Lattice Duality: The Origin of Probability and Entropy

Kevin H. Knuth

NASA Ames Research Center, Mail Stop 269-3,
Moffett Field CA 94035-1000, USA

Abstract

Bayesian probability theory is an inference calculus, which originates from a generalization of inclusion on the Boolean lattice of logical assertions to a degree of inclusion represented by a real number. Dual to this lattice is the distributive lattice of questions constructed from the ordered set of down-sets of assertions, which forms the foundation of the calculus of inquiry—a generalization of information theory. In this paper we introduce this novel perspective on these spaces in which machine learning is performed and discuss the relationship between these results and several proposed generalizations of information theory in the literature.

Key words: probability, entropy, lattice, information theory, Bayesian inference, inquiry
PACS:

1 Introduction

It has been known for some time that probability theory can be derived as a generalization of Boolean implication to degrees of implication represented by real numbers [11,12]. Straightforward consistency requirements dictate the form of the sum and product rules of probability, and Bayes' theorem [11,12,47,46,20,34], which forms the basis of the inferential calculus, also known as inductive inference. However, in machine learning applications it is often times more useful to rely on information theory [45] in the design of an algorithm. On the surface, the connection between information theory and probability theory seems clear—information depends on entropy and entropy

Email address: kevin.h.knuth@nasa.gov (Kevin H. Knuth).

is a logarithmically-transformed version of probability. However, as I will describe, there is a great deal more structure lying below this seemingly placid surface.

Great insight is gained by considering a set of logical statements as a Boolean lattice. I will show how this lattice of logical statements gives rise to a dual lattice of possible questions that can be asked. The lattice of questions has a measure, analogous to probability, which I will demonstrate is a generalized entropy. This generalized entropy not only encompasses information theory, but also allows for new quantities and relationships, several of which already have been suggested in the literature.

A problem can be solved in either the space of logical statements or in the space of questions. By better understanding the fundamental structures of these spaces, their relationships to one another, and their associated calculi we can expect to be able to use them more effectively to perform automated inference and inquiry.

In §2, we provide an overview of order theory, specifically partially-ordered sets and lattices. I will introduce the notion of extending inclusion on a finite lattice to degrees of inclusion effectively extending the algebra to a calculus, the rules of which are derived in the appendix. These ideas are used to recast the Boolean algebra of logical statements and to derive the rules of the inferential calculus (probability theory) in §3. I will focus on finite spaces of statements rather than continuous spaces. In §4, I will use order theory to generate the lattice of questions from the lattice of logical statements. I will discuss how consistency requirements lead to a generalized entropy and the inquiry calculus, which encompasses information theory. In §5 I discuss the use of these calculi and their relationships to several proposed generalizations of information theory.

2 Partially-Ordered Sets and Lattices

2.1 Order Theory and Posets

In this section, I introduce some basic concepts of order theory that are necessary in this development to understand the spaces of logical statements and questions. Order theory works to capture the notion of ordering elements of a set. The central idea is that one associates a set with a *binary ordering relation* to form what is called a *partially-ordered set*, or a *poset* for short. The ordering relation, generically written $\leq$, satisfies reflexivity, antisymmetry, and

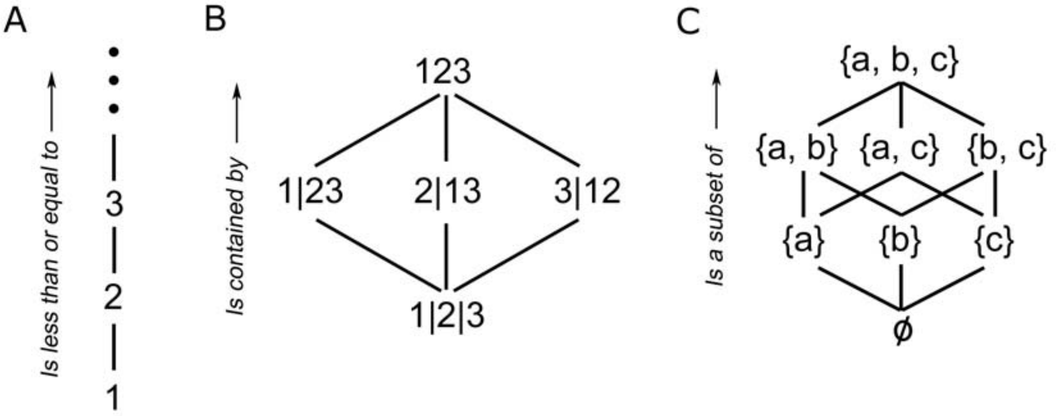

Fig. 1. Diagrams of posets described in the text. A. Natural numbers ordered by 'less than or equal to', B. $\mathbf{\Pi}_3$ the lattice of partitions of three elements ordered by 'is contained by', C. $\mathbf{2}^3$ the lattice of all subsets of three elements $\{a, b, c\}$ ordered by 'is a subset of'.

transitivity, so that for elements a, b, and c we have

$$
\begin{array}{ll}
P1. \text{ For all } \ a, \quad a \le a & (\textit{Reflexivity}) \\
P2. \text{ If } \ a \le b \ \text{ and } \ b \le a, \ \text{ then } \ a = b & (\textit{Antisymmetry}) \\
P3. \text{ If } \ a \le b \ \text{ and } \ b \le c, \ \text{ then } \ a \le c & (\textit{Transitivity})
\end{array}
$$

The ordering $a \le b$ is generally read 'b includes a'. In cases where $a \le b$ and $a \ne b$, we write $a < b$. If it is true that $a < b$, but there does not exist an element x in the set such that $a < x < b$, then we write $a \prec b$, read 'b covers a', indicating that b is a direct successor to a in the hierarchy induced by the ordering relation.

This concept of covering can be used to construct diagrams of posets. If an element b includes an element a then it is drawn higher in the diagram. If b covers a then they are connected by a line. These poset diagrams (or Hasse diagrams) are useful in visualizing the order induced on a set by an ordering relation. Figure 1 shows three posets. The first is the natural numbers ordered by the usual 'is less than or equal to'. The second is $\mathbf{\Pi}_3$ the lattice of partitions of three elements. A partition y includes a partition x, $x \le y$, when every cell of x is contained in a cell of y. The third poset, denoted $\mathbf{2}^3$, is the powerset of the set of three elements $\mathcal{P}(\{a, b, c\})$, ordered by set inclusion $\subseteq$. The orderings in Figures 1b and c are called *partial orders* since some elements are incomparable with respect to the ordering relation. For example, since it is neither true that $\{a\} \le \{b\}$ or that $\{b\} \le \{a\}$, the elements $\{a\}$ and $\{b\}$ are incomparable, written $\{a\}||\{b\}$. In contrast, the ordering in Figure 1a is a *total order*, since all pairs of elements are comparable with respect to the ordering relation.

A poset P possesses a greatest element if there exists an element $\top \in P$, called the *top*, where $x \leq \top$ for all $x \in P$. Dually, the least element $\bot \in P$, called the *bottom*, exists when $\bot \leq x$ for all $x \in P$. For example, the top of $\mathbf{\Pi}_3$ is the partition 123 where all elements are in the same cell. The bottom of $\mathbf{\Pi}_3$ is the partition 1|2|3 where each element is in its own cell. The elements that cover the bottom are called *atoms*. For example, in $\mathbf{2}^3$ the atoms are the singleton sets $\{a\}$, $\{b\}$, and $\{c\}$.

Given a pair of elements x and y, their *upper bound* is defined as the set of all $z \in P$ such that $x \leq z$ and $y \leq z$. In the event that a unique *least upper bound* exists, it is called the *join*, written $x \vee y$. Dually, we can define the *lower bound* and the *greatest lower bound*, which if it exists, is called the *meet*, $x \wedge y$. Graphically the join of two elements can be found by following the lines upward until they first converge on a single element. The meet can be found by following the lines downward. In the lattice of subsets of the powerset $\mathbf{2}^3$, the join $\vee$, corresponds to the set union $\cup$, and the meet $\wedge$ corresponds to the set intersection $\cap$. Elements that cannot be expressed as a join of two elements are called *join-irreducible elements*. In the lattice $\mathbf{2}^3$, these elements are the atoms.

Last, the dual of a poset P, written P^∂ can be formed by reversing the ordering relation, which can be visualized by flipping the poset diagram upside-down. This action exchanges joins and meets and is the reason that their relations come in pairs, as we will see below. There are different notions of duality and the notion after which this paper is titled will be discussed later.

2.2 Lattices

A *lattice* L is a poset where the join and meet exist for every pair of elements. We can view the lattice as a set of objects ordered by an ordering relation, with the join $\vee$ and meet $\wedge$ describing the hierarchical structure of the lattice. This is a structural viewpoint. However, we can also view the lattice from an operational viewpoint as an algebra on the space of elements. The algebra is defined by the operations $\vee$ and $\wedge$ along with any other relations induced by the structure of the lattice. Dually, the operations of the algebra uniquely determine the ordering relation, and hence the lattice structure. Viewed as operations, the join and meet obey the following properties for all $x, y, z \in \mathcal{L}$

$L1.$ $x \vee x = x, \quad x \wedge x = x$ (*Idempotency*)

$L2.$ $x \vee y = y \vee x, \quad x \wedge y = y \wedge x$ (*Commutativity*)

$L3.$ $x \vee (y \vee z) = (x \vee y) \vee z, \quad x \wedge (y \wedge z) = (x \wedge y) \wedge z$ (*Associativity*)

$L4.$ $x \vee (x \wedge y) = x \wedge (x \vee y) = x$ (*Absorption*)

The fact that lattices are algebras can be seen by considering the consistency relations, which express the relationship between the ordering relation and the join and meet operations.

$$x \leq y \quad \Leftrightarrow \quad \begin{array}{c} x \wedge y = x \\ \\ x \vee y = y \end{array} \qquad (\mathit{Consistency\ Relations})$$

Lattices that obey the distributivity relation

$$D1. \quad x \wedge (y \vee z) = (x \wedge y) \vee (x \wedge z) \quad (\mathit{Distributivity\ of} \wedge \mathit{over} \vee)$$

and its dual

$$D2. \quad x \vee (y \wedge z) = (x \vee y) \wedge (x \vee z) \quad (\mathit{Distributivity\ of} \vee \mathit{over} \wedge)$$

are called *distributive lattices*. All distributive lattices can be expressed in terms of elements consisting of sets ordered by set inclusion.

A lattice is *complemented* if for every element x in the lattice, there exists a unique element $\sim x$ such that

$$\begin{array}{ll} C1. & x \vee \sim x = \top \\ C2. & x \wedge \sim x = \bot \end{array} \qquad (\mathit{Complementation})$$

Note that the lattice $\mathbf{2}^3$ (Fig. 1c) is complemented, whereas the lattice $\mathbf{\Pi}_3$ (Fig. 1b) is not.

2.3 Inclusion and the Incidence Algebra

Inclusion on a poset can be quantified by a function called the *zeta function*

$$\zeta(x, y) = \begin{cases} 1 \text{ if } & x \leq y \\ 0 \text{ if } & x \not\leq y \end{cases} \qquad (\mathit{zeta\ function}) \tag{1}$$

which describes whether the element y includes the element x. This function belongs to a class of real-valued functions $f(x, y)$ of two variables defined on a poset, which are non-zero only when $x \leq y$. This set of functions comprises the *incidence algebra* of the poset [42]. The sum of two functions $f(x, y) + g(x, y)$ in the incidence algebra is defined the usual way by

$$h(x, y) = f(x, y) + g(x, y), \tag{2}$$

as is multiplication by a scalar $h(x, y) = \lambda f(x, y)$. However, the product of two functions is found by taking the convolution over the interval of elements in the poset

$$h(x, y) = \sum_{x \leq z \leq y} f(x, z) g(z, y). \tag{3}$$

To invert functions in the incidence algebra, one must rely on the *Möbius function* $\mu(x, y)$, which is the inverse of the zeta function [44,42,3]

$$\sum_{x \leq z \leq y} \zeta(x, z) \mu(z, y) = \delta(x, y), \tag{4}$$

where $\delta(x, y)$ is the Kronecker delta function. These functions are the generalized analogues of the familiar Riemann zeta function and the Möbius function in number theory, where the poset is the set of natural numbers ordered by 'divides'. We will see that they play an important role both in inferential reasoning as an extension of inclusion on the Boolean lattice of logical statements, and in the quantification of inquiry.

2.4 Degrees of Inclusion

It is useful to generalize this notion of inclusion on a poset. I first introduce the dual of the zeta function, $\zeta^{\partial}(x, y)$, which quantifies whether x includes y, that is

$$\zeta^{\partial}(x, y) = \begin{cases} 1 \text{ if } & x \geq y \\ 0 \text{ if } & x \ngeq y \end{cases} \qquad (\textit{dual of the zeta function}) \tag{5}$$

Note that the dual of the zeta function on a poset P is equivalent to the zeta function defined on its dual P^{∂}, since the ordering relation is simply reversed. I will generalize inclusion by introducing the function $z(x, y)$, [1]

$$z(x, y) = \begin{cases} 1 \text{ if } & x \geq y \\ 0 \text{ if } & x \wedge y = \bot \\ z \text{ otherwise, where } & 0 < z < 1. \end{cases} \qquad (\textit{degrees of inclusion}) \tag{6}$$

where inclusion on the poset is generalized to degrees of inclusion represented by real numbers. [2] This new function quantifies the *degree to which x includes y*. This generalization is asymmetric in the sense that the condition where $\zeta^{\partial}(x, y) = 1$ is preserved, whereas the condition where $\zeta^{\partial}(x, y) = 0$ has been modified. The motivation here is that, if we are certain that x includes y then we want to indicate this knowledge. However, if we know that x does not

[1] I have dropped the ∂ symbol since the definition is clear.

[2] This function need not be normalized to unity, as we will see later.

include y, then we can quantify the *degree* to which x includes y. In this sense, the algebra is extended to a calculus. Later, I will demonstrate the utility of such a generalization.

The values of the function z must be consistent with the poset structure. In the case of a lattice, when the arguments are transformed using the algebraic manipulations of the lattice, the corresponding values of z must be consistent with these transformations. By enforcing this consistency, we can derive the rules by which the degrees of inclusion are to be manipulated. This method of requiring consistency with the algebraic structure was first used by Cox to prove that the sum and product rules of probability theory are the only rules consistent with the underlying Boolean algebra [11,12]. The rules for the distributive lattices I will describe below are derived in the appendix, and the general methodology is discussed in greater detail elsewhere [34].

Consider a distributive lattice $\mathcal{D}$ and elements $x, y, t \in \mathcal{D}$. Given the degree to which x includes t, $z(x, t)$, and the degree to which y includes t, $z(y, t)$, we would like to be able to determine the degree to which the join $x \vee y$ includes t, $z(x \vee y, t)$. In the appendix, I show that consistency with associativity of the join requires that

$$z(x \vee y, t) = z(x, t) + z(y, t) - z(x \wedge y, t). \tag{7}$$

For a join of multiple elements $x_1, x_2, \ldots, x_n$, this degree is found by

$$z(x_1 \vee x_2 \vee \cdots \vee x_n, t) = \\ \sum_i z(x_i, t) - \sum_{i<j} z(x_i \wedge x_j, t) + \sum_{i<j<k} z(x_i \wedge x_j \wedge x_k, t) - \cdots, \tag{8}$$

which I will call the *sum rule for distributive lattices.* This sum rule exhibits Gian-Carlo Rota's *inclusion-exclusion principle*, where terms are added and subtracted to avoid double-counting of the join-irreducible elements in the join [28,42,3]. The inclusion-exclusion principle is a consequence of the Möbius function for distributive lattices, which leads to an alternating sum and difference as one sums down the interval in the lattice. This demonstrates that the form of the sum rule is inextricably tied to the underlying lattice structure [34].

For the meet of two elements $x \wedge y$, it is clear that we can use (7) to obtain

$$z(x \wedge y, t) = z(x, t) + z(y, t) - z(x \vee y, t). \tag{9}$$

However, another useful form can be obtained by requiring consistency with distributivity. In the appendix, I show that this consistency constraint leads to

$$z(x \wedge y, t) = C z(x, t) z(y, x \wedge t), \tag{10}$$

which is the *product rule for distributive lattices.* The constant C acts as a normalization factor, and is necessary when these degrees are normalized to values other than unity.

Last, requiring consistency with commutativity of the meet leads to the analog of Bayes' Theorem for distributive lattices

$$z(y, x \wedge t) = \frac{z(y,t)z(x, y \wedge t)}{z(x,t)}. \tag{11}$$

One does not think typically of Bayes' Theorem outside of the context of probability theory, however, it is a general rule that is applicable to all distributive lattices. As I will demonstrate, it can be used in computing probabilities among logical assertions, as well as in working with questions.

2.5 Measures and Valuations

The fact that one can define functions that take lattice elements to real numbers was utilized by Rota, who used this to develop and promote the field of *geometric probability* [43,28]. The more familiar term *measure* typically refers to a function μ defined on a Boolean lattice $\mathcal{B}$, which takes elements of a Boolean lattice to a real number. For example, given $x \in \mathcal{B}$, $\mu : x \mapsto \mathbb{R}$. The term *valuation* is a more general term that takes a lattice element $x \in \mathcal{L}$ to a real number, $v : x \mapsto \mathbb{R}$, or more generally to an element of a commutative ring with identity [44], $v : x \mapsto \mathbb{A}$. The function z, introduced above (6), is a *bi-valuation* since it takes two lattice elements $x, y \in \mathcal{L}$ as its arguments, $z : x, y \mapsto \mathbb{R}$. When applied to a Boolean lattice, the function z is also a measure.

3 Logical Statements

George Boole [8,9] was the first to understand the algebra of logical statements, which I will interchangeably call logical assertions. Boole's algebra is so familiar, that I will spend little effort in describing it. In this algebra, there are two binary relations called conjunction (AND) and disjunction (OR), and one unary operation called complementation (NOT). The binary operations are commutative, associative, and distributive.

Let us now adopt a different perspective, where we view this Boolean structure as a set of logical statements ordered by logical implication. A statement x includes a statement y, $y \leq x$, when y implies x, written $y \rightarrow x$. Thus the ordering relation $\leq$ is represented by $\rightarrow$. Logical implication as an ordering

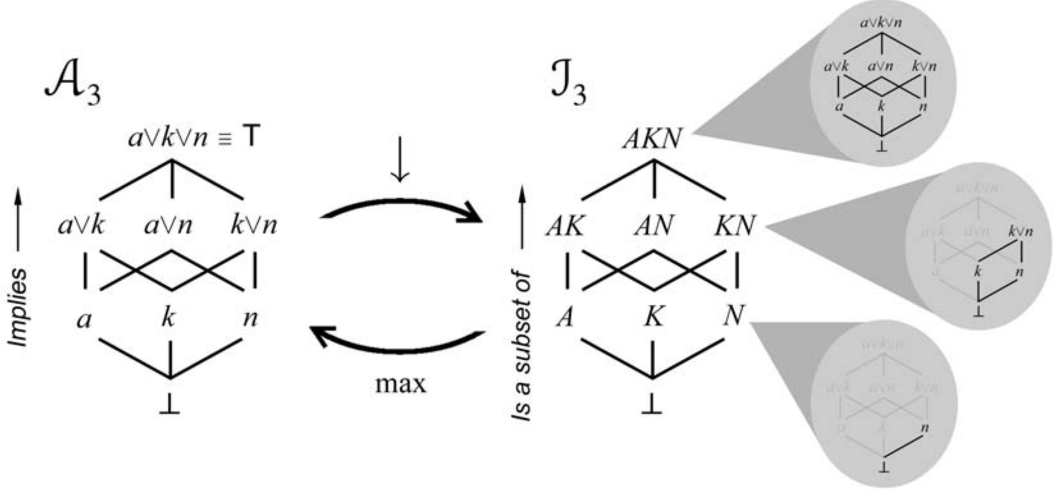

Fig. 2. The lattice of assertions $\mathcal{A}_3$ generated by three atomic assertions a, k, and n. The ideals of this lattice form the lattice of ideal questions $\mathfrak{I}_3$ ordered by set inclusion $\subseteq$, which is isomorphic to $\mathcal{A}_3$. The the maximum of the set of statements in any ideal maps back to the assertion lattice. The statements comprising three ideals are highlighted on the right.

relation among a set of logical assertions sets up a partial order on the set and forms a *Boolean lattice.* The join and meet are identified with the disjunction and conjunction, respectively. In this case the order-theoretic notation for the join and the meet conveniently matches the logical notation for the disjunction and conjunction. However, one must remember that the join and meet describe different operations in different lattices. A Boolean lattice follows L1-L4, D1, D2, C1 and C2, which is neatly summarized by saying that it is a *complemented distributive lattice.*

To better picture this, consider a simple example [33] concerning the matter of 'Who stole the tarts made by the Queen of Hearts?' For the sake of this example, let us say that there are three mutually exclusive statements, one of which answers this question:

$$a = \textit{`Alice stole the tarts!'}$$
$$k = \textit{`The Knave of Hearts stole the tarts!'}$$
$$n = \textit{`No one stole the tarts!'}$$

The lattice $\mathcal{A}_3$ generated by these assertions is shown in Figure 2. The bottom element of the lattice is called the *logical absurdity.* It is always false, and as such, it implies every other statement in the lattice. The three statements a, k, and n are the atoms which cover the bottom. All other logical statements in this space can be generated from joins of these three statements. For example, the statement $a \vee k$ is the statement '*Either Alice or the Knave stole the tarts!*' The top element $\top = a \vee k \vee n$ is called the *truism*, since it is trivially true that '*Either Alice, the Knave, or nobody stole the tarts!*'. The truism is implied by every other statement in the lattice. Since the lattice is Boolean, each element in the lattice has a unique complement (C1, C2). The statement $a =$'*Alice stole the tarts!*' has as its complement the statement $\sim a = k \vee n =$'*Either the Knave or no one stole the tarts!*' This statement $\sim a$ is equivalent to '*Alice did not steal the tarts!*' Last, note that this lattice (Fig. 2) is isomorphic to the lattice of powersets $\mathbf{2}^3$ (Fig. 1c), therefore Boolean algebra describes the operations on a powerset as well as implication on a set of logical statements.

3.1 The Origin of Probability Theory

Deductive reasoning describes the act of using the Boolean lattice structure to determine whether one logical statement implies another given partial knowledge of relations among a set of logical statements. From the perspective of posets, this equates to determining whether one element of a poset includes another element given some partial knowledge of inclusion among a set of poset elements. Since inclusion on a poset is encoded by the zeta function ζ and its dual ζ^∂ (5), either of these functions can be used to quantify implication on $\mathcal{A}$ and perform deductive reasoning.

Inductive reasoning or *inference* is different from deductive reasoning in the sense that it incorporates a notion of uncertainty not found in the Boolean lattice structure. Just as ζ^∂ quantifies deductive reasoning, its generalization z quantifies inductive reasoning. Probability[3] is simply this function z (6) defined on the Boolean lattice $\mathcal{A}$,

$$p(x|y) = z(x, y), \tag{12}$$

so that implication on the lattice is generalized to degrees of implication represented by real numbers

$$p(y|x) = \begin{cases} 1 \text{ if } \ x \rightarrow y \\ 0 \text{ if } \ x \wedge y = \bot \\ p \text{ otherwise, where } \ 0 < p < 1. \end{cases} \qquad (\textit{probability}) \tag{13}$$

To make this more concrete, consider the example in Figure 2. Clearly, $\top \geq a$, which is equivalent to $a \rightarrow \top$, so that $\zeta^\partial(\top, a) = 1$ and $p(\top|a) = 1$. Now, $\top \nrightarrow a$ and $a \wedge \top = a$, therefore $\zeta^\partial(a, \top) = 0$ and $p(a|\top) = p$ where $0 < p < 1$. While the truism, $\top =$'*Either Alice or the Knave or no one stole the tarts!*',

[3] I could call this degree of implication *plausibility* or perhaps by a new term, however we will see that this quantity follows all of the rules of probability theory. Since there is neither an operational nor a mathematical difference between this degree of implication and probability, I see no need to indicate a difference semantically.

does not imply that $a =$'*Alice stole the tarts!*', the *degree* to which the premise implies that Alice stole the tarts, $p(a|\top)$, is a very useful quantity. This is the essence of inductive reasoning.

Since probability is the function z defined on the Boolean lattice $\mathcal{A}$, the rules by which probabilities may be manipulated derive directly from requiring that probability be consistent with the underlying Boolean algebra. Since Boolean algebras are distributive, we have already shown that there are three rules for manipulating probabilities: the *sum rule of probability*

$$p(x \vee y|t) = p(x|t) + p(y|t) - p(x \wedge y|t), \tag{14}$$

which is equivalent to (7), the *product rule of probability*

$$p(x \wedge y|t) = p(x|t)p(y|x \wedge t), \tag{15}$$

which is equivalent to (10) with $C = 1$, and *Bayes' theorem*

$$p(y|x \wedge t) = \frac{p(y|t)p(x|y \wedge t)}{p(x|t)}, \tag{16}$$

which is equivalent to (11). These three rules constitute the inferential calculus, which is a generalization of the Boolean algebra of logical assertions. There are several very important points to be made here. Probabilities are functions of pairs of logical statements and quantify the degree to which one logical statement implies another. For this reason, they are necessarily conditional. Since the rules by which probabilities are to be manipulated derive directly from consistency with the underlying Boolean algebra, probability theory is *literally* an extension of logic, as argued so effectively by E.T. Jaynes [25].

3.2 Join-Irreducible Statements and Prior Probabilities

The join-irreducible elements of the Boolean lattice of logical statements are the atoms that cover the absurdity $\perp$. This set of atomic statements $\{a_i\}$ comprises the exhaustive set of mutually exclusive assertions that form the basis of this lattice. All other statements in the lattice can be found by taking joins of these atoms. Given assignments of the prior probabilities for the set of $\{a_i\}$, (eg. $p(a_i|\top)$), the prior probabilities for all other statements in the lattice can be computed using the sum rule of the inferential calculus. This was proven by Gian-Carlo Rota [44, Theorem 1, Corollary 2, p.35] who showed that:

Theorem 1 (Rota, Assigning Valuations [44]) *A valuation in a finite distributive lattice is uniquely determined by the values it takes on the set of join-irreducibles of* $\mathfrak{L}$*, and these values can be arbitrarily assigned.*

Furthermore, given the prior probability of the degree to which $\top$ implies an assertion x, the degree to which any other element of the lattice implies x can be found via the product rule

$$p(x|y) = p(x|y \wedge \top) = \frac{p(x \wedge y|\top)}{p(y|\top)}. \tag{17}$$

Thus by assigning the prior probability that the truism implies each of the atoms, all the other probabilities can be uniquely determined using the inferential calculus.

What is even more remarkable here, is that Rota proved that the values of the prior probabilities can be *arbitrarily assigned.* This means that there are no constraints imposed by the lattice structure, or equivalently the Boolean algebra, on the values of the prior probabilities. Thus the inferential calculus tells us nothing about assigning prior probabilities. Objective assignments can only be made by relying on additional consistency principles, such as symmetry, constraints, and consistency with other aspects of the problem at hand. Examples of useful principles are Jaynes' *Principle of Maximum Entropy* [23] and his *Principle of Group Invariance* [22], which is a generalization of the *Principle of Indifference* [6,35,27]. Once these assignments are made, the inferential calculus, induced by consistency with order-theoretic principles, dictates the remaining probabilities.

3.3 Remarks on Lattice Products

Two spaces of logical statements can be combined by taking the lattice product, which can be written as the Cartesian product of the lattice elements. By equating the bottom elements of the two spaces, we get a distributive lattice. Such products of lattices are very important in inference, since it is exactly what one does when one takes a lattice of hypotheses and combines them with data. The product rule and Bayes' theorem are extremely useful in these situations where the prior probabilities are assigned on the two lattices forming the product. These issues are discussed in more detail elsewhere [34].

4 Questions

In his last published work exploring the relationships between inference and inquiry [13], Cox defined a question as the set of all logical statements that answer it. At first glance, this definition is strikingly simple. However, with further thought one sees that it captures the essence of a question and does so in a form that is accessible to mathematical investigation.

In the previous section on logical statements, the modern viewpoint of lattice theory may seem like overkill. Its heavy mathematics are not necessary to reach the same conclusions that one reaches by simply working with Boole's algebra. In addition, while some of new insight is gained, there is little there to change how one uses probability theory to solve inference problems. Here however, we will find lattice theory to be of great advantage by enabling us to visualize relations among sets of assertions that comprise the sets of answers to questions.

4.1 Down-sets and Ideals

If a logical statement x answers a question, then any statement y such that $y \rightarrow x$, or equivalently in order-theoretic notation, $y \leq x$, answers the same question. Thus a question is not defined by just any set of logical statements, it is defined by a set that is closed when going down the assertion lattice. Such a set is called a *down-set* [14]

Definition 1 (Down-set) *A down-set is a subset J of an ordered set L, written $J = \downarrow L$, where if $a \in J$, $x \in L$, $x \leq a$ then $x \in J$.*

Let us begin exploring questions by considering the down-set formed from a set containing a single element $\{x\}$, which we write[4] as $X = \downarrow\{x\} \equiv \downarrow x$. Given any logical statement x in the Boolean lattice of assertions, we can consider the down-set formed from that assertion x

$$X = \downarrow x = \{y \mid y \rightarrow x \; \forall \; x, y \in \mathcal{A}\} \tag{18}$$

Such a down-set is called an *ideal* [7,14], and to emphasize this I have called these questions *ideal questions* to denote the fact that they are ideals of the assertion lattice $\mathcal{A}$ [32].

We are now in a position to compare questions. Two questions are equivalent if they ask the same thing—or equivalently when they are answered by the same set of assertions. The questions '*Is it raining?*' and '*Is it not raining?*' are both answered by either the statement '*It is raining!*' or the statement '*It is not raining!*' and all the statements that imply them. Thus our two questions ask the same thing and are therefore equivalent. Furthermore, if one question X is a subset of another question Y in a space of questions $\mathcal{Q}$, then answering the question X will necessarily answer the question Y. This means that we can use the binary ordering relation 'is a subset of' to implement the ordering relation 'answers' and therefore order the set of questions.

[4] Note that I am using lowercase letters to represent logical statements, uppercase letters to represent questions, and script letters to represent an entire lattice.

The set of ideal questions $\mathcal{I}$ ordered by set inclusion forms a lattice (Fig. 2) isomorphic to the original assertion lattice [7]. Thus there is a one-to-one onto mapping from each statement $x \in \mathcal{A}$ to its ideal question $X \in \mathcal{Q}$. The atomic assertions map to atomic questions, each of which has only two possible answers. For example, the statement k from our previous example maps to

$$K = \downarrow k = \{k, \perp\}, \tag{19}$$

which is answered by either '*The knave stole the tarts!*' or the absurdity $\perp$. Robert Fry calls these atomic questions *elementary questions* [17], since you basically receive either exactly what you asked or no useful answer. The non-atomic statements map to more complex questions, such as

$$KN = \downarrow k \vee n = \{k \vee n, k, n, \perp\}, \tag{20}$$

where the symbol KN is considered to be a single symbol representing a single question formed from the down-set of the join of the statements k and n. Similarly, I will use AKN to represent the question $AKN = \downarrow a \vee k \vee n$. The lattice of ideal questions $\mathcal{I}$ can be mapped back to the lattice of assertions $\mathcal{A}$ by selecting the maximum element in the set.

4.2 The Lattice of Questions

We can construct more complex questions by considering down-sets, which are set unions of the ideals of the assertion lattice. For example, the question $T =$'*Who stole the tarts?*' is formed from the union of the three elementary questions

$$T = A \cup K \cup N. \tag{21}$$

Since

$$\begin{aligned} A &= \downarrow a = \{a, \perp\} \\ K &= \downarrow k = \{k, \perp\} \\ N &= \downarrow n = \{n, \perp\} \end{aligned} \tag{22}$$

the question $T = A \cup K \cup N$ can be written as

$$T = \{a, k, n, \perp\}. \tag{23}$$

In this way, the question T is defined by its set of possible answers, including the absurdity. We could also ask the binary question $B =$'*Did or did not Alice steal the tarts?*'. This question can be written as the down-set formed from $a =$'*Alice stole the tarts!*', and its complement $\sim a =$'*Either the Knave or no*

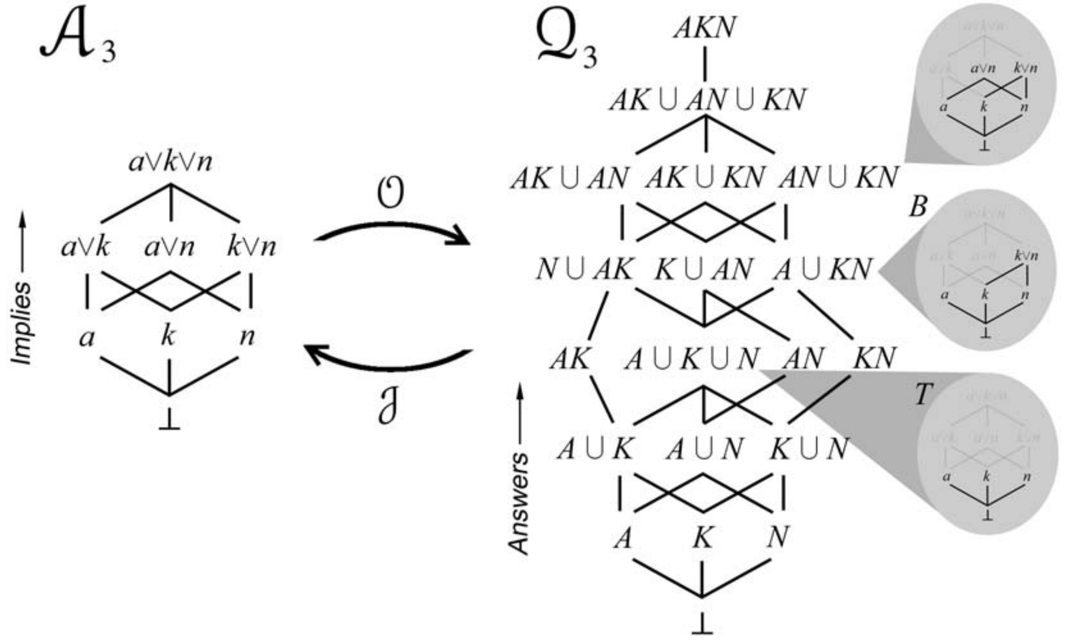
$\mathcal{A}_3$
$a \vee k \vee n$
$a \vee k$ $a \vee n$ $k \vee n$
a k n
$\perp$
Implies
$\mathcal{O}$
$\mathcal{J}$
$\mathcal{Q}_3$
AKN
$AK \cup AN \cup KN$
$AK \cup AN$ $AK \cup KN$ $AN \cup KN$
$N \cup AK$ $K \cup AN$ $A \cup KN$
AK $A \cup K \cup N$ AN KN
$A \cup K$ $A \cup N$ $K \cup N$
A K N
$\perp$
Answers
B
T

Fig. 3. The ordered set of down-sets of the lattice of assertions $\mathcal{A}$ results in the lattice of all possible questions $\mathcal{Q} = \mathcal{O}(\mathcal{A})$ ordered by $\subseteq$. The join-irreducible elements of $\mathcal{Q}$ are the ideal questions, which are isomorphic to the lattice of assertions $\mathcal{A} \sim \mathcal{J}(\mathcal{Q})$. The down-sets corresponding to several questions, including the questions T and B discussed in the text, are illustrated on the right.

one stole the tarts!'

$$\begin{aligned} B &= \downarrow a \quad \cup \quad \downarrow \sim a \\ &= \downarrow a \quad \cup \quad \downarrow k \vee n \\ &= \{a, \perp\} \cup \{k \vee n, k, n, \perp\} \\ &= \{k \vee n, a, k, n, \perp\}, \end{aligned} \tag{24}$$

where any one of the statements in the set will answer the question B. We can write B compactly as $B = A \cup KN$.

This construction produces every possible question given the lattice of assertions $\mathcal{A}$, see Figure 3. Since the questions are sets, the set of questions ordered by set inclusion forms a poset ordered by $\subseteq$, which is a distributive lattice [7,14]. More specifically, this construction results in the ordered set of down-sets of $\mathcal{A}$, which is written $\mathcal{O}(\mathcal{A})$. Thus the Boolean lattice $\mathbf{2}^N$ is mapped to the *free distributive lattice* $FD(N)$. Even though $FD(N)$ is a lattice of sets and is distributive, it is not complemented [32]. Thus questions in general have no complements.

The question lattice $\mathcal{Q}$ is closed under set union and set intersection, which correspond to the join and the meet, respectively. Therefore, $T = A \cup K \cup N \equiv A \vee K \vee N$. Unfortunately, the terminology introduced by Cox is at odds with the order-theoretic terminology, since the *joint question* is formed from the meet of two questions, and it asks what the two questions ask jointly, whereas the join of two questions, the *common question*, asks what the two questions ask in common.

Consider the two questions T and B. Since $T \subseteq B$, the question T necessarily answers the question B. Thus asking '*Who stole the tarts?*' will resolve '*Did or did not Alice steal the tarts?*' The converse is not true, since if one asks '*Did or did not Alice steal the tarts?*', the reply could be '*Either the Knave or no one stole the tarts!*', which still does not answer '*Who stole the tarts?*' Thus the question T lies below the question B in the lattice $\mathcal{Q}_3$ indicating that T answers B.

The consistency relations (discussed in §2.2) can be used to better visualize these relationships. Consider again the two questions $T =$'*Who stole the tarts?*' and $B =$'*Did or did not Alice steal the tarts?*'. The join of these two questions $T \vee B$ asks what they ask in common, which is '*Did or did not Alice steal the tarts?*'. Whereas, their meet $T \wedge B$ asks what they ask jointly, which is '*Who stole the tarts?*'. So we have that

$$\begin{aligned} T \vee B &= B \\ T \wedge B &= T \end{aligned}$$

which, by the consistency relations, implies that

$$T \subseteq B.$$

This can also be determined by taking the set union for the join and the set intersection for the meet, and working with expressions for the sets defining T and B.

4.3 The Central Issue

Just as statements can be true or false, questions can be real or vain. Cox defined a *real question* as a question that is answered by at least one true statement, whereas a *vain question* is a question that is answered by no true statement [13]. In Lewis Carroll's *Alice's Adventures in Wonderland*, it turned out that no one stole the tarts. Thus, any question not allowing for that possibility is a vain question—there does not exist a true answer that will resolve the issue.

When the truth values of the statements are not known, a question Q is only assured to be a real question when it is answered by every one of the atomic statements of $\mathcal{A}$, or equivalently when $Q \wedge Q_i \neq \bot$ for all elementary questions $Q_i \in \mathcal{Q}$. Put simply, all possibilities must be accounted for. Previously, I called these questions *assuredly real questions* [32], which I will shorten here to *real questions*. The set of real questions is a sublattice $\mathcal{R}$ of the lattice $\mathcal{Q}$. That is, it is closed under joins and meets.

The bottom of $\mathcal{R}$ is the smallest real question, and it answers all other questions in $\mathcal{R}$. It is formed from the join of all of the elementary questions, and as such it does not accept an ambiguous answer. For this reason, I call it the *central issue*. In our example, the central issue is the question $T =$'*Who stole the tarts?*'. Resolving the central issue will answer all the other real questions. Recall that by answering $T =$'*Who stole the tarts?*', we necessarily will have answered $B =$'*Did or did not Alice steal the tarts?*' As one ascends the real sublattice, the questions become more and more ambiguous. For example, the question $AN \cup KN$ will narrow down the inquiry, resolving whether it was Alice or the Knave, but not necessarily ruling out that no one stole the tarts.

4.4 Duality between the Assertions and Questions

The question lattice $\mathcal{Q}$ is formed by taking the ordered set of down-sets of the assertion lattice, which can be represented by the map $\mathcal{A} \mapsto \mathcal{O}(\mathcal{A})$, so that $\mathcal{Q} = \mathcal{O}(\mathcal{A})$. The join-irreducible questions $\mathcal{J}(\mathcal{Q})$ are the ideal questions, which by themselves form a lattice that is isomorphic to the assertion lattice $\mathcal{A}$, which can be represented by the map $\mathcal{Q} \mapsto \mathcal{J}(\mathcal{Q})$. Thus we have two isomorphisms $\mathcal{Q} \sim \mathcal{O}(\mathcal{A})$ and $\mathcal{A} \sim \mathcal{J}(\mathcal{Q})$. This correspondence, called *Birkhoff's Representation Theorem* [14], holds for all finite ordered sets $\mathcal{A}$. The lattice $\mathcal{Q}$ is called the *dual* of $\mathcal{J}(\mathcal{Q})$, and the lattice $\mathcal{A}$ is called the *dual* of $\mathcal{O}(\mathcal{A})$. This is of course a different notion of duality than I introduced earlier. What is surprising is that the join-irreducible map takes lattice products to sums of lattices, so that the map $\mathcal{J}$ acts like a logarithm, whereas the map $\mathcal{O}$ acts like the exponential function [14].

4.5 The Geometry of Questions

There are some interesting relationships between the lattice of questions and geometric constructs based on simplexes. A *simplex* is the simplest possible polytope in a space of given dimension. In zero dimensions, a 0-simplex is a point. A 1-simplex is a line segment, which consists of two 0-simplexes connected by a line. A 2-simplex is a triangle consisting of three 0-simplexes joined by three 1-simplexes, in conjunction with the filled in interior. The 3-simplex is a tetrahedron. Finally the n-simplex is an n-hypertetrahedron.

Since, an $n-1$ simplex can be used to construct an n-simplex, we can order these simplexes with an ordering relation 'contains'. For example, if two 0-

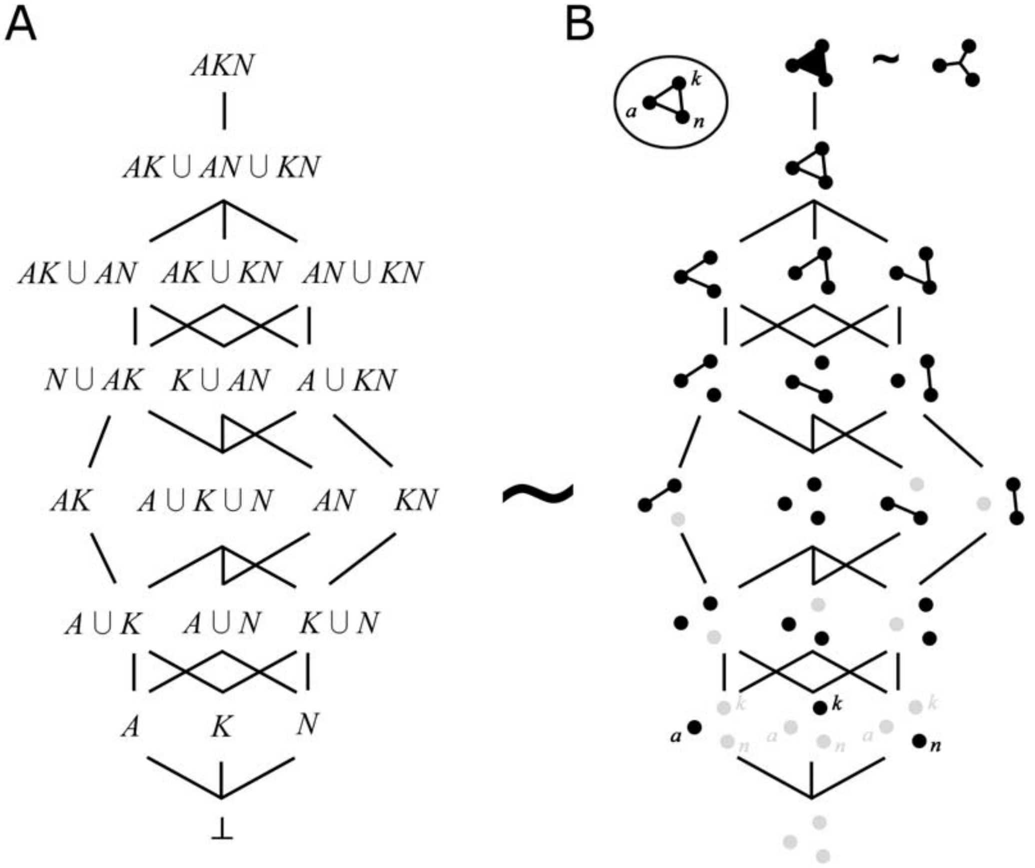

Fig. 4. The lattice of questions (A) is isomorphic to the lattice of simplicial complexes (B). The atomic questions A, K, and N are isomorphic to the three 0-simplexes. The real questions are isomorphic to the simplicial complexes that include every 0-simplex in the space. Questions are not only related to these geometric constructs, they are also isomorphic to hypergraphs. Since low-order hypergraphs look like low-order simplicial complexes, the lattice of hypergraphs with three generators is almost identical. The only exception is that instead of the 2-simplex at the top, we have the hypergraph connecting the three points (see top of B).

simplexes, A and B, are used to create a 1-simplex AB, we write $A \leq AB$ and $B \leq AB$. We can also define a join of a m-simplex with an n-simplex as a geometric object akin to the set union of the two simplexes. Such an object is called a *simplicial complex.* The set of all simplicial complexes formed from N distinct 0-simplexes forms the free distributive lattice $FD(N)$ [28]. We can identify each n-simplex with an ideal question in the question lattice formed by taking the down-set of the join of n assertions. This allows us to set up a one-to-one correspondence between the set of questions and the set of simplicial complexes. The lattice of questions is thus isomorphic to the lattice of simplicial complexes (Figure 4).

Another interesting isomorphism can be established. Hypergraphs are graphs with generalized edges that can connect more than a single node. By identifying each n-simplex with an n-hypergraph, a one-to-one correspondence can be made between simplicial complexes and hypergraphs. Thus, a lattice of hypergraphs can be constructed, which is isomorphic to both the lattice of simplicial complexes and the lattice of questions. The relationship between hypergraphs and information theory was noted by Tony Bell [4]. I will show that the lattice on which Bell's co-information is a valuation is precisely the question lattice [32].

4.6 The Inquiry Calculus

The algebra of questions provides us with the operations with which questions can be manipulated. Given two questions, we can form the common question and the joint question using the join and meet respectively. Inclusion on the lattice $\mathfrak{Q}$ indicates whether one question is answered by another. We now extend this algebra to a calculus by generalizing inclusion on this lattice to degrees of inclusion represented by real numbers just as we did on the Boolean lattice. Consider two questions, one of which I will call an outstanding *issue* I, and the other an *inquiry* Q. The degree to which the issue I is resolved (or answered) by the inquiry Q is a measure of the *relevance* of the inquiry to the issue. This is expressed mathematically by defining,

$$d(I|Q) = z(I,Q), \tag{25}$$

which is explicitly written as

$$d(I|Q) = \begin{cases} 1 \text{ if } \; I \geq Q \\ 0 \text{ if } \; I \wedge Q = \bot \\ d \text{ otherwise, where } \; 0 < d < 1. \end{cases} \qquad (\textit{relevance}) \tag{26}$$

If the degree is low, then the inquiry has little relevance to the issue. If it is zero, the inquiry does not resolve the issue, and thus is not relevant. For this reason, I call this degree the relevance [5], which I will write as $d(I|Q) = z(I,Q)$. This can be read as the degree to which I is answered by Q, which is the same as the degree to which I includes Q on the lattice. In practice one would most likely work with real questions and compute quantities like $d(T|B)$, which is

[5] This degree was called 'bearing' by Cox, and Robert Fry adopted the symbol b, which is an upside-down p to reflect the relationship with probability. Ariel Caticha suggested the name 'relevance' since its Latin origin would make it more accessible to speakers whose native language was not English. I have chosen to define the relevance as d, which reverses the order of the arguments to reflect the relationship between the relevance d and the probability p as valuations equivalent to the function z defined on their respective lattices. Thus $d(I|Q) \equiv b(Q|I)$.

the degree to which asking '*Did or did not Alice steal the tarts?*' resolves '*Who stole the tarts?*'. This quantity $d(T|B)$ measures the relevance of the question B to the issue T.

The rules of the calculus are straightforward, since they were developed earlier for distributive lattices in general and applied to the assertions in the form of probability. There is the sum rule for the relevance of a question Q to the join of two questions $X \vee Y$

$$d(X \vee Y|Q) = d(X|Q) + d(Y|Q) - d(X \wedge Y|Q), \tag{27}$$

and its generalization

$$d(X_1 \vee X_2 \vee \cdots \vee X_n|Q) = \\ \sum_i d(X_i|Q) - \sum_{i<j} d(X_i \wedge X_j|Q) + \sum_{i<j<k} d(X_i \wedge X_j \wedge X_k|Q) - \cdots , \tag{28}$$

the product rule

$$d(X \wedge Y|Q) = Cd(X|Q)d(Y|X \wedge Q), \tag{29}$$

and a Bayes' theorem analog

$$d(Y|X \wedge Q) = \frac{d(Y|Q)d(X|Y \wedge Q)}{d(X|Q)}, \tag{30}$$

where the constant C in the product rule is the value of the relevance $d(\top|X)$. Relevances, like probabilities, need not be normalized to one.

With the rules of the inquiry calculus in hand, and with Rota's theorem [44, Theorem 1, Corollary 2, p.35], we can take relevances assigned to the join-irreducible elements $\mathfrak{J}(\mathfrak{Q})$ (ideal questions) and compute the relevances between all pairs of questions on the lattice. However, we need an objective means by which to assign these relevances for the ideal questions.

4.7 Entropy from Consistency

To assign relevances, we must maintain consistency with both the algebraic properties of the question lattice $\mathfrak{Q}$ and the probability assignments on the Boolean lattice $\mathcal{A}$. While, I will outline how this is done below, the detailed proofs will be published elsewhere. Clearly from Rota's theorem, we need only determine the relevances of the ideal questions. Once those are assigned, the rest follow from the inquiry calculus.

To determine the form of the relevance, I make a single assumption. That is the degree to which the top question $\top$ answers a join-irreducible question X

depends only on the probability of the assertion x from which the question X was generated. That is, given the ideal question $X = \downarrow x$

$$d(X|\top) = H(p(x|\top)), \tag{31}$$

where H is a function to be determined. In this way, the relevance of the ideal question is required to be consistent with the probability assignments on $\mathcal{A}$. I will outline how the form of the function H can then be determined completely from the properties of the inquiry calculus.

The lattice structure and the inquiry calculus imposes important restrictions on the behavior of the relevance. Given three questions $X, Y, Q \in \mathcal{Q}$ the relevance is *additive* only when $X \wedge Y = \bot$

$$d(X \vee Y|Q) = d(X|Q) + d(Y|Q), \quad \text{iff} \quad X \wedge Y = \bot. \tag{32}$$

However, in general the result is *subadditive*

$$d(X \vee Y|Q) \leq d(X|Q) + d(Y|Q). \tag{33}$$

This is a result of the generalized sum rule, which includes additional terms to avoid double-counting the overlap between the two questions. Commutativity of the join requires that

$$d(X_1 \vee X_2 \vee \cdots \vee X_n|Q) = d(X_{\pi(1)} \vee X_{\pi(2)} \vee \cdots \vee X_\pi(n)|Q) \tag{34}$$

for all permutations $(\pi(1), \pi(2) \cdots, \pi(n))$ of $(1, 2, \cdots, n)$. Thus the relevance must be *symmetric* with respect to the order of the joins.

Last, we consider what happens when an assertion f, known to be false is added to the system. Associated with this assertion f is a question $F = \downarrow f \in \mathcal{Q}$. Now consider the relevance $d(X_1 \vee X_2 \vee \cdots \vee X_n \vee F|Q)$. Since f is known to be false, it can be identified with the absurdity $\bot$, and the lattice $\mathcal{A}$ collapses from $\mathbf{2}^{n+1}$ to $\mathbf{2}^n$. The associated question F is then identified with $F = \downarrow\bot = \bot$, where it is understood that the first $\bot$ refers to the bottom of the lattice $\mathcal{A}$ and the second refers to the bottom of the lattice $\mathcal{Q}$. Since $X \vee \bot = X$, we require, for consistency,

$$d(X_1 \vee X_2 \vee \cdots \vee X_n \vee F|Q) = d(X_1 \vee X_2 \vee \cdots \vee X_n|Q). \tag{35}$$

This requirement is called *expansibility.*

I now define a *partition question* as a real question where its set of answers are neatly partitioned. More specifically

Definition 2 (Partition Question) *A partition question is a real question $P \in \mathcal{R}$ formed from the join of a set of ideal questions $P = \bigvee_{i=1}^{n} X_i$ where $\forall\, X_j, X_k \in \mathcal{J}(\mathcal{Q})$, $X_j \wedge X_k = \bot$.*

I will denote the set of partition questions by $\mathcal{P}$. There are five partition questions in our earlier example: AKN, $N \cup AK$, $K \cup AN$, $A \cup KN$, and $A \cup K \cup N$, which form a lattice isomorphic to the partition lattice $\mathbf{\Pi}_3$ in Figure 1b.

For partition questions, the relevance can be easily computed using (32)

$$d(\bigvee_{i=1}^{n} X_i|\top) = \sum_{i=1}^{n} H(p(q_i|\top)). \tag{36}$$

Writing this as a function K_n of the n probabilities, we get

$$d(\bigvee_{i=1}^{n} X_i|\top) = K_n(p_1, p_2, \cdots, p_n)), \tag{37}$$

where I have written $p_i = p(q_i|\top)$. An important result from Aczél et al. [2] states that if this function K_n satisfies additivity (32), subadditivity (33), symmetry (34), and expansibility (35), then the function can be written as a linear combination of the Shannon and Hartley entropies

$$K_n(p_1, p_2, \cdots, p_n) = a\ H_m(p_1, p_2, \cdots, p_n) + b\ {}_oH_m(p_1, p_2, \cdots, p_n), \tag{38}$$

where a, b are arbitrary non-negative constants, the Shannon entropy [45] is defined as

$$H_m(p_1, p_2, \cdots, p_n) = -\sum_{i=1}^{n} p_i \log_2 p_i, \tag{39}$$

and the Hartley entropy [21] is defined as

$${}_oH_m(p_1, p_2, \cdots, p_n) = \log_2 N(P), \tag{40}$$

where $N(P)$ is the number of non-zero arguments p_i. An additional condition suggested by Aczél states that the Shannon entropy is the only solution if the result is to be small for small probabilities [2]. That is, that the relevance varies continuously as a function of the probability. For the remainder of this work, I will assume that this is the case.

Given these results, it is straightforward to show that the relevance of an ideal question (31) can be written as

$$d(X|\top) = -ap(x|\top) \log_2 p(x|\top), \tag{41}$$

which is proportional to the probability-weighted surprise. With this result in hand, the inquiry calculus enables us to calculate all the other relevances of pairs of questions in the lattice. By requiring consistency with the lattice structure and assuming that the relevance of an ideal question is a continuous function of the probability of its corresponding assertion, we have found that the relevance of a partition question is equal to the Shannon entropy. Thus

$$d(A \vee K \vee N|\top) \propto -p_a \log_2 p_a - p_k \log_2 p_k - p_n \log_2 p_n, \tag{42}$$

where $p_a \equiv p(a|\top), \cdots$. Whereas,

$$d(A \vee KN|\top) \propto -p_a \log_2 p_a - p_{k\vee n} \log_2 p_{k\vee n}, \tag{43}$$

where $p_{k\vee n} \equiv p(k \vee n|\top)$.

The inquiry calculus allows us to compute the degree to which the question T =‘*Who stole the tarts?*’ is answered by the question B =‘*Did or did not Alice steal the tarts?*’ by

$$\begin{aligned} d(T|B) &= d(T|B \wedge \top) \qquad (44)\\ &= d(B|T \wedge \top)\frac{d(B|\top)}{d(T|\top)} \\ &= d(B|T)\frac{d(B|\top)}{d(T|\top)} \\ &= C\frac{d(B|\top)}{d(T|\top)}, \end{aligned}$$

where C is the chosen normalization constant for the relevance. By assigning probabilities to the different cases, this is easily computed using the equations (42) and (43) above.

The relevance of questions such as $AN \cup KN \equiv AN \vee KN$ is even more interesting, since this must be computed using the sum rule

$$d(AN \vee KN|Q) = d(AN|Q) + d(KN|Q) - d(AN \wedge KN|Q), \tag{45}$$

which is equivalent to the mutual information between AN and KN

$$I(AN; KN) = H(AN) + H(KN) - H(AN, KN), \tag{46}$$

although the information-theoretic notation obscures the conditionality of these measures. Thus the relevance of the common question is related to the mutual information, which describes what information is shared by the two questions. The term $d(AN \wedge KN|Q)$ is then identified as the join entropy. In the context of information theory, Cox's choice in naming the common question and joint question is more clear.

However, the inquiry calculus holds new possibilities. The relevance of questions comprised of the join of multiple questions must then be computed using the generalized sum rule (28), which is related to the sum rule via the Möbius function for the lattice. Combined with the Shannon entropy for relevance, this leads to the *generalized entropy* conjectured by Cox as the appropriate measure [12,13]. Furthermore, one can see that this is also the *co-information*, rediscovered by Bell [4], and the lattice on which they are a valuation is precisely the question lattice [29]. We now have a well-founded generalization of information theory, where the commonalities among a set of any number of questions can be quantified.

5 Discussion

There are some significant deviations in this work from Cox's initial explorations. First, the question algebra is the free distributive algebra—not a Boolean algebra as Cox suggested. Cox actually first believed (correctly) that the algebra could not possibly be Boolean [12][pp. 52-3], but then later assumed that complements of questions exist [13][pp. 151-2], which led to the false conclusion that the algebra must be Boolean. This belief, that questions follow a Boolean algebra and therefore possess complements, spread to several early papers following Cox, including two of my own. The second major deviation is that the ordering relation that I have used for question is reversed from the one implicitly adopted by Cox. This led to a version of the consistency relations in Cox's work that is reversed from the consistency relations used in order theory, where joins and meets of questions are swapped. This is related to the third deviation, where I have adopted a notation for relevance that is consistent with the function z from which it and probability both derive. Thus, I use $d(I|Q)$ to describe the degree to which the issue I is resolved by the question Q, which is in keeping with the notation for probability where $p(x|t)$ is the degree to which the statement x is implied by the statement t. This function $d(I|Q)$ is equivalent to Fry's notation $b(Q|I)$. This is a difficult decision, but I feel that mathematical consistency is more important than historical consistency—especially at the early stages of development. Nevertheless, I would like to stress that Cox's achievement in penetrating the realm of questions is remarkable—especially considering the historical focus on the logic of assertions and the surprising lack of attention paid to questions. One notable exception is a paper by Felix Klein titled 'What is a Question?' [29].

Cox's method for deriving probability theory from consistency has both supporters and critics. It is my experience that many criticisms originate from a lack of understanding of how physicists use symmetries, constraints and consistency to narrow down the form of an unknown function in a problem. In fortunate circumstances, this approach leads to a unique solution, or at least a useful solution, with perhaps an arbitrary constant. In more challenging situations, such as using symmetries to understand the behavior of the strong nuclear force, this approach may only exclude possibilities. Other criticisms seem to focus on details related to probability and plausibility. By deriving these rules for degrees of inclusion defined on distributive lattices, I have taken any disagreement to a wider arena—specifically one where probability and plausibility are no longer the issue. The fact that the application of this methodology of deriving laws from ordering relations [34] leads to probability theory, information theory, geometric probability [43,28], quantum mechanics [10], and perhaps to measures in new spaces [34] suggests that there are important insights to be gained here.

It is remarkable that Cox's insight led him to correctly conjecture the relationship between relevance and probability as a generalized entropy [12,13]. Generalizations to information theory go back to its inception [40], and appeared recently in the form of Bell's co-informations, which he realized were related to lattice structures [4]. Bell's co-information is precisely Cox's generalized entropy, and in this paper I have shown that the lattice on which they operate is precisely the question lattice.

In retrospect, the relationship between questions and information theory is not surprising. For example, the design of a communication channel can be viewed as designing a question to be asked of a transmitter. Experimental design, which is also a form of question-asking, has relied on entropy [36,15,38]. Active learning has made experimental design an active process and entropy has found a role here as well alongside Bayesian inference [39,37]. Question-asking is also important when searching a space for an optimal solution [24,41]. Generalizations of information theory have found use in searching for causal interactions among a set of variables. Transfer entropy is designed to extend mutual information to address the asymmetric issue of causality [26]. Given two time-series, X and Y, the transfer entropy can be neatly expressed in terms of the relevance of the common question $X_{i+1} \vee Y_i$ minus relevance of $X_i \vee X_{i+1} \vee Y_i$, where $X_i =$'*What is the value of the time series X at the time i?*' Last, Robert Fry has been working to extend the inquiry calculus to cybernetic control [18] and neural network design [16,19]. Given the scope of these applications, it will be interesting to see the implications that a detailed understanding of the inquiry calculus will have on future developments.

It is already known that some problems can be solved in both the space of assertions and the space of questions. For example, the Infomax Independent Component Analysis (ICA) algorithm [5] is a machine learning algorithm that was originally derived using information theory. However, by considering a source separation problem in terms of the logical statements describing the physical situation, one can derive ICA using probability theory [30]. The information-theoretic derivation can be interpreted in terms of maximizing the relevance of the common question $X \vee Y$, where $X =$'*What are the recorded signals?*' and $Y =$'*How well have we modelled the source activity?*' This is accomplished by using the prior probability distribution of the source amplitudes to *encode* how well the sources have been modelled [31].[6] This notion of encoding answers to questions is important to inductive inquiry.

In this paper, I have laid out the relationship between the algebra of a finite set of logical statements and the algebra of the corresponding questions. The Boolean lattice of assertions $\mathcal{A}$ gives rise to the free distributive lattice of questions $\mathcal{Q}$ as the ordered set of down-sets of the assertion lattice. The join-irreducible elements of the question lattice form a lattice that is isomorphic to the original assertion lattice. Thus the assertion lattice $\mathcal{A}$ is dual to the question lattice $\mathcal{Q}$ in the sense of Birkhoff's representation theorem. Furthermore, the I showed that the question lattice is isomorphic to both the lattice of simplicial complexes and the lattice of hypergraphs, which connects questions to geometric constructs. By generalizing the zeta function on each lattice, I have demonstrated that their algebras can be generalized to calculi, which effectively enable us to *measure* statements and questions. Probability theory is the calculus on $\mathcal{A}$, and as such it is *literally* an extension of Boolean logic. Cox's generalized entropies, which are called co-informations by Bell, quantify the relevance of a question on an issue. Traditional information theory is thus only a portion of the inquiry calculus, which now offers new possibilities. This formalism must now be extended to continuous spaces. An understanding of these fundamental relationships will be essential to utilizing the full power of these mathematical constructs.

References

[1] Aczél J. *Lectures on Functional Equations and Their Applications.* New York:Academic Press, 1966.

[2] Aczél J., Forte B. and Ng C.T. Why the Shannon and Hartley entropies are 'natural'. *Adv. Appl. Prob.*, Vol. 6, pp. 131–146, 1974.

[3] Barnabei M. and Pezzoli E. Gian-Carlo Rota on combinatorics. In *Möbius functions* (ed. J.P.S. Kung), Boston:Birkhauser, pp. 83–104, 1995.

[4] Bell A.J. The co-information lattice. *Proceedings of the Fifth International Workshop on Independent Component Analysis and Blind Signal Separation: ICA 2003* (eds. S. Amari, A. Cichocki, S. Makino and N. Murata), 2003.

[5] Bell A.J. and Sejnowski T.J. An information maximisation approach to blind separation and blind deconvolution, *Neural Computation*, Vol.7, No. 6, pp. 1129–1159, 1996.

[6] Note that complements of questions are used erroneously and unnecessarily in this paper, and that the notation differs from that introduced in this present work as described above.

[6] Bernoulli J. *Ars conjectandi*, Basel:Thurnisiorum, 1713.

[7] Birkhoff G.D. *Lattice Theory*, Providence:American Mathematical Society, 1967.

[8] Boole G. The calculus of logic. *Dublin Mathematical Journal*, Vol. 3, pp. 183–198, 1848.

[9] Boole G. *An investigation of the laws of thought.* London:Macmillan, 1854.

[10] Caticha A. Consistency, amplitudes and probabilities in quantum theory. *Phys. Rev. A*, Vol. 57, pp. 1572–1582, 1998.

[11] Cox R.T. Probability, frequency, and reasonable expectation. *Am. J. Physics*, Vol. 14, pp. 1–13, 1946.

[12] Cox, R.T. *The algebra of probable inference.* Baltimore:Johns Hopkins Press, 1961.

[13] Cox R.T. Of inference and inquiry. In *The Maximum Entropy Formalism* (eds. R. D. Levine & M. Tribus). pp. 119–167, Cambridge:MIT Press, 1979.

[14] Davey B.A. & Priestley H.A. *Introduction to lattices and order.* Cambridge:Cambridge Univ. Press, 2002.

[15] Fedorov V.V. *Theory of Optimal Experiments*, New York:Academic Press, 1972.

[16] Fry R.L. Observer-participant models of neural processing. *IEEE Trans. Neural Networks*, Vo. 6, pp. 918–928, 1995.

[17] Fry R.L. Maximum entropy and Bayesian methods. Electronic course notes (525.475), Johns Hopkins University, 1999.

[18] Fry R.L. The engineering of cybernetic systems. In *Bayesian Inference and Maximum Entropy Methods in Science and Engineering, Baltimore MD, USA, August 2001* (ed. R. L. Fry). New York:AIP, pp. 497–528, 2002.

[19] Fry R.L., Sova R.M. A logical basis for neural network design. In *Implementation Techniques: Neural Network Systems Techniques and Applications*, Vol. 3, (ed. C. Leondes). London:Academic Press, 1998.

[20] Garrett A.J.M Whence the laws of probability? *Maximum Entropy and Bayesian Methods. Boise, Idaho, USA, 1997* (eds. G. J. Erickson, J. T. Rychert & C. R. Smith), Dordrecht:Kluwer Academic Publishers, pp. 71–86, 1998.

[21] Hartley R.V. Transmission of information. *Bell System Tech. J.*, Vol. 7, pp. 535–563, 1928.

[22] Jaynes E.T. Prior Probabilities. *IEEE Trans. Syst. Sci. Cyb.* Vol. SSC-4, pp. 227–, 1968.

[23] Jaynes E.T. Where do we stand on maximum entropy? In *The Maximum Entropy Formalism* (eds. R. D. Levine & M. Tribus), pp. 15–118, Cambridge:MIT Press, 1979.

[24] Jaynes E.T. Entropy and search theory. In *Maximum Entropy and Bayesian Methods in Inverse Problems* (eds. C. R. Smith & W. T. Grandy, Jr.), Dordrecht:Reidel, pp. 443–, 1985.

[25] Jaynes E.T. *Probability theory: the logic of science.* Cambridge:Cambridge Univ. Press, 2003.

[26] Kaiser A., Schreiber T. Information transfer in continuous processes, *Physica D*, Vol. 166, pp. 43–62, 2002.

[27] Keynes J.M. *A treatise on probability.* London:Macmillan, 1921.

[28] Klain D.A. & Rota G.-C. *Introduction to geometric probability.* Cambridge:Cambridge Univ. Press, 1997.

[29] Klein F. What is a question?, *The Monist*, Vol. 39, pp. 350–364, 1929.

[30] Knuth K.H. A Bayesian approach to source separation. In *Proceedings of the First International Workshop on Independent Component Analysis and Signal Separation: ICA'99* (eds. J.-F. Cardoso, C. Jutten and P. Loubaton), Aussois, France, pp. 283–288, 1999.

[31] Knuth K.H. Source separation as an exercise in logical induction. In *Bayesian Inference and Maximum Entropy Methods in Science and Engineering, Paris 2000* (ed. A. Mohammad-Djafari), AIP Conference Proceedings Vol. 568, Melville NY:American Institute of Physics, pp. 340–349, 2001.

[32] Knuth K.H. What is a question? In *Bayesian Inference and Maximum Entropy Methods in Science and Engineering, Moscow ID, USA, August 2002* (ed. C. Williams). AIP Conference Proceedings Vol. 659, Melville NY:AIP, , pp. 227–242, 2002.

[33] Knuth K.H. Intelligent machines in the 21st century: foundations of inference and inquiry, *Phil. Trans. Roy. Soc. Lond. A*, Vol. 361, No. 1813, pp. 2859-2873, 2003.

[34] Knuth K.H. Deriving laws from ordering relations. In press: *Bayesian Inference and Maximum Entropy Methods in Science and Engineering, Jackson Hole WY, USA, August 2003* (ed. G. J. Erickson). AIP Conference Proceedings Vol. 707, Melville NY:American Institute of Physics, 2004.

[35] Laplace P.S. *Théorie analytique des probabilités.* Paris:Courcier Imprimeur, 1812.

[36] Lindley D.V. On the measure of information provided by an experiment. *Ann. Math. Statist.* Vol. 27, pp. 986–1005, 1956.

[37] Loredo T.J. Bayesian adaptive exploration. In press: *Bayesian Inference and Maximum Entropy Methods in Science and Engineering, Jackson Hole WY, USA, August 2003* (ed. G. J. Erickson). AIP Conference Proceedings Vol. 707, Melville NY:AIP, 2004.

[38] LUTTRELL S.P. The use of transinformation in the design of data sampling schemes for inverse problems. *Inverse Problems* Vol. 1, pp. 199–218, 1985.

[39] MACKAY D.J.C. Information-based objective functions for active data selection. *Neural Computation* Vol. 4 No. 4, pp. 589–603, 1992.

[40] MCGILL W.J. Multivariate information transmission. *IEEE Trans. Info. Theory*, Vol. 4, pp. 93–111, 1955.

[41] PIERCE J.G. A new look at the relation between information theory and search theory. In *The Maximum Entropy Formalism* (eds. R. D. Levine & M. Tribus), pp. 339–402, Cambridge:MIT Press, 1979.

[42] ROTA G.-C. On the foundations of combinatorial theory I. Theory of Möbius functions. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, Vol. 2, pp. 340–368, 1964.

[43] ROTA G.-C. Geometric probability. *The Mathematical Intelligencer*, Vol. 20, pp. 11–16, 1998.

[44] ROTA G.-C. On the combinatorics of the Euler characteristic. In *Studies in Pure Mathematics (Presented to Richard Rado)*, London:Academic Press, pp. 221–233, 1971.

[45] SHANNON C.E. AND WEAVER W. *A mathematical theory of communication.* Chicago:Univ. of Illinois Press, 1949.

[46] SMITH C.R. AND ERICKSON G.J. Probability theory and the associativity equation. In *Maximum Entropy and Bayesian Methods*, (ed. P. Fougere). Dordrecht:Kluwer, pp. 17–30, 1990.

[47] TRIBUS M. *Rational Descriptions, Decisions and Designs*, New York:Pergamon Press, 1969.

APPENDIX: Deriving the rules of the calculi

In this appendix, I derive the sum rule, product rule, and Bayes' theorem analog for distributive lattices. These rules are equally applicable to probability on the Boolean lattice of assertions $\mathcal{A}$, to relevance on the free distributive lattice of questions $\mathcal{Q}$, and any other distributive lattice. The first derivation of these rules was by Cox for complemented distributive lattices (Boolean lattices) [11,12]. The derivations rely on maintaining consistency between the proposed calculus and the properties of the algebra. In Cox's derivations, he relied on consistency with complementation to obtain the sum rule, and consistency with associativity of the meet to obtain the product rule. The derivation of Bayes' theorem is, in contrast, well-known since it follows directly from commutativity of the meet. An interesting variation of Cox's derivation for

Boolean algebra relying on a single algebraic operation (NAND) was introduced by Anthony Garrett [20].

Below, I expound on the derivations introduced by Ariel Caticha, which rely on associativity and distributivity [10]. The implications of Caticha's derivation are profound, since his results imply that the sum rule, the product rule, and Bayes' theorem are consistent with distributive lattices in general. These implications are discussed in greater detail elsewhere [34].

Consistency with Associativity

Consider a distributive lattice $\mathcal{D}$, two join-irreducible elements, $a, b \in \mathcal{J}(\mathcal{D})$, where $a \wedge b = \bot$, and a third element $t \in \mathcal{D}$ such that $a \wedge t \neq \bot$ and $b \wedge t \neq \bot$. We begin by introducing a degree of inclusion (see Eqn. 6) represented by the function ϕ, so that the degree to which a includes t is given by $\phi(a, t)$. We would like to be able to compute the degree to which the join $a \vee b$ includes t. In terms of probability, this is the degree to which t implies $a \vee b$.

Since $a \wedge b = \bot$, this degree of inclusion can only be a function of $\phi(a, t)$ and $\phi(b, t)$, which can be written as

$$\phi(a \vee b, t) = S(\phi(a, t), \phi(b, t)). \tag{A-1}$$

The function S, will tell us how to use $\phi(a, t)$ and $\phi(a, t)$ to compute $\phi(a \vee b, t)$. The hope is that the consistency constraint will be sufficient to identify the form of S—we will see that this is the case.

The function S must maintain consistency with the distributive algebra $\mathcal{D}$. Consider another join-irreducible element $c \in \mathcal{J}(\mathcal{D})$ where $a \wedge c = \bot$, $b \wedge c = \bot$, and form the element $(a \vee b) \vee c$. We can use associativity of the lattice to write this element two ways

$$(a \vee b) \vee c = a \vee (b \vee c). \tag{A-2}$$

Consistency requires that each expression gives the same result when the degree of inclusion is calculated

$$S(\phi(a \vee b, t), \phi(c, t)) = S(\phi(a, t), \phi(b \vee c, t)). \tag{A-3}$$

Applying S to the arguments $\phi(a \vee b, t)$ and $\phi(b \vee c, t)$ above, we get

$$S(S(\phi(a, t), \phi(b, t)), \phi(c, t)) = S(\phi(a, t), S(\phi(b, t), \phi(c, t))). \tag{A-4}$$

This can be further simplified by letting $u = \phi(a, t)$, $v = \phi(b, t)$, and $w = \phi(c, t)$ resulting in a *functional equation* for the function S, which Aczél called *the associativity equation* [1, pp. 253-273].

$$S(S(u,v),w) = S(u,S(v,w)). \tag{A-5}$$

The general solution [1], is

$$S(u,v) = f(f^{-1}(u) + f^{-1}(v)), \tag{A-6}$$

where f is an arbitrary function. This is simplified by letting $g = f^{-1}$

$$g(S(u,v)) = g(u) + g(v). \tag{A-7}$$

Writing this in terms of the original expressions we get,

$$g(\phi(a \vee b, t)) = g(\phi(a,t)) + g(\phi(b,t)), \tag{A-8}$$

which reveals that there exists a function $g : \mathbb{R} \to \mathbb{R}$ re-mapping these numbers to a more convenient representation. Defining $z(a,t) \equiv g(\phi(a,t))$ we get

$$z(a \vee b, t) = z(a,t) + z(b,t), \tag{A-9}$$

which is the sum rule for the join of two join-irreducible elements.

This rule can be extended to all elements in $\mathcal{D}$ by using the Möbius function for the lattice, or equivalently the *inclusion-exclusion relation*, which avoids double-counting the elements in the calculation [28,42,3,34]. This leads to the generalized sum rule for the join of two arbitrary elements

$$z(a \vee b, t) = z(a,t) + z(b,t) - z(a \wedge b, t), \tag{A-10}$$

and

$$\begin{aligned} &z(x_1 \vee x_2 \vee \cdots \vee x_n, t) = \\ &\qquad \sum_i z(x_i,t) - \sum_{i<j} z(x_i \wedge x_j, t) + \sum_{i<j<k} z(x_i \wedge x_j \wedge x_k, t) - \cdots \end{aligned} \tag{A-11}$$

for the join of multiple arbitrary elements $x_1, x_2, \ldots, x_n$ [34].

Consistency with Distributivity

Given $x, y, t \in \mathcal{D}$, we would like to be able to compute the degree to which the meet $x \wedge y$ includes t, written $z(x \wedge y, t)$. We can easily use (A-10) to obtain

$$z(x \wedge y, t) = z(x,t) + z(y,t) - z(x \vee y, t). \tag{A-12}$$

However, another form can be found by requiring consistency with distributivity D1. Following Cox [11,12], and relying on the consistency arguments given by Jaynes [25], Tribus [47], and Smith and Erickson [46], this degree can be written two ways as a function P of two arguments

$$z(x \wedge y, t) = P(z(x,t), z(y, x \wedge t)) = P(z(y,t), z(x, y \wedge t)), \tag{A-13}$$

where the function P will tell us how to do the calculation. The two expressions on the right are a consequence of commutativity, which we will address later.

We will focus for now on the first expression of P, and consider five elements $a, b, r, s, t \in \mathcal{D}$ where $a \wedge b = \bot$ and $r \wedge s = \bot$. By considering distributivity D1 of the meet over the join, we can write $a \wedge (r \vee s)$ two ways

$$a \wedge (r \vee s) = (a \wedge r) \vee (a \wedge s). \tag{A-14}$$

Consistency with distributivity D1 requires that their relevances calculated these two ways are equal. Using the sum rule (A-10) and the form of P (A-13), distributivity requires that

$$P(z(a,t), z(r \vee s, a \wedge t)) = z(a \wedge r, t) + z(a \wedge s, t), \tag{A-15}$$

which simplifies to

$$\begin{aligned} &P(z(a,t), z(r, a \wedge t) + z(s, a \wedge t)) = \\ &\qquad P(z(a,t), z(r, a \wedge t)) + P(z(a,t), z(s, a \wedge t)). \end{aligned} \tag{A-16}$$

Defining $u = z(a,t)$, $v = z(r, a \wedge t)$, and $w = z(s, a \wedge t)$, the equation above can be written as

$$P(u, v + w) = P(u, v) + P(u, w). \tag{A-17}$$

This functional equation for P captures the essence of distributivity D1.

We will now show that $P(u, v + w)$ is linear in its second argument. Defining $k = w + v$, and writing (A-17) as

$$P(u, k) = P(u, v) + P(u, w), \tag{A-18}$$

we can compute the second derivative with respect to x. Using the chain rule we find that

$$\frac{\partial}{\partial v} = \frac{\partial k}{\partial v}\frac{\partial}{\partial k} = \frac{\partial}{\partial k} = \frac{\partial k}{\partial w}\frac{\partial}{\partial k} = \frac{\partial}{\partial w}, \tag{A-19}$$

so that the second derivative with respect to k can be written as

$$\frac{\partial^2}{\partial k^2} = \frac{\partial}{\partial v}\frac{\partial}{\partial w}. \tag{A-20}$$

The second derivative of $P(u, k)$ with respect to k is then easily shown to be

$$\frac{\partial^2}{\partial k^2} P(u, k) = 0, \tag{A-21}$$

which implies that the function P is linear in its second argument

$$P(u, v) = A(u)v + B(u), \tag{A-22}$$

where A and B are functions to be determined. Substitution of (A-22) into (A-17) gives $B(u) = 0$.

Now we consider $(a \vee b) \wedge r$, which using D1 can be written as

$$(a \vee b) \wedge r = (a \wedge r) \vee (b \wedge r), \tag{A-23}$$

gives a similar functional equation

$$P(v + w, u) = P(v, u) + P(w, u), \tag{A-24}$$

where $u = z(r, t)$, $v = z(a, r \wedge t)$, $w = z(b, r \wedge t)$. Following the approach above, we see that P is also linear in its first argument

$$P(u, v) = A(v)u. \tag{A-25}$$

Together with (A-22), the general solution is

$$P(u, v) = Cuv, \tag{A-26}$$

where C is an arbitrary constant. Thus we have the *product rule*

$$z(x \wedge y, t) = Cz(x, t)z(y, x \wedge t), \tag{A-27}$$

which tells us the degree to which t includes the meet $x \wedge y$. The constant C acts as a normalization factor, and is necessary when these valuations are normalized to values other than unity. It should be noted that this only satisfies the distributivity of the meet over the join D1. There are reasons why D1 is preferred over D2 related to the lattice product, which are discussed elsewhere [34].

Consistency with Commutativity

Commutativity of the meet is the reason that there are two forms for the function P, the product rule (A-13)

$$z(x \wedge y, t) = Cz(x, t)z(y, x \wedge t) \tag{A-28}$$

and

$$z(y \wedge x, t) = Cz(y, t)z(x, y \wedge t). \tag{A-29}$$

Equating the degrees (A-28) and (A-29) results in

$$Cz(x, t)z(y, x \wedge t) = Cz(y, t)z(x, y \wedge t), \tag{A-30}$$

which leads to Bayes' theorem

$$z(y, x \wedge t) = \frac{z(y,t)z(x, y \wedge t)}{z(x,t)}. \tag{A-31}$$

This demonstrates that there is a sum rule, a product rule, and a Bayes' Theorem analog for bi-valuations on all distributive lattices. This realization clears up the mystery as to why some quantities in science act like probabilities, but clearly are not probabilities [34].