this measure is equivalent to  $I_{\cup}^{VK}$  [15] and to  $I_{\cup}^{BROJA}$  [14], in the sense that the three of them achieve the same optimum value [21]. The multivariate extension of  $I_{\cup}^{BROJA}$  was proposed by Griffith and Koch [15], defined as

$$I_{\cup}^{BROJA}(A_1, \dots, A_m \rightarrow T) := \min_{\tilde{A}_1, \dots, \tilde{A}_m} I(\tilde{A}_1, \dots, \tilde{A}_m; T) \text{ such that } \forall i P(\tilde{A}_i, T) = P(A_i, T),$$

which we present because it makes it clear what conditions are enforced upon the marginals. There is a relation between  $I_{\cup}^{BROJA}(A_1, \dots, A_m) = I_{\cup}^d(A_1, \dots, A_m)$  and  $I_{\cup}^{CI}(A_1, \dots, A_m)$  whenever the sources  $A_1, \dots, A_m$  are singletons. In this case, and only in this case, does the set  $\mathcal{Q}$  involved in the computation of  $I_{\cup}^{CI}(A_1, \dots, A_m)$  has only one element:  $q(t, a_1, \dots, a_m) = p(t)p(a_1|t)\dots p(a_m|t)$ . Since this distribution, as well as the original distribution  $p$ , are both admissible points in  $I_{\cup}^d$ , we have that  $I_{\cup}^d \leq I_{\cup}^{CI}$ , which implies that  $S^d \geq S^{CI}$ . On the other hand, if there is at least one source  $A_1, \dots, A_m$  that is not a singleton, the measures are not trivially comparable. For example, suppose we wish to compute  $I_{\cup}((Y_1, Y_2), (Y_2, Y_3))$ . We know that the solution of  $I_{\cup}^d((Y_1, Y_2), (Y_2, Y_3))$  is a distribution  $p^*$  whose marginals  $p^*(y_1, y_2, t)$  and  $p^*(y_2, y_3, t)$  must coincide with the marginals under the original  $p$ . However, in the computation of  $I_{\cup}^{CI}((Y_1, Y_2), (Y_2, Y_3))$ , it may be the case that the solution  $p^*$  of  $I_{\cup}^d((Y_1, Y_2), (Y_2, Y_3))$  is not in the set  $\mathcal{Q}$ , involved in the computation of  $I_{\cup}^{CI}$ , and it achieves a lower mutual information with  $T$ . That is, it might be the case that  $I_{p^*}(Y; T) < I_q(Y; T)$ , for all  $q \in \mathcal{Q}$ . In such a case, we would have  $I_{\cup}^d > I_{\cup}^{CI}$ .

It is convenient to be able to upper-bound certain measures with other measures. For example, Gomes and Figueiredo [25] (see that paper for the definitions of these measures) showed that for any source  $(A_1, \dots, A_m), m \geq 1$ ,

$$I_{\cap}^d(A_1, \dots, A_m) \leq I_{\cap}^{ln}(A_1, \dots, A_m) \leq I_{\cap}^{mc}(A_1, \dots, A_m).$$

However, we argue that the inability to draw such strong conclusions (or bounds) is a positive aspect of PID. This is because there are many different ways to define the information (be it redundant, unique, union, etc) that one wishes to capture. If one could trivially relate all measures, it would mean that it would be possible to know *a priori* how those measures would behave. Consequently, this would imply the absence of variability/freedom in how to measure different information concepts, as those measures would capture, not equivalent, but similar types of information, as they would all be ordered. It is precisely because one cannot order different measures of information trivially that PID provides a rich and complex framework to distinguish different types of information, although we believe PID is still in its infancy.

James *et al.* [16] introduced a measure of unique information, which we recall now. In the bivariate case – *i.e.*, consider  $p(y_1, y_2, t)$  – let  $q$  be the maximum entropy distribution that preserves the marginals  $p(y_1, t)$  and  $p(y_2, t)$ , and let  $r$  be the maximum entropy distribution that preserves the marginals  $p(y_1, t)$ ,  $p(y_2, t)$ , and  $p(y_1, y_2)$ . Although there is no closed form for  $r$ , which has to be computed using

an iterative algorithm [37], it may be shown that the solution for  $q$  is  $q(y_1, y_2, t) = p(t)p(y_1|t)p(y_2|t)$ . This is the same distribution  $q$  that we consider for the bivariate decomposition (3). James *et al.* [16] suggest defining unique information  $U_i$  as the least change (in sources-target mutual information) that involves the addition of the  $(Y_i, T)$  marginal constraint, that is

$$U_1 = \min\{I_q(Y_1; T|Y_2), I_r(Y_1; T|Y_2)\}, \quad (11)$$

and analogously for  $U_2$ . They show that their measure yields a nonnegative decomposition for the bivariate case. Since  $I(Y_1; T|Y_2) = S + U_1$ , some algebra leads to

$$S^{\text{dep}} = I(Y; T) - \min\{I_q(Y; T), I_r(Y; T)\}, \quad (12)$$

where  $S^{\text{dep}}$  is the synergy resulting from the decomposition of James *et al.* [16] in the bivariate case. Recall that our measure of synergy for the bivariate case is given by

$$S^{\text{CI}} = I(Y; T) - \min\{I_q(Y; T), I_p(Y; T)\}. \quad (13)$$

The similarity is striking. Computing  $S^{\text{dep}}$  for the bivariate distributions in Table 9 shows that it coincides with the decomposition given by our measure, except for the AND distribution, where we obtained  $S^{\text{dep}} \approx 0.27072$  and  $S^{\text{CI}} \approx 0.27042$ . We could not obtain  $S^{\text{dep}}$  for the RDNUNQXOR distribution because the algorithm that computes  $r$  did not finish in the allotted time of 10 minutes. James *et al.* [16] showed that, for whichever bivariate distribution,  $I_r(Y; T) \leq I_p(Y; T)$ , therefore for the bivariate case we have  $S^{\text{CI}} \leq S^{\text{dep}}$ . Unfortunately, the measure of unique information proposed by James *et al.* [16], unlike the usual proposals of intersection or union information, does not allow for the computation of the partial information atoms in the complete redundancy lattice if  $n > 2$ . The authors also comment that it is not clear if their measure satisfies monotonicity when  $n > 2$ . Naturally, our measure is not the same as  $S^{\text{dep}}$ , so it doesn't retain the operational interpretation of unique information  $U_i$  being the least amount that influences  $I(Y; T)$  when the marginal constraint  $(Y_i, T)$  is added to the resulting maximum entropy distributions. Given the form of  $S^{\text{dep}}$ , one could define  $I_{\cup}^{\text{dep}} := \min\{I_q(Y; T), I_r(Y; T)\}$  and study its properties. Clearly, it does not satisfy the self-redundancy axiom, but we wonder if it could be adjusted so that it satisfies all of the proposed axioms. The  $n = 2$  decomposition retains the operational interpretation of the original measure, but it is not clear whether this is true for  $n > 2$ . For the latter case, the maximum entropy distributions that we wrote as  $q$  and  $r$  have different definitions [16]. We leave this for future work.

**Table 9.** Application of the measures reviewed in Griffith and Koch [15], ( $S^{\text{WB}}$ ,  $S^{\text{WMS}}$  and  $S^{\Delta I}$ ),  $S^{\text{SD}}$  introduced by Rosas *et al.* [36],  $S^{\text{d}}$  introduced by Kolchinsky [21] and our measure of synergy  $S^{\text{CI}}$  to different distributions. The bottom four distributions are trivariate. We write DNF to mean that a specific computation did not finish within 10 minutes.

Example	$S^{\text{WB}}$	$S^{\text{WMS}}$	$S^{\Delta I}$	$S^{\text{d}}$	$S^{\text{SD}}$	$S^{\text{CI}}$
XOR	1	1	1	1	1	1
AND	0.5	0.189	0.104	0.5	0.311	0.270
COPY	1	0	0	0	1	0
RDNXOR	1	0	1	1	1	1
RDNUNQXOR	2	0	1	1	DNF	1
XORDUPLICATE	1	1	1	1	1	1
ANDDUPLICATE	0.5	-0.123	0.038	0.5	0.311	0.270
XORLOSES	0	0	0	0	0	0
XORMULTICOAL	1	1	1	1	DNF	1

## 6. Conclusion and Future Work

In this paper, we introduced a new measure of *union information* for the *partial information decomposition* (PID) framework, based on the channel perspective, which quantifies synergy as the information that is beyond conditional independence of the sources, given the target. This measure has a clear interpretation and is very easy to compute, unlike most measures of union information or synergy, which require solving an optimization problem. The main contributions and conclusions of the paper can be summarized as follows.

- We introduced new measures of union information and synergy for the PID framework, which thus far was mainly developed based on measures of redundant or unique information. We provided its operational interpretation and defined it for an arbitrary number of sources.
- We proposed an extension of the Williams-Ber axioms for measures of union information and showed our proposed measure satisfies them.
- We reviewed, commented on and rejected some of the previously proposed properties for measures of union information and synergy in the literature.
- We showed that measures of union information that satisfy the extension of the Williams-Ber axioms necessarily satisfy a few other appealing properties, as well as the derived measures of synergy.
- We reviewed previous measures of union information and synergy, critiqued them and compared them with our proposed measure.
- The new measure is easy to compute. For example in the bivariate case, if the supports of  $Y_1, Y_2$ , and  $T$  have size  $S_1, S_2$ , and  $S_T$ , respectively, the computation time of our measure grows like  $\mathcal{O}(S_1 S_2 S_T)$ .
- We provide code for the computation of our measure for the bivariate case and for source  $\{\{Y_1\}, \{Y_2\}, \{Y_3\}\}$  in the trivariate case.

Finally, we believe this paper opens several avenues for future research, thus we point out several directions to be pursued in upcoming work:

- We saw that the synergy yielded by the measure of James *et al.* [16] is given by  $S^{\text{dep}} = I(Y; T) - \min\{I_q(Y; T), I_r(Y; T)\}$ . Given its analytical expression, one could start by defining a measure of union information as  $I_{\cup}(Y_1, Y_2 \rightarrow T) = \min\{I_q(Y; T), I_r(Y; T)\}$ , possibly tweak it so it satisfies the WB axioms, study its properties and possibly extend it to the multivariate case.
- Our proposed measure may ignore conditional dependencies that are present in  $p$  in favor of maximizing mutual information, as we commented in section 3.3. This is a compromise so that the measure satisfies monotonicity. We believe this is a potential drawback of our measure, and we suggest the investigation of a measure similar to ours, but that doesn't ignore conditional dependencies that it has access to.
- Implementing our measure in the `dit` package [38].
- This paper reviewed measures of union information and synergy, as well as properties that were suggested throughout the literature. Sometimes we did so by providing examples where the suggested properties fail, and other times simply by commenting. We suggest something similar be done for measures of redundant information.

## 7. Code Availability

The code will be publicly available.

**Funding:** This research was partially funded by: FCT – *Fundação para a Ciência e a Tecnologia*, under grants number SFRH/BD/145472/2019 and UIDB/50008/2020; Instituto de Telecomunicações; Portuguese Recovery and Resilience Plan, through project C645008882-00000055 (NextGenAI, CenterforResponsibleAI).

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

## Appendix A

In this appendix, we present the remaining distributions for which we computed different measures of synergy. For these distributions each outcome has the same probability, so we don't present their probabilities.

**Table A1.** RDNXOR (left), XORLOSES (center) and XORMULTICOAL (right).

T	Y <sub>1</sub>	Y <sub>2</sub>	
0	0	0	
1	0	1	
1	1	0	
0	1	1	
2	2	2	
3	2	3	
3	3	2	
2	3	3	
T	Y <sub>1</sub>	Y <sub>2</sub>	Y <sub>3</sub>
0	0	0	0
1	0	1	1
1	1	0	1
0	1	1	0
T	Y <sub>1</sub>	Y <sub>2</sub>	Y <sub>3</sub>
0	0	0	0
0	1	1	1
0	2	2	2
0	3	3	3
1	2	1	0
1	3	0	1
1	0	3	2
1	1	2	3

Table A2. RDNUNQXOR.

$T$	$Y_1$	$Y_2$
0	0	0
1	0	1
1	1	0
0	1	1
2	0	2
3	0	3
3	1	2
2	1	3
4	2	0
5	2	1
5	3	0
4	3	1
6	2	2
7	2	3
7	3	2
6	3	3
$T$	$Y_1$	$Y_2$
8	4	4
9	4	5
9	5	4
8	5	5
10	4	6
11	4	7
11	5	6
10	5	7
12	6	4
13	6	5
13	7	4
12	7	5
14	6	6
15	6	7
15	7	6
14	7	7

## References

- Williams, P.; Beer, R. Nonnegative decomposition of multivariate information. *arXiv preprint arXiv:1004.2515* **2010**.
- Lizier, J.; Flecker, B.; Williams, P. Towards a synergy-based approach to measuring information modification. 2013 IEEE Symposium on Artificial Life (ALIFE). IEEE, 2013, pp. 43–51.
- Wibral, M.; Finn, C.; Wollstadt, P.; Lizier, J.T.; Priesemann, V. Quantifying information modification in developing neural networks via partial information decomposition. *Entropy* **2017**, *19*, 494.
- Rauh, J. Secret sharing and shared information. *Entropy* **2017**, *19*, 601.
- Vicente, R.; Wibral, M.; Lindner, M.; Pipa, G. Transfer entropy—a model-free measure of effective connectivity for the neurosciences. *Journal of computational neuroscience* **2011**, *30*, 45–67.
- Ince, R.; Van Rijsbergen, N.; Thut, G.; Rousselet, G.; Gross, J.; Panzeri, S.; Schyns, P. Tracing the flow of perceptual features in an algorithmic brain network. *Scientific reports* **2015**, *5*, 1–17.
- Gates, A.; Rocha, L. Control of complex networks requires both structure and dynamics. *Scientific reports* **2016**, *6*, 1–11.
- Faber, S.; Timme, N.; Beggs, J.; Newman, E. Computation is concentrated in rich clubs of local cortical networks. *Network Neuroscience* **2019**, *3*, 384–404.
- James, R.; Ayala, B.; Zakirov, B.; Crutchfield, J. Modes of information flow. *arXiv preprint arXiv:1808.06723* **2018**.
- Arellano-Valle, R.; Contreras-Reyes, J.; Genton, M. Shannon Entropy and Mutual Information for Multivariate Skew-Elliptical Distributions. *Scandinavian Journal of Statistics* **2013**, *40*, 42–62.

11. Cover, T. *Elements of information theory*; John Wiley & Sons, 1999.
12. Harder, M.; Salge, C.; Polani, D. Bivariate measure of redundant information. *Physical Review E* **2013**, *87*, 012130.
13. Gutknecht, A.; Wibral, M.; Makkeh, A. Bits and pieces: Understanding information decomposition from part-whole relationships and formal logic. *Proceedings of the Royal Society A* **2021**, *477*, 20210110.
14. Bertschinger, N.; Rauh, J.; Olbrich, E.; Jost, J.; Ay, N. Quantifying unique information. *Entropy* **2014**, *16*, 2161–2183.
15. Griffith, V.; Koch, C. Quantifying synergistic mutual information. In *Guided self-organization: inception*; Springer, 2014; pp. 159–190.
16. James, R.; Emenheiser, J.; Crutchfield, J. Unique information via dependency constraints. *Journal of Physics A: Mathematical and Theoretical* **2018**, *52*, 014002.
17. Chicharro, D.; Panzeri, S. Synergy and redundancy in dual decompositions of mutual information gain and information loss. *Entropy* **2017**, *19*, 71.
18. Bertschinger, N.; Rauh, J.; Olbrich, E.; Jost, J. Shared information—New insights and problems in decomposing information in complex systems. *Proceedings of the European conference on complex systems 2012*. Springer, 2013, pp. 251–269.
19. Rauh, J.; Banerjee, P.; Olbrich, E.; Jost, J.; Bertschinger, N.; Wolpert, D. Coarse-graining and the Blackwell order. *Entropy* **2017**, *19*, 527.
20. Ince, R. Measuring multivariate redundant information with pointwise common change in surprisal. *Entropy* **2017**, *19*, 318.
21. Kolchinsky, A. A Novel Approach to the Partial Information Decomposition. *Entropy* **2022**, *24*, 403.
22. Barrett, A. Exploration of synergistic and redundant information sharing in static and dynamical Gaussian systems. *Physical Review E* **2015**, *91*, 052802.
23. Griffith, V.; Chong, E.; James, R.; Ellison, C.; Crutchfield, J. Intersection information based on common randomness. *Entropy* **2014**, *16*, 1985–2000.
24. Griffith, V.; Ho, T. Quantifying redundant information in predicting a target random variable. *Entropy* **2015**, *17*, 4644–4653.
25. Gomes, A.F.; Figueiredo, M.A. Orders between Channels and Implications for Partial Information Decomposition. *Entropy* **2023**, *25*, 975.
26. Pearl, J. *Causality*; Cambridge university press, 2009.
27. Colenbier, N.; Van de Steen, F.; Uddin, L.Q.; Poldrack, R.A.; Calhoun, V.D.; Marinazzo, D. Disambiguating the role of blood flow and global signal with partial information decomposition. *NeuroImage* **2020**, *213*, 116699.
28. Sherrill, S.P.; Timme, N.M.; Beggs, J.M.; Newman, E.L. Partial information decomposition reveals that synergistic neural integration is greater downstream of recurrent information flow in organotypic cortical cultures. *PLoS computational biology* **2021**, *17*, e1009196.
29. Sherrill, S.P.; Timme, N.M.; Beggs, J.M.; Newman, E.L. Correlated activity favors synergistic processing in local cortical networks in vitro at synaptically relevant timescales. *Network Neuroscience* **2020**, *4*, 678–697.
30. Proca, A.M.; Rosas, F.E.; Luppi, A.I.; Bor, D.; Crosby, M.; Mediano, P.A. Synergistic information supports modality integration and flexible learning in neural networks solving multiple tasks. *arXiv preprint arXiv:2210.02996* **2022**.
31. Kay, J.W.; Schulz, J.M.; Phillips, W.A. A comparison of partial information decompositions using data from real and simulated layer 5b pyramidal cells. *Entropy* **2022**, *24*, 1021.
32. Liang, P.P.; Cheng, Y.; Fan, X.; Ling, C.K.; Nie, S.; Chen, R.; Deng, Z.; Mahmood, F.; Salakhutdinov, R.; Morency, L.P. Quantifying & modeling feature interactions: An information decomposition framework. *arXiv preprint arXiv:2302.12247* **2023**.
33. Hamman, F.; Dutta, S. Demystifying Local and Global Fairness Trade-offs in Federated Learning Using Partial Information Decomposition. *arXiv preprint arXiv:2307.11333* **2023**.
34. Gutknecht, A.J.; Makkeh, A.; Wibral, M. From Babel to Boole: The Logical Organization of Information Decompositions. *arXiv preprint arXiv:2306.00734* **2023**.
35. Quax, R.; Har-Shemesh, O.; Sloot, P.M. Quantifying synergistic information using intermediate stochastic variables. *Entropy* **2017**, *19*, 85.
36. Rosas, F.E.; Mediano, P.A.; Rassouli, B.; Barrett, A.B. An operational information decomposition via synergistic disclosure. *Journal of Physics A: Mathematical and Theoretical* **2020**, *53*, 485001.

37. Krippendorff, K. Ross Ashby's information theory: a bit of history, some solutions to problems, and what we face today. *International journal of general systems* **2009**, *38*, 189–212.
38. James, R.; Ellison, C.; Crutchfield, J. "dit": a Python package for discrete information theory. *Journal of Open Source Software* **2018**, *3*, 738.

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.