

Information Dependencies

Mehmet M. Dalkilic
Indiana University Computer Science
Lindley Hall 215
Bloomington, Indiana 47405 USA
dalkilic@cs.indiana.edu

Edward L. Roberston
Indiana University Computer Science
Lindley Hall 215
Bloomington, Indiana 47405 USA
edrbtsn@cs.indiana.edu

ABSTRACT

This paper uses the tools of information theory to examine and reason about the information content of the attributes within a relation instance. For two sets of attributes X and Y , an *information dependency measure* (InD measure) characterizes the uncertainty remaining about the values for the set Y when the values for the set X are known. A variety of arithmetic inequalities (InD *inequalities*) are shown to hold among InD measures; InD inequalities hold in any relation instance. Numeric constraints (InD *constraints*) on InD measures, consistent with the InD inequalities, can be applied to relation instances. Remarkably, functional and multivalued dependencies correspond to setting certain constraints to zero, with Armstrong's axioms shown to be consequences of the arithmetic inequalities applied to constraints. As an analog of completeness, for any set of constraints consistent with the inequalities, we may construct a relation instance that approximates these constraints within any positive ϵ . InD measures suggest many valuable applications in areas such as data mining.

Categories and Subject Descriptors

H.1.1 [Information Systems]: Models and Principles—*information theory*; H.2.1 [Information Systems]: Database Management; H.2.8 [Information Systems]: Database Management—*data mining*; H.2.m [Information Systems]: Miscellaneous

General Terms

Design, Theory, Measurement

Keywords

information dependency, entropy, functional dependency, Armstrong's Axioms, Multivalued dependency

1. INTRODUCTION

THAT the well-developed discipline of *information theory* seemed to have so little to say about *information systems* is a long-standing conundrum. Attempts to use information

theory to “measure” the information content of a relation are blocked by the inability to accurately characterize the underlying domain. An answer to this mystery is that we have been looking in the wrong place. The tools of information theory, dealing closely with representation issues, apply within a relation instance and between the various attributes of that instance.

The traditional approach to information theory is based upon communication via a *channel*. In each instance there is a fixed set of messages $M = \{v_1, \dots, v_n\}$; when one of these is transmitted from the *sender* to the *receiver* (via the channel), the receiver gains a certain amount of information. The less likely a message is to be sent, the more meaningful is its receipt. This is formalized by assigning to each message v_i a probability p_i (subject to the natural constraint that $\sum_{i=1}^n p_i = 1$) and defining the information content of v_i to be $\log 1/p_i$ (all logarithms in this paper are base 2).

Another way of viewing this measure is that the amount of information in a message is related to how “surprising” the message is—a weather report during the month of July contains little information if the prediction is “hot,” but a prediction of “snow” carries a lot of information. The issue of surprise is also related to the recipient's “state of knowledge.” In the weather report example, the astonishment of the report “snow” was directly related to the knowledge that it was July; in January the information content of the two reports would be vastly different. Thus the in- or interdependence of two sets of messages is highly significant. If two message sets are independent (in the intuitive and the statistical sense), receipt of a message from one set does not alter the information content of the other (*e.g.* temperature and wind speed). If two message sets are not independent, receipt of a message from the first set may greatly alter the likelihood of receipt, and hence information content, of messages from the second set (*e.g.* temperature and form of precipitation).

A central concept in information theory is the *entropy* \mathcal{H} of a set of messages, the weighted average of the message information:

Definition 1. Entropy. Given a set $M = \{v_1, \dots, v_n\}$ of messages with probabilities $P_M = \{p_1, \dots, p_n\}$, the entropy of M is $\mathcal{H}(M) = \mathcal{H}(p_1, \dots, p_n) = \sum_{i=1}^n p_i \log 1/p_i$.

Entropy is closely related to encoding of messages, in that

encoding each v_i using $\log 1/p_i$ bits gives the minimal number of expected bits for transmitting messages of M .

Remark 1. Suppose for messages of M , no probability is 0 and $\mathcal{H}(M) = 0$. Then $M = \{v_1\}$, i.e. M contains a single message.

In a database context, information content is measured in terms of selection (specification of a specific value) rather than transmission. This avoids the thorny problem which seems to say that, since the database is stored on site and no transmission occurs, there is no information. In particular, the model looks at an instance of a single relation and at values for some arbitrarily selected tuple. Because of the assumption that all tuples are equally likely, the information required to specify one particular tuple from a relation instance with n tuples is, of course, $\log n$ and the minimal cost of uniform encoding requires $\log n$ bits. Thus a single value carries $\log n$ bits only if it is drawn from an attribute which has n distinct values, that is, when the attribute is a key.

The major results of this paper use the common definition of information to characterize *information dependencies*. This characterization has three steps. The first extends the use of entropy as a measure of information to be an *information dependency measure* (Section 4). The second derives a number of *arithmetic inequalities* which must always hold between particular measures in a relation instance (Section 5). The third investigates the consequences of placing *numeric constraints* on some or all measures of a relation instance. Most significantly, functional and multi-valued dependency result from constraining certain particular measures (or their differences) to zero (Section 6).

In addition, the measure/constraint formulation exhibits an analog of completeness in that, for any set of numeric constraints consistent with the arithmetic inequalities and any positive ϵ , there is a relation instance that achieves those constraints within ϵ (Section 7).

This characterization of information dependency has many important theoretic and practical implications. It allows us to more carefully investigate notions of approximate functional dependency. It can help with normalization. It opens up whole realms of data mining approaches.

2. PRELIMINARIES

2.1 Notations

Here are the notations and conventions.

Relations All relation instances are non-empty multi-sets. \mathbf{r}, \mathbf{s} denote relation instances. Operators π, σ do not filter for distinctiveness.

Attributes R is schema for instance \mathbf{r} and $X, Y, Z, V, W \subseteq R$. XY denotes $X \cup Y$ and A is equivalent to $\{A\}$ for $A \in R$. X, Y, Z partition R .

Values Null values are not considered here. v is equivalent to $\langle v \rangle$ when $\langle v \rangle \in \pi_A(\mathbf{r})$. $\ell = \mathbf{count-distinct}(\pi_X(\mathbf{r}))$. x_i

enumerates the instances of $\mathbf{distinct}(\pi_X(\mathbf{r}))$, so $1 \leq i \leq \ell$, similarly for m and y_j wrt Y , and n and z_k wrt Z .

Probabilities $P(S = v) = \frac{\mathbf{count}(\sigma_{S=v}(\mathbf{r}))}{\mathbf{count}(\mathbf{r})}$ for $S \subseteq R$. $p_i = P(X = x_i)$ (note use of i is consistent with above), similarly $p_j = P(Y = y_j)$, $p_k = P(Z = z_k)$, $p_{ij} = P(X = x_i \& Y = y_j)$, and so forth. $\sum_i^n p_i = \sum_{i=1}^n p_i$ and likewise for p_j, p_k , etc.

Definition 2. Conditional Probability. The *conditional probability* of $Y = y_j$ given $X = x_i$, written $P(Y = y_j | X = x_i)$, is $P(Y = y_j)$ in the instance $\sigma_{X=x_i}(\mathbf{r})$. In symbols, $P(Y = y_j | X = x_i) = P(X = x_i \& Y = y_j) / P(X = x_i)$.

Definition 3. Independence. X, Y are *independent* if $P(Y = y_j) = P(Y = y_j | X = x_i)$.

In this paper, there are log function expressions of the form $\log(1/0)$. By convention (continuity arguments), $0 \log(1/0) = 0$.

LEMMA 2.1. Let $P = \{p_1, \dots, p_n\}$ be a probability distribution and $Q = \{q_1, \dots, q_n\}$ such that $\sum_i^n q_i \leq 1$ and $(\forall i) 1 \leq i \leq n, 0 \leq q_i \leq 1$. Then $\sum_i^n p_i \log 1/p_i \leq \sum_i^n p_i \log 1/q_i$. Shown in [12, page 22].

2.2 Entropy

Definition 4. Entropy in a relation instance. For a relation instance \mathbf{r} with schema $R \supseteq X$, the entropy of \mathbf{r} over X , written $\mathcal{H}_X(\mathbf{r})$, is

$$\mathcal{H}_X(\mathbf{r}) = \sum_i^n p_i \log 1/p_i$$

3. THE BOUNDS ON ENTROPY

To ease notation, we write \mathcal{H}_X for $\mathcal{H}_X(\mathbf{r})$. From now on, we understand that \mathcal{H} is always associated with a non-empty instance \mathbf{r} . When \mathbf{r} is not clear from context, we write $\mathcal{H}_X^{\mathbf{r}}$.

In the remainder of this section, we establish upper and lower bounds on the entropy function.

LEMMA 3.1. Upper and Lower Bounds on Entropy. $0 \leq \mathcal{H}_X \leq \log \ell$. Proved in [14, pp.14-40].

A consequence of our notation allows us to find the joint entropy of sets $X, Y \subseteq R$. The joint entropy of X, Y , written \mathcal{H}_{XY} , is $\mathcal{H}_{XY} = \mathcal{H}(p_{1,1}, \dots, p_{\ell,m}) = \sum_i^\ell \sum_j^m p_{i,j} \log 1/p_{i,j}$.

LEMMA 3.2. Bounds on Joint Entropy. $X, Y \subseteq R$,

$$\mathcal{H}_X + \mathcal{H}_Y \geq \mathcal{H}_{XY} \geq \mathbf{max}(\mathcal{H}_X, \mathcal{H}_Y)$$

with $\mathcal{H}_X + \mathcal{H}_Y = \mathcal{H}_{XY}$ if X, Y are independent.

PROOF. First inequality: This follows in the spirit of [14, page 28]. \square

PROOF. Second inequality: Observe that $p_i = \sum_j p_{i,j}$. Let $q_i = \max\{p_{i,j} | 1 \leq j \leq m\}$. Then for any j , $p_i \geq q_i \geq p_{i,j}$ and consequently, $\log 1/p_{i,j} \geq \log 1/q_i$ and thus,

$$\begin{aligned} \mathcal{H}_X &= \sum_i^\ell p_i \log 1/p_i \leq \sum_i^\ell p_i \log 1/q_i \text{ \textbf{Lm 2.1}} \\ &= \sum_i^\ell \sum_j^m p_{i,j} \log 1/q_i \leq \sum_i^\ell \sum_j^m p_{i,j} \log 1/p_{i,j} \\ &= \mathcal{H}_{XY} \end{aligned}$$

and symmetrically for \mathcal{H}_Y as well. \square

4. InD MEASURES

An *information dependency measure* (InD measure) between X and Y , for $X, Y \subseteq R$, attempts to answer the question ‘‘How much do we *not* know about Y provided we know X ?’’ Using the notation of **Section 2**, if we know that $X = x_i$, then we are possibly more informed about $Y = y_j$ and therefore, can recalculate the entropy of Y as

$$\begin{aligned} \mathcal{H}(Y|X = x_i) &= \\ \mathcal{H}(p_{i,1}/p_i, \dots, p_{i,m}/p_i) &= \sum_j^m \frac{p_{i,j}}{p_i} \log \frac{p_i}{p_{i,j}} \end{aligned}$$

Amortizing this over each of the ℓ different X values according to the respective probabilities p_i gives the entropy of Y dependent on X , resulting in the following definition of an information dependency measure. Note that these are measures, not metrics.

Definition 5. Information Dependency Measure. The *information dependency measure* (InD measure) of Y given X is $\mathcal{H}_{X \rightarrow Y}$,

$$\sum_i^\ell p_i \cdot \mathcal{H}(Y|X = x_i) = \sum_i^\ell \sum_j^m p_{i,j} \log \frac{p_i}{p_{i,j}}$$

We will now normally drop the word ‘‘entropy’’ when referring to these measures, but it is important to keep in mind that this value is not a declaration of dependency (as is the case with FDs) but a *measure* of dependency. We now characterize an InD measure $\mathcal{H}_{X \rightarrow Y}$ in terms of InD measures \mathcal{H}_X and \mathcal{H}_{XY} (traditional information theory has similar results *e.g.*, [14]).

LEMMA 4.1. $\mathcal{H}_{X \rightarrow Y} = \mathcal{H}_{XY} - \mathcal{H}_X$.

PROOF.

$$\begin{aligned} \mathcal{H}_{X \rightarrow Y} &= \sum_i^\ell \sum_j^m p_{i,j} \log \frac{p_i}{p_{i,j}} \\ &= \sum_i^\ell \sum_j^m p_{i,j} \log 1/p_{i,j} - \sum_i^\ell \sum_j^m p_{i,j} \log 1/p_i \\ &= \mathcal{H}_{XY} - \sum_i^\ell p_i \log 1/p_i \\ &= \mathcal{H}_{XY} - \mathcal{H}_X \end{aligned}$$

\square

r		InD measures	
A	B		
a	e	\mathcal{H}_A	= 7/4
a	f	\mathcal{H}_B	= 3/2
a	e	\mathcal{H}_{AB}	= 9/4
a	f	\mathcal{H}_{A-B}	= 1/2
b	g	\mathcal{H}_{B-A}	= 3/4
b	g		
c	g		
d	g		

Figure 1: (left) An instance r . (right) InD measures of r . Observe that $\mathcal{H}_{A-B} = \mathcal{H}_{AB} - \mathcal{H}_A$ and $\mathcal{H}_{A-B} = \sum_i^4 p_{a_i} \cdot \mathcal{H}(B|a_i) = 1/2 \mathcal{H}(1/2, 1/2, 0) + 1/4 \mathcal{H}(0, 0, 1) + 2(1/8) \mathcal{H}(0, 0, 1) = 1/2 + 0 + 0 = 1/2$.

A	AB	B
	$\square 0$	00 :f
a : 0	$\square 1$	01 :e
b : 10	$\square 0$	
c : 110	$\square 1 0$	1 :g
d : 111	$\square 1 1$	

Figure 2: Encodings of A, overlap of A and B, and A from Fig.1. The \square contains the portion of the bit string that encodes A, \sqsubset similarly for B. The surprise after receiving $A=a$ is witnessed by the fact that, although we know we will receive the first bit of $B=e$ or $B=f$, *i.e.* 0, we need an additional bit for both the second bit of $B=e$ and $B=f$. Receipt of $A=b, A=c$, or $A=d$, on the other hand, poses no further surprise since $B=g$ is completely contained therein.

Note that $\mathcal{H}_{X \rightarrow Y}$ is a measure of the information needed to represent Y given that X is known, not the information that X contains about Y . This latter quantity of course is measured by

$$\mathcal{H}_Y - \mathcal{H}_{X \rightarrow Y} = \mathcal{H}_Y + \mathcal{H}_X - \mathcal{H}_{XY} = \mathcal{H}_X - \mathcal{H}_{XY}$$

yielding the observation that Y contains exactly as much information about X that X has about Y .

5. InD MEASURE INEQUALITIES

The relationships among InD measures are characterized by inequalities and expressions involving the various measures. Of these formulae, several are named according to the corresponding functional dependency inference rules, which they characterize under special circumstances.

LEMMA 5.1. *Reflexivity.* $\mathcal{H}_{X \rightarrow Y} = 0$, for $Y \subseteq X \subseteq R$.

PROOF. Let $Z = YX$. Then by **Lm 4.1** $\mathcal{H}_{Z \rightarrow Y} = \mathcal{H}_{ZY} - \mathcal{H}_Z = \mathcal{H}_Z - \mathcal{H}_Z = 0$. \square

LEMMA 5.2. $\mathcal{H}_{XZ \rightarrow YZ} = \mathcal{H}_{XZ \rightarrow Y}$.

PROOF.

$$\begin{aligned}
& \mathcal{H}_{XZ \rightarrow YZ} \\
&= \mathcal{H}_{XYZ} - \mathcal{H}_{XZ} && \text{Lm 4.1} \\
&= \mathcal{H}_{XZ \rightarrow Y} + \mathcal{H}_{XZ} - \mathcal{H}_{XZ} && \text{Lm 4.1} \\
&= \mathcal{H}_{XZ \rightarrow Y}
\end{aligned}$$

□

As might be expected, InD relationships may be expressed in general or significantly tighter special cases. The following three lemmas illustrate the situation: two InDs may interact little so they combine to sum their InDs, or they may interact strongly, so their combination yields total dependencies. Putting restrictions on the left- or right-hand sides constrains the interactions and hence tightens the InD relationships.

LEMMA 5.3. *Union (right).* $\mathcal{H}_{X \rightarrow Y} + \mathcal{H}_{X \rightarrow Z} \geq \mathcal{H}_{X \rightarrow YZ}$ with equality if $p_{j|i}$ and $p_{k|i}$ are independent.

PROOF.

$$\begin{aligned}
& \mathcal{H}_{X \rightarrow Y} + \mathcal{H}_{X \rightarrow Z} \\
&= \sum_i^n \sum_j^m p_{i,j} \log 1/p_{j|i} + \sum_i^n \sum_k^q p_{i,k} \log 1/p_{k|i} \\
&= \sum_i^n p_i \sum_j^m \sum_k^q p_{j,k|i} \log 1/p_{j|i} p_{k|i} \\
&\geq \sum_i^n p_i \sum_j^m \sum_k^q p_{j,k|i} \log 1/p_{j,k|i} \quad (\forall i) 1 \leq i \leq n, \text{ Lm 2.1}
\end{aligned}$$

If $p_{j|i}$ and $p_{k|i}$ are independent then $p_{j,k|i} = p_{j|i} \cdot p_{k|i}$ and then

$$\begin{aligned}
& \mathcal{H}_{X \rightarrow Y} + \mathcal{H}_{X \rightarrow Z} \\
&= \sum_i^n p_i \sum_j^m \sum_k^q p_{j,k|i} \log 1/p_{j|i} p_{k|i} \\
&= \sum_i^n p_i \sum_j^m \sum_k^q p_{j,k|i} \log 1/p_{j,k|i} \\
&= \mathcal{H}_{X \rightarrow YZ}
\end{aligned}$$

□

LEMMA 5.4. $\mathcal{H}_{X \rightarrow YZ} = \mathcal{H}_{X \rightarrow Y} + \mathcal{H}_{XY \rightarrow Z} \geq \max(\mathcal{H}_{X \rightarrow Y}, \mathcal{H}_{XY \rightarrow Z})$.

PROOF.

$$\begin{aligned}
& \mathcal{H}_{X \rightarrow YZ} \\
&= \mathcal{H}_{XYZ} - \mathcal{H}_X && \text{Lm 4.1} \\
&= \mathcal{H}_{XY \rightarrow Z} + \mathcal{H}_{XY} - \mathcal{H}_X && \text{Lm 4.1} \\
&= \mathcal{H}_{XY \rightarrow Z} + \mathcal{H}_{X \rightarrow Y} && \text{Lm 4.1}
\end{aligned}$$

□

LEMMA 5.5. $\mathcal{H}_{XY \rightarrow Z} \leq \mathcal{H}_{X \rightarrow Z}$.

PROOF.

$$\begin{aligned}
& \mathcal{H}_{XY \rightarrow Z} \\
&= \mathcal{H}_{X \rightarrow YZ} - \mathcal{H}_{X \rightarrow Y} && \text{Lm 5.4} \\
&\leq \mathcal{H}_{X \rightarrow Y} + \mathcal{H}_{X \rightarrow Z} - \mathcal{H}_{X \rightarrow Y} && \text{Lm 5.3} \\
&\leq \mathcal{H}_{X \rightarrow Z}
\end{aligned}$$

□

LEMMA 5.6. *Union (left).* $\min(\mathcal{H}_{X \rightarrow Z}, \mathcal{H}_{Y \rightarrow Z}) \geq \mathcal{H}_{XY \rightarrow Z}$.

PROOF.

$$\begin{aligned}
& \mathcal{H}_{X \rightarrow Z} \geq \mathcal{H}_{XY \rightarrow Y} && \text{Lm 5.5} \\
& \mathcal{H}_{Y \rightarrow Z} \geq \mathcal{H}_{XY \rightarrow Z} && \text{Lm 5.5} \\
& \min(\mathcal{H}_{X \rightarrow Z}, \mathcal{H}_{Y \rightarrow Z}) \geq \mathcal{H}_{XY \rightarrow Z}
\end{aligned}$$

□

LEMMA 5.7. *Augmentation (1).* $\mathcal{H}_{XZ \rightarrow YZ} \leq \mathcal{H}_{X \rightarrow Y}$.

PROOF.

$$\begin{aligned}
& \mathcal{H}_{XZ \rightarrow YZ} \\
&= \mathcal{H}_{XZ \rightarrow Y} && \text{Lm 5.2} \\
&\leq \mathcal{H}_{X \rightarrow Y} && \text{Lm 5.5}
\end{aligned}$$

□

LEMMA 5.8. *Transitivity.* $\mathcal{H}_{X \rightarrow Y} + \mathcal{H}_{Y \rightarrow Z} \geq \mathcal{H}_{X \rightarrow Z}$.

PROOF.

$$\begin{aligned}
& \mathcal{H}_{X \rightarrow Y} + \mathcal{H}_{Y \rightarrow Z} \\
&\geq \mathcal{H}_{X \rightarrow XY} + \mathcal{H}_{XY \rightarrow XZ} && \text{Lm 5.7} \\
&= \mathcal{H}_{XY} - \mathcal{H}_X + \mathcal{H}_{XYZ} - \mathcal{H}_{XY} && \text{Lm 4.1} \\
&= \mathcal{H}_{XYZ} - \mathcal{H}_X \\
&\geq \mathcal{H}_{XZ} - \mathcal{H}_X && \text{Lm 3.2} \\
&= \mathcal{H}_{X \rightarrow Z} && \text{Lm 4.1}
\end{aligned}$$

□

LEMMA 5.9. *Union (full).* $\mathcal{H}_{X \rightarrow Y} + \mathcal{H}_{W \rightarrow Z} \geq \mathcal{H}_{XW \rightarrow YZ}$

PROOF.

$$\begin{aligned}
& \mathcal{H}_{X \rightarrow Y} + \mathcal{H}_{W \rightarrow Z} \\
&\geq \mathcal{H}_{XW \rightarrow YW} + \mathcal{H}_{WY \rightarrow ZY} && \text{Lm 5.7} \\
&\geq \mathcal{H}_{XW \rightarrow YZ} && \text{Lm 5.8}
\end{aligned}$$

□

LEMMA 5.10. *Decomposition.* if $Z \subseteq Y$, then $\mathcal{H}_{X \rightarrow Y} \geq \mathcal{H}_{X \rightarrow Z}$.

PROOF.

$$\begin{aligned}
& \mathcal{H}_{Y \rightarrow Z} = 0 && \text{Lm 5.1} \\
& \mathcal{H}_{X \rightarrow Y} + \mathcal{H}_{Y \rightarrow Z} \geq \mathcal{H}_{X \rightarrow Z} && \text{Lm 5.8} \\
& \mathcal{H}_{X \rightarrow Y} \geq \mathcal{H}_{X \rightarrow Z}
\end{aligned}$$

□

LEMMA 5.11. *Pseudotransitivity.* $\mathcal{H}_{X \rightarrow Y} + \mathcal{H}_{WY \rightarrow Z} \geq \mathcal{H}_{XW \rightarrow Z}$.

PROOF.

$$\begin{aligned}
& \mathcal{H}_{X \rightarrow Y} + \mathcal{H}_{WY \rightarrow Z} \\
&\geq \mathcal{H}_{XW \rightarrow YW} + \mathcal{H}_{WY \rightarrow Z} && \text{Lm 5.7} \\
&\geq \mathcal{H}_{XW \rightarrow Z} && \text{Lm 5.8}
\end{aligned}$$

□

LEMMA 5.12. *For $XYZ = R$, if $\mathcal{H}_{X \rightarrow Y} + \mathcal{H}_{X \rightarrow Z} = \mathcal{H}_{X \rightarrow YZ}$, then $\mathcal{H}_{WX \rightarrow YV} + \mathcal{H}_{WX \rightarrow Z-V} = \mathcal{H}_{WX \rightarrow YZ}$.*

PROOF. By Lm 5.2 we may assume $w \log V \subseteq W \subseteq Y \cup Z$. Let $\check{Y} = W \cap Y$ and $\check{Z} = W \cap Z$.

$$\begin{aligned}
& \mathcal{H}_{X \rightarrow Y} + \mathcal{H}_{X \rightarrow Z} - \mathcal{H}_{X \rightarrow YZ} \\
&= \mathcal{H}_{XY} - \mathcal{H}_X + \mathcal{H}_{XZ} - \mathcal{H}_X - \mathcal{H}_{XYZ} + \mathcal{H}_X \\
&\geq \mathcal{H}_{XY\check{Y}} - \mathcal{H}_{X\check{Z}} + \mathcal{H}_{XZ\check{Z}} - \mathcal{H}_{XYZ} \\
&\geq \mathcal{H}_{XY\check{Y}\check{Z}} + \mathcal{H}_{XZ\check{Y}\check{Z}} - \mathcal{H}_{X\check{Y}\check{Z}} - \mathcal{H}_{XYZ} \\
&= \mathcal{H}_{XYWV} + \mathcal{H}_{X(Z-V)W} - \mathcal{H}_{XW} - \mathcal{H}_{WYZW} \\
&= \mathcal{H}_{WX \rightarrow YV} + \mathcal{H}_{WX \rightarrow (Z-V)} - \mathcal{H}_{WX \rightarrow YZ}
\end{aligned}$$

□

6. FUNCTIONAL DEPENDENCIES, ARMSTRONG'S AXIOMS, AND MULTIVALUED DEPENDENCIES

6.1 Functional dependencies

Functional dependencies (FDs) are long-known and well-studied [13, 15]. For $X, Y \subseteq R$, X *functionally determines* Y , written $X \rightarrow Y$, if any X value yields a single Y value.

LEMMA 6.1. $X \rightarrow Y$ holds iff $\mathcal{H}_{X \rightarrow Y} = 0$.

PROOF. \Rightarrow : Recasting the FD in terms of probabilities, given any $x_i \in X$, there is a single $y_j \in Y$ such that $p_{i,j} > 0$, and consequently $p_{i,j} = p_i$, and $\mathcal{H}(Y|X = x_i) = 0$ for any x_i . \Leftarrow : Since $\mathcal{H}_{XY} = \mathcal{H}_X$, $\mathcal{H}(Y|x_i) = 0$, for any x_i . Further, $p_i > 0$ for all i . By Remark 1.1, $Y|x_i$ is a singleton; hence, $X \rightarrow Y$. \square

6.2 Armstrong's axioms

Armstrong's axioms [13] are important for functional dependency theory because they provide the basis for a dependency inferencing system. There are commonly three rules given as the Armstrong Axioms, which are merely specializations of the above inequalities.

1. **Reflexivity** If $Y \subseteq X$ then $X \rightarrow Y$
2. **Augmentation** $X \rightarrow Y \Rightarrow XZ \rightarrow YZ$
3. **Transitivity** $X \rightarrow Y \ \& \ Y \rightarrow Z \Rightarrow X \rightarrow Z$

THEOREM 6.1. *The Armstrong axioms follow directly from InD inequalities.*

PROOF. (Augmentation) $0 \leq \mathcal{H}_{XZ \rightarrow YZ}$ by Lm 3.1 and by Lm 5.7 $\mathcal{H}_{XZ \rightarrow YZ} \leq \mathcal{H}_{X \rightarrow Y}$. From the assumption $X \rightarrow Y$ and by Lm 6.1, $\mathcal{H}_{X \rightarrow Y} = 0$; thus, $\mathcal{H}_{XZ \rightarrow YZ} \leq 0$ and by Lm 6.1, $XY \rightarrow YZ$. Reflexivity and transitivity follow similarly using applications of Lm 5.1 and Lm 5.8, respectively. \square

An additional three rules derived from the axioms are often cited as fundamental: union, psuedotransitivity, decomposition. These also follow from Lm 5.3, Lm 5.11, and Lm 5.10 respectively. Interestingly, a critical distinction between Armstrong's axioms and InD inequalities is that in the former, union can be derived from the original three axioms, whereas the latter union must be derived from first principles.

6.3 Multivalued dependencies

In this section, X, Y, Z partition R . Multivalued dependencies (MVDs) arise naturally in database design and are intimately related to the (natural) join operator \bowtie . A multivalued dependency, written $X \twoheadrightarrow Y$ (or $X \twoheadrightarrow Y|Z$ when Z must be made explicit) is defined to hold whenever $\mathbf{r} = \pi_{XY}(\mathbf{r}) \bowtie \pi_{XZ}(\mathbf{r})$.

In almost every case, this definition makes sense only when \mathbf{r} is a set (as opposed to multiset) and that π is set project (with an implicit **distinct** operator). If π is given bag semantics (no duplicate elimination), the number of occurrences of each tuple in the joins the square of the number

of times it occurs in the original \mathbf{r} . Working directly with InD measures allows a more encompassing definition.

Definition 6. A (generalized) multivalued dependency $X \twoheadrightarrow Y$ (or $X \twoheadrightarrow Y|Z$) holds whenever

$$\mathcal{H}_{X \twoheadrightarrow YZ} = \mathcal{H}_{X \twoheadrightarrow Y} + \mathcal{H}_{X \twoheadrightarrow Z}$$

At first glance this appears to introduce two conflicting notations for “ \twoheadrightarrow .” We will see shortly that these are infact the same. Until then, “ \twoheadrightarrow ” should be considered according to the generalized definition of multivalued dependency. Also, we are back to bag (multiset or non-duplicate elimination) semantics for all relational operators except **distinct**.

PROPOSITION 6.1. $X \twoheadrightarrow Y|Z$ holds in \mathbf{r} iff there exists $\mathbf{r}', \mathbf{r}''$, over XY, XZ , respectively such that

$$\mathbf{r} = \mathbf{r}' \bowtie \mathbf{r}''$$

PROOF. \Rightarrow : The instances $\mathbf{r}', \mathbf{r}''$ are constructed in pieces for each separate X value and then unioned together. Hence, in the following we consider just one arbitrary, but fixed, x_i . Also, we use the shorthand notation

$$\#_{j,k} = \text{count}(\sigma_{X=x_i \ \& \ Y=y_j \ \& \ Z=z_k}(\mathbf{r}))$$

and use “ $_$ ” to indicate a “don't care” for j or k . That is, $\#_{j,_} = \sum_k \#_{j,k}$ and similarly for $\#_{_,k}$. Lastly, $\# = \#_{_,_}$

Since $X \twoheadrightarrow Y|Z$, by definition $\mathcal{H}_{X \twoheadrightarrow YZ} = \mathcal{H}_{X \twoheadrightarrow Y} + \mathcal{H}_{X \twoheadrightarrow Z}$, and by Lm 5.3 and Definition 3, $p_{jk|i} = p_{j|i} \cdot p_{k|i}$. Using our shorthand and simplifying we have $\#_{j,k} = (\#_{j,_} \cdot \#_{_,k}) / \#$. Now let

$$\begin{aligned} c' &= \text{gcd}(\#_{1,_}, \dots, \#_{m,_}) \\ c'' &= \text{gcd}(\# / c', \#_{_,1}, \dots, \#_{_,n}) \end{aligned}$$

We need to show that $c'' = \# / c'$. Assume the contrary, i.e. $c'' \cdot f = \# / c'$ for some $f \neq 1$. Form the instance with $\#_{j,_} / c' \times \#_{_,k} / c''$ copies of $\langle x_i, y_j, z_k \rangle$. This has a total size of $f \cdot \#$ and since it has the same distribution, each cell indexed by j, k has $f \cdot \#_{j,k}$ entries. Thus, f divides $\#_{j,_} / c' \times \#_{_,k} / c''$. Since this must hold for all j, k , it must be that either f divides all $\#_{j,_} / c'$ or all $\#_{_,k} / c''$, contrary to the definitions of c' and c'' . Finally, defining \mathbf{r}' to have $\#_{j,_} / c'$ copies of $\langle x_i, y_j \rangle$ and \mathbf{r}'' to have $\#_{_,k} / c''$ copies of $\langle x_i, z_k \rangle$ suffices.

\Rightarrow : A simple consequence of Lm 5.3 and Definition 3. \square

COROLLARY 6.1. *If \mathbf{r} is a set instance and $X \twoheadrightarrow Y|Z$ holds in \mathbf{r} , then $\mathbf{r} = \pi_{XY}(\mathbf{r}) \bowtie \pi_{XZ}(\mathbf{r})$.*

Figure 3 shows an example of a bag instance \mathbf{r} (where X, Y , and Z are single attributes) and its decomposition into

\mathbf{r}		
X	Y	Z
a	b	d
a	b	e
a	c	d
a	c	e

\mathbf{r}'	
X	Y
a	b
a	c

8
4
4
2

4
2

\mathbf{r}''	
X	Z
a	d
a	e

2
1

Figure 3: Example MVD decomposition

\mathbf{r}' and \mathbf{r}'' as provided by **Proposition 6.1**. The number to the right of each tuple indicates its multiplicity in the respective relation instance. Note that another decomposition is obtained by halving the multiplicities for \mathbf{r}' and doubling them for \mathbf{r}'' (the Proposition always maximizes \mathbf{r}' whenever more than one decomposition is possible).

Since acyclic join dependencies can be characterized by a set of MVDs, it is clear that InD inequalities can characterize them as well, though the “work” is really done by the characterization of the set of MVDs.

6.4 Additional InD inference rules

There are three standard rules of MVD inference:

1. **Complementation** If $X \twoheadrightarrow Y$, then $X \twoheadrightarrow (R - XY)$
2. **Augmentation** For $V \subseteq W$, if $X \twoheadrightarrow Y$ then $XW \twoheadrightarrow YV$
3. **Transitivity** If $X \twoheadrightarrow Y$ and $Y \twoheadrightarrow Z$, then $X \twoheadrightarrow (Z - Y)$

Both complementation and augmentation trivially true under InD inequalities. The last rule, transitivity, is rather interesting. For its proof, we find an alternative characterization of MVDs.

LEMMA 6.2. $X \twoheadrightarrow Y$ iff $\mathcal{H}_{X \rightarrow Z} = \mathcal{H}_{XY \rightarrow Z}$

PROOF.

$$\begin{aligned} \mathcal{H}_{X \rightarrow Y} + \mathcal{H}_{X \rightarrow Z} &= \mathcal{H}_{X \rightarrow YZ} && \text{Definition 3} \\ \mathcal{H}_{X \rightarrow Y} + \mathcal{H}_{X \rightarrow Z} &= \mathcal{H}_{X \rightarrow Y} + \mathcal{H}_{XY \rightarrow Z} && \text{Lm 5.4} \\ \mathcal{H}_{X \rightarrow Z} &= \mathcal{H}_{XY \rightarrow Z} \end{aligned}$$

□

Interestingly, this is an alternative characterization of MVDs. In this case, Y does not contribute any information about Z .

LEMMA 6.3. $\mathcal{H}_{X \rightarrow VW} - \mathcal{H}_{XY \rightarrow WV} \geq \mathcal{H}_{X \rightarrow V} - \mathcal{H}_{XY \rightarrow V}$.

PROOF.

$$\begin{aligned} \mathcal{H}_{XV \rightarrow W} &\geq \mathcal{H}_{XYV \rightarrow W} && \text{Lm 5.5} \\ \mathcal{H}_{XV \rightarrow W} - \mathcal{H}_{XYV \rightarrow W} + &&& \\ \mathcal{H}_{X \rightarrow X} + \mathcal{H}_{XY \rightarrow XY} &\geq 0 && \text{Lm 5.1} \\ \mathcal{H}_{VWX} - \mathcal{H}_X - \mathcal{H}_{XYVW} + \mathcal{H}_{XY} - &&& \\ \mathcal{H}_{XV} + \mathcal{H}_X + \mathcal{H}_{XYV} - \mathcal{H}_{XY} &\geq 0 && \text{Lm 4.1} \\ \mathcal{H}_{X \rightarrow VW} - \mathcal{H}_{XY \rightarrow VW} &\geq \mathcal{H}_{X \rightarrow V} - \mathcal{H}_{XY \rightarrow V} && \text{Lm 4.1} \end{aligned}$$

□

LEMMA 6.4. As a consequence of **Lm 6.3**, $\mathcal{H}_{X \rightarrow VW} = \mathcal{H}_{XY \rightarrow WV}$, then $\mathcal{H}_{X \rightarrow V} = \mathcal{H}_{XY \rightarrow V}$.

LEMMA 6.5. If $Y \twoheadrightarrow W|VX$, then $XY \twoheadrightarrow W|V$ by **Lm 5.12**.

LEMMA 6.6. Let $XYWV = R$. If $X \twoheadrightarrow Y|WV$ and $Y \twoheadrightarrow W|XV$, then $\mathcal{H}_{X \rightarrow Y} + \mathcal{H}_{X \rightarrow V} + \mathcal{H}_{X \rightarrow W} = \mathcal{H}_{X \rightarrow R}$.

PROOF.

$$\begin{aligned} \mathcal{H}_{X \rightarrow R} &= \mathcal{H}_{X \rightarrow Y} + \mathcal{H}_{X \rightarrow WV} && \\ &= \mathcal{H}_{X \rightarrow Y} + \mathcal{H}_{XY \rightarrow WV} && \text{Lm 6.4} \\ &= \mathcal{H}_{X \rightarrow Y} + \mathcal{H}_{XY \rightarrow W} + \mathcal{H}_{XY \rightarrow V} && \text{Lm 6.5} \\ &= \mathcal{H}_{X \rightarrow Y} + \mathcal{H}_{X \rightarrow W} + \mathcal{H}_{X \rightarrow V} && \text{Lms 6.4, 6.3} \end{aligned}$$

□

LEMMA 6.7. *Transitivity for MVDs.*

PROOF.

$$\begin{aligned} \mathcal{H}_{X \rightarrow R} &= \mathcal{H}_{X \rightarrow Y} + \mathcal{H}_{X \rightarrow W} + \mathcal{H}_{X \rightarrow V} && \text{Lm 6.6} \\ &\geq \mathcal{H}_{X \rightarrow W} + \mathcal{H}_{X \rightarrow YV} && \text{Lm 5.3} \\ &\geq \mathcal{H}_{X \rightarrow WYV} = \mathcal{H}_{X \rightarrow R} && \text{Lm 5.3} \end{aligned}$$

□

6.5 Rules involving both FDs and MVDs

There are a pair of rules that allow mixing of FDs and MVDs:

1. **Conversion** $X \rightarrow Y \Leftrightarrow X \twoheadrightarrow Y$
2. **Interaction** $X \twoheadrightarrow Y \ \& \ XY \twoheadrightarrow Z \Leftrightarrow X \twoheadrightarrow Z$

The rule for conversion is trivial. Interaction follows from **Lm 6.4**.

In Section 6.2, we stated a critical difference between Armstrong axioms and InD inequalities was the distinction between what were axioms and derivable rules. Additionally, there appear to be other fundamental differences between FDs and MVDs, and InD inequalities. For example, consider the following problem (after [6]). Let R be a schema and $F = \{X \rightarrow Y|X, Y \subseteq R\}$ a set of FDs over R . Let $\mathcal{I}(R, F)$ be the set of all relation instances over R that satisfy F . For $X \subseteq R$, let $\Pi_X(\mathcal{I}(R, F)) = \{\pi_X(\mathbf{r}) | \mathbf{r} \in \mathcal{I}(R, F)\}$. The question is whether there exists a set G of FDs over X such that $\Pi_X(\mathcal{I}(R, F)) = \mathcal{I}(X, G)$. It is known that in general such a G does not exist. Further, a similar negative result holds for MVDs. InD measures are a broader class than FDs and MVDs, and the expectation is that a theorem holds: it does, trivially since all relation instances satisfy any set of InD inequalities.

7. InD MEASURE CONSTRAINTS

To summarize the previous sections, we have defined InD measures on an instance, values that reflect how much information is additionally required about a second set of

attributes given a first set. We have proved a number of arithmetic equalities and inequalities between various InD measures for a given schema; these (in)equalities must hold for any instance of that schema. And we have shown that constraining certain InD measures, or simple expressions involving InD measures, to 0 imposes functional or multi-valued dependences on the instances. We now generalize this last step by considering arbitrary numeric constraints upon InD measures, *e.g.*, $\mathcal{H}_{X \rightarrow Y} \geq 4/9$. A relation instance \mathbf{r} over $R \supseteq \{X, Y\}$ is a solution to this constraint if $\mathcal{H}_{X \rightarrow Y}^{\mathbf{r}} \geq 4/9$ by standard arithmetic. Formally,

Definition 7. An InD constraint system over schema R is an $m \times n$ linear system

$$\begin{aligned} a_{11}\mathcal{H}_{X_1} + a_{12}\mathcal{H}_{X_2} + \dots + a_{1n}\mathcal{H}_{X_n} &\geq b_1 \\ a_{21}\mathcal{H}_{X_1} + a_{22}\mathcal{H}_{X_2} + \dots + a_{2n}\mathcal{H}_{X_n} &\geq b_2 \\ &\vdots \\ a_{m1}\mathcal{H}_{X_1} + a_{m2}\mathcal{H}_{X_2} + \dots + a_{mn}\mathcal{H}_{X_n} &\geq b_m \end{aligned}$$

where $X_i \in 2^R$, $a_{ij}, b_i \in \mathbb{Q}$.

The constraint system is characterized by $\mathbf{A} = [a_{ij}]$, $\mathbf{b} = (b_1, \dots, b_m)$, and $\mathbf{X} = (X_1, \dots, X_n)$ and will be written as $\mathbf{A}\mathbf{H}_{\mathbf{X}} \geq \mathbf{b}$, where $\mathbf{H}_{\mathbf{X}} = (\mathcal{H}_{X_1}, \dots, \mathcal{H}_{X_n})^{\text{Transpose}}$. Observe that Definition 7 is sufficient to describe any InD measure or inequality. InD constraint systems can be as simple as requiring a single FD or as extensive as specifying the entropies of all subsets of R . However, not every \mathbf{A} , \mathbf{b} , and \mathbf{X} make sense as applied to a relation instance. Either the \mathbf{A} and \mathbf{b} may admit no solutions (*e.g.* $\mathcal{H}_X - \mathcal{H}_Y > 5, \mathcal{H}_Y - \mathcal{H}_X > 7$) or the solutions may violate the InD measure constraints for \mathbf{X} (*e.g.* $\mathcal{H}_{X \rightarrow Y} = 3, \mathcal{H}_{Y \rightarrow Z} = 1, \mathcal{H}_{X \rightarrow Z} = 5$ violates **Lm 5.8**).

Definition 8. An InD constraint system $\mathbf{A}, \mathbf{b}, \mathbf{X}$ is *feasible* provided that the linear system \mathbf{A}, \mathbf{b} plus all InD measure constraints inferable from \mathbf{X} is solvable.

Note that this use of “feasible” is borrowed from linear algebra. Also a solution to this extended system involves finding values for each \mathcal{H}_X , $X \subseteq R$ because the constraints involve all such subsets.

7.1 Instances for feasible constraint systems

The question naturally arises whether an instance always exists for a feasible constraint system. The affirmative answer to this question, whose proof is sketched below, provides InD measures with an analog to completeness.

Before venturing into the proof of the theorem itself, we prove a simple result merely for the sake of providing intuition for what comes after. There are two things to be observed while reading the following proof: first, the duality between instance counts and approximate probabilities, and, second, the way interpolation occurs.

LEMMA 7.1. *Given a rational $c \geq 0$, there exists a relation instance \mathbf{r} over a single attribute A such that $|\mathcal{H}_A^{\mathbf{r}} - c| < \epsilon$ for any $0 < \epsilon$.*

PROOF. Let $k = \lfloor \log(c) \rfloor$. Form an instance with $k + 1$ values v_1, \dots, v_{k+1} and associate probabilities with each v_i as follows: $v_{k+1} = x/(k + 1)$ and, for $1 \leq i \leq k$, $v_i = (k + 1 - x)/(k \times (k + 1))$. Thus the instance \mathbf{r} is parameterized by x . For such an instance, $h(x) = \mathcal{H}_A^{\mathbf{r}}$. This h is a continuous function such that $h(0) = 2^k$ and $h(1) = 2^{k+1}$, that is $h(0) \leq c \leq h(1)$. So by the intermediate value theorem [1] there is some x_0 such that $h(x_0) = c$. Truncate the associated probabilities $p_{k+1} = x_0/(k + 1)$ and $p_i = (k + 1 - x_0)/(k \times (k + 1))$ to n decimal places so that $|c - \mathcal{H}_A| \leq \epsilon$. Now form \mathbf{r} with $p_i \times 10^n$ copies of v_i , $1 \leq i \leq k$, and $p_{k+1} \times 10^n$ copies of v_{k+1} . \square

While this proof is non-constructive, we can find a suitable x by, for example, binary search.

THEOREM 7.1. *Instance existence. For any feasible constraint system \mathbf{A}, \mathbf{b} , and \mathbf{X} , and any $\epsilon > 0$, there is a relation instance \mathbf{r} that satisfies \mathbf{A}, \mathbf{b} , and \mathbf{X} within ϵ .*

PROOF. 1. Using the observation from Definition 7.2, solve \mathbf{A}, \mathbf{b} , and \mathbf{X} for fixed values for \mathcal{H}_{A_1}, \dots

2. Pick $m > 1/\epsilon$

3. Give every attribute a value with large probability, namely $1 - (1/2m)^k$, where k is the number of attributes. Note that these highly probable attributes contribute a negligible amount to any entropy since their probabilities are so close to 1.

4. The remaining probabilities for each attribute A_i will be divided among b_i equal size buckets. Thus, $\mathcal{H}_{A_i} = (1/m^k)(\log(1/(m^k \times b_i))) \simeq 1/m^k \times \log b_i$. Find b_i such that $\alpha_i - 1/m^k \log(b_i) < \epsilon$

Remark 2. Wlog, the A_i are ordered in decreasing entropy. Hence $b_i \geq b_{i+1}$.

We will add attributes in order A_1, A_2, \dots ,

5. At stage $i + 1$, construction has included A_1, \dots, A_i , and we are adding A_{i+1} ; that is, we already have p_{j_1, \dots, j_i} and want to construct $p_{j_1, \dots, j_{i+1}}$. We also have a single distribution q corresponding to A_{i+1} . We actually construct two distributions p^ℓ and p^u , for “p lower” and “p upper”.

(a) The upper case is simple: A_{i+1} is independent from A_1, \dots, A_i : $p_{j_1, \dots, j_i, j_{i+1}}^u = p_{j_1, \dots, j_i} \times q_{j_{i+1}}$

(b) The lower case is found by allocating the q_j among the various p 's. Because $b_i \geq b_{i+1}$, there are more than enough i buckets to go around. With some small error, each non-zero p will correspond to a unique $q \neq 0$.

The error $\mathcal{H}_{A_i A_{i+1}}^{p^\ell} - \mathcal{H}_{A_i}^{p^\ell} < \epsilon^k$ and by induction $\mathcal{H}_{A_n A_{i+1}}^{p^\ell} - \mathcal{H}_{A_n}^{p^\ell} < \epsilon^{k-m}$, for $1 \leq m \leq i$. Interpolate between p^ℓ and p^u to match other entropies

□

This is conceptually similar to **Lm** 7.1, but relies upon the unusual structure of pu caused by the almost-unity cases of p and q and another iteration.

8. APPLICATIONS AND EXTENSIONS

We have presented a formal foundation incorporating information theory in relational databases. There are many interesting and valuable applications and extensions of this work that we are already pursuing.

8.1 Computation of InD Measures

The ability to compute InD measurements is obviously critical to the application of those measures. Because counting is central to our interpretation of \mathcal{H} and because that counting takes place over the lattice of subsets of R , the well-known datacube operation [7], which performs counts over subsets of R , is the obvious first choice for an implementation. Anything that people have thought about with respect to datacube should be thought about with respect to InDs.

Application of datacube enhancements has its challenges. For example, one promising approach [2] prunes the computation when the size of a partition falls below a given threshold. However, the entropy measure depends upon the contribution of lots of small pieces. Hence the ability to prune depends upon a more sophisticated combination of InD values; the inequalities that we have already established should provide a substantial step in this direction.

Another approach to computing \mathcal{H} is to use a hash-based counting technique that we have developed [11]. This technique is designed to efficiently evaluate counts of any arbitrary subset of the attributes. It also lends itself to efficient incremental updates.

8.2 Datamining

Datamining [5], the search for interesting patterns in large databases, motivated our initial work, our goal was clarifying what it means to be “interesting.” One candidate for interesting patterns is certainly the ϵ -approximate functional dependencies, or ϵ -FDs—that is cases where $\mathcal{H}_{X \rightarrow Y} \leq \epsilon$. The search for ϵ -FDs in \mathbf{r} takes place upon the lattice of $\langle 2^R, \subseteq \rangle$, where $\mathcal{H}_{X \rightarrow Y} \leq \epsilon$ is checked for every $X \subsetneq Y$. The InD inequalities facilitate this search.

Kivinen *et al.* [8], considers finding approximate FDs. The central notion is that of *violating pair*; for an instance \mathbf{r} over R and $X, Y \subseteq R$, a pair of tuples $s, t \in \mathbf{r}$ *violates* $X \rightarrow Y$ if $s.X = t.X \Rightarrow s.Y \neq t.Y$. They define three normalized measures g_1, g_2, g_3 are based upon the number of violating pairs, the number of violating tuples, and the number of violating tuples removed to achieve a dependency, respectively. The authors state that problematically the measures give very different values for some particular relations, and therefore, choosing which measure is the best—if any are—is difficult. We feel that the InD measure can shed some light upon the metrics. The connection between these measures and InD measures is illustrated with three instances $\mathbf{r} = \{\langle a, 1 \rangle, \langle a, 2 \rangle, \langle b, 1 \rangle, \langle c, 1 \rangle, \langle c, 2 \rangle\}$, $\mathbf{s} = \mathbf{r} - \{\langle c, 2 \rangle\} \cup \{\langle a, 3 \rangle\}$ and $\mathbf{t} = \mathbf{s} \cup \{\langle a, 4 \rangle, \langle a, 5 \rangle, \langle a, 6 \rangle, \langle d, 1 \rangle, \langle d, 1 \rangle\}$

	\mathcal{H}_X	$\mathcal{H}_{X \rightarrow Y}$	g_1	g_2	g_3
\mathbf{r}	1.52	.80	.16	.8	.4
\mathbf{s}	1.37	.95	.36	.8	.4
\mathbf{t}	1.77	1.55	.36	.8	.4

This example shows that $\mathcal{H}_{X \rightarrow Y}$ can sometimes make finer distinctions than the g_i s. On the applications side, Kivinen *et al* have done substantial work related to approximate FDs as in [8]. The paper is important not only for the notion of approximate dependency, but also a brief discussion about how the errors can be cast into Armstrong Axiom-like inequalities.

8.3 Other Metrics

Rather than considering what information X lacks about Y , we may look at the information X contains about Y , that is $\hat{\mathcal{I}}_{X \rightarrow Y} = \mathcal{H}_Y - \mathcal{H}_{X \rightarrow Y}$ and its normalized form $\mathcal{I}_{X \rightarrow Y} = \hat{\mathcal{I}}/\mathcal{H}_Y$. Some interesting results about \mathcal{I} and $\hat{\mathcal{I}}$ are $\max(\mathcal{I}_{X \rightarrow Y}, \mathcal{I}_{X \rightarrow Z}) \geq \mathcal{I}_{X \rightarrow YZ} \geq \min(\mathcal{I}_{X \rightarrow Y}, \mathcal{I}_{X \rightarrow Z})$; $0 \leq \mathcal{I}_{X \rightarrow Y} \leq 1$; $\hat{\mathcal{I}}_{X \rightarrow Y} = \hat{\mathcal{I}}_{Y \rightarrow X}$. While \mathcal{I} makes the specification of FDs more natural ($X \rightarrow Y$ iff $\mathcal{I}_{X \rightarrow Y} = 1$), it cannot be used to characterize MVDs. Another interesting measure that uses additional notions from information theory is *rate* of the language $s = \mathcal{H}_X^s / \text{count}(\mathbf{r})$ which is the average number of bits required for each tuple projected on X . The absolute rate is $s_{ab} = \log(\text{count}(\mathbf{r}))$. The difference $s_{ab} - s$ indicates the redundancy. As X approaches R , the average tuple entropy increases, reducing redundancy. This is pertinent especially to the following section.

8.4 Connections to relational algebra

Examining how InDs behave with relational operators. For example,

LEMMA 8.1. *Let $R = \{X, Y, Z\}$ and \mathbf{r} be an instance of R . if $\mathbf{r}' = \pi_{XY}(\mathbf{r}) \bowtie \pi_{XZ}(\mathbf{r})$, then $\mathcal{H}_{YZ}^{\mathbf{r}'} = \mathcal{H}_Y^{\mathbf{r}'} + \mathcal{H}_Z^{\mathbf{r}'}$.*

For instance, when employing a lossless decomposition, how will both the InD measures and rates (from above) change to indicate the decomposition was indeed lossless.

9. RELATED WORK

There is a dearth of literature in this area, marrying information theory to information systems. The closest work seems to be Piatetsky-Shapiro in [4] who proposes a generalization of functional dependencies, called *probabilistic dependency* ($pdep$). The author begins with the $pdep1(X) = \sum_i p_i^2$ (using our notation). To relate two sets of attributes X, Y , $pdep(X, Y) = \sum_i p_i \sum_j p_j^2$. Observe that $pdep$ approaches 1 as X comes closer to functionally determining Y . Since $pdep$ is itself inadequate, the author normalizes it using proportion in variation, resulting in the known statistical measure $\tau(X, Y) = (pdep(X, Y) - pdep1(Y)) / (1 - pdep1(Y))$. If $\tau(X, Y) > \tau(Y, X)$, then $X \rightarrow Y$ is a better FD than $Y \rightarrow X$ (and vice versa). The author describes the expectation of both $pdep$ efficiently sample for these values.

In the area of artificial intelligence, an algorithm developed to create decision trees, a means of classification, by Quinlan, notably ID3 [9] and C4.5 [10] uses entropy to dictate

how the building should proceed. In this case of supervised learning, an attribute A is selected as the target, and the remaining attributes $R - \{A\}$ the classifier. The algorithm works by progressively selecting attributes from the initial set $R - \{A\}$, measuring be classified properly.

The rare uses of information theory in databases include recovering base sequence data when only sums of sequence intervals are available[3]. In this case, the sequence is comparable to a message source.

10. ACKNOWLEDGEMENTS

The authors would like to thank the readers for their instructive and helpful comments. Additionally the authors would like to thank Dennis Groth, Dirk Van Gucht, Chris Giannella, Richard Martin, and others.

11. REFERENCES

- [1] R. G. Bartle. *The Elements of Real Analysis Second Edition*. John Wiley & Sons, Inc., New York, New York, 1976.
- [2] K. Beyer and R. Ramakrishnan. Bottom-up computation of sparse and iceberg CUBEs. In *Proceedings of the Eighteenth ACM SIGMOD Conference*. ACM, May 1999.
- [3] C. Faloutsos, H. V. Jagadish, and N. D. Sidiropoulos. Recovering information from summary data. In *Proceedings of the 23rd VDLB Conference*, 1997.
- [4] G. Piatetsky-Shapiro. Probabilistic data dependencies. In *Machine Discovery Workshop (Aberdeen, Scotland)*, 1992.
- [5] G. Piatetsky-Shapiro, U. Fayyad, and P. Smith, editors. *From data mining to knowledge discovery: An overview*. AAAI/MIT Press, 1996.
- [6] S. Ginsburg and R. Hull. Characterizations for functional dependency and boyce-codd normal form families. In *Theoretical Computer Science*, volume 26, pages 243–286, 1983.
- [7] J. Gray, S. Chaudhuri, A. Bosworth, A. Layman, D. Reichart, and M. Venkatrao. Data cube: A relational aggregation operator generalizing gro up-by, cross-tab and sub-totals. In *Data Mining and Knowledge Discovery*, volume 1, 1997.
- [8] Jyrki Kivinen and Heikki Mannila. Approximate inference of functional dependencies from relations. *Theoretical Computer Science*, 149:129–149, 1995.
- [9] J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, 1986.
- [10] J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufman, San Mateo, CA, 1993.
- [11] S. Rao, A. Badia, and D. Van Gucht. Providing better support for quantified query processing. In *Proceedings of the Fifteenth ACM SIGMOD Conference*. ACM, 1996.
- [12] S. Roman. *Coding and Information Theory*. Springer-Verlag, New York, New York, 1992.
- [13] Serge Abiteboul, Richard Hull, and Victor Vianu. *Foundations of Databases*. Addison-Wesley Publishing Company, New York, New York, 1995.
- [14] Thomas Cover and Joy Thomas. *Elements of Information Theory*. John Wiley & Sons, Inc., New York, New York, 1991.
- [15] J. D. Ullman. *Principles of Database and Knowledge-Base Systems Vol. 1*. Computer Science Press, Rockville, Maryland, 1988.