ISSN PRINT 1367-0751
ISSN ONLINE 1368-9894

LOGIC JOURNAL of the IGPL

INTEREST GROUP IN PURE AND APPLIED LOGICS

THE INTERNATIONAL FEDERATION FOR COMPUTATIONAL LOGIC IFCoLog

OXFORD
UNIVERSITY PRESS

Logical information theory: new logical foundations for information theory

DAVID ELLERMAN[*], *Philosophy Department, University of California at Riverside.*

Abstract

There is a new theory of information based on logic. The definition of Shannon entropy as well as the notions on joint, conditional and mutual entropy as defined by Shannon can all be *derived* by a uniform transformation from the corresponding formulas of logical information theory. Information is first defined in terms of *sets* of distinctions without using any probability measure. When a probability measure is introduced, the logical entropies are simply the values of the (product) probability measure on the sets of distinctions. The compound notions of joint, conditional and mutual entropies are obtained as the values of the measure, respectively, on the union, difference and intersection of the sets of distinctions. These compound notions of logical entropy satisfy the usual Venn diagram relationships (e.g. inclusion–exclusion formulas) since they are values of a measure (in the sense of measure theory). The uniform transformation into the formulas for Shannon entropy is linear so it explains the long-noted fact that the Shannon formulas satisfy the Venn diagram relations—as an analogy or mnemonic—since Shannon entropy is not a measure (in the sense of measure theory) on a given set. What is the logic that gives rise to logical information theory? Partitions are dual (in a category-theoretic sense) to subsets, and the logic of partitions was recently developed in a dual/parallel relationship to the Boolean logic of subsets (the latter being usually misspecified as the special case of 'propositional logic'). Boole developed logical probability theory as the normalized counting measure on subsets. Similarly the normalized counting measure on partitions is logical entropy—when the partitions are represented as the set of distinctions that is the complement to the equivalence relation for the partition. In this manner, logical information theory provides the set-theoretic and measure-theoretic foundations for information theory. The Shannon theory is then derived by the transformation that replaces the counting of distinctions with the counting of the number of binary partitions (bits) it takes, on average, to make the *same* distinctions by uniquely encoding the distinct elements—which is why the Shannon theory perfectly dovetails into coding and communications theory.

Keywords: Partition logic, logical entropy, Shannon entropy.

1 Introduction

This article develops the logical theory of information-as-distinctions. It can be seen as the application of the logic of partitions [15] to information theory. Partitions are dual (in a category-theoretic sense) to subsets. George Boole developed the notion of logical probability [7] as the normalized counting measure on subsets in his logic of subsets. This article develops the normalized counting measure on partitions as the analogous quantitative treatment in the logic of partitions. The resulting measure is a new logical derivation of an old formula measuring diversity and distinctions, e.g., Corrado Gini's index of mutability or diversity [19], that goes back to the early 20th century. In view of the idea of information as being based on distinctions (see next section), I refer to this logical measure of distinctions as 'logical entropy'.

This raises the question of the relationship of logical entropy to Claude Shannon's entropy ([40], [41]). The entropies are closely related since they are both ultimately based on the concept of information-as-distinctions—but they represent two different way to quantify distinctions. Logical entropy directly counts the distinctions (as defined in partition logic) whereas Shannon entropy, in effect, counts the minimum number of binary partitions (or yes/no questions) it takes, on average,

[*] E-mail: david@ellerman.org

Vol. 25 No. 5, © The Author 2017. Published by Oxford University Press. All rights reserved.
For Permissions, please email: journals.permissions@oup.com
doi:10.1093/jigpal/jzx022 Advance Access published 7 August 2017

to uniquely determine or designate the distinct entities. Since that gives (in standard examples) a binary code for the distinct entities, the Shannon theory is perfectly adapted for applications to the theory of coding and communications.

The logical theory and the Shannon theory are also related in their compound notions of joint entropy, conditional entropy and mutual information. Logical entropy is a measure in the mathematical sense, so as with any measure, the compound formulas satisfy the usual Venn diagram relationships. The compound notions of Shannon entropy were *defined* so that they also satisfy similar Venn diagram relationships. However, as various information theorists, principally Lorne Campbell, have noted [9], Shannon entropy is not a measure (outside of the standard example of 2^n equiprobable distinct entities where it is the count n of the number of yes/no questions necessary to unique determine or encode the distinct entities)—so one can conclude only that the 'analogies provide a convenient mnemonic' [9, p. 112] in terms of the usual Venn diagrams for measures. Campbell wondered if there might be a 'deeper foundation' [9, p. 112] to clarify how the Shannon formulas can be defined to satisfy the measure-like relations in spite of not being a measure. That question is addressed in this article by showing that there is a transformation of formulas that transforms each of the logical entropy compound formulas into the corresponding Shannon entropy compound formula, and the transform preserves the Venn diagram relationships that automatically hold for measures. This 'dit-bit transform' is heuristically motivated by showing how average counts of distinctions ('dits') can be converted in average counts of binary partitions ('bits').

Moreover, Campbell remarked that it would be 'particularly interesting' and 'quite significant' if there was an entropy measure of sets so that joint entropy corresponded to the measure of the union of sets, conditional entropy to the difference of sets, and mutual information to the intersection of sets [9, p. 113]. Logical entropy precisely satisfies those requirements.

2 Logical information as the measure of distinctions

There is now a widespread view that information is fundamentally about differences, distinguishability and distinctions. As Charles H. Bennett, one of the founders of quantum information theory, put it:

> So information really is a very useful abstraction. It is the notion of distinguishability abstracted away from what we are distinguishing, or from the carrier of information. [5, p. 155]

This view even has an interesting history. In James Gleick's book, *The Information: A History, A Theory, A Flood*, he noted the focus on differences in the 17th century polymath, John Wilkins, who was a founder of the Royal Society. In 1641, the year before Isaac Newton was born, Wilkins published one of the earliest books on cryptography, *Mercury or the Secret and Swift Messenger*, which not only pointed out the fundamental role of differences but noted that any (finite) set of different things could be encoded by words in a binary code.

> For in the general we must note, That whatever is capable of a competent Difference, perceptible to any Sense, may be a sufficient Means whereby to express the Cogitations. It is more convenient, indeed, that these Differences should be of as great Variety as the Letters of the Alphabet; but it is sufficient if they be but twofold, because Two alone may, with somewhat more Labour and Time, be well enough contrived to express all the rest. [47, Chap. XVII, p. 69]

Wilkins explains that a five letter binary code would be sufficient to code the letters of the alphabet since $2^5 = 32$.

> Thus any two Letters or Numbers, suppose *A.B.* being transposed through five Places, will yield Thirty Two Differences, and so consequently will superabundantly serve for the Four and twenty Letters ... .[47, Chap. XVII, p. 69]

As Gleick noted:

> Any difference meant a binary choice. Any binary choice began the expressing of cogitations. Here, in this arcane and anonymous treatise of 1641, the essential idea of information theory poked to the surface of human thought, saw its shadow, and disappeared again for [three] hundred years. [20, p. 161]

Thus *counting distinctions* [12] would seem the right way to measure information,[1] and that is the measure which emerges naturally out of partition logic—just as finite logical probability emerges naturally as the measure of *counting elements* in Boole's subset logic.

Although usually named after the special case of 'propositional' logic, the general case is Boole's logic of subsets of a universe U (the special case of $U = 1$ allows the propositional interpretation since the only subsets are 1 and ∅ standing for truth and falsity). Category theory shows there is a duality between sub-sets and quotient-sets (= partitions = equivalence relations), and that allowed the recent development of the dual logic of partitions ([13], [15]). As indicated in the title of his book, *An Investigation of the Laws of Thought on which are founded the Mathematical Theories of Logic and Probabilities* [7], Boole also developed the normalized counting measure on subsets of a finite universe U which was finite logical probability theory. When the same mathematical notion of the normalized counting measure is applied to the partitions on a finite universe set U (when the partition is represented as the complement of the corresponding equivalence relation on $U \times U$) then the result is the formula for logical entropy.

In addition to the philosophy of information literature [4], there is a whole sub-industry in mathematics concerned with different notions of 'entropy' or 'information' ([2]; see [45] for a recent 'extensive' analysis) that is long on formulas and 'intuitive axioms' but short on interpretations. Out of that plethora of definitions, logical entropy is the *measure* (in the technical sense of measure theory) of information that arises out of partition logic just as logical probability theory arises out of subset logic.

The logical notion of information-as-distinctions supports the view that the notion of information is independent of the notion of probability and should be based on finite combinatorics. As Andrey Kolmogorov put it:

> Information theory must precede probability theory, and not be based on it. By the very essence of this discipline, the foundations of information theory have a finite combinatorial character. [27, p. 39]

Logical information theory precisely fulfills Kolmogorov's criterion.[2] It starts simply with a set of distinctions defined by a partition on a finite set U, where a distinction is an ordered pair of elements

[1] This article is about what Adriaans and van Benthem call 'Information B: Probabilistic, information-theoretic, measured quantitatively', not about 'Information A: knowledge, logic, what is conveyed in informative answers' where the connection to philosophy and logic is built-in from the beginning. Likewise, the paper is not about Kolmogorov-style 'Information C: Algorithmic, code compression, measured quantitatively' [4, p. 11].

[2] Kolmogorov had something else in mind such as a combinatorial development of Hartley's $\log(n)$ on a set of n equiprobable elements [28].

of U in distinct blocks of the partition. Thus the 'finite combinatorial' object is the *set of distinctions* ('ditset') or *information set* ('infoset') associated with the partition, i.e., the complement in $U \times U$ of the equivalence relation associated with the partition. To get a quantitative measure of information, any probability distribution on U defines a product probability measure on $U \times U$, and the *logical entropy* is simply that probability measure of the information set.

3 Duality of subsets and partitions

Logical entropy is to the logic of partitions as logical probability is to the Boolean logic of subsets. Hence we will start with a brief review of the relationship between these two dual forms of mathematical logic.

Modern category theory shows that the concept of a subset dualizes to the concept of a quotient set, equivalence relation, or partition. F. William Lawvere called a subset or, in general, a subobject a 'part' and then noted: 'The dual notion (obtained by reversing the arrows) of 'part' is the notion of *partition*.' [31, p. 85] That suggests that the Boolean logic of subsets (usually named after the special case of propositions as 'propositional' logic) should have a dual logic of partitions ([13], [15]).

A *partition* $\pi = \{B_1, ..., B_m\}$ on U is a set of subsets, called cells or blocks, B_i that are mutually disjoint and jointly exhaustive ($\cup_i B_i = U$). In the duality between subset logic and partition logic, the dual to the notion of an 'element' of a subset is the notion of a 'distinction' of a partition, where $(u, u') \in U \times U$ is a *distinction* or *dit* of π if the two elements are in different blocks, i.e., the 'dits' of a partition are dual to the 'its' (or elements) of a subset. Let $\mathrm{dit}(\pi) \subseteq U \times U$ be the set of distinctions or *ditset* of π. Thus the *information set* or *infoset* associated with a partition π is ditset $\mathrm{dit}(\pi)$. Similarly an *indistinction* or *indit* of π is a pair $(u, u') \in U \times U$ in the same block of π. Let $\mathrm{indit}(\pi) \subseteq U \times U$ be the set of indistinctions or *inditset* of π. Then $\mathrm{indit}(\pi)$ is the equivalence relation associated with π and $\mathrm{dit}(\pi) = U \times U - \mathrm{indit}(\pi)$ is the complementary binary relation that has been called an *apartness relation* or a *partition relation*.

4 Classical subset logic and partition logic

The algebra associated with the subsets $S \subseteq U$ is the Boolean algebra $\wp(U)$ of subsets of U with the inclusion of elements as the partial order. The corresponding algebra of partitions π on U is the *partition algebra* $\prod(U)$ defined as follows:

- the *partial order* $\sigma \preceq \pi$ of partitions $\sigma = \{C, C', ...\}$ and $\pi = \{B, B', ...\}$ holds when π *refines* σ in the sense that for every block $B \in \pi$ there is a block $C \in \sigma$ such that $B \subseteq C$, or, equivalently, using the element-distinction pairing, the partial order is the inclusion of distinctions: $\sigma \preceq \pi$ if and only if (iff) $\mathrm{dit}(\sigma) \subseteq \mathrm{dit}(\pi)$;
- the minimum or bottom partition is the *indiscrete partition* (or blob) $\mathbf{0} = \{U\}$ with one block consisting of all of U;
- the maximum or top partition is the *discrete partition* $\mathbf{1} = \{\{u_j\}\}_{j=1,...,n}$ consisting of singleton blocks;
- the *join* $\pi \vee \sigma$ is the partition whose blocks are the non-empty intersections $B \cap C$ of blocks of π and blocks of σ, or, equivalently, using the element-distinction pairing, $\mathrm{dit}(\pi \vee \sigma) = \mathrm{dit}(\pi) \cup \mathrm{dit}(\sigma)$;
- the *meet* $\pi \wedge \sigma$ is the partition whose blocks are the equivalence classes for the equivalence relation generated by: $u_j \sim u_{j'}$ if $u_j \in B \in \pi$, $u_{j'} \in C \in \sigma$, and $B \cap C \neq \emptyset$; and

- $\sigma \Rightarrow \pi$ is the *implication partition* whose blocks are: (1) the singletons $\{u_j\}$ for $u_j \in B \in \pi$ if there is a $C \in \sigma$ such that $B \subseteq C$, or (2) just $B \in \pi$ if there is no $C \in \sigma$ with $B \subseteq C$, so that trivially: $\sigma \Rightarrow \pi = \mathbf{1}$ iff $\sigma \preceq \pi$.[3]

The logical partition operations can also be defined in terms of the corresponding logical operations on subsets. A ditset $\text{dit}(\pi)$ of a partition on U is a subset of $U \times U$ of a particular kind, namely the complement of an equivalence relation. An *equivalence relation* is reflexive, symmetric and transitive. Hence the ditset complement, i.e., a partition relation (or apartness relation), is a subset $P \subseteq U \times U$ that is:

(1) irreflexive (or anti-reflexive), $P \cap \Delta = \emptyset$ (where $\Delta = \{(u,u) : u \in U\}$ is the *diagonal*), i.e., no element $u \in U$ can be distinguished from itself;
(2) symmetric, $(u,u') \in P$ implies $(u',u) \in P$, i.e., if u is distinguished from u', then u' is distinguished from u; and
(3) anti-transitive (or co-transitive), if $(u,u'') \in P$ then for any $u' \in U$, $(u,u') \in P$ or $(u',u'') \in P$, i.e., if u is distinguished from u'', then any other element u' must be distinguished from u or u'' because if u' was equivalent to both, then by transitivity of equivalence, u would be equivalent to u'' contrary to them being distinguished.

That is how distinctions work at the logical level, and that is why the ditset of a partition is the ‘probability-free’ notion of an information set or infoset in the logical theory of information-as-distinctions.

Given any subset $S \subseteq U \times U$, the *reflexive-symmetric-transitive (rst) closure* $\overline{S^c}$ of the complement S^c is the smallest equivalence relation containing S^c, so its complement is the largest partition relation contained in S, which is called the *interior* $\text{int}(S)$ of S. This usage is consistent with calling the subsets that equal their rst-closures *closed subsets* of $U \times U$ (so closed subsets = equivalence relations) so the complements are the *open subsets* (= partition relations). However it should be noted that the rst-closure is not a topological closure since the closure of a union is not necessarily the union of the closures, so the ‘open’ subsets do not form a topology on $U \times U$.

The interior operation $\text{int} : \wp(U \times U) \rightarrow \wp(U \times U)$ provides a universal way to define logical operations on partitions from the corresponding logical subset operations in Boolean logic:

apply the subset operation to the ditsets and then, if necessary,
take the interior to obtain the ditset of the partition operation.

Since the same operations can be defined for subsets and partitions, one can interpret a formula $\Phi(\pi,\sigma,\ldots)$ either way as a subset or a partition. Given either subsets or partitions of U substituted for the variables π, σ,..., one can apply, respectively, subset or partition operations to evaluate the whole formula. Since $\Phi(\pi,\sigma,\ldots)$ is either a subset or a partition, the corresponding proposition is ‘u is an element of $\Phi(\pi,\sigma,\ldots)$’ or ‘(u,u') is a distinction of $\Phi(\pi,\sigma,\ldots)$’. And then the definitions of a valid formula are also parallel, namely, no matter what is substituted for the variables, the whole formula evaluates to the top of the algebra. In that case, the subset $\Phi(\pi,\sigma,\ldots)$ contains all elements of U, i.e., $\Phi(\pi,\sigma,\ldots) = U$, or the partition $\Phi(\pi,\sigma,\ldots)$ distinguishes all pairs (u,u') for distinct elements of U, i.e., $\Phi(\pi,\sigma,\ldots) = \mathbf{1}$. The parallelism between the dual logics is summarized in the following Table 1.

[3] There is a general method to define operations on partitions corresponding to the Boolean operations on subsets ([13], [15]) but the lattice operations of join and meet, and the implication are sufficient to define a partition algebra $\prod(U)$ parallel to the familiar powerset Boolean algebra $\wp(U)$.

TABLE 1. Duality between subset logic and partition logic

Table 1	Subset logic	Partition logic
'Elements' (its or dits)	Elements u of S	Dits (u,u') of π
Inclusion of 'elements'	Inclusion $S \subseteq T$	Refinement: $\text{dit}(\sigma) \subseteq \text{dit}(\pi)$
Top of order = all 'elements'	U all elements	$\text{dit}(\mathbf{1}) = U^2 - \Delta$, all dits
Bottom of order = no 'elements'	$\emptyset$ no elements	$\text{dit}(\mathbf{0}) = \emptyset$, no dits
Variables in formulas	Subsets S of U	Partitions π on U
Operations: $\vee, \wedge, \Rightarrow, \ldots$	Subset ops.	Partition ops.
Formula $\Phi(x,y,\ldots)$ holds	u element of $\Phi(S,T,\ldots)$	(u,u') dit of $\Phi(\pi,\sigma,\ldots)$
Valid formula	$\Phi(S,T,\ldots) = U$, $\forall S,T,\ldots$	$\Phi(\pi,\sigma,\ldots) = \mathbf{1}$, $\forall \pi,\sigma,\ldots$

5 Classical logical probability and logical entropy

George Boole [7] extended his logic of subsets to finite logical probability theory where, in the equiprobable case, the probability of a subset S (event) of a finite universe set (outcome set or sample space) $U = \{u_1,\ldots,u_n\}$ was the number of elements in S over the total number of elements: $\Pr(S) = \frac{|S|}{|U|} = \sum_{u_j \in S} \frac{1}{|U|}$. Pierre-Simon Laplace's classical finite probability theory [30] also dealt with the case where the outcomes were assigned real point probabilities $p = \{p_1,\ldots,p_n\}$ so rather than summing the equal probabilities $\frac{1}{|U|}$, the point probabilities of the elements were summed: $\Pr(S) = \sum_{u_j \in S} p_j = p(S)$–where the equiprobable formula is for $p_j = \frac{1}{|U|}$ for $j = 1,\ldots,n$. The conditional probability of an event $T \subseteq U$ given an event S is $\Pr(T|S) = \frac{p(T \cap S)}{p(S)}$.

In Gian-Carlo Rota's Fubini Lectures [38] (and in his lectures at MIT), he has remarked in view of duality between partitions and subsets that, quantitatively, the 'lattice of partitions plays for information the role that the Boolean algebra of subsets plays for size or probability' [29, p. 30] or symbolically:

information : partitions :: probability : subsets.

Since 'Probability is a measure on the Boolean algebra of events' that gives quantitatively the 'intuitive idea of the size of a set', we may ask by 'analogy' for some measure to capture a property for a partition like 'what size is to a set.' Rota goes on to ask:

> How shall we be led to such a property? We have already an inkling of what it should be: it should be a measure of information provided by a random variable. Is there a candidate for the measure of the amount of information? [38, p. 67]

Our claim is quite simple; the analogue to the size of a subset is the size of the ditset or information set, the set of distinctions, of a partition.[4] The normalized size of a subset is the logical probability of the event, and the normalized size of the ditset of a partition is, in the sense of measure theory, 'the measure of the amount of information' in a partition. Thus we define the *logical entropy* of a partition $\pi = \{B_1,\ldots,B_m\}$, denoted $h(\pi)$, as the size of the ditset $\text{dit}(\pi) \subseteq U \times U$ normalized by the size of $U \times U$:

$$h(\pi) = \frac{|\text{dit}(\pi)|}{|U \times U|} = \sum_{(u_j,u_k) \in \text{dit}(\pi)} \frac{1}{|U|} \frac{1}{|U|}$$

Logical entropy of π (equiprobable case).

[4] The lattice of partitions on U is isomorphically represented by the lattice of partition relations or ditsets on $U \times U$ ([13], [15]), so in that sense, the size of the ditset of a partition is its 'size'.

This is just the product probability measure of the equiprobable or uniform probability distribution on U applied to the information set or ditset $\mathrm{dit}(\pi)$. The inditset of π is $\mathrm{indit}(\pi) = \cup_{i=1}^{m}(B_i \times B_i)$ so where $p(B_i) = \frac{|B_i|}{|U|}$ in the equiprobable case, we have:

$$h(\pi) = \frac{|\mathrm{dit}(\pi)|}{|U \times U|} = \frac{|U \times U| - \sum_{i=1}^{m}|B_i \times B_i|}{|U \times U|} = 1 - \sum_{i=1}^{m}\left(\frac{|B_i|}{|U|}\right)^2 = 1 - \sum_{i=1}^{m} p(B_i)^2.$$

The corresponding definition for the case of point probabilities $p = \{p_1, \ldots, p_n\}$ is to just add up the probabilities of getting a particular distinction:

$$h_p(\pi) = \sum_{(u_j, u_k) \in \mathrm{dit}(\pi)} p_j p_k$$

Logical entropy of π with point probabilities p.

Taking $p(B_i) = \sum_{u_j \in B_i} p_j$, the logical entropy with point probabilities is:

$$h_p(\pi) = \sum_{(u_j, u_k) \in \mathrm{dit}(\pi)} p_j p_k = \sum_{i \neq i'} p(B_i) p(B_{i'}) = 2 \sum_{i < i'} p(B_i) p(B_{i'}) = 1 - \sum_{i=1}^{m} p(B_i)^2.$$

Instead of being given a partition $\pi = \{B_1, \ldots, B_m\}$ on U with point probabilities p_j defining the finite probability distribution of block probabilities $\{p(B_i)\}_i$, one might be given only a finite probability distribution $p = \{p_1, \ldots, p_m\}$. Then substituting p_i for $p(B_i)$ gives the:

$$h(p) = 1 - \sum_{i=1}^{m} p_i^2 = \sum_{i \neq j} p_i p_j$$

Logical entropy of a finite probability distribution.[5]

Since $1 = \left(\sum_{i=1}^{n} p_i\right)^2 = \sum_i p_i^2 + \sum_{i \neq j} p_i p_j$, we again have the logical entropy $h(p)$ as the probability $\sum_{i \neq j} p_i p_j$ of drawing a distinction in two independent samplings of the probability distribution p.

That two-draw probability interpretation follows from the important fact that logical entropy is always the value of a probability measure. The product probability measure on the subsets $S \subseteq U \times U$ is:

$$\mu(S) = \sum \left\{ p_i p_j : \left(u_i, u_j\right) \in S \right\}$$

Product measure on $U \times U$.

Then the logical entropy $h(p) = \mu(\mathrm{dit}(\mathbf{1}_U))$ is just the product measure of the information set or ditset $\mathrm{dit}(\mathbf{1}_U) = U \times U - \Delta$ of the discrete partition $\mathbf{1}_U$ on U.

There are also parallel 'element ↔ distinction' probabilistic interpretations:

- $\Pr(S) = \sum_{u_i \in S} p_i$ is the probability that a single draw from U gives a *element* u_j of S, and
- $h_p(\pi) = \mu(\mathrm{dit}(\pi)) = \sum_{(u_j, u_k) \in \mathrm{dit}(\pi)} p_j p_k$ is the probability that two independent (with replacement) draws from U gives a *distinction* $\left(u_j, u_k\right)$ of π.

The duality between logical probabilities and logical entropies based on the parallel roles of '*its*' (elements of subsets) and '*dits*' (distinctions of partitions) is summarized in Table 2.

This concludes the argument that logical information theory arises out of partition logic just as logical probability theory arises out of subset logic. Now we turn to the formulas of logical information theory and the comparison to the formulas of Shannon information theory.

[5] The formula $1 - \sum_i p_i^2$ is quite old as a measure of diversity and goes back at least to Gini's index of mutability in 1912 [19]. For the long history of the formula, see [12] or [14].

TABLE 2. Classical logical probability theory and classical logical information theory

Table 2	Logical probability theory	Logical information theory
'Outcomes'	Elements $u \in U$ finite	Dits $\left(u,u'\right) \in U \times U$ finite
'Events'	Subsets $S \subseteq U$	Ditsets $\operatorname{dit}(\pi) \subseteq U \times U$
Equiprobable points	$\Pr(S) = \frac{\|S\|}{\|U\|}$	$h(\pi) = \frac{\|\operatorname{dit}(\pi)\|}{\|U \times U\|}$
Point probabilities	$\Pr(S) = \sum \left\{p_j : u_j \in S\right\}$	$h(\pi) = \sum \left\{p_j p_k : \left(u_j,u_k\right) \in \operatorname{dit}(\pi)\right\}$
Interpretation	$\Pr(S) =$ 1-draw prob. of S-element	$h(\pi) =$ 2-draw prob. of π-distinction

6 Entropy as a *measure* of information

For a partition $\pi = \{B_1,...,B_m\}$ with block probabilities $p(B_i)$ (obtained using equiprobable points or with point probabilities), the *Shannon entropy of the partition* (using logs to base 2) is:

$$H(\pi) = -\sum_{i=1}^{m} p(B_i)\log(p(B_i)).$$

Or if given a finite probability distribution $p = \{p_1,...,p_m\}$, the *Shannon entropy of the probability distribution* is:

$$H(p) = -\sum_{i=1}^{m} p_i \log(p_i).$$

Shannon entropy and the many other suggested 'entropies' are routinely called 'measures' of information [2]. The formulas for mutual information, joint entropy and conditional entropy are defined so these Shannon entropies satisfy Venn diagram formulas (e.g. [1, p. 109]; [35, p. 508]) that would follow automatically if Shannon entropy were a measure in the technical sense. As Lorne Campbell put it:

> Certain analogies between entropy and measure have been noted by various authors. These analogies provide a convenient mnemonic for the various relations between entropy, conditional entropy, joint entropy, and mutual information. It is interesting to speculate whether these analogies have a deeper foundation. It would seem to be quite significant if entropy did admit an interpretation as the measure of some set. [9, p. 112]

For any finite set X, a *measure* μ is a function $\mu : \wp(X) \to \mathbb{R}$ such that:

(1) $\mu(\emptyset) = 0$,
(2) for any $E \subseteq X$, $\mu(E) \geq 0$, and
(3) for any disjoint subsets E_1 and E_2, $\mu(E_1 \cup E_2) = \mu(E_1) + \mu(E_2)$.

Considerable effort has been expended to try to find a framework in which Shannon entropy would be a measure in this technical sense and thus would satisfy the desiderata:

> that $H(\alpha)$ and $H(\beta)$ are measures of sets, that $H(\alpha,\beta)$ is the measure of their union, that $I(\alpha,\beta)$ is the measure of their intersection, and that $H(\alpha|\beta)$ is the measure of their difference. The possibility that $I(\alpha,\beta)$ is the entropy of the "intersection" of two partitions is particularly interesting. This "intersection," if it existed, would presumably contain the information common to the partitions α and β. [9, p. 113]

But these efforts have not been successful beyond special cases such as 2^n equiprobable elements where, as Campbell notes, the Shannon entropy is just the counting measure n of the minimum number of binary partitions it takes to distinguish all the elements. In general, Shannon entropy is *not* a measure on a set.[6]

In contrast, it is 'quite significant' that logical entropy *is* a measure such as the normalized counting measure on the ditset $\mathrm{dit}(\pi)$ representation of a partition π as a subset of the set $U \times U$. Thus all of Campbell's desiderata are true when:

- 'sets' = ditsets, the set of distinctions of partitions (or, in general, information sets of distinctions), and
- 'entropies' = normalized counting measure of the ditsets (or, in general, product probability measure on the infosets), i.e. the logical entropies.

7 The dit-bit transform

The logical entropy formulas for various compound notions (e.g. conditional entropy, mutual information and joint entropy) stand in certain Venn diagram relationships *because* logical entropy is a measure. The Shannon entropy formulas for these compound notions, e.g. $H(\alpha,\beta)=H(\alpha)+H(\beta)-I(\alpha,\beta)$, are *defined* so as to satisfy the Venn diagram relationships *as if* Shannon entropy was a measure when it is not. How can that be? Perhaps there is some 'deeper foundation' [9, p. 112] to explain why the Shannon formulas still satisfy those measure-like Venn diagram relationships.

Indeed, there is such a connection, the dit-bit transform. This transform can be heuristically motivated by considering two ways to treat the standard set U_n of n elements with the equal probabilities $p_0 = \frac{1}{n}$. In that basic case of an equiprobable set, we can derive the dit-bit connection, and then by using a probabilistic average, we can develop the Shannon entropy, expressed in terms of bits, from the logical entropy, expressed in terms of (normalized) dits, or vice versa.

Given U_n with n equiprobable elements, the number of dits (of the discrete partition on U_n) is $n^2 - n$ so the normalized dit count is:

$$h(p_0) = h\left(\tfrac{1}{n}\right) = \tfrac{n^2-n}{n^2} = 1 - \tfrac{1}{n} = 1 - p_0 \text{ normalized dits.}$$

That is the dit-count or logical measure of the information in a set of n distinct elements (think of it as the logical entropy of the discrete partition on U_n with equiprobable elements).

But we can also measure the information in the set by the number of binary partitions it takes (on average) to distinguish the elements, and that bit-count is [23]:

$$H(p_0) = H\left(\tfrac{1}{n}\right) = \log(n) = \log\left(\tfrac{1}{p_0}\right) \text{ bits.}$$

Shannon-Hartley entropy for an equiprobable set U of n elements

The *dit-bit connection* is that the Shannon-Hartley entropy $H(p_0) = \log\left(\frac{1}{p_0}\right)$ will play the same role in the Shannon formulas that $h(p_0) = 1 - p_0$ plays in the logical entropy formulas—when both are formulated as probabilistic averages or expectations.

[6] Perhaps, one should say that Shannon entropy is not the measure of any independently defined set. The fact that the Shannon formulas 'act like a measure on a set' can, of course, be formalized by formally associating an (indefinite) 'set' with each random variable X and then *defining* the measure value on the 'set' as $H(X)$. But since there is no independently defined 'set' with actual members and this 'measure' is defined by the Shannon entropy values (rather than the other way around), nothing is added to the already-known fact that the Shannon entropies act like a measure in the Venn diagram relationships. This formalization exercise seems to have been first carried out by Guo Ding Hu [25] but was also noted by Imre Csiszar and Janos Körner [11], and redeveloped by Raymond Yeung ([48], [49]).

The common thing being measured is an equiprobable U_n where $n=\frac{1}{p_0}$.[7] The dit-count for U_n is $h(p_0)=1-p_0$ and the bit-count for U is $H(p_0)=\log\left(\frac{1}{p_0}\right)$, and the dit-bit transform converts one count into the other. Using this dit-bit transform between the two different ways to quantify the 'information' in U_n, each entropy can be developed from the other. Nevertheless, this dit-bit connection should not be interpreted as if it was just converting a length using centimeters to inches or the like. Indeed, the (average) bit-count is a 'coarser-grid' that loses some information in comparison to the (exact) dit-count as shown by the analysis (below) of mutual information. There is no bit-count mutual information between independent probability distributions but there is always dit-count information even between two (non-trivial) independent distributions (see below the proposition that non-empty supports always intersect).

We start with the logical entropy of a probability distribution $p=\{p_1,...,p_n\}$:

$$h(p)=\sum_{i=1}^{n}p_i h(p_i)=\sum_i p_i(1-p_i).$$

It is expressed as the probabilistic average of the dit-counts or logical entropies of the sets U_{1/p_i} with $\frac{1}{p_i}$ equiprobable elements. But if we switch to the binary-partition bit-counts of the information content of those same sets U_{1/p_i} of $\frac{1}{p_i}$ equiprobable elements, then the bit-counts are $H(p_i)=\log\left(\frac{1}{p_i}\right)$ and the probabilistic average is the Shannon entropy: $H(p)=\sum_{i=1}^{n}p_i H(p_i)=\sum_i p_i\log\left(\frac{1}{p_i}\right)$. Both entropies have the mathematical form as a probabilistic average or expectation:

$$\sum_i p_i\left(\text{amount of 'information' in } U_{1/p_i}\right)$$

and differ by using either the dit-count or bit-count conception of information in U_{1/p_i}.

The graph of the dit-bit transform is familiar in information theory from the natural log inequality: $\ln p_i \leq p_i - 1$. Taking negatives of both sides gives the graph (Figure 1) of the dit-bit transform for natural logs: $1-p_i \rightsquigarrow \ln\left(\frac{1}{p_i}\right)$.

The dit-bit connection carries over to all the compound notions of entropy so that the Shannon notions of conditional entropy, mutual information, cross-entropy and divergence can all be developed from the corresponding notions for logical entropy. Since the logical notions are the values of a probability measure, the compound notions of logical entropy have the usual Venn diagram relationships. And then by the dit-bit transform, those Venn diagram relationships carry over to the compound Shannon formulas since the dit-bit transform preserves sums and differences (i.e. is, in that sense, linear). *That* is why the Shannon formulas can satisfy the Venn diagram relationships even though Shannon entropy is not a measure.

The logical entropy formula $h(p)=\sum_i p_i(1-p_i)$ (and the corresponding compound formulas) are put into that form of an average or expectation to apply the dit-bit transform. But logical entropy is the exact measure of the information set $S_{i\neq i'}=\{(i,i'):i\neq i'\}\subseteq\{1,...,n\}\times\{1,...,n\}$ for the product probability measure $\mu:\wp\left(\{1,...,n\}^2\right)\rightarrow[0,1]$ where for $S\subseteq\{1,...,n\}^2$, $\mu(S)=\sum\{p_i p_{i'}:(i,i')\in S\}$, i.e., $h(p)=\mu\left(S_{i\neq i'}\right)$.

[7] Note that $n=1/p_0$ need not be an integer. We are following the usual practice in information theory where an implicit 'on average' interpretation is assumed since actual 'binary partitions' or 'binary digits' (or 'bits') only come in integral units. The 'on average' provisos are justified by the 'Noiseless Coding Theorem' covered in the later section on the statistical interpretation of Shannon entropy.

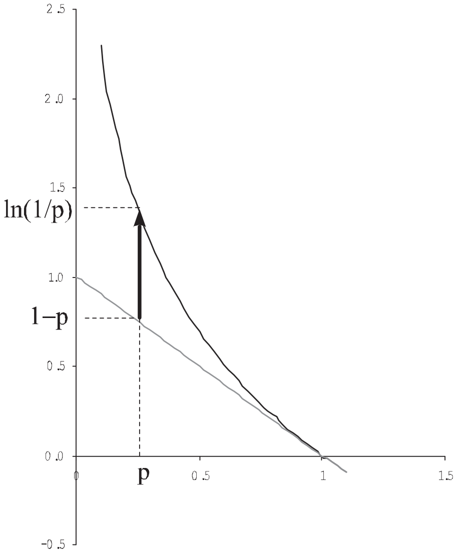

FIG. 1. The dit-bit transform $1-p \rightsquigarrow \ln\left(\frac{1}{p}\right)$ (natural logs).

8 Information algebras and joint distributions

Consider a joint probability distribution $\{p(x,y)\}$ on the finite sample space $X \times Y$ (where to avoid trivialities, assume $|X|,|Y| \geq 2$), with the marginal distributions $\{p(x)\}$ and $\{p(y)\}$ where $p(x) = \sum_{y \in Y} p(x,y)$ and $p(y) = \sum_{x \in X} p(x,y)$. For notational simplicity, the entropies can be considered as functions of the random variables or of their probability distributions, e.g., $h(\{p(x,y)\}) = h(X,Y)$, $h(\{p(x)\}) = h(X)$, and $h(\{p(y)\}) = h(Y)$. For the joint distribution, we have the:

$$h(X,Y) = \sum_{x \in X, y \in Y} p(x,y)[1-p(x,y)] = 1 - \sum_{x,y} p(x,y)^2$$

Logical entropy of the joint distribution

which is the probability that two samplings of the joint distribution will yield a pair of *distinct* ordered pairs (x,y), $(x',y') \in X \times Y$, i.e., with an X-distinction $x \neq x'$ *or* a Y-distinction $y \neq y'$(since ordered pairs are distinct if distinct on one or more of the coordinates). The logical entropy notions for the probability distribution $\{p(x,y)\}$ on $X \times Y$ are all product probability measures $\mu(S)$ of certain subsets $S \subseteq (X \times Y)^2$. These *information sets* or *infosets* are defined solely in terms of equations and inequations (the 'calculus of identity and difference') independent of any probability distributions.

For the logical entropies defined so far, the infosets are:

$$S_X = \{((x,y),(x',y')) : x \neq x'\},$$
$$h(X) = \mu(S_X) = 1 - \sum_x p(x)^2;$$
$$S_Y = \{((x,y),(x',y')) : y \neq y'\},$$
$$h(Y) = \mu(S_Y) = 1 - \sum_y p(y)^2; \text{ and}$$
$$S_{X \vee Y} = \{((x,y),(x',y')) : x \neq x' \vee y \neq y'\} = S_X \cup S_Y,$$
$$h(X,Y) = \mu(S_{X \vee Y}) = \mu(S_X \cup S_Y) = 1 - \sum_{x,y} p(x,y)^2.$$

TABLE 3. Truth table for atoms in the information algebra

Atoms	$x\neq x'$	$y\neq y'$	$X\not\equiv Y$	$X\supset Y$
$S_{X\wedge Y}$	T	T	F	T
$S_{X\wedge\neg Y}$	T	F	T	F
$S_{Y\wedge\neg X}$	F	T	T	T
$S_{\neg X\wedge\neg Y}$	F	F	F	T

The infosets S_X and S_Y, as well as their complements

$$S_{\neg X}=\{((x,y),(x',y')):x=x'\}=(X\times Y)^2-S_X \text{ and}$$
$$S_{\neg Y}=\{((x,y),(x',y')):y=y'\}=(X\times Y)^2-S_Y,[8]$$

generate a Boolean subalgebra $\mathcal{I}(X\times Y)$ of $\wp((X\times Y)\times(X\times Y))$ which will be called the *information algebra of* $X\times Y$. It is defined independently of any probability measure $\{p(x,y)\}$ on $X\times Y$, and any such measure defines the product measure μ on $(X\times Y)\times(X\times Y)$, and the corresponding logical entropies are the product measures on the infosets in $\mathcal{I}(X\times Y)$.

The four atoms $S_X\cap S_Y=S_{X\wedge Y}$, $S_X\cap S_{\neg Y}=S_{X\wedge\neg Y}$, $S_{\neg X}\cap S_Y=S_{\neg X\wedge Y}$, and $S_{\neg X}\cap S_{\neg Y}=S_{\neg X\wedge\neg Y}$ in the information Boolean algebra are non-empty and correspond to the four rows in the truth table (Table 3) for the two propositions $x\neq x'$ and $y\neq y'$(and to the four disjoint areas in the Venn diagram of Figure 2).

For $n=2$ variables X and Y, there are $2^{(2^n)}=16$ ways to fill in the T's and F's to define all the possible Boolean combinations of the two propositions so there are 16 subsets in the information algebra $\mathcal{I}(X\times Y)$. The 15 non-empty subsets in $\mathcal{I}(X\times Y)$ are defined in disjunctive normal form by the union of the atoms that have a T in their row. For instance, the set $S_{X\not\equiv Y}$ corresponding to the symmetric difference or inequivalence $(x\neq x')\not\equiv(y\neq y')$ is $S_{X\not\equiv Y}=S_{X\wedge\neg Y}\cup S_{Y\wedge\neg X}=(S_X-S_Y)\cup(S_Y-S_X)$.

The information algebra $\mathcal{I}(X\times Y)$ is a finite combinatorial structure defined solely in terms of $X\times Y$ using only the distinctions and indistinctions between the elements of X and Y. Any equivalence between Boolean expressions that is a tautology, e.g., $x\neq x'\equiv(x\neq x'\wedge\neg(y\neq y'))\vee(x\neq x'\wedge y\neq y')$, gives a set identity in the information Boolean algebra, e.g., $S_X=(S_X\cap S_{\neg Y})\cup(S_X\cap S_Y)$. Since that union is disjoint, any probability distribution on $X\times Y$ gives the logically necessary identity $h(X)=h(X|Y)+m(X,Y)$ (see below).

In addition to the logically necessary relationships between the logical entropies, other relationships may hold depending on the particular probability distribution on $X\times Y$. Even though all the 15 subsets in the information algebra aside from the empty set $\emptyset$ are always non-empty, some of the logical entropies can still be 0. Indeed, $h(X)=0$ iff the marginal distribution on X has $p(x)=1$ for some $x\in X$. These more specific relationships will depend not just on the infosets but also on their *positive supports* (which depend on the probability distribution):

$$Supp(S_X)=\{((x,y),(x',y')):x\neq x',p(x,y)p(x',y')>0\}\subseteq(X\times Y)^2$$
$$Supp(S_Y)=\{((x,y),(x',y')):y\neq y',p(x,y)p(x',y')>0\}\subseteq(X\times Y)^2.$$

Now $Supp(S_X)\subseteq S_X$ and $Supp(S_Y)\subseteq S_Y$, and for the product probability measure μ on $(X\times Y)^2$, the sets $S_X-Supp(S_X)$ and $S_Y-Supp(S_Y)$ are of measure 0 so:

$$\mu(Supp(S_X))=\mu(S_X)=h(X)$$
$$\mu(Supp(S_Y))=\mu(S_Y)=h(Y).$$

[8] Note that $S_{\neg X}$ and $S_{\neg Y}$ intersect in the diagonal $\Delta\subseteq(X\times Y)^2$.

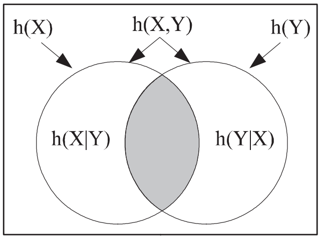

FIG. 2. $h(X,Y)=h(X|Y)+h(Y)=h(Y|X)+h(X)$.

Consider $S_{X\supset Y}=S_{X\wedge Y}\cup S_{Y\wedge\neg X}\cup S_{\neg X\wedge\neg Y}$ and suppose that the probability distribution gives $\mu(S_{X\supset Y})=1$ so that $\mu(S_{X\wedge\neg Y})=0$. That means in a double draw of (x,y) and (x',y'), if $x\neq x'$, then there is zero probability that $y=y'$, so $x\neq x'$ implies (probabilistically) $y\neq y'$. In terms of the Venn diagram, the $h(X)$ area is a subset of the $h(Y)$ area. i.e., $Supp(S_X)\subseteq Supp(S_Y)$ in terms of the underlying sets.

9 Conditional entropies

9.1 Logical conditional entropy

All the compound notions for Shannon and logical entropy could be developed using either partitions (with point probabilities) or probability distributions of random variables as the given data. Since the treatment of Shannon entropy is most often in terms of probability distributions, we will stick to that case for both types of entropy. The formula for the compound notion of logical entropy will be developed first, and then the formula for the corresponding Shannon compound entropy will be obtained by the dit-bit transform.

The general idea of a conditional entropy of a random variable X given a random variable Y is to measure the information in X when we take away the information contained in Y, i.e., the set difference operation in terms of information sets.

For the definition of the conditional entropy $h(X|Y)$, we simply take the product measure of the set of pairs (x,y) and (x',y') that give an X-distinction but not a Y-distinction. Hence we use the inequation $x\neq x'$ for the X-distinction and negate the Y-distinction $y\neq y'$ to get the infoset that is the difference of the infosets for X and Y:

$$\begin{gathered}S_{X\wedge\neg Y}=\{((x,y),(x',y')):x\neq x'\wedge y=y'\}=S_X-S_Y \text{ so}\\ h(X|Y)=\mu(S_{X\wedge\neg Y})=\mu(S_X-S_Y).\end{gathered}$$

Since $S_{X\vee Y}$ can be expressed as the disjoint union $S_{X\vee Y}=S_{X\wedge\neg Y}\uplus S_Y$, we have for the measure μ:

$$h(X,Y)=\mu(S_{X\vee Y})=\mu(S_{X\wedge\neg Y})+\mu(S_Y)=h(X|Y)+h(Y),$$

which is illustrated in the Venn diagram Figure 2.
In terms of the probabilities:

$$\begin{gathered}h(X|Y)=h(X,Y)-h(Y)=\sum_{x,y}p(x,y)(1-p(x,y))-\sum_y p(y)(1-p(y))\\ =\sum_{x,y}p(x,y)[(1-p(x,y))-(1-p(y))]\end{gathered}$$

Logical conditional entropy of X given Y.

Also of interest is the:

$$d(X,Y)=h(X|Y)+h(Y|X)=\mu(S_X\not\equiv S_Y),$$

Logical distance metric

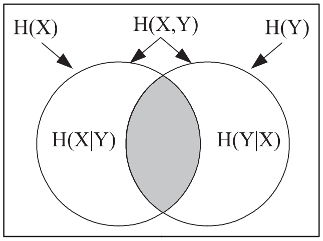

FIG. 3. $H(X,Y)=H(X|Y)+H(Y)=H(Y|X)+H(X)$.

where $\not\equiv$ is the inequivalence (symmetric difference) operation on sets. This logical distance is a Hamming-style distance function [34, p. 66] based on the difference between the random variables. Unlike the Kullback–Leibler divergence (see below), this logical distance is a distance metric.

9.2 Shannon conditional entropy

Given the joint distribution $\{p(x,y)\}$ on $X\times Y$, the conditional probability distribution for a specific $y_0\in Y$ is $p(x|y_0)=\frac{p(x,y_0)}{p(y_0)}$ which has the Shannon entropy: $H(X|y_0)=\sum_x p(x|y_0)\log\left(\frac{1}{p(x|y_0)}\right)$. Then the Shannon conditional entropy $H(X|Y)$ is usually defined as the *average* of these entropies:

$$H(X|Y)=\sum_y p(y)\sum_x \frac{p(x,y)}{p(y)}\log\left(\frac{p(y)}{p(x,y)}\right)=\sum_{x,y}p(x,y)\log\left(\frac{p(y)}{p(x,y)}\right)$$

Shannon conditional entropy of X given Y.

All the Shannon notions can be obtained by the dit-bit transform of the corresponding logical notions. Applying the transform $1-p\rightsquigarrow\log\left(\frac{1}{p}\right)$ to the logical conditional entropy expressed as an average of '$1-p$' expressions: $h(X|Y)=\sum_{x,y}p(x,y)[(1-p(x,y))-(1-p(y))]$, yields the Shannon conditional entropy:

$$H(X|Y)=\sum_{x,y}p(x,y)\left[\log\left(\frac{1}{p(x,y)}\right)-\log\left(\frac{1}{p(y)}\right)\right]=\sum_{x,y}p(x,y)\log\left(\frac{p(y)}{p(x,y)}\right).$$

Since the dit-bit transform preserves sums and differences, we will have the same sort of Venn diagram formula for the Shannon entropies and this can be illustrated in the analogous 'mnemonic' Venn diagram (Figure 3).

10 Mutual information

10.1 Logical mutual information

Intuitively, the mutual logical information $m(X,Y)$ in the joint distribution$\{p(x,y)\}$ would be the probability that a sampled pair of pairs (x,y) and (x',y') would be distinguished in *both* coordinates, i.e., a distinction $x\neq x'$ of $p(x)$ *and* a distinction $y\neq y'$ of $p(y)$. In terms of subsets, the subset for the mutual information is intersection of infosets for X and Y:

$$S_{X\wedge Y}=S_X\cap S_Y \text{ so } m(X,Y)=\mu(S_{X\wedge Y})=\mu(S_X\cap S_Y).$$

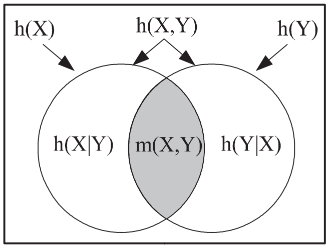

FIG. 4. $h(X,Y)=h(X|Y)+h(Y|X)+m(X,Y)$.

In terms of disjoint unions of subsets:

$$S_{X\vee Y}=S_{X\wedge\neg Y}\uplus S_{Y\wedge\neg X}\uplus S_{X\wedge Y}$$

so

$$h(X,Y)=\mu(S_{X\vee Y})=\mu(S_{X\wedge\neg Y})+\mu(S_{Y\wedge\neg X})+\mu(S_{X\wedge Y})$$
$$=h(X|Y)+h(Y|X)+m(X,Y) \text{ (as in Figure 4),}$$

or:

$$m(X,Y)=h(X)+h(Y)-h(X,Y).$$

Expanding $m(X,Y)=h(X)+h(Y)-h(X,Y)$ in terms of probability averages gives:

$$m(X,Y)=\sum_{x,y}p(x,y)[[1-p(x)]+[1-p(y)]-[1-p(x,y)]]$$

Logical mutual information in a joint probability distribution.

Since $S_Y=S_{Y\wedge\neg X}\cup S_{Y\wedge X}=(S_Y-S_X)\cup(S_Y\cap S_X)$ and the union is disjoint, we have the formula:

$$h(Y)=h(Y|X)+m(X,Y)$$

which can be taken as the basis for a logical analysis of variation (ANOVA) for categorical data. The total variation in Y, $h(Y)$, is equal to the variation in Y 'within' X (i.e. with no variation in X), $h(Y|X)$, plus the variation 'between' Y and X (i.e. variation in both X and Y), $m(X,Y)$.

It is a non-trivial fact that two non-empty partition ditsets *always* intersect [12]. The same holds for the positive supports of the basic infosets S_X and S_Y.

PROPOSITION 1 (Two non-empty supports always intersect)
If $h(X)h(Y)>0$, then $m(X,Y)>0$.

PROOF. Assuming $h(X)h(Y)>0$, the support $Supp(S_X)$ is non-empty and thus there are two pairs (x,y) and (x',y') such that $x\neq x'$ and $p(x,y)p(x',y')>0$. If $y\neq y'$ then $((x,y),(x',y'))\in Supp(S_Y)$ as well and we are finished, i.e., $Supp(S_X)\cap Supp(S_Y)\neq\emptyset$. Hence assume $y=y'$. Since $Supp(S_Y)$ is also non-empty and thus $p(y)\neq 1$, there is another y'' such that for some x'', $p(x'',y'')>0$. Since x'' cannot be equal to both x and x' (by the anti-transitivity of distinctions), at least one of the pairs $((x,y),(x'',y''))$ or $((x',y),(x'',y''))$ is in both $Supp(S_X)$ and $Supp(S_Y)$, and thus the product measure on $S_{\wedge\{X,Y\}}=\{((x,y),(x',y')):x\neq x'\wedge y\neq y'\}$ is positive, i.e., $m(X,Y)>0$. ■

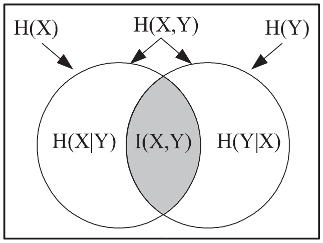

FIG. 5. $H(X,Y)=H(X|Y)+H(Y|X)+I(X,Y)$.

10.2 Shannon mutual information

Applying the dit-bit transform $1-p \rightsquigarrow \log\left(\frac{1}{p}\right)$ to the logical mutual information formula

$$m(X,Y)=\sum_{x,y}p(x,y)[[1-p(x)]+[1-p(y)]-[1-p(x,y)]]$$

expressed in terms of probability averages gives the corresponding Shannon notion:

$$\begin{aligned} I(X,Y)&=\sum_{x,y}p(x,y)\left[\left[\log\left(\tfrac{1}{p(x)}\right)\right]+\left[\log\left(\tfrac{1}{p(y)}\right)\right]-\left[\log\left(\tfrac{1}{p(x,y)}\right)\right]\right] \\ &=\sum_{x,y}p(x,y)\log\left(\tfrac{p(x,y)}{p(x)p(y)}\right) \end{aligned}$$

Shannon mutual information in a joint probability distribution.

Since the dit-bit transform preserves sums and differences, the logical formulas for the Shannon entropies gives the mnemonic Figure 5:

$$I(X,Y)=H(X)+H(Y)-H(X,Y)=H(X,Y)-H(X|Y)-H(Y|X).$$

This is the usual Venn diagram for the Shannon entropy notions that needs to be explained—since the Shannon entropies are not measures. Of course, one could just say the relationship holds for the Shannon entropies because that is how they were defined. It may seem a happy accident that the Shannon definitions all satisfy the measure-like Venn diagram formulas, but as one author put it: ‘Shannon carefully contrived for this “accident” to occur’ [39, p. 153]. As noted above, Campbell asked if ‘these analogies have a deeper foundation’ [9, p. 112] and the dit-bit transform answers that question.

11 Independent joint distributions

A joint probability distribution $\{p(x,y)\}$ on $X \times Y$ is *independent* if each value is the product of the marginals: $p(x,y)=p(x)p(y)$.

For an independent distribution, the Shannon mutual information

$$I(X,Y)=\sum_{x\in X,y\in Y}p(x,y)\log\left(\tfrac{p(x,y)}{p(x)p(y)}\right)$$

is immediately seen to be zero so we have:

$$H(X,Y)=H(X)+H(Y)$$

Shannon entropies for independent $\{p(x,y)\}$.

TABLE 4. Logical entropy relationships under independence

Atomic areas	X	Y
$m(X,Y)=$	$h(X)\times$	$h(Y)$
$h(X\vert Y)=$	$h(X)\times$	$[1-h(Y)]$
$h(Y\vert X)=$	$[1-h(X)]\times$	$h(Y)$
$1-h(X,Y)=$	$[1-h(X)]\times$	$[1-h(Y)]$

For the logical mutual information $m(X,Y)$, independence gives:

$$\begin{aligned} m(X,Y) &= \sum_{x,y} p(x,y)[1-p(x)-p(y)+p(x,y)] \\ &= \sum_{x,y} p(x)p(y)[1-p(x)-p(y)+p(x)p(y)] \\ &= \sum_x p(x)[1-p(x)]\sum_y p(y)[1-p(y)] \\ &= h(X)h(Y) \end{aligned}$$

Logical entropies for independent $\{p(x,y)\}$.

Independence means the joint probability $p(x,y)$ can always be separated into $p(x)$ times $p(y)$. This carries over to the standard two-draw probability interpretation of logical entropy. Thus independence means that in two draws, the probability $m(X,Y)$ of getting distinctions in both X and Y is equal to the probability $h(X)$ of getting an X-distinction times the probability $h(Y)$ of getting a Y-distinction. Similarly, Table 4 shows that, under independence, the four atomic areas in Figure 4 can each be expressed as the four possible products of the areas $\{h(X),1-h(X)\}$ and $\{h(Y),1-h(Y)\}$ that are defined in terms of one variable.

The non-empty-supports-always-intersect proposition shows that $h(X)h(Y)>0$ implies $m(X,Y)>0$, and thus that logical mutual information $m(X,Y)$ is still positive for independent distributions when $h(X)h(Y)>0$, in which case $m(X,Y)=h(X)h(Y)$. This is a striking difference between the average bit-count Shannon entropy and the dit-count logical entropy. Aside from the waste case where $h(X)h(Y)=0$, there are always positive probability mutual distinctions for X and Y, and that dit-count information is not recognized by the coarser-grained average bit-count Shannon entropy.

12 Cross-entropies and divergences

Given two probability distributions $p=\{p_1,...,p_n\}$ and $q=\{q_1,...,q_n\}$ on the same sample space $U=\{1,...,n\}$, we can again consider the drawing of a pair of points but where the first drawing is according to p and the second drawing according to q. The probability that the points are distinct would be a natural and more general notion of logical entropy that would be the:

$$h(p\|q)=\sum_i p_i(1-q_i)=1-\sum_i p_i q_i$$

Logical cross entropy of p and q

which is symmetric. Adding subscripts to indicate which probability measures are being used, the value of the product probability measure μ_{pq} on any $S\subseteq U^2$ is $\mu_{pq}(S)=\sum\{p_i q_{i'}:(i,i')\in S\}$. Thus on the standard information set $S_{i\neq i'}=\left\{(i,i')\in U^2:i\neq i'\right\}=\text{dit}(\mathbf{1}_U)$, the value is:

$$h(p||q)=\mu_{pq}\left(S_{i\neq i'}\right).$$

The logical cross entropy is the same as the logical entropy when the distributions are the same, i.e., if $p=q$, then $h(p\|q)=h(p)=\mu_p\left(S_{i\neq i'}\right)$.

Although the logical cross entropy formula is symmetrical in p and q, there are two different ways to express it as an average to apply the dit-bit transform: $\sum_i p_i(1-q_i)$ and $\sum_i q_i(1-p_i)$. The two transforms are the two asymmetrical versions of Shannon cross entropy:

$$H(p\|q)=\sum_i p_i \log\left(\frac{1}{q_i}\right) \text{ and } H(q||p)=\sum_i q_i \log\left(\frac{1}{p_i}\right)$$

which is not symmetrical due to the asymmetric role of the logarithm, although if $p=q$, then $H(p\|p)=H(p)$. When the logical cross entropy is expressed as an average in a symmetrical way: $h(p||q)=\frac{1}{2}\left[\sum_i p_i(1-q_i)+\sum_i q_i(1-p_i)\right]$, then the dit-bit transform is a *symmetrized Shannon cross entropy*:

$$H_s(p||q)=\frac{1}{2}[H(p||q)+H(q||p)].$$

The *Kullback–Leibler divergence* (or *relative entropy*) $D(p\|q)=\sum_i p_i \log\left(\frac{p_i}{q_i}\right)$ is defined as a 'measure' of the distance or divergence between the two distributions where $D(p\|q)=H(p\|q)-H(p)$. A basic result is the:

$$D(p\|q)\geq 0 \text{ with equality if and only if } p=q$$
Information inequality [10, p. 26].

A *symmetrized Kullback–Leibler divergence* is:

$$D_s(p||q)=D(p||q)+D(q||p)=2H_s(p||q)-[H(p)+H(q)].$$

But starting afresh, one might ask: 'What is the natural notion of distance between two probability distributions $p=\{p_1,...,p_n\}$ and $q=\{q_1,...,q_n\}$ that would always be non-negative, and would be zero if and only if they are equal?' The (Euclidean) distance metric between the two points in $\mathbb{R}^n$ would seem to be the logical answer—so we take that distance squared as the definition of the:

$$d(p\|q)=\sum_i (p_i-q_i)^2$$
Logical divergence (or *logical relative entropy*)[9]

which is symmetric and we trivially have:

$$d(p||q)\geq 0 \text{ with equality iff } p=q$$
Logical information inequality.

We have component-wise:

$$0\leq (p_i-q_i)^2=p_i^2-2p_iq_i+q_i^2=2\left[\frac{1}{n}-p_iq_i\right]-\left[\frac{1}{n}-p_i^2\right]-\left[\frac{1}{n}-q_i^2\right]$$

so that taking the sum for $i=1,...,n$ gives:

$$\begin{aligned} d(p\|q)&=\sum_i (p_i-q_i)^2 \\ &=2\left[1-\sum_i p_iq_i\right]-\left[\left(1-\sum_i p_i^2\right)+\left(1-\sum_i q_i^2\right)\right] \\ &=2h(p\|q)-[h(p)+h(q)] \\ &=2\mu_{pq}\left(S_{i\neq i'}\right)-\left[\mu_p\left(S_{i\neq i'}\right)+\mu_q\left(S_{i\neq i'}\right)\right]. \end{aligned}$$

Logical divergence

[9] In a context where the logical distance $d(X,Y)=h(X|Y)+h(Y|X)$ and the logical divergence $d(p||q)$ are both defined, e.g., two partitions π and σ on the same set U, then the two concepts are the same, i.e., $d(\pi,\sigma)=d(\pi||\sigma)$.

TABLE 5. Comparisons between Shannon and logical entropy formulas

	Shannon entropy	Logical entropy
Entropy	$H(p)=\sum p_i\log(1/p_i)$	$h(p)=\sum p_i(1-p_i)$
Mutual Info.	$I(X,Y)=H(X)+H(Y)-H(X,Y)$	$m(X,Y)=h(X)+h(Y)-h(X,Y)$
Cond. entropy	$H(X\|Y)=H(X)-I(X,Y)$	$h(X\|Y)=h(X)-m(X,Y)$
Independence	$I(X,Y)=0$	$m(X,Y)=h(X)h(Y)$
Indep. Relations	$H(X\|Y)=H(X)$	$h(X\|Y)=h(X)(1-h(Y))$
Cross entropy	$H(p\\|q)=\sum p_i\log(1/q_i)$	$h(p\\|q)=\sum p_i(1-q_i)$
Divergence	$D(p\\|q)=\sum_i p_i\log\left(\frac{p_i}{q_i}\right)$	$d(p\\|q)=\sum_i(p_i-q_i)^2$
Relationships	$D(p\\|q)=H(p\\|q)-H(p)$	$d(p\\|q)=2h(p\\|q)-[h(p)+h(q)]$
Info. Inequality	$D(p\\|q)\geq 0$ with $=$ iff $p=q$	$d(p\\|q)\geq 0$ with $=$ iff $p=q$

Aside from a scale factor, the logical divergence is the same as the *Jensen difference* [36, p. 25] $J(p,q)=h(p||q)-\frac{h(p)+h(q)}{2}$. Then the information inequality implies that the logical cross-entropy is greater than or equal to the average of the logical entropies, i.e., the non-negativity of the Jensen difference:

$$h(p||q)\geq\frac{h(p)+h(q)}{2} \text{ with equality iff } p=q.$$

The half-and-half probability distribution $\frac{p+q}{2}$ that mixes p and q has the logical entropy of

$$h\left(\frac{p+q}{2}\right)=\frac{h(p\|q)}{2}+\frac{h(p)+h(q)}{4}=\frac{1}{2}\left[h(p||q)+\frac{h(p)+h(q)}{2}\right]$$

so that:

$$h(p||q)\geq h\left(\frac{p+q}{2}\right)\geq\frac{h(p)+h(q)}{2} \text{ with equality iff } p=q.$$

Mixing different p and q increases logical entropy.

The logical divergence can be expressed in the symmetrical form of averages to apply the dit-bit transform:

$$d(p\|q)=\left[\sum_i p_i(1-q_i)+\sum_i q_i(1-p_i)\right]-\left[\left(\sum_i p_i(1-p_i)\right)+\left(\sum_i q_i(1-q_i)\right)\right]$$

so the dit-bit transform is:

$$\begin{aligned}&\left[\sum_i p_i\log\left(\frac{1}{q_i}\right)+\sum_i q_i\log\left(\frac{1}{p_i}\right)-\sum_i p_i\log\left(\frac{1}{p_i}\right)-\sum_i q_i\log\left(\frac{1}{q_i}\right)\right]\\ &=\left[\sum_i p_i\log\left(\frac{p_i}{q_i}\right)+\sum_i q_i\log\left(\frac{q_i}{p_i}\right)\right]=D(p||q)+D(q||p)\\ &=D_s(p||q).\end{aligned}$$

Since the logical divergence $d(p||q)$ is symmetrical, it develops via the dit-bit transform to the *symmetrized* version $D_s(p||q)$ of the Kullback–Leibler divergence. The logical divergence $d(p||q)$ is clearly a distance function (or metric) on probability distributions, but even the symmetrized Kullback–Leibler divergence $D_s(p||q)$ may fail to satisfy the triangle inequality [11, p. 58] so it is not a distance metric.

13 Summary of formulas and dit-bit transforms

The Table 5 summarizes the concepts for the Shannon and logical entropies.

Table 6. The dit-bit transform from logical entropy to Shannon entropy

	The dit-bit Transform: $1-p_i \rightsquigarrow \log\left(\frac{1}{p_i}\right)$
$h(p)=$	$\sum_i p_i(1-p_i)$
$H(p)=$	$\sum_i p_i \log(1/p_i)$
$h(X \mid Y)=$	$\sum_{x,y} p(x,y)[(1-p(x,y))-(1-p(y))]$
$H(X \mid Y)=$	$\sum_{x,y} p(x,y)\left[\log\left(\frac{1}{p(x,y)}\right)-\log\left(\frac{1}{p(y)}\right)\right]$
$m(X,Y)=$	$\sum_{x,y} p(x,y)[[1-p(x)]+[1-p(y)]-[1-p(x,y)]]$
$I(X,Y)=$	$\sum_{x,y} p(x,y)\left[\log\left(\frac{1}{p(x)}\right)+\log\left(\frac{1}{p(y)}\right)-\log\left(\frac{1}{p(x,y)}\right)\right]$
$h(p \Vert q)=$	$\frac{1}{2}\left[\sum_i p_i(1-q_i)+\sum_i q_i(1-p_i)\right]$
$H_s(p \Vert q)=$	$\frac{1}{2}\left[\sum_i p_i \log\left(\frac{1}{q_i}\right)+\sum_i q_i \log\left(\frac{1}{p_i}\right)\right]$
$d(p \Vert q)=$	$2h(p \Vert q)-\left[\left(\sum_i p_i(1-p_i)\right)+\left(\sum_i q_i(1-q_i)\right)\right]$
$D_s(p \Vert q)=$	$2H_s(p \Vert q)-\left[\sum_i p_i \log\left(\frac{1}{p_i}\right)+\sum_i q_i \log\left(\frac{1}{q_i}\right)\right]$

The Table 6 summarizes the dit-bit transforms.

14 Entropies for multivariate joint distributions

Let $\{p(x_1,...,x_n)\}$ be a probability distribution on $X_1 \times ... \times X_n$ for finite X_i's. Let S be a subset of $(X_1 \times ... \times X_n)^2$ consisting of certain ordered pairs of ordered n-tuples $\left((x_1,...,x_n),\left(x_1',...,x_n'\right)\right)$ so the product probability measure on S is:

$$\mu(S)=\sum\left\{p(x_1,...,x_n)p\left(x_1',...,x_n'\right):\left((x_1,...,x_n),\left(x_1',...,x_n'\right)\right)\in S\right\}.$$

Then all the logical entropies for this n-variable case are given as the product measure of certain infosets S. Let $I,J \subseteq N$ be subsets of the set of all variables $N=\{X_1,...,X_n\}$ and let $x=(x_1,...,x_n)$ and $x'=\left(x_1',...,x_n'\right)$.

Since two ordered n-tuples are different if they differ in some coordinate, the *joint logical entropy* of all the variables is: $h(X_1,...,X_n)=\mu(S_{\vee N})$ where:

$$\begin{gathered} S_{\vee N}=\left\{(x,x'):\vee_{i=1}^{n}\left(x_i \neq x_i'\right)\right\}=\cup\left\{S_{X_i}:X_i \in N\right\} \text{ where} \\ S_{X_i}=S_{x_i \neq x_i'}=\left\{(x,x'):x_i \neq x_i'\right\} \end{gathered}$$

(where $\vee$ represents the disjunction of statements). For a non-empty $I \subseteq N$, the joint logical entropy of the variables in I could be represented as $h(I)=\mu(S_{\vee I})$ where:

$$S_{\vee I}=\left\{(x,x'):\vee\left(x_i \neq x_i'\right) \text{ for } X_i \in I\right\}=\cup\left\{S_{X_i}:X_i \in I\right\}$$

so that $h(X_1,...,X_n)=h(N)$.

As before, the information algebra $\mathcal{I}(X_1 \times ... \times X_n)$ is the Boolean subalgebra of $\wp\left((X_1 \times ... \times X_n)^2\right)$ generated by the basic infosets S_{X_i} for the variables and their complements $S_{\neg X_i}$.

For the conditional logical entropies, let $I,J \subseteq N$ be two non-empty disjoint subsets of N. The idea for the conditional entropy $h(I|J)$ is to represent the information in the variables I given by the defining condition: $\vee\left(x_i \neq x_i'\right)$ for $X_i \in I$, after taking away the information in the variables J which is defined by the condition: $\vee\left(x_j \neq x_j'\right)$ for $X_j \in J$. 'After the bar |' means 'negate' so we negate that

TABLE 7. Abramson's example giving negative Shannon mutual information $I(X,Y,Z)$

X	Y	Z	$p(x,y,z)$	$p(x,y),p(x,z),p(y,z)$	$p(x),p(y),p(z)$
0	0	0	1/4	1/4	1/2
0	0	1	0	1/4	1/2
0	1	0	0	1/4	1/2
0	1	1	1/4	1/4	1/2
1	0	0	0	1/4	1/2
1	0	1	1/4	1/4	1/2
1	1	0	1/4	1/4	1/2
1	1	1	0	1/4	1/2

condition $\vee\left(x_j \neq x_j'\right)$ for $X_j \in J$ and add it to the condition for I to obtain the conditional logical entropy as $h(I|J)=h(\vee I|\vee J)=\mu(S_{\vee I|\vee J})$ (where $\wedge$ represents the conjunction of statements):

$$\begin{aligned} S_{\vee I|\vee J} &= \left\{(x,x'): \vee\left(x_i \neq x_i'\right) \text{ for } X_i \in I \text{ and} \wedge\left(x_j = x_j'\right) \text{ for } X_j \in J\right\} \\ &= \cup\left\{S_{X_i} : X_i \in I\right\} - \cup\left\{S_{X_j} : X_j \in J\right\} = S_{\vee I} - S_{\vee J}. \end{aligned}$$

The general rule is that the sets satisfying the after-the-bar condition are subtracted from the sets satisfying the before-the-bar condition:

$$\begin{aligned} S_{\vee I|\vee J} &= \cup\left\{S_{X_i} : X_i \in I\right\} - \cup\left\{S_{X_j} : X_j \in J\right\} = \left\{(x,x') : \left(\vee_I x_i \neq x_i'\right) \wedge \left(\wedge_J x_j = x_j'\right)\right\} \\ S_{\vee I|\wedge J} &= \cup\left\{S_{X_i} : X_i \in I\right\} - \cap\left\{S_{X_j} : X_j \in J\right\} = \left\{(x,x') : \left(\vee_I x_i \neq x_i'\right) \wedge \left(\vee_J x_j = x_j'\right)\right\} \\ S_{\wedge I|\vee J} &= \cap\left\{S_{X_i} : X_i \in I\right\} - \cup\left\{S_{X_j} : X_j \in J\right\} = \left\{(x,x') : \left(\wedge_I x_i \neq x_i'\right) \wedge \left(\wedge_J x_j = x_j'\right)\right\} \\ S_{\wedge I|\wedge J} &= \cap\left\{S_{X_i} : X_i \in I\right\} - \cap\left\{S_{X_j} : X_j \in J\right\} = \left\{(x,x') : \left(\wedge_I x_i \neq x_i'\right) \wedge \left(\vee_J x_j = x_j'\right)\right\}. \end{aligned}$$

For the mutual logical information of a non-empty set of variables I, $m(I)=m(\wedge I)=\mu(S_{\wedge I})$ where:

$$S_{\wedge I} = \left\{(x,x') : \wedge_I x_i \neq x_i'\right\}.$$

For the conditional mutual logical information, let $I,J \subseteq N$ be two non-empty disjoint subsets of N so that $m(I|J)=m(\wedge I|\vee J)=\mu\left(S_{\wedge I|\vee J}\right)$ where:

$$S_{\wedge I|\vee J} = \left\{(x,x') : \left(\wedge_I x_i \neq x_i'\right) \wedge \left(\wedge_J x_j = x_j'\right)\right\}.$$

For the logical analysis of variation (ANOVA) of categorical data, the logical entropies in the multivariate case divide up the variation into the natural parts. For instance, suppose that two explanatory variables X_1 and X_2 affect a third response variable Y according to a probability distribution $\{p(x_1,x_2,y)\}$ on $X_1 \times X_2 \times Y$. The logical division in the information sets is:

$$S_Y = S_{Y\wedge\neg X_1\wedge\neg X_2} \cup S_{Y\wedge\neg X_1\wedge X_2} \cup S_{Y\wedge X_1\wedge\neg X_2} \cup S_{Y\wedge X_1\wedge X_2},$$

where $S_{Y\wedge\neg X_1\wedge\neg X_2}$ represents the variation in Y when X_1 and X_2 don't vary, $S_{Y\wedge\neg X_1\wedge X_2}$ is the variation in Y when X_1 does not vary but X_2 does, and so forth. The union is disjoint so the formula for the multivariate logical analysis of variation is:

$$h(Y) = h(Y|X_1,X_2) + m(Y,X_2|X_1) + m(Y,X_1|X_2) + m(Y,X_1,X_2),$$

with the obvious generalization to more explanatory variables $X_1,\ldots,X_n$. Figure 6 (with $X_1 = X$ and $X_2 = Z$) gives the Venn diagram.

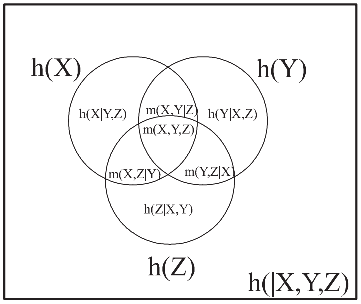

FIG. 6. Venn diagram for logical entropies.

And finally by expressing the logical entropy formulas as averages, the dit-bit transform will give the corresponding versions of Shannon entropy. Consider an example of a joint distribution $\{p(x,y,z)\}$ on $X\times Y\times Z$. The mutual logical information $m(X,Y,Z)=\mu\left(S_{\wedge\{X,Y,Z\}}\right)$ where:

$$S_{\wedge\{X,Y,Z\}}=\{((x,y,z),(x',y',z')):x\neq x'\wedge y\neq y'\wedge z\neq z'\}=S_X\cap S_Y\cap S_Z.$$

From the Venn diagram for $h(X,Y,Z)$, we have (using a variation on the inclusion–exclusion principle)[10]:

$$m(X,Y,Z)=h(X)+h(Y)+h(Z)-h(X,Y)-h(X,Z)-h(Y,Z)+h(X,Y,Z).$$

Substituting the averaging formulas for the logical entropies gives:

$$m(X,Y,Z)= \\ \sum_{x,y,z}p(x,y,z)\left[\begin{array}{c}[1-p(x)]+[1-p(y)]+[1-p(z)]\\ -[1-p(x,y)]-[1-p(x,z)]-[1-p(y,z)]+[1-p(x,y,z)]\end{array}\right].$$

Then applying the dit-bit transform gives the corresponding formula for the multivariate 'Shannon' mutual information:[11]

$$I(X,Y,Z)= \\ \sum_{x,y,z}p(x,y,z)\left[\begin{array}{c}\log\left(\frac{1}{p(x)}\right)+\log\left(\frac{1}{p(y)}\right)+\log\left(\frac{1}{p(z)}\right)\\ -\log\left(\frac{1}{p(x,y)}\right)-\log\left(\frac{1}{p(x,z)}\right)-\log\left(\frac{1}{p(y,z)}\right)+\log\left(\frac{1}{p(x,y,z)}\right)\end{array}\right] \\ =\sum_{x,y,z}p(x,y,z)\left[\log\left(\frac{p(x,y)p(x,z)p(y,z)}{p(x)p(y)p(z)p(x,y,z)}\right)\right]\text{(e.g. [17, p. 57] or [1, p. 129]).}$$

To emphasize that Venn-like diagrams are only a mnemonic analogy, Norman Abramson gives an example [1, pp. 130–1] where the Shannon mutual information of three variables is negative.[12]

Consider the joint distribution $\{p(x,y,z)\}$on $X\times Y\times Z$ where $X=Y=Z=\{0,1\}$. Suppose two dice are thrown, one after the other. Then $X=1$ if the first die came up odd, $Y=1$ if the second die came

[10] The usual version of the inclusion–exclusion principle would be: $h(X,Y,Z)=h(X)+h(Y)+h(Z)-m(X,Y)-m(X,Z)-m(Y,Z)+m(X,Y,Z)$ but $m(X,Y)=h(X)+h(Y)-h(X,Y)$ and so forth, so substituting for $m(X,Y)$, $m(X,Z)$, and $m(Y,Z)$ gives the formula.

[11] The multivariate generalization of the 'Shannon' mutual information was developed not by Shannon but by William J. McGill [33] and Robert M. Fano ([16], [17]) at MIT in the early 1950s and independently by Nelson M. Blachman [6]. The criterion for it being the 'correct' generalization seems to be that it satisfied the generalized Venn diagram formulas that are automatically satisfied by any measure and are thus also obtained *from* the multivariate logical mutual information using the dit-bit transform—as is done here.

[12] Fano had earlier noted that, for three or more variables, the mutual information could be negative [17, p. 58].

up odd, and $Z=1$ if $X+Y$ is odd [18, Exercise 26, p. 143]. Then the probability distribution is in Table 7.

Since the logical mutual information $m(X,Y,Z)$ is the measure $\mu\left(S_{\wedge\{X,Y,Z\}}\right)$, it is always non-negative and in this case is 0:

$$\begin{aligned} m(X,Y,Z) &= h(X)+h(Y)+h(Z)-h(X,Y)-h(X,Z)-h(Y,Z)+h(X,Y,Z) \\ &= \tfrac{1}{2}+\tfrac{1}{2}+\tfrac{1}{2}-\tfrac{3}{4}-\tfrac{3}{4}-\tfrac{3}{4}+\tfrac{3}{4}=\tfrac{3}{2}-\tfrac{6}{4}=0. \end{aligned}$$

All the simple and compound notions of logical entropy have a direct interpretation as a two-draw probability. The logical mutual information $m(X,Y,Z)$ is the probability that in two independent samples of $X\times Y\times Z$, the outcomes would differ in all coordinates. This means the two draws would have the form (x,y,z) and $(1-x,1-y,1-z)$ for the binary variables, but it is easily seen by inspection that $p(x,y,z)=0$ or $p(1-x,1-y,1-z)=0$, so the products of those two probabilities are all 0 as computed—and thus there is no three-way overlap. The two-way overlaps are $m(X,Y)=h(X)+h(Y)-h(X,Y)=\frac{1}{2}+\frac{1}{2}-\frac{3}{4}=\frac{1}{4}$ or since each pair of variables is independent, $m(X,Y)=h(X)h(Y)=\frac{1}{2}\times\frac{1}{2}=\frac{1}{4}$, and similarly for the other pairs of variables. The non-empty-supports-always-intersect result holds for any two variables, but the example shows that there is no necessity in having a three-way overlap, i.e., $h(X)h(Y)h(Z)>0$ does not imply $m(X,Y,Z)>0$.[13]

The Venn diagram like formula for $m(X,Y,Z)$ carries over to $I(X,Y,Z)$ by the dit-bit transform, but the transform does not preserve non-negativity. In this case, the 'area' $I(X,Y,Z)$ is negative:

$$\begin{aligned} I(X,Y,Z) &= H(X)+H(Y)+H(Z)-H(X,Y)-H(X,Z)-H(Y,Z)+H(X,Y,Z) \\ &= 1+1+1-2-2-2+2=3-4=-1. \end{aligned}$$

It is unclear how that can be interpreted as the mutual information contained in the three variables or how the corresponding 'Venn diagram' (Figure 7) can be anything more than a mnemonic. Indeed, as Imre Csiszar and Janos Körner remark:

> The set-function analogy might suggest to introduce further information quantities corresponding to arbitrary Boolean expressions of sets. E.g., the 'information quantity' corresponding to $\mu(A\cap B\cap C)=\mu(A\cap B)-\mu((A\cap B)-C)$ would be $I(X,Y)-I(X,Y|Z)$; this quantity has, however, no natural intuitive meaning. [11, pp. 53–4]

Of course, all this works perfectly well in logical information theory for the 'arbitrary Boolean expressions of sets' in the information algebra $\mathcal{I}(X\times Y\times Z)$, e.g.,

$$m(X,Y,Z)=\mu(S_X\cap S_Y\cap S_Z)=\mu(S_X\cap S_Y)-\mu((S_X\cap S_Y)-S_Z)=m(X,Y)-m(X,Y|Z),$$

which also is a (two-draw) probability measure and thus always non-negative.

Note how the supposed 'intuitiveness' of independent random variables giving disjoint or at least 'zero overlap' Venn diagram areas in the two-variable Shannon case comes at the cost of possibly having 'no natural intuitive meaning' and negative 'areas' in the multivariate case. In probability theory, for a joint probability distribution of 3 or more random variables, there is a distinction between the variables being pair-wise independent and being mutually independent. In any counterexample where three variables are pairwise but not mutually independent [18, p. 127], the Venn diagram areas for $H(X)$, $H(Y)$ and $H(Z)$ have to have pairwise zero overlaps, but since they are not mutually independent, all three areas have a non-zero overlap. The only way that can happen is for the pairwise

[13] The simplest example suffices. There are three non-trivial partitions on a set with three elements and those partitions have no dits in common.

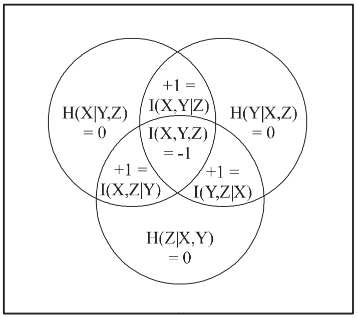

FIG. 7. Negative 'area' $I(X,Y,Z)$ in 'Venn diagram'.

overlaps such as $I(X,Y)=0$ between $H(X)$ and $H(Y)$ to have a positive part $I(X,Y|Z)$ (always non-negative [49, Theorem 2.34, p. 23]) and a negative part $I(X,Y,Z)$ that add to 0 as in Figure 7.

15 Logical entropy and some related notions

The Taylor series for $\ln(x+1)$ around $x=0$ is:

$$\ln(x+1)=\ln(1)+x-\frac{1}{2!}x^2+\frac{1}{3!}x^3 2-\ldots=x-\frac{x^2}{2}+\frac{x^3}{3}-\ldots$$

so substituting $x=p_i-1$ (with $p_i>0$) gives a version of the Newton–Mercator series:

$$-\ln(p_i)=\ln\left(\frac{1}{p_i}\right)=1-p_i+\frac{(p_i-1)^2}{2}-\frac{(p_i-1)^3}{3}+\ldots.$$

Then multiplying by p_i and summing yields:

$$\begin{aligned}H_e(p)&=-\sum_i p_i\ln(p_i)=\sum_i p_i(1-p_i)+\sum_i\frac{p_i(p_i-1)^2}{2}-\ldots\\&=h(p)+\sum_i\frac{p_i(p_i-1)^2}{2}-\ldots.\end{aligned}$$

A similar relationship holds in the quantum case between the von Neumann entropy $S(\rho)=-\mathrm{tr}[\rho\ln(\rho)]$ and the *quantum logical entropy* $h(\rho)=\mathrm{tr}[\rho(1-\rho)]=1-\mathrm{tr}\left[\rho^2\right]$ which is defined by having a density matrix ρ replace the probability distribution p and the trace replace the sum.

This relationship between the Shannon/von Neumann entropies and the logical entropies in the classical and quantum cases is responsible for presenting the logical entropy as a 'linear' approximation to the Shannon or von Neumann entropies since $1-p_i$ is the linear term in the series for $-\ln(p_i)$ [*before* the multiplication by p_i to make the term quadratic!]. And $h(p)=1-\sum_i p_i^2$ or it quantum counterpart $h(\rho)=1-\mathrm{tr}\left[\rho^2\right]$ are even called 'linear entropy' (e.g. [8]) even though the formulas are obviously quadratic.[14] Another name for the quantum logical entropy found in the literature is 'mixedness' [26, p. 5] which at least does not call a quadratic formula 'linear.' It is even called 'impurity' since the complement $1-h(\rho)=\mathrm{tr}\left[\rho^2\right]$ (i.e. the quantum version of Alan Turing's repeat rate $\sum_i p_i^2$ [21]) is called the 'purity.'

[14] Sometimes the misnomer 'linear entropy' is applied to the rescaled logical entropy $\frac{n}{n-1}h(\pi)$. The maximum value of the logical entropy is $h(\mathbf{1})=1-\frac{1}{n}=\frac{n-1}{n}$ so the rescaling gives a maximum value of 1. In terms of the partition-logic derivation of the logical entropy formula, this amounts to sampling without replacement and normalizing $|\mathrm{dit}(\pi)|$ by the number of possible distinctions $|U\times U-\Delta|=n^2-n$ (where $\Delta=\{(u,u):u\in U\}$ is the diagonal) instead of $|U\times U|=n^2$ since: $\frac{|\mathrm{dit}(\pi)|}{|U\times U-\Delta|}=\frac{|\mathrm{dit}(\pi)|}{n(n-1)}=\frac{n}{n-1}\frac{|\mathrm{dit}(\pi)|}{n^2}=\frac{n}{n-1}h(\pi)$.

Quantum logical entropy is beyond the scope of this article but it might be noted that some quantum information theorists have been using that concept to rederive results previously derived using the von Neumann entropy such as the Klein inequality, concavity, and a Holevo-type bound for Hilbert–Schmidt distance ([42], [43]). Moreover, the logical derivation of the logical entropy formulas using the notion of distinctions gives a certain naturalness to the notion of quantum logical entropy.

> We find this framework of partitions and distinction most suitable (at least conceptually) for describing the problems of quantum state discrimination, quantum cryptography and in general, for discussing quantum channel capacity. In these problems, we are basically interested in a distance measure between such sets of states, and this is exactly the kind of knowledge provided by logical entropy ([12], [42]).

There are many older results derived under the misnomer 'linear entropy' or derived for the quadratic special case of the Tsallis–Havrda–Charvat entropy ([24], [44], [45]). Those parameterized families of entropy formulas are sometimes criticized for lacking a convincing interpretation, but we have seen that the quadratic case is interpreted simply as a two-draw probability of a 'dit' of the partition—just as in the dual case, the normalized counting measure of a subset is the one-draw probability of an 'it' in the subset.

In accordance with its quadratic nature, logical entropy is the logical special case of C. R. Rao's quadratic entropy [36]. Two elements from $U=\{u_1,...,u_n\}$ are either identical or distinct. Gini [19] introduced d_{ij} as the 'distance' between the i^{th} and j^{th} elements where $d_{ij}=1$ for $i\neq j$ and $d_{ii}=0$—which might be considered the 'logical distance function' $d_{ij}=1-\delta_{ij}$, so logical distance is complement of the Kronecker delta. Since $1=(p_1+...+p_n)(p_1+...+p_n)=\sum_i p_i^2+\sum_{i\neq j}p_ip_j$, the logical entropy, i.e., Gini's index of mutability, $h(p)=1-\sum_i p_i^2=\sum_{i\neq j}p_ip_j$, is the average logical distance between distinct elements. But in 1982, C. R. Rao [36] generalized this as quadratic entropy by allowing other distances $d_{ij}=d_{ji}$ for $i\neq j$ (but always $d_{ii}=0$) so that $Q=\sum_{i\neq j}d_{ij}p_ip_j$ would be the average distance between distinct elements from U.

Rao's treatment also includes (and generalizes) the natural extension of logical entropy to continuous probability density functions $f(x)$ for a random variable X: $h(X)=1-\int f(x)^2\,dx$. It might be noted that the natural extension of Shannon entropy to continuous probability density functions $f(x)$ through the limit of discrete approximations contains terms $1/\log(\Delta x_i)$ that blow up as the mesh size Δx_i goes to zero (see [34, pp. 34–38]).[15] Hence the definition of Shannon entropy in the continuous case is defined not by the limit of the discrete formula but by the *analogous* formula $H(X)=-\int f(x)\log(f(x))dx$ which, as Robert McEliece notes, 'is not in any sense a measure of the randomness of X' [34, p. 38] in addition to possibly having negative values [46, p. 74].

16 The statistical interpretation of Shannon entropy

Shannon, like Ralph Hartley [23] before him, starts with the question of how much 'information' is required to single out a designated element from a set U of equiprobable elements. Alfréd Rényi formulated this in terms of the search [37] for a hidden designated element like the answer in a Twenty Questions game. But being able to always find the designated element is equivalent to being able to distinguish all elements from one another.

[15] For expository purposes, we have restricted the treatment to finite sample spaces U. For some countably infinite discrete probability distributions, the Shannon entropy blows up to infinity [49, Example 2.46, p. 30], while the countable logical infosets are always well-defined and the logical entropy is always in the half-open interval $[0,1)$.

One might quantify 'information' as the minimum number of yes-or-no questions in a game of Twenty Questions that it would take in general to *distinguish* all the possible 'answers' (or 'messages' in the context of communications). This is readily seen in the standard case where $|U|=n=2^m$, i.e., the size of the set of equiprobable elements is a power of 2. Then following the lead of Wilkins over three centuries earlier, the 2^m elements could be encoded using words of length m in a binary code such as the digits $\{0,1\}$ of binary arithmetic (or $\{A,B\}$ in the case of Wilkins). Then an efficient or minimum set of yes-or-no questions needed to single out the hidden element is the set of m questions:

'Is the j^{th} digit in the binary code for the hidden element a 1?'

for $j=1,...,m$. Each element is distinguished from any other element by their binary codes differing in at least one digit. The information gained in finding the outcome of an equiprobable binary trial, like flipping a fair coin, is what Shannon calls a *bit*. Hence the information gained in distinguishing all the elements out of 2^m equiprobable elements is:

$$m=\log_2(2^m)=\log_2(|U|)=\log_2\left(\frac{1}{p_0}\right) \text{ bits,}$$

where $p_0=\frac{1}{2^m}$ is the probability of any given element (all logs to base 2).[16]

In the more general case where $|U|=n$ is not a power of 2, Shannon and Hartley extrapolate to the definition of $H(p_0)$ where $p_0=\frac{1}{n}$ as:

$$H(p_0)=\log\left(\frac{1}{p_0}\right)=\log(n)$$

Shannon-Hartley entropy for an equiprobable set U of n elements.

The Shannon formula then extrapolates further to the case of different probabilities $p=\{p_1,...,p_n\}$ by taking the average:

$$H(p)=\sum_{i=1}^{n} p_i \log\left(\frac{1}{p_i}\right).$$

Shannon entropy for a probability distribution $p=\{p_1,...,p_n\}$

How can that extrapolation and averaging be made rigorous to offer a more convincing interpretation? Shannon uses the law of large numbers. Suppose that we have a three-letter alphabet $\{a,b,c\}$ where each letter was equiprobable, $p_a=p_b=p_c=\frac{1}{3}$, in a multi-letter message. Then a one-letter or two-letter message cannot be exactly coded with a binary $0,1$ code with equiprobable 0's and 1's. But any probability can be better and better approximated by longer and longer representations in the binary number system. Hence we can consider longer and longer messages of N letters along with better and better approximations with binary codes. The long-run behaviour of messages $u_1u_2...u_N$ where $u_i\in\{a,b,c\}$ is modelled by the law of large numbers so that the letter a on average occur $p_aN=\frac{1}{3}N$ times and similarly for b and c. Such a message is called *typical*.

The probability of any one of those typical messages is:

$$p_a^{p_aN}p_b^{p_bN}p_c^{p_cN}=\left[p_a^{p_a}p_b^{p_b}p_c^{p_c}\right]^N$$

or, in this case,

$$\left[\left(\frac{1}{3}\right)^{1/3}\left(\frac{1}{3}\right)^{1/3}\left(\frac{1}{3}\right)^{1/3}\right]^N=\left(\frac{1}{3}\right)^N.$$

Hence the number of such typical messages is 3^N.

[16] This is the special case where Campbell [9] noted that Shannon entropy acted as a measure to count that number of binary partitions.

If each message was assigned a unique binary code, then the number of 0, 1's in the code would have to be X where $2^X = 3^N$ or $X = \log\left(3^N\right) = N\log(3)$. Hence the number of equiprobable binary questions or bits needed per letter (i.e. to distinguish each letter) of a typical message is:

$$N\log(3)/N = \log(3) = 3 \times \tfrac{1}{3}\log\left(\tfrac{1}{1/3}\right) = H(p).$$

This example shows the general pattern.

In the general case, let $p = \{p_1, ..., p_n\}$ be the probabilities over a n-letter alphabet $A = \{a_1, ..., a_n\}$. In an N-letter message, the probability of a particular message $u_1u_2...u_N$ is $\Pi_{i=1}^{N}\Pr(u_i)$ where u_i could be any of the symbols in the alphabet so if $u_i = a_j$ then $\Pr(u_i) = p_j$.

In a *typical* message, the i^{th} symbol will occur p_iN times (law of large numbers) so the probability of a typical message is (note change of indices to the letters of the alphabet):

$$\Pi_{k=1}^{n} p_k^{p_kN} = \left[\Pi_{k=1}^{n} p_k^{p_k}\right]^N.$$

Thus the probability of a typical message is P^N where it is as if each letter in a typical message was equiprobable with probability $P = \Pi_{k=1}^{n} p_k^{p_k}$. No logs have been introduced into the argument yet, so we have an interpretation of the base-free *'numbers-equivalent'*[17] *(or 'anti-log' or exponential) Shannon entropy*:

$$E(p) = P^{-1} = \Pi_{k=1}^{n}\left(\frac{1}{p_i}\right)^{p_i} = 2^{H(p)},$$

it is as if each letter in a typical message is being draw from an alphabet with $E(p) = 2^{H(p)}$ equiprobable letters. Hence the number of N-letter messages from the equiprobable alphabet is then $E(p)^N$. The choice of base 2 means assigning a unique binary code to each typical message requires X bits where $2^X = E(p)^N$ where:

$$X = \log\left\{E(p)^N\right\} = N\log[E(p)] = NH(p).$$

Dividing by the number N of letters gives the average bit-count interpretation of the Shannon entropy; $H(p) = \log[E(p)] = \sum_{k=1}^{n} p_k \log\left(\frac{1}{p_k}\right)$ is the *average number of bits necessary to distinguish, i.e., uniquely encode, each letter in a typical message.*

This result, usually called the *Noiseless Coding Theorem*, allows us to *conceptually* relate the logical and Shannon entropies (the dit-bit transform gives the quantitative relationship). In terms of the simplest case for partitions, the Shannon entropy $H(\pi) = \sum_{B\in\pi} p_B \log_2(1/p_B)$ is a *requantification* of the logical measure of information $h(\pi) = \frac{|\text{dit}(\pi)|}{|U\times U|} = 1 - \sum_{B\in\pi} p_B^2$. Instead of directly counting the distinctions of π, the idea behind Shannon entropy is to count the (minimum) number of binary partitions needed to make all the distinctions of π. In the special case of π having 2^m equiprobable blocks, the number of binary partitions β_i needed to make the distinctions $\text{dit}(\pi)$ of π is m. Represent each block by an m-digit binary number so the i^{th} binary partition β_i just distinguishes those blocks with i^{th} digit 0 from those with i^{th} digit 1.[18] Thus there are m binary partitions β_i such that $\vee_{i=1}^{m}\beta_i = \pi$ ($\vee$ is here the partition join) or, equivalently, $\cup_{i=1}^{m}\text{dit}(\beta_i) = \text{dit}\left(\vee_{i=1}^{m}\beta_i\right) = \text{dit}(\pi)$. Thus m is the exact number of binary partitions it takes to make the distinctions of π. In the general case, Shannon gives

[17] When an event or outcome has a probability p_i, it is intuitive to think of it as being drawn from a set of $\frac{1}{p_i}$ equiprobable elements (particularly when $\frac{1}{p_i}$ is an integer) so $\frac{1}{p_i}$ is called the *numbers-equivalent of the probability* p_i [3]. For a development of $E(p)$ from scratch, see [12], [14].

[18] Thus as noted by John Wilkins in 1641, five letter words in a two-letter code would suffice to distinguish $2^5 = 32$ distinct entities [47].

the above statistical interpretation so that $H(\pi)$ is the minimum *average* number of binary partitions or bits needed to make the distinctions of π.

Note the difference in emphasis. Logical information theory is only concerned with counting the distinctions between distinct elements, not with uniquely designating the distinct entities. By requantifying to count the number of binary partitions it takes to make the same distinctions, the emphasis shifts to the length of the binary code necessary to uniquely designate the distinct elements. Thus the Shannon information theory perfectly dovetails into coding theory and is often presented today as the unified theory of information and coding (e.g. [34] or [22]). It is that shift to not only making distinctions but uniquely coding the distinct outcomes that gives the Shannon theory of information, coding and communication such importance in applications.

It might be noted that Shannon formula is often connected to (and even sometimes identified with) the Boltzmann–Gibbs entropy in statistical mechanics–which was the source for Shannon's nomenclature. But that connection is only a numerical approximation, not an identity in functional form, where the natural logs of factorials in the Boltzmann formula are approximated using the first two terms in the Stirling approximation [14]. Indeed, as pointed out by David J. C. MacKay, one can use the next term in the Stirling approximation to give a 'more accurate approximation' [32, p. 2] to the entropy of statistical mechanics–but no one would suggest using such a formula in information theory. While the use of the 'entropy' terminology is here to stay in information theory, the Shannon Noiseless Coding Theorem gives the basis to interpret the Shannon formula, not numerical approximations to the Boltzmann-Gibbs entropy in statistical mechanics.

17 Concluding remarks

Logical information theory is based on the notion of information-as-distinctions. It starts with the finite combinatorial information sets which are the ditsets of partitions on a finite U or the infosets S_X and S_Y associated with a finite $X \times Y$—and that calculus of identities and differences is expressed in the information Boolean algebra $\mathcal{I}(U)$ or $\mathcal{I}(X \times Y)$. No probabilities are involved in the definition of the information sets of distinctions. But when a probability distribution is defined on U or on $X \times Y$, then the product probability distribution is determined on U^2 or $(X \times Y)^2$ respectively. The quantitative logical entropy of an information set is the value of the product probability measure on the set.

Since conventional information theory has heretofore been focused on the original notion of Shannon entropy (and quantum information theory on the corresponding notion of von Neumann entropy), much of the article has compared the logical entropy notions to the corresponding Shannon entropy notions.

Logical entropy, like logical probability, is a measure, while Shannon entropy is not. The compound Shannon entropy concepts nevertheless satisfy the measure-like Venn diagram relationships that are automatically satisfied by a measure. This can be explained by the dit-bit transform so that by putting a logical entropy notion into the proper form as an average of dit-counts, one can replace a dit-count by a bit-count and obtain the corresponding Shannon entropy notion—which shows a deeper relationship behind the Shannon compound entropy concepts.

In sum, the logical theory of information-as-distinctions is the ground level logical theory of information stated first in terms of sets of distinctions and then in terms of two-draw probability measures on the sets. The Shannon information theory is a higher-level theory that requantifies distinctions by counting the minimum number of binary partitions (bits) that are required, on average, to make all the same distinctions, i.e., to encode the distinguished elements—and is thus well-adapted for the theory of coding and communication.

References

[1] N. Abramson. *Information Theory and Coding*. McGraw-Hill, 1963.

[2] J. Aczel and Z. Daroczy. *On Measures of Information and Their Characterization*. Academic Press, 1975.

[3] M. A. Adelman. Comment on the H concentration measure as a numbers-equivalent. *Review of Economics and Statistics*, **51**, 99–101, 1969.

[4] P. Adriaans and J. van Benthem, (eds). *Philosophy of Information. Vol. 8. Handbook of the Philosophy of Science*. North-Holland, 2008.

[5] C. H. Bennett. Quantum information: qubits and quantum error correction. *International Journal of Theoretical Physics* **42**, 153–76, 2003.

[6] N. M. Blachman. A generalization of mutual information. *Proceedings of IRE*, **49**, 1331–32, 1961.

[7] G. Boole. *An Investigation of the Laws of Thought on which are Founded the Mathematical Theories of Logic and Probabilities*. Macmillan and Co, 1854.

[8] F. Buscemi, P. Bordone and A. Bertoni. Linear entropy as an entanglement measure in two-fermion systems. *ArXiv.org*. 2, 2007.

[9] L. L. Campbell. Entropy as a measure. *IEEE Transactions on Information Theory*, IT-11, 112–114, 1965.

[10] T. Cover and J. Thomas. *Elements of Information Theory*. John Wiley, 1991.

[11] I. Csiszar and J. Körner. *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Academic Press, 1981.

[12] D. Ellerman. Counting distinctions: on the conceptual foundations of Shannon's information theory. *Synthese*, **168**, 119–49, 2009.

[13] D. Ellerman. The logic of partitions: introduction to the dual of the logic of subsets. *Review of Symbolic Logic*, **3**, 287–350, 2010.

[14] D. Ellerman. An introduction to logical entropy and its relation to shannon entropy. *International Journal of Semantic Computing*, **7**, 121–45, 2013.

[15] D. Ellerman. An introduction of partition logic. *Logic Journal of the IGPL* **22**, 94–125, 2014.

[16] R. M. Fano. The transmission of information II. In *Research Laboratory of Electronics Report 149*. MIT Press, 1950.

[17] R. M. Fano. *Transmission of Information*. MIT Press, 1961.

[18] W. Feller. *An Introduction to Probability Theory and Its Applications Vol. 1*. 3rd ed. John Wiley, 1968.

[19] C. Gini. *Variabilità e mutabilità*. Tipografia di Paolo Cuppini, 1912.

[20] J. Gleick. *The Information: A History, A Theory, A Flood*. Pantheon, 2011.

[21] I. J. Good. A.M. Turing's statistical work in World War II. *Biometrika*, **66**, 393–396, 1979.

[22] R. W. Hamming. *Coding and Information Theory*. Prentice-Hall, 1980.

[23] R. V. L. Hartley. Transmission of information. *Bell System Technical Journal*, **7**, 535–563, 1928.

[24] J. Havrda and F. Charvat. Quantification methods of classification processes: concept of structural α-entropy. *Kybernetika*, (Prague), **3**, 30–35, 1967.

[25] G. D. Hu. On the amount of information (in Russian). *Teoriya Veroyatnostei I Primenen*, **4**, 447–55, 1962.

[26] G. Jaeger. *Quantum Information: An Overview*. Springer Science+Business Media, 2007.

[27] A. N. Kolmogorov. Combinatorial foundations of information theory and the calculus of probabilities. *Russian Mathematical Surveys*, **38**, 29–40, 1983.

[28] A. N. Kolmogorov. Three approaches to the definition of the notion of amount of information. In *Selected Works of A. N. Kolmogorov: Vol. III Information Theory and the Theory of Algorithms*, A. N. Shiryayev, ed., pp. 184–93. Springer Science+Business Media, 1993.

[29] J. P. S. Kung, G.-C. Rota and C. H. Yan. *Combinatorics: The Rota Way*. Cambridge University Press, 2009.

[30] P.-S. Laplace. (1825). *Philosophical Essay on Probabilities*, trans. A. I. Dale. ed. Springer Verlag, 1995.

[31] F. W. Lawvere and R. Rosebrugh. *Sets for Mathematics*. Cambridge University Press, 2003.

[32] D. J. C. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.

[33] W. J. McGill. Multivariate information transmission. *Psychometrika*, **19**, 97–116, 1954.

[34] R. J. McEliece. *The Theory of Information and Coding: A Mathematical Framework for Communication (Encyclopedia of Mathematics and Its Applications, Vol. 3)*. Addison-Wesley, 1977.

[35] M. Nielsen and I. Chuang. *Quantum Computation and Quantum Information*. Cambridge University Press, 2000.

[36] C. R. Rao. Diversity and dissimilarity coefficients: a unified approach. *Theoretical Population Biology*, **21**, 24–43, 1982.

[37] A. Rényi. *Probability Theory*, Trans. Laszlo Vekerdi, ed. North-Holland, 1970.

[38] G.-C. Rota. Twelve problems in probability no one likes to bring up. In *Algebraic Combinatorics and Computer Science*, Henry Crapo and Domenico Senato, eds, pp. 57–93. Springer, 2001.

[39] W. W. Rozeboom. The theory of abstract partials: an introduction. *Psychometrika*, **33**, 133–67, 1968.

[40] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, **27**, 379–423; 623–56, 1948.

[41] C. E. Shannon and W. Weaver. *The Mathematical Theory of Communication*. University of Illinois Press, 1964.

[42] B. Tamir and E. Cohen. Logical entropy for quantum states. *ArXiv.org*. December. http://de.arxiv.org/abs/1412.0616v2, 2014.

[43] B. Tamir and E. Cohen. A Holevo-type bound for a Hilbert Schmidt distance measure. *Journal of Quantum Information Science*, **5**, 127–33, 2015.

[44] C. Tsallis. Possible generalization for Boltzmann-Gibbs statistics. *Journal of Statistical Physics*, **52**, 479–87, 1988.

[45] C. Tsallis. *Introduction to Nonextensive Statistical Mechanics*. Springer Science+Business Media, 2009.

[46] J. Uffink. *Measures of Uncertainty and the Uncertainty Principle*. (PhD Thesis), University of Utrecht, 1990.

[47] J. Wilkins. (1641). *Mercury or the Secret and Swift Messenger*. John Nicholson, London, 1707.

[48] R. W. Yeung. A new outlook on Shannon's information measures. *IEEE Transactions on Information Theory*, **37**, 466–74, 1991.

[49] R. W. Yeung. *A First Course in Information Theory*. Springer Science+Business Media, 2002.

Received 12 September 2016