A New Outlook on Shannon's Information Measures

Raymond W. Yeung, *Member, IEEE*

***Abstract*—Let X_i, $i=1,\cdots,n$, be discrete random variables, and $\tilde{X}_i$ be a set variable corresponding to X_i. Define the universal set Ω to be $\bigcup_{i=1}^{n}\tilde{X}_i$ and let $\mathcal{F}$ be the σ-field generated by $\{\tilde{X}_i, i=1,\cdots,n\}$. It is shown that Shannon's information measures on the random variables X_i, $i=1,\cdots,n$, constitute a unique measure μ^* on $\mathcal{F}$, which is called the I-Measure. In other words, the Shannon information measure (i.e., Shannon's information measures as a whole) *is* a measure on $\mathcal{F}$, thus establishing the analogy between information theory and set theory. Therefore each information theoretic operation can formally be viewed as a set theoretic operation, and vice versa. This point of view, which we believe is of fundamental importance, has apparently been overlooked in the past by information theorists. As a consequence the I-Diagram is introduced, which is a geometrical representation of the relationship among the information measures. The I-Diagram is analogous to the Venn Diagram in set theory. The use of the I-Diagram is discussed; some applications of which reveal results that may otherwise be difficult to discover. A formula is also derived for the value of the I-Measure of the atoms of $\mathcal{F}$ and its sub-σ-fields generated by some subsets of the basic set variables.**

***Index Terms*—Shannon's information measures, I-Measure, I-Diagram.**

I. Preliminaries and Motivation

THE random variables in this paper are discrete. For two random variables X and Y, Shannon's information measures ([1], also c.f. [2], [3], [6]–[9]) are $H(X)$, $H(Y)$, $H(X|Y)$, $H(Y|X)$, and $I(X;Y)$. Let $\tilde{X}$ and $\tilde{Y}$ be set variables corresponding to X and Y, respectively. By letting $\tilde{X}\cup\tilde{Y}$ be the universal set Ω, we can define a real measure[1] μ^* on the σ-field

$$\mathcal{F}=\{(\tilde{X}\cup\tilde{Y}),\tilde{X},\tilde{Y},(\tilde{X}\cap\tilde{Y}),(\tilde{X}\cap\tilde{Y}^c),(\tilde{X}^c\cap\tilde{Y}),(\tilde{X}\cap\tilde{Y})^c,\phi\}$$

by

$$\begin{aligned}\mu^*(\tilde{X}\cup\tilde{Y})&=H(X,Y)\\ \mu^*(\tilde{X})&=H(X)\\ \mu^*(\tilde{Y})&=H(Y)\\ \mu^*(\tilde{X}\cap\tilde{Y})&=I(X;Y)\\ \mu^*(\tilde{X}-\tilde{Y})&=H(X|Y)\qquad(\tilde{X}-\tilde{Y}=\tilde{X}\cap\tilde{Y}^c)\\ \mu^*(\tilde{Y}-\tilde{X})&=H(Y|X)\\ \mu^*\big((\tilde{X}\cap\tilde{Y})^c\big)&=H(X|Y)+H(Y|X)\end{aligned}$$

and

$$\mu^*(\phi)=0. \tag{1.1}$$

We observe that the left sides of the first six equalities in (1.1) can be obtained from the right sides (which are Shannon's information measures on X and Y) via the following substitution of symbols:

$$\begin{aligned}H/I &\rightarrow \mu\\ , &\rightarrow \cup\\ ; &\rightarrow \cap\\ | &\rightarrow -.\end{aligned}$$

We shall refer to this as the *formal substitution of symbols*. (There is no substantial difference between entropy and mutual information; entropy is sometimes referred to as self-information.) Thus for two random variables X and Y, Shannon's information measures can formally be regarded as a measure on the σ-field $\mathcal{F}$. We shall refer to μ^* as the *I-Measure* for the random variables X and Y. It is easy to see that for any information theoretic identity on X and Y, we can obtain a corresponding set theoretic identity via the formal substitution of symbols. For example, for the information theoretic identity

$$H(X,Y)\equiv H(X)+H(Y)-I(X;Y),$$

there is the set theoretic identity

$$\mu(\tilde{X}\cup\tilde{Y})\equiv\mu(\tilde{X})+\mu(\tilde{Y})-\mu(\tilde{X}\cap\tilde{Y}).$$

Manuscript received June 20, 1989; revised September 21, 1990. This work was presented in part at the 1990 IEEE International Symposium on Information Theory, San Diego, CA, January 14–19, 1990.

The author is with AT&T Bell Laboratories, Room 3M317, Crawfords Corner Road, Holmdel, NJ 07733-1988.

IEEE Log Number 9041820.

[1]Here the word "measure" is used in the sense of measure theory.

0018-9448/91/0500-0466$01.00 ©1991 IEEE

The latter, of course, is a special case of the inclusion–exclusion formula.

Since the distinction between a random variable X and the corresponding set variable $\tilde{X}$ is apparent from the context in most cases, we shall use X to denote both the random variable X and the set variable $\tilde{X}$. For example, $X \cup Y$ obviously means $\tilde{X} \cup \tilde{Y}$. We shall specify whether we are referring to the random variable or the set variable when necessary.

In light of the above analogy between information theory and set theory for two random variables, it is natural to ask whether this analogy can be generalized. To be precise, we raise the following two questions for any finite number of random variables.

1) For any information theoretic identity, is there a corresponding set theoretic identity via the formal substitution of symbols?
2) For any set theoretic identity, is there a corresponding "information theoretic" identity? If so, in what sense?

It was proved by Hu Guo Ding [4] (also see [8, p. 51]) that the proposition in 1) is true. This result, although fundamental in nature, may be less useful from the application point of view. From this point of view, the result asserts that one can discover a set theoretic identity by first discovering an information theoretic identity. This, however, is not a very good approach to discover a set theoretic identity, because there is a much richer set of operations in set theory than in information theory. After all, we are more interested in discovering information theoretic identities than set theoretic identities. (In Hu Guo Ding's framework, not every set theoretic identity has an information theoretic interpretation. For example, the set identity $\mu(X)+\mu(X^c)\equiv\mu(\Omega)$ has no information theoretic interpretation because Ω is not defined. Therefore it is not clear from their work how all the set theoretic operations can be applied in information theory.)

In this paper we present a new approach to understand the underlying mathematical structure of Shannon's information measures, which provides answers to 1) and 2) on the same footing. In Section II, we construct the I-Measure μ^* for any finite number of random variables on a properly defined σ-field $\mathscr{F}$, and show that it is the unique measure on $\mathscr{F}$ that is consistent with Shannon's information measures. Therefore the Shannon information measure (Shannon's information measure as a whole) *is* a measure on $\mathscr{F}$. This point of view, which we believe is of fundamental importance, has apparently been overlooked in the past by information theorists. As a consequence of this result, the use of a diagram similar to a Venn Diagram to represent the relation among the information measures becomes valid. We call such a diagram an I-Diagram. Section III is a formal discussion of the use of the I-diagram, some applications of which reveal results that may otherwise be difficult to discover. The use of diagrams to represent the relation among Shannon's information measures has been suggested by Reza [2], Abramson [3], Dyckman [5], and Papoulis [15]. In Section IV, we discuss some properties of the *mutual information* among three random variables. A formula for the value of the I-Measure on the atoms of $\mathscr{F}$ and its sub-σ-fields generated by some subsets of the basic set variables is derived in Section V. In Section VI we present an interpretation of our results. In Section VII, we conclude by addressing several open issues.

II. Construction of the I-Measure

Let $\mathscr{F}$ be the σ-field generated by $W=\{X_i, i=1,\cdots,n\}$ with $\Omega=\bigcup_{i=1}^n X_i$ being the universal set, where each X_i denotes a random variable as well as the corresponding set variable. We shall refer to the X_i's as the *basic* set variables. An element $A\in\mathscr{F}$ is called an atom of $\mathscr{F}$ if $A=\bigcap_{i=1}^n Y_i$, where Y_i is either X_i or X_i^c. Let $\mathscr{A}\subset\mathscr{F}$ be the set of all the atoms of $\mathscr{F}$ except for $\bigcap_{i=1}^n X_i^c$, which is ϕ because

$$\bigcap_{i=1}^n X_i^c=\left(\bigcup_{i=1}^n X_i\right)^c=\Omega^c=\phi.$$

Then a measure μ on $\mathscr{F}$ is completely specified by *any* set of values $\mu(A)$, $A\in\mathscr{A}$. Let $\|\cdot\|$ be the cardinality of a set. Note that there are 2^n atoms in $\mathscr{F}$, so $\|\mathscr{A}\|=2^n-1$. Since each element of $\mathscr{F}$ is the union of a collection of elements in $\mathscr{A}$, $\|\mathscr{F}\|=2^{\|\mathscr{A}\|}=2^{(2^n-1)}$.

Theorem 1: Let

$$\mathscr{B}=\left\{B\in\mathscr{F}\colon B=\bigcup_{X\in G}X \text{ for some } G\subset W, G\neq\phi\right\}.$$

Then a measure μ on $\mathscr{F}$ is completely specified by *any* set of values $\mu(B)$, $B\in\mathscr{B}$.

Proof: Each element of $\mathscr{B}$ is the union of a nonempty collection of elements of W. Therefore

$$\begin{aligned}\|\mathscr{B}\|&=\sum_{r=1}^n\binom{n}{r}\\&=\sum_{r=0}^n\binom{n}{r}-\binom{n}{0}\\&=2^n-1\end{aligned}$$

using the binomial formula. Thus $\|\mathscr{A}\|=\|\mathscr{B}\|=2^n-1$. Denote (2^n-1) by k. Let $\boldsymbol{u}$ be a column k-vector of $\mu(A)$, $A\in\mathscr{A}$, and $\boldsymbol{v}$ be a column k-vector of $\mu(B)$, $B\in\mathscr{B}$. It is clear that for each $B\in\mathscr{B}$, $\mu(B)$ can be expressed as the sum of some of the elements of $\boldsymbol{u}$. Thus

$$\boldsymbol{v}\equiv\boldsymbol{C}\boldsymbol{u}, \tag{2.1}$$

where $\boldsymbol{C}$ is a $k\times k$ matrix. Note that $\boldsymbol{C}$ is unique. On the other hand, for each $A\in\mathscr{A}$, it can be shown by induction (see Appendix) that $\mu(A)$ can be expressed as a linear combination of $\mu(B)$, $B\in\mathscr{B}$ using the following identi-

ties:

$$\mu(X \cap Y - Z) \equiv \mu(X - Z) + \mu(Y - Z) - \mu(X \cup Y - Z) \quad (2.2)$$

$$\mu(X - Y) \equiv \mu(X \cup Y) - \mu(Y). \quad (2.3)$$

(The existence of this expansion does not imply its uniqueness, but we shall see shortly that this is the case.) Thus

$$\boldsymbol{u} \equiv \boldsymbol{D}\boldsymbol{v} \quad (2.4)$$

for some $k \times k$ matrix $\boldsymbol{D}$. Substituting (2.1) into (2.4) we obtain $\boldsymbol{u} \equiv (\boldsymbol{D}\boldsymbol{C})\boldsymbol{u}$, which implies $\boldsymbol{D}$ is the inverse of $\boldsymbol{C}$, so $\boldsymbol{D}$ is *unique*. Therefore $\mu(A)$, $A \in \mathscr{A}$ are determined once $\mu(B)$, $B \in \mathscr{B}$ are specified. Hence μ is completely specified by *any* set of values $\mu(B)$, $B \in \mathscr{B}$. □

We now prove two identities that will be used shortly.

Lemma 1a:

$$\mu(X \cap Y - Z) \equiv \mu(X \cup Z) + \mu(Y \cup Z) - \mu(X \cup Y \cup Z) - \mu(Z).$$

Proof: This is immediate from (2.2) and (2.3). □

Lemma 1b:

$$I(X;Y|Z) \equiv H(X,Z) + H(Y,Z) - H(X,Y,Z) - H(Z).$$

Proof:

$$\begin{aligned} I(X;Y|Z) &\equiv H(X|Z) - H(X|Y,Z) \\ &\equiv H(X,Z) - H(Z) - [H(X,Y,Z) - H(Y,Z)] \\ &\equiv H(X,Z) + H(Y,Z) - H(X,Y,Z) - H(Z). \end{aligned}$$

□

Note that Lemmas 1a and 1b are related by the formal substitution of symbols. We now construct the I-Measure μ^* on $\mathscr{F}$ using Theorem 1 by defining $\mu^*(\cup_{X \in G} X) \equiv H(X, X \in G)$, for all nonempty $G \subset W$. Shannon's information measures include the entropy, the conditional entropy, and the mutual information and conditional mutual information between two groups of variables. The I-Measure μ^* must be consistent with these information measures in order to be meaningful. In other words, the following identities must hold for all nonempty G, G', $G'' \subset W$:

1) $\mu^*(\cup_{X \in G} X) \equiv H(X, X \in G)$
2) $\mu^*((\cup_{X \in G} X) - (\cup_{Y \in G'} Y)) \equiv H(X, X \in G | Y, Y \in G')$
3) $\mu^*((\cup_{X \in G} X) \cap (\cup_{Y \in G'} Y)) \equiv I(X, X \in G; Y, Y \in G')$
4) $\mu^*((\cup_{X \in G} X) \cap (\cup_{Y \in G'} Y) - (\cup_{Z \in G''} Z)) \equiv I(X, X \in G; Y, Y \in G' | Z, Z \in G'')$.

1) is obviously true by definition. We now show that 4) is true:

$$\begin{aligned} &\mu^*\left(\left(\bigcup_{X \in G} X\right) \cap \left(\bigcup_{Y \in G'} Y\right) - \left(\bigcup_{Z \in G''} Z\right)\right) \\ &\quad \equiv \mu^*\left(\left(\bigcup_{X \in G} X\right) \cup \left(\bigcup_{Z \in G''} Z\right)\right) \\ &\qquad + \mu^*\left(\left(\bigcup_{Y \in G'} Y\right) \cup \left(\bigcup_{Z \in G''} Z\right)\right) \\ &\qquad - \mu^*\left(\left(\bigcup_{X \in G} X\right) \cup \left(\bigcup_{Y \in G'} Y\right) \cup \left(\bigcup_{Z \in G''} Z\right)\right) \\ &\qquad - \mu^*\left(\bigcup_{Z \in G''} Z\right) \quad \text{by Lemma 1a} \\ &\quad \equiv H((X, X \in G), (Z, Z \in G'')) \\ &\qquad + H((Y, Y \in G'), (Z, Z \in G'')) \\ &\qquad - H((X, X \in G), (Y, Y \in G'), (Z, Z \in G'')) \\ &\qquad - H(Z, Z \in G'') \\ &\quad \equiv I(X, X \in G; Y, Y \in G' | Z, Z \in G'') \end{aligned}$$

by Lemma 1b. 2) and 3) can be proved likewise. Therefore μ^* is consistent with Shannon's information measures. It is also clear that μ^* is the unique measure on $\mathscr{F}$ which is consistent with Shannon's information measures, because for such a measure, 1) must be satisfied. Therefore the Shannon information measure (Shannon's information measures as a whole) *is* a measure on $\mathscr{F}$. We point out, however, that there exists $A \in \mathscr{F}$ such that $\mu^*(A)$ does not correspond to a Shannon's information measure (e.g., $\mu^*(X \cap Y \cap Z)$). Nevertheless, we call $\mu^*(A)$ for all $A \in \mathscr{F}$ information measures.

Now, for any information theoretic identity, we can obtain the corresponding set theoretic identity by the direct substitutions in (1)–(4). Thus the proposition in question 1) is true. On the other hand, for any set theoretic identity for a measure μ on $\mathscr{F}$, the identity is still valid if μ is replaced by μ^*. (Note that a set theoretic identity is invariant with the measure, but this is not true for a set theoretic inequality.) Then the set theoretic identity is also an information theoretic identity in the sense that μ^* is uniquely defined by Shannon's information measures, thus answering question 2). Hence the analogy between information theory and set theory is established.

For the rest of the paper we shall assume that $\mathscr{F}$ is the σ-field generated by the set variables corresponding to all the random variables involved in the discussion, and Ω is the union of all these set variables. We shall also use the formal substitution of symbols in both directions for μ^* on the atoms of $\mathscr{F}$ and its sub-σ-fields generated by some subsets of the basic set variables, which include all Shannon's information measures (see examples in Section V).

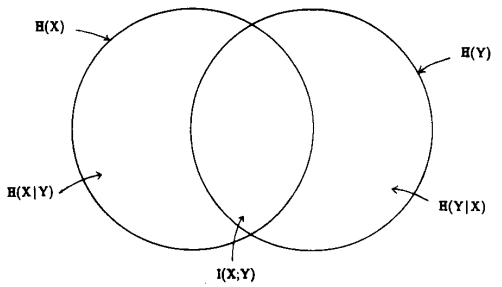

Fig. 1. I-Diagram for X and Y.

For example, $I(X;Y;Z)$ is the same as $\mu^*(X \cap Y \cap Z)$. We call this quantity the *mutual information* among the three random variables X, Y, and Z. Similarly, we have the mutual information among any finite number of random variables. Some properties of the quantity $I(X;Y;Z)$ will be discussed in Section IV. We, however, shall not use the formal substitution of symbols for μ^* on an element in $\mathcal{F}$ that is not an atom of either $\mathcal{F}$ or its sub-σ-fields generated by some subsets of the basic set variables (e.g., $\mu^*((X_1 \cap X_2)^c)$).

III. The I-Diagram

We have established that the Shannon information measure is a measure on $\mathcal{F}$. Therefore it is valid to use a diagram similar to a Venn Diagram to represent the relation among the information measures. We shall refer to such a diagram as an I-Diagram. The I-Diagram for two random variables X and Y is shown in Fig. 1.

A Venn Diagram involving more than three set variables is less easy to visualize, because it in general cannot be illustrated in two dimensions. This is also true for an I-Diagram involving more than three variables. In this section we first discuss the use of the I-Diagram involving three random variables. We then discuss the use of the I-Diagram involving four random variables that form a Markov chain.

In an I-Diagram, the "area" of a region represents the value of μ^* on the corresponding subset of Ω in $\mathcal{F}$. However, it is not in general true that μ^* is nonnegative (see Section IV). Therefore the area of a region in an I-Diagram can represent a negative value. As a consequence, when two random variables X and Y are independent, it only implies that the sum of the area of the regions representing $I(X;Y)$ in the I-Diagram vanishes. It was incorrectly pointed out in [2] that when two random variables are independent, the corresponding set variables are disjoint. We, however, contract a particular region in the I-Diagram if the measure of the corresponding subset of Ω vanishes.

A. Random Variables X, Y, Z

The I-Diagram in the general form for random variables X, Y, and Z is shown in Fig. 2.

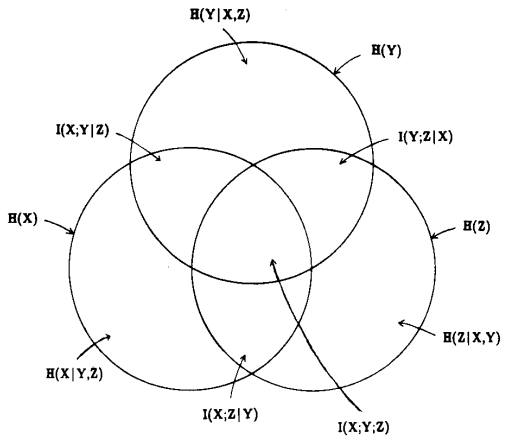

Fig. 2. I-Diagram for X, Y, and Z.

Example 1: We first point out that $I(X;Y;Z)$ is symmetrical in X, Y, and Z. We see from Fig. 2 that

$$\begin{aligned} I(X;Y) - I(X;Y|Z) &\equiv I(Y;Z) - I(Y;Z|X) \\ &\equiv I(X;Z) - I(X;Z|Y) \\ &\quad (\equiv I(X;Y;Z)). \end{aligned}$$

This identity is not well known although it is simple.

Example 2: Let X and Z be independent. Then

$$I(X;Z) = I(X;Z|Y) + I(X;Y;Z) = 0.$$

Since $I(X;Z|Y)$ is nonnegative, $I(X;Y;Z)$ must be nonpositive. Therefore

$$I(X;Y) = I(X;Y|Z) + I(X;Y;Z) \le I(X;Y|Z).$$

This is readily obtained by inspection of Fig. 2.

Example 3: If X, Y, and Z are pairwise independent, then

$$I(X;Y) = I(Y;Z) = I(X;Z) \quad (=0).$$

Then it can be seen by inspection of Fig. 2 that

$$I(X;Y|Z) = I(Y;Z|X) = I(X;Z|Y).$$

Example 4: Suppose $X \multimap Y \multimap Z$ is a Markov chain, that is $I(X;Z|Y) = 0$. By contracting the region corresponding to $I(X;Z|Y)$ in Fig. 2, we obtain the I-Diagram in Fig. 3. Note that in this I-Diagram the area of all the regions are nonnegative because they correspond to Shannon's information measures. In other words, μ^* is a nonnegative measure on $\mathcal{F}$. Then the following can be obtained by inspection:

a) $I(X;Y) \ge I(X;Z)$
b) $H(X|Y) \le H(X|Z)$
c) $I(X;Y) \ge (\le) I(Y;Z)$ iff $I(X;Y|Z) \ge (\le) I(Y;Z|X)$.

The idea of c) is that when we compare $I(X;Y)$ and $I(Y;Z)$, we can eliminate the quantity commonly "possessed" by both, that is, $I(X;Y;Z)$. It was mentioned in

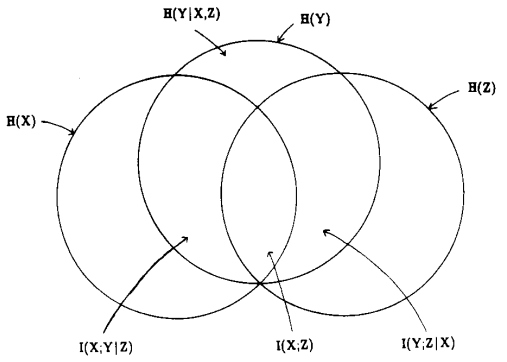

Fig. 3. I-Diagram for $X-\circ-Y-\circ-Z$.

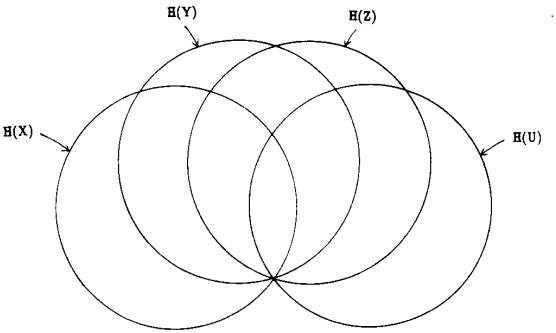

Fig. 4. I-Diagram for $X-\circ-Y-\circ-Z-\circ-U$.

Csiszar and Körner [8] that this quantity has no intuitive meaning. We, however, recognize its *mathematical* significance because it is a quantity commonly possessed by the set variables X, Y, and Z, although this quantity may be negative.

B. Random Variables X, Y, Z, U that Form a Markov Chain

It is in general not possible to illustrate the I-Diagram for four random variables in two dimensions. However, this is possible if $X-\circ-Y-\circ-Z-\circ-U$ is a Markov chain. Using the Markov subchains, we have

a) $X-\circ-Y-\circ-Z$ implies

$$I(X;Z;U|Y)+I(X;Z|Y,U)=I(X;Z|Y)=0. \quad (3.1)$$

b) $X-\circ-Y-\circ-U$ implies

$$I(X;Z;U|Y)+I(X;U|Y,Z)=I(X;U|Y)=0. \quad (3.2)$$

c) $X-\circ-Z-\circ-U$ implies

$$I(X;Y;U|Z)+I(X;U|Y,Z)=I(X;U|Z)=0. \quad (3.3)$$

d) $Y-\circ-Z-\circ-U$ implies

$$I(X;Y;U|Z)=I(Y;U|X,Z)=I(Y;U|Z)=0. \quad (3.4)$$

e) $(X,Y)-\circ-Z-\circ-U$ implies

$$\begin{aligned} &I(X;Y;U|Z)+I(X;U|Y,Z)+I(Y;U|X,Z)\\ &=I(X;U|Z)+I(Y;U|X,Z)\\ &=I(X,Y;U|Z)\\ &=0. \end{aligned} \quad (3.5)$$

Now (3.1) and (3.2) imply

$$I(X;U|Y,Z)=I(X;Z|Y,U). \quad (3.6)$$

(3.3) and (3.6) imply

$$I(X;Y;U|Z)=-I(X;Z|Y,U). \quad (3.7)$$

(3.4) and (3.7) imply

$$I(Y;U|X,Z)=I(X;Z|Y,U). \quad (3.8)$$

We then substitute (3.6), (3.7), and (3.8) in (3.5) to obtain $I(X;Z|Y,U)=0$. From (3.1), (3.6), (3.7), and (3.8) this implies $I(X;Z;U|Y)$, $I(X;U|Y,Z)$, $I(X;Y;U|Z)$ and $I(Y;U|X,Z)$ all vanish. By contracting the corresponding regions, the I-Diagram for X, Y, Z, and U is shown in Fig. 4. Note that the I-Diagrams in Figs. 3 and 4 have the common property that all the circles in the diagram intersect at one point.

Again the area of all the regions in Fig. 4 correspond to Shannon's information measures, thus μ^* is nonnegative. The following can then be obtained by inspection:

a) $I(Y;Z)\geq I(X;U)$
b) $I(Y;Z)=I(X;U)+I(X;Z|U)+I(Y;U|X)+I(Y;Z|X,U)$
c) $H(Y,U|X,Z)=H(Y|X,Z)+H(U|Z)$
d) $H(Y|X,U)=H(Y|X,Z)+I(Y;Z|X,U)$.

Note that a) is the celebrated Data Processing Theorem ([2], [3], [6]–[9]). These relations can of course be obtained using the chain rule and the Markov conditions, but they can be *visualized* with the I-Diagram. In an upcoming paper [14] we shall discuss the general structure of the I-Measure of a Markov chain.

IV. Characterization of $I(X;Y;Z)$

In this section we discuss some properties of the quantity $I(X;Y;Z)$, which appears to be of fundamental interest.

Theorem 2:

$$-\min\{I(X;Y|Z),I(Y;Z|X),I(X;Z|Y)\} \leq I(X;Y;Z)\leq \min\{I(X;Y),I(Y;Z),I(X;Z)\}.$$

Proof: We first prove the lower bound,

$$\begin{aligned} I(X;Y;Z)&=I(X;Y)-I(X;Y|Z)\\ &\geq -I(X;Y|Z). \end{aligned} \quad (4.1)$$

By symmetry of $I(X;Y;Z)$, it is also greater than or equal to $-(Y;Z|X)$ and $-I(X;Z|Y)$, proving the lower bound. The upper bound can also be obtained trivially. From (4.1) we have $I(X;Y;Z)\leq I(X;Y)$. Again we establish the upper bound by symmetry, completing the proof. □

We now give a classical example (see Gallager [6], Problem 2.20) in which the quantity $I(X;Y;Z)$ is strictly

negative. Let X and Y be independent binary random variables taking values in $\{0,1\}$, with $p[X=0]=p[Y=0]=\frac{1}{2}$, and Z be their modulo 2 sum. Then $I(X;Y)=0$, and

$$\begin{aligned} I(X;Y|Z) &= H(X|Z)-H(X|Y,Z) \\ &= 1-0 \\ &= 1. \end{aligned}$$

Therefore

$$\begin{aligned} I(X;Y;Z) &= I(X;Y)-I(X;Y|Z) \\ &= -1. \end{aligned}$$

Recall that $I(X;Y;Z)$ is symmetrical in X, Y, and Z. The interpretation of $I(X;Y;Z)$ being negative is that the mutual information between *any* two of the random variables X, Y, and Z increases when the other random variable is given. Similar interpretations can be made for $I(X;Y;Z)$ being positive and zero.

A trivial sufficient condition for $I(X;Y;Z)$ to vanish is X, Y, and Z being mutually independent. A less trivial sufficient condition is given in the next theorem.

Theorem 3: $I(X;Y;Z)$ vanishes if

1) $X-\circ-Y-\circ-Z$ is a Markov chain and X and Z are independent, or
2) $X-\circ-Z-\circ-Y$ is a Markov chain and X and Y are independent, or
3) $Y-\circ-X-\circ-Z$ is a Markov chain and Y and Z are independent.

Proof: It suffices to prove 1) because $I(X;Y;Z)$ is symmetrical in X, Y, and Z. The Markov chain $X-\circ-Y-\circ-Z$ implies

$$p(X,Y,Z)=p(X)p(Y|X)p(Z|Y).$$

The independence of X and Z implies

$$p(X,Z)=p(X)p(Z).$$

Thus

$$\begin{aligned} & I(X;Y;Z) \\ &= \boldsymbol{E}\log\frac{p(X,Y)p(Y,Z)p(X,Z)}{p(X)p(Y)p(Z)p(X,Y,Z)} \\ &= \boldsymbol{E}\log\frac{p(X,Y)p(Y,Z)p(X)p(Z)}{p(X)p(Y)p(Z)p(X)p(Y|X)p(Z|Y)} \\ &= \boldsymbol{E}\log\frac{p(X,Y)p(Y,Z)}{p(Y)p(X)p(Y|X)p(Z|Y)} \\ &= \boldsymbol{E}\log\frac{p(Y|X)p(Z|Y)}{p(Y|X)p(Z|Y)} \\ &= \boldsymbol{E}\log 1 \\ &= 0. \end{aligned}$$

□

It is not apparent that, for example, condition 1) in Theorem 3 is possible if neither X and Y nor Y and Z are independent of each other. The following is such an example: Let X and Z be independent of each other, and choose Y to be (X,Z) (also see examples in Hekstra and Willems [12]). This has recently been studied by Berger and Yeung, in [10], [11], where they tackled a multiterminal source coding problem. They introduced the notion of *weak independence* and their results are as follows.

Definition: X is weakly independent of Y if the rows of the stochastic matrix $\boldsymbol{P}_{X|Y}=[p(x|y)]$ are linearly dependent.

Theorem 4: For jointly distributed random variables X and Y, there exists a random variable Z satisfying

1) $X-\circ-Y-\circ-Z$ is a Markov chain,
2) X and Z are independent,
3) Y and Z are not independent,

if and only if X is weakly independent of Y.

The conditions in Theorem 3, however, are not necessary. The following is a counterexample. For binary random variables X, Y, and Z, let $p[X=i,Y=j,Z=k]$ be denoted by p_{ijk}, $i,j,k\in\{0,1\}$. Then $I(X;Y;Z)$ vanishes for the distribution

$$\begin{aligned} &p_{000}=0.0625 \quad p_{001}=0.07719 \quad p_{010}=0.0625 \\ &p_{011}=0.0625 \quad p_{100}=0.0625 \quad p_{101}=0.1103 \\ &p_{110}=0.1875 \quad p_{11}=0.0375 \end{aligned}$$

while none of the conditions in Theorem 3 is satisfied.

V. A Formula for the Value of μ^*

In this section we derive a formula for the value of μ^* on the atoms of $\mathscr{F}$ and its sub-σ-fields generated by some subsets of the basic set variables, which include all the Shannon's information measures. (For example, suppose $\mathscr{F}$ is generated by $\{X_1,X_2,X_3\}$. Then $H(X_1|X_2)$ (i.e., $\mu^*(X_1-X_2)$) is an atom of the sub-σ-field of $\mathscr{F}$ generated by $\{X_1,X_2\}$.)

Let $G=\{X_1,\cdots,X_n\}$, $Q_e(G)$ be the set of the joint probabilities of an even number of elements in G, and $Q_o(G)$ be the set of the joint probabilities of an odd number of elements in G. For example, for $G=\{X_1,X_2,X_3\}$,

$$Q_e(G)=\{p(X_1,X_2),p(X_2,X_3),p(X_3,X_1)\}$$

and

$$Q_o(G)=\{p(X_1),p(X_2),p(X_3),p(X_1,X_2,X_3)\}. \quad (5.1)$$

For any set Q such that each element in Q is the joint probability of some finite number of variables, define

$$\pi(Q)=\prod_{p\in Q}p,$$

where $\pi(\phi)=1$ by convention, and for any finite collection of variables S, define the set $J(Q,Y\in S)$ such that it is obtained from Q by replacing $p(\cdot)$ with $p(\cdot,Y\in S)$ in the elements of Q. For example, following (5.1) and $S=\{Y_1,Y_2\}$,

$$\begin{aligned} J(Q_e(G),Y\in S)=\{&p(X_1,X_2,Y_1,Y_2), \\ &p(X_2,X_3,Y_1,Y_2),p(X_3,X_1,Y_1,Y_2)\}. \end{aligned}$$

If $S=\phi$, we adopt the convention $J(Q,Y\in S)=Q$. Simi-

larly, we define the set $K(Q, Y \in S)$ such that it is obtained from Q by replacing $p(\cdot)$ with $p(\cdot|Y \in S)$ in the elements of Q. Thus

$$K(Q_e(G), Y \in S) = \{p(X_1, X_2|Y_1, Y_2), p(X_2, X_3|Y_1, Y_2), p(X_3, X_1|Y_1, Y_2)\}.$$

Theorem 5: Let $G = \{X_1, \cdots, X_n\}$, $n \geq 1$, and $S = \{Y_1, \cdots, Y_m\}$, $m \geq 0$. Then

$$\mu^*\left(\left(\bigcap_{X \in G} X\right) - \left(\bigcup_{Y \in S} Y\right)\right) = E \log \frac{\pi(K(Q_e(G), Y \in S))}{\pi(K(Q_o(G), Y \in S))}.$$

Instead of proving this theorem directly, which we believe is very difficult, we first prove the following lemma.

Lemma 2: Following the notation in Theorem 5,

$$\mu^*\left(\left(\bigcap_{X \in G} X\right) \cup \left(\bigcup_{Y \in S} Y\right)\right) = E \log \frac{\pi(J(Q_e(G), Y \in S))}{\pi(J(Q_o(G), Y \in S))}. \tag{5.2}$$

Proof: We prove (5.2) by induction on n. For $n = 1$ and for all $m \geq 0$,

$$\begin{aligned}\mu^*\left(X_1 \cup \left(\bigcup_{Y \in S} Y\right)\right) &= H(X_1, Y \in S) \\ &= E \log \frac{1}{p(X_1, Y \in S)} \\ &= E \log \frac{\pi(J(Q_e(G), Y \in S))}{\pi(J(Q_o(G), Y \in S))},\end{aligned}$$

where $Q_e(G)$ is empty and $Q_o(G) = \{p(X_1)\}$ for this case. Thus (5.2) is true for $n = 1$. Assume for some $n \geq 1$ (5.2) is true for all $m \geq 0$, and consider $G' = \{X_1, \cdots, X_n, X_{n+1}\}$. Then

$$\begin{aligned}&\mu^*\left(\left(\bigcap_{X \in G'} X\right) \cup \left(\bigcup_{Y \in S} Y\right)\right) \\ &= \mu^*\left(\left(\left(\bigcap_{X \in G} X\right) \cap X_{n+1}\right) \cup \left(\bigcup_{Y \in S} Y\right)\right) \\ &= \mu^*\left(\left(\left(\bigcap_{X \in G} X\right) \cup \left(\bigcup_{Y \in S} Y\right)\right) \cap \left(X_{n+1} \cup \left(\bigcup_{Y \in S} Y\right)\right)\right) \\ &= \mu^*\left(\left(\bigcap_{X \in G} X\right) \cup \left(\bigcup_{Y \in S} Y\right)\right) + \mu^*\left(X_{n+1} \cup \left(\bigcup_{Y \in S} Y\right)\right) \\ &\quad - \mu^*\left(\left(\bigcap_{X \in G} X\right) \cup \left(X_{n+1} \cup \left(\bigcup_{Y \in S} Y\right)\right)\right) \quad \text{by Lemma 1a} \\ &= E \log \frac{\pi(J(Q_e(G), Y \in S))}{\pi(J(Q_o(G), Y \in S))} + E \log \frac{1}{p(X_{n+1}, Y \in S)} - E \log \frac{\pi(J(Q_e(G), X_{n+1}, Y \in S))}{\pi(J(Q_o(G), X_{n+1}, Y \in S))} \\ &= E \log \frac{\pi(J(Q_e(G), Y \in S))\pi(J(Q_o(G), X_{n+1}, Y \in S))}{\pi(J(Q_o(G), Y \in S))\pi(J(Q_e(G), X_{n+1}, Y \in S))p(X_{n+1}, Y \in S)} \\ &= E \log \frac{\pi(J(Q_e(G) \cup J(Q_o(G), X_{n+1}), Y \in S))}{\pi(J(Q_o(G) \cup J(Q_e(G), X_{n+1}) \cup \{p(X_{n+1})\}, Y \in S))}.\end{aligned}$$

Close examination of $Q_e(G \cup \{X_{n+1}\})$ and $Q_o(G \cup \{X_{n+1}\})$ reveals that

$$Q_e(G \cup \{X_{n+1}\}) = Q_e(G) \cup J(Q_o(G), X_{n+1})$$

and

$$Q_o(G \cup \{X_{n+1}\}) = Q_o(G) \cup J(Q_e(G), X_{n+1}) \cup \{p(X_{n+1})\}.$$

Thus

$$\begin{aligned}&\mu^*\left(\left(\bigcap_{X \in G'} X\right) \cup \left(\bigcup_{Y \in S} Y\right)\right) \\ &\quad = E \log \frac{\pi(J(Q_e(G \cup \{X_{n+1}\}), Y \in S))}{\pi(J(Q_o(G \cup \{X_{n+1}\}), Y \in S))} \\ &\quad = E \log \frac{\pi(J(Q_e(G'), Y \in S))}{\pi(J(Q_o(G'), Y \in S))},\end{aligned}$$

completing the proof. □

Proof of Theorem 5: We have

$$\|Q_e(G)\| = \sum_{\substack{2 \leq k \leq n \\ k \text{ even}}} \binom{n}{k}$$

$$\|Q_o(G)\| = \sum_{\substack{1 \leq k \leq n \\ k \text{ odd}}} \binom{n}{k}.$$

By substituting $a = 1$ and $b = -1$ in the binomial theorem

$$(a + b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k},$$

we see that

$$\|Q_e(G)\| = 2^{n-1} - 1$$

and

$$\|Q_o(G)\| = 2^{n-1}.$$

Now,

$$\mu^*\left(\left(\bigcap_{X\in G'} X\right) - \left(\bigcup_{Y\in S} Y\right)\right)$$
$$= \mu^*\left(\left(\bigcap_{X\in G'} X\right) \cup \left(\bigcup_{Y\in S} Y\right)\right) - \mu^*\left(\bigcup_{Y\in S} Y\right).$$

Therefore

$$\mu^*\left(\left(\bigcap_{X\in G'} X\right) - \left(\bigcup_{Y\in S} Y\right)\right)$$
$$= E\log\frac{\pi(J(Q_e(G), Y\in S))}{\pi(J(Q_o(G), Y\in S))} - E\log\frac{1}{p(Y\in S)}$$
$$= E\log\frac{\pi(J(Q_e(G), Y\in S))p(Y\in S)}{\pi(J(Q_o(G), Y\in S))}$$
$$= E\log\frac{\pi(K(Q_e(G), Y\in S))}{\pi(K(Q_o(G), Y\in S))},$$

where the last equality is obtained by dividing the numerator and denominator by $p(Y\in S)^{2^{n-1}}$, proving the theorem. □

VI. Discussion

We adopt the standard interpretation that the entropy of a random variable is the amount of *uncertainty* about that random variable. For two random variables X and Y, why should the amount of uncertainty about X reduced when Y is given be the same as the amount of uncertainty about Y reduced when X is given (i.e., $H(X) - H(X|Y)$ be identical to $H(Y) - H(Y|X)$)? More generally, why should Shannon's information measures possess the structure of a measure? While it may be difficult to explain this result for the general case on an intuitive level, we shall see that it is transparent in some special cases.

Let Y_i, $i = 1, \cdots, 7$ be independent random variables, and X_j, $j = 1,2,3$ be random variables defined by

$$X_1 = (Y_1, Y_2, Y_3, Y_4),$$
$$X_2 = (Y_2, Y_4, Y_6, Y_7),$$
$$X_3 = (Y_3, Y_4, Y_5, Y_6).$$

The relationship among X_1, X_2, and X_3 is illustrated in Fig. 5. In this particular example, Shannon information measures on the random variables X_1, X_2, and X_3 *naturally* possess the structure of a measure. It is intuitively clear, for example, that

$$H(X_1) - H(X_1|X_2) = H(X_2) - H(X_2|X_1),$$

since both sides of this equation are equal to the sum of the entropies of Y_2 and Y_4. Note that in this example μ^* is nonnegative.

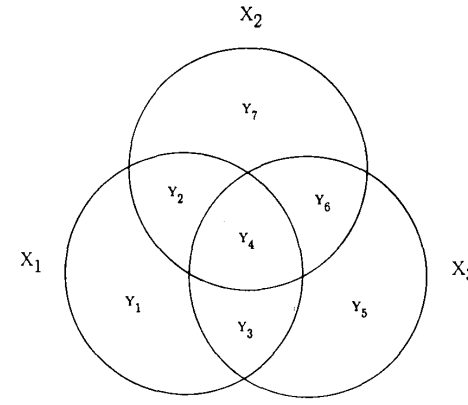

Fig. 5. Relationship among X_1, X_2, and X_3.

Our results show that Shannon's information measures on any finite number of random variables *always* possess the structure of a measure. This may be viewed as an extension of our interpretation in this example, although the nonnegativity of μ^* is not necessarily preserved in the general case.

VII. Future Work

We now address some issues for further investigations.

a) We have constructed a real measure μ^* on $\mathscr{F}$, which we call the *I*-Measure, from the joint distribution of the random variables involved. It should be pointed out that not every real measure μ on $\mathscr{F}$ is an *I*-Measure. For μ to be an *I*-Measure, it is necessary that the value of μ on the elements of $\mathscr{F}$, which correspond to Shannon's information measures are nonnegative. However, given such a measure, it is not clear whether we can always find a joint distribution for the random variables such that Shannon's information measures on these random variables agree with the value of μ on the corresponding elements of $\mathscr{F}$. This is a very fundamental question to be answered.

b) The value of μ^* on the elements of $\mathscr{F}$ that correspond to Shannon's information measures are always nonnegative. A question of interest is: What are the elements of $\mathscr{F}$ on which the value of μ^* are always nonnegative? The more general question of what linear combinations of entropies are always nonnegative was raised by Te Sun Han [13].

Acknowledgment

The author thanks both reviewers for their valuable comments. They contributed much to the structure of the paper and to the discussion in Section VI.

Appendix
A Variation of the Inclusion–Excursion Formula

In this appendix we derive a formula that has the same spirit as the inclusion–exclusion formula. This formula is stated in the following theorem.

Theorem A1: For a set-additive function μ,

$$\mu\left(\bigcap_{r=1}^{n} A_k - B\right) = \sum_{1\le i\le n} \mu(A_i - B) - \sum_{1\le i<j\le n} \mu(A_i \cup A_j - B) + \sum_{1\le i<j<k\le n} \mu(A_i \cup A_j \cup A_k - B) + \cdots + (-1)^{n+1}\mu(A_1 \cup \cdots \cup A_n - B). \tag{A.1}$$

Proof: The theorem will be proved by induction. First (A.1) is obviously true for $n = 1$. Assume (A.1) is true for some $n \ge 1$. We now consider

$$\begin{aligned}
&\mu\left(\bigcap_{r=1}^{n+1} A_k - B\right) \\
&= \mu\left(\left(\bigcap_{r=1}^{n} A_k\right) \cap A_{n+1} - B\right) \\
&= \mu\left(\bigcap_{r=1}^{n} A_k - B\right) + \mu(A_{n+1} - B) \\
&\quad - \mu\left(\left(\bigcap_{r=1}^{n} A_k\right) \cup A_{n+1} - B\right) \qquad \text{by (2.2)} \\
&= \left[\sum_{1\le i\le n} \mu(A_i - B) - \sum_{1\le i<j\le n} \mu(A_i \cup A_j - B) \right. \\
&\quad + \sum_{1\le i<j<k\le n} \mu(A_i \cup A_j \cup A_k - B) \\
&\quad \left. + \cdots + (-1)^{n+1}\mu(A_1 \cup \cdots \cup A_n - B)\right] \\
&\quad + \mu(A_{n+1} - B) - \mu\left(\bigcap_{r=1}^{n} (A_k \cup A_{n+1}) - B\right) \\
&= \left[\sum_{1\le i\le n} \mu(A_i - B) - \sum_{1\le i<j\le n} \mu(A_i \cup A_j - B) \right. \\
&\quad + \sum_{1\le i<j<k\le n} \mu(A_i \cup A_j \cup A_k - B) + \cdots \\
&\quad \left. + (-1)^{n+1}\mu(A_1 \cup \cdots \cup A_n - B)\right] + \mu(A_{n+1} - B) \\
&\quad - \left[\sum_{1\le i\le n} \mu(A_i \cup A_{n+1} - B) \right. \\
&\quad - \sum_{1\le i<j\le n} \mu(A_i \cup A_j \cup A_{n+1} - B) \\
&\quad + \sum_{1\le i<j<k\le n} \mu(A_i \cup A_j \cup A_k \cup A_{n+1} - B) \\
&\quad \left. + \cdots + (-1)^{n+1}\mu(A_1 \cup \cdots \cup A_n \cup A_{n+1} - B)\right] \\
&= \sum_{1\le i\le n+1} \mu(A_i - B) - \sum_{1\le i<j\le n+1} \mu(A_i \cup A_j - B) \\
&\quad + \sum_{1\le i<j<k\le n+1} \mu(A_i \cup A_j \cup A_k - B) \\
&\quad + \cdots + (-1)^{n+2}\mu(A_1 \cup \cdots \cup A_{n+1} - B),
\end{aligned}$$

where we have used the induction hypothesis to expand $\mu(\bigcap_{r=1}^{n} A_k - B)$ and $\mu(\bigcap_{r=1}^{n}(A_k \cup A_{n+1}) - B)$. Thus the theorem is proved. □

References

[1] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, 623–656, 1948.

[2] F. M. Reza, *An Introduction to Information Theory*. New York: McGraw-Hill, 1961.

[3] N. M. Abramson, *Information Theory and Coding*. New York: McGraw-Hill, 1963.

[4] Hu Guo Ding, "On the amount of information," *Teor. Veroyatnost. i Primenen. 4*, pp. 447–455, 1962 (in Russian).

[5] H. L. Dyckman, private communication, 1987.

[6] R. G. Gallagher, *Information Theory and Reliable Communication*. New York: Wiley, 1968.

[7] T. Berger, *Rate Distortion Theory: A Mathematical Basis for Data Compression*. Englewood Cliffs, NJ: Prentice Hall, 1971.

[8] I. Csiszar and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. New York: Academic Press, and Budapest: Akademiai Kiado, 1981.

[9] R. E. Blahut, *Principles and Practice of Information Theory*. New York: Addison-Wesley, 1987.

[10] T. Berger and R. W. Yeung, "Multiterminal source coding with encoder breakdown," *IEEE Trans. Inform. Theory*, vol. 35, no. 2, pp. 237–244, Mar. 1989.

[11] R. W. Yeung, "Some results on multiterminal source coding," Ph.D. dissertation, School of Elect. Eng., Cornell Univ., Ithaca, NY, May 1988.

[12] A. P. Hekstra and F. M. J. Willems, "Dependence balance bounds for single-output two-way channel," *IEEE Trans. Inform. Theory*, vol. 35, no. 1, pp. 44–53, Jan. 1989.

[13] Te Sun Han, "Nonnegative entropy measures of multivariate symmetric correlations," *Inform. Contr.*, vol. 36, pp. 133–156, 1978.

[14] R. W. Yeung, "The structure of the *I*-Measure of a Markov chain," submitted to *IEEE Trans. Inform. Theory*.

[15] A. Papoulis, *Probability, Random Variables and Stochastic Processes*, 2nd ed. New York: McGraw Hill, 1984.