On an Information Theoretic Approximation Measure for Functional Dependencies

Chris Giannella and Edward Robertson*

Computer Science Department, Indiana University, Bloomington, IN 47405, USA
{cgiannel,edrbtsn}@cs.indiana.edu

Abstract. We investigate the problem of defining an approximation measure for functional dependencies (FDs). For fixed sets of attributes, X and Y, an approximation measure is a function which maps relation instances to real numbers. The number to which an instance is mapped, intuitively, describes the strength of the dependency, $X \rightarrow Y$, in that instance. We define an approximation measure for FDs based on a connection between Shannon's information theory and relational database theory. Our measure is normalized to lie between zero and one (inclusive), and maps a relation instance to zero if and only if $X \rightarrow Y$ holds in the instance. Hence, the smaller the number to which an instance is mapped, the "closer" $X \rightarrow Y$ is to being an FD in the instance.
To put our measure in context, we compare it to a slight variation of a measure previously defined by Kivinen and Mannila, g_3. We denote the variation as $\hat{g_3}$, although, our results, essentially, apply unchanged to g_3. The purpose of comparing our measure with $\hat{g_3}$ is to develop a deeper understanding of not only our measure, but also, $\hat{g_3}$. Moreover, we gain a deeper understanding of the natural intuitive notion of an approximate FD. We observe that our measure and $\hat{g_3}$ agree at their extremes but are quite different in-between. As a result, we conclude that our measure and $\hat{g_3}$ are significantly different. An interesting question emerges from this conclusion: is there a rigorous way to determine when one measure better captures the meaning of the degree to which an FD is approximate?

1 Introduction

Over approximately the last ten years, a new research direction has emerged involving functional dependencies (FDs). Researchers have been addressing the problem of finding all of the FDs which hold in a given relation instance ([4], [5], [6], [7], [9], [10], [12], [15]). We call this *FD discovery research*. The primary motivation for FD discovery research is different than that for the original FD research in the 70s. The research in the 70s was primarily motivated by database design (*e.g.* schema normal forms). The primary motivation for FD discovery research is not database design. Instead, it is knowledge discovery. FDs represent valuable knowledge of the "structure" of the relation instance.

* Both authors were supported by National Science Foundation grant 82407.

In some cases, an FD may not hold because of a few tuples. This FD can be thought to *approximately* hold. For example ([4]), `first name` $\rightarrow$ `gender` may approximately hold. Approximate functional dependencies (AFDs) also represent valuable knowledge of the structure of the relation instance. The discovery of such knowledge can be valuable to domain experts (whose data is contained in the database). Paraphrasing from [4] page 100: an AFD in a database of chemical compounds relating various structural attributes to carcinogenicity could provide valuable hints to biochemists for potential causes of cancer (but cannot be taken as a fact without further analysis by domain specialists).

AFD discovery research consists of three primary parts: (1) defining an approximation measure for AFDs, (2) developing methods for applying AFDs to pre-existing problems, (3) developing algorithms for efficiently computing AFDs. Huhtala *et al.* [4] address (3) by developing an algorithm, *TANE*, for discovering all AFDs which hold in a relation instance. TANE uses an approximation measure, g_3, proposed in [6] to define when an AFD is deemed to hold. In this paper we consider a slight variant of g_3 defined as the minimum number of tuples that need be removed for the FD to hold divided by the size of the relation instance *minus one*. In [4] and [6], g_3 is defined as the minimum number of tuples divided by the size of the relation (*i.e.* the "minus one" is dropped). We change the denominator of g_3, so that the range of our measure includes one. This change makes essentially no difference in our analysis, but, for the sake of clarity, we refer to the changed measure as $\hat{g_3}$.

We are interested in addressing area (1). To do so, we define an approximation measure for AFDs based on a connection between information theory and "classical" relational database theory [1], [2], [11]. This measure, intuitively, is the amount of "information" the left-hand side of the AFD contains about the right-hand side normalized to lie between zero and one (inclusive). The amount of information is quantified in terms of *information dependencies* [2]. We then compare this information theoretic measure with $\hat{g_3}$.

2 Purpose and Primary Contributions

The primary purpose of this paper is to investigate the use of information dependencies as a means for defining an approximation measure for AFDs. In doing so, we gain a deeper understanding of the natural intuitive notion of an AFD. To achieve this purpose we compare our information theoretic measure with $\hat{g_3}$. Doing so not only puts our new measure in a previously established context, but, develops a deeper understanding of not only our new measure, but also, $\hat{g_3}$.

Our primary contributions are the following: (i) a new approximation measure for AFDs based on information theory, (ii) a rigorous comparison between the measure and $\hat{g_3}$. We observe that at the extremes (zero and one) the measures correspond, but, in-between they are quite different. Moreover, in the limit, the measures do not correspond at the extremes (*i.e.* one measure may approach an extreme while the other does not).

Our information theoretic measure and $\hat{g_3}$ both can be used as an approximation measure for AFDs. We have shown that these two measures behave quite differently. An interesting question emerges: is there a rigorous way to determine when one measure better captures the meaning of the degree to which an FD is approximate?

2.1 Related Work

Cavallo and Pittarelli [1] propose an information theoretic measure for AFDs (similar ideas were described previously by Malvestuto [8] and Nambiar [11] but a measure was not proposed). Their measure (and ours) is based on the concept of an information dependency. However, their measure is normalized differently than ours. This difference causes their measure to behave quite differently than our measure and $\hat{g_3}$. These differences are described in section 5.

Piatetski-Shapiro defines *probabilistic data dependencies* in [13]. Based on probabilistic data dependencies he goes on to define a normalized measure of association which corresponds to the τ association measure proposed previously by Goodman and Kruskal [3]. τ can be used to define an approximation measure for AFDs (although not discussed in [3] or [13]). This measure behaves quite differently than our information theoretic measure and $\hat{g_3}$. These differences are described in section 5.

Three different approximation measures for AFDs are defined in [6] (and called *error measures*). The definitions of these three measures are based directly on the definition of an FD holding in a relation instance. The third of these measures is g_3. In [6] it is shown that these three measures give very different outputs for some relations. The authors conclude that it is not clear which, if any, of these measures is the most natural measure of the degree to which an FD is approximate. Our measure is defined on fundamentally different principals than the three measures of [6]. Hence a rigorous comparison of its behavior and $\hat{g_3}$ is valuable.

2.2 Paper Layout

In section 3, two approximation measures for AFDs are defined. The first is based solely in standard relational database theory and the second is based on a connection between information theory and relational database theory. In section 4, the two measures are compared. Part of the comparison involves studying the limiting behavior of the measures. In section 5, two additional measures from the literature are discussed. Comparisons are made between these measures and measures defined in section 3. Finally, in section 6, future work is described.

3 Two Approximation Measures for AFDs

In this section we define two approximation measures for AFDs. The first is based on standard relational database theory. The second is based on a connection between information theory and relational database theory, namely, *Information*

Dependencies [2]. We assume that the reader is familiar with the standard definitions from relational database theory and do not state them here (see [14]).

Let $\mathcal{S}$ be a relation schema, X and Y be non-empty, disjoint subsets of $\mathcal{S}$ and r be an instance over $\mathcal{S}$. There are several ways of defining an AFD approximation measure. We use a slight variant of the approximation measure g_3 in [4] and [6].

Definition 1

$$\hat{g_3}_{X \to Y}(r) = \frac{min\{|s| : s \subseteq r, \forall t_1, t_2 \in r - s, t_1[X] = t_2[X] \Rightarrow t_1[Y] = t_2[Y]\}}{|r| - 1}.$$

$\hat{g_3}$ is normalized to lie between zero and one (inclusive), and equals zero if and only if $X \to Y$ holds. Take note that, $\hat{g_3}_{X \to Y}(r)$ can be computed in time $O(|r|lg(|r|))$. $\hat{g_3}$ differs from g_3 only in that the denominator of g_3 is $|r|$. This difference is insignificant in our analysis to follow.

A	B	C
1	3	1
2	2	0
3	1	1
1	3	0
2	2	1
3	1	0

Fig. 1.

To define the approximation measure based on information dependencies, some basic definitions from [2] are needed. They are included here for the sake of being self-contained. The relational algebra operators Π and σ are assumed to be duplicate removing (*i.e.* they return sets).

Let $\Pi_X(r) = \{x_1, \ldots, x_p\}$ and $c(x_i) = |\sigma_{X=x_i}(r)|$ for $1 \leq i \leq p$ (clearly $\sum_{i=1}^{p} c(x_i) = |r|$). Likewise, let $\Pi_Y(r) = \{y_1, \ldots, y_q\}$ and $c(y_j) = |\sigma_{Y=y_j}(r)|$ for $1 \leq j \leq q$. Finally, let $c(x_i, y_j) = |\sigma_{X=x_i, Y=y_j}(r)|$ for $1 \leq i \leq p, 1 \leq j \leq q$.

Definition 2 *Entropy ([2]).*

1. *The* entropy of Y over r, *written* $\mathcal{H}_Y(r)$, *is*[1] $\sum_{j=1}^{q} \frac{c(y_j)}{|r|} lg(\frac{|r|}{c(y_j)})$.
2. *The* entropy of Y over r given $X = x_i$, *written* $\mathcal{H}_{Y|X=x_i}(r)$, *is*[2] $\sum_{j=1}^{q} \frac{c(x_i,y_j)}{c(x_i)} lg(\frac{c(x_i)}{c(x_i,y_j)})$.
3. *The* information dependency (InD) *measure of* Y *given* X, *written* $\mathcal{H}_{X \to Y}(r)$, *is* $\sum_{i=1}^{p} \frac{c(x_i)}{|r|} \mathcal{H}_{Y|X=x_i}(r)$.

Example 1. Let s be the relation instance over schema A, B, C seen in Fig. 1. Careful inspection shows that $\mathcal{H}_A(s) = \frac{2}{6}lg(\frac{6}{2}) + \frac{2}{6}lg(\frac{6}{2}) + \frac{2}{6}lg(\frac{6}{2}) = lg(3)$ and that $\mathcal{H}_{A \to B}(s) = \frac{1}{3}\mathcal{H}_{B|A=1}(s) + \frac{1}{3}\mathcal{H}_{B|A=2}(s) + \frac{1}{3}\mathcal{H}_{B|A=3}(s) = \frac{1}{3}0 + \frac{1}{3}0 + \frac{1}{3}0 = 0$.

The following facts about entropy and the InD measure will be quite useful.

Fact 1 *1.* **Alternate characterization of the InD measure:** $\mathcal{H}_{X \to Y}(r) = \mathcal{H}_{X \cup Y}(r) - \mathcal{H}_X(r)$.
2. $0 \leq \mathcal{H}_X(r) \leq lg(|r|)$.

[1] "lg" denotes the logarithm base 2.

[2] $c(x_i, y_j)$ may be zero. In this event we define $\frac{c(x_i,y_j)}{c(x_i)} lg(\frac{c(x_i)}{c(x_i,y_j)})$ to be zero.

3. $|\Pi_X(r)| = 1$ *if and only if* $\mathcal{H}_X(r) = 0$.
4. $|\Pi_X(r)| = |r|$ *if and only if* $\mathcal{H}_X(r) = lg(|r|)$.
5. *If* $|\Pi_X(r)| = 1$, *then* $\mathcal{H}_{X \cup Y}(r) = \mathcal{H}_Y(r)$.

Proof: 1., 2. See [2]. For the rest of the proof recall the definition $\mathcal{H}_X(r) = \sum_{i=1}^{p} \frac{c(x_i)}{|r|} lg(\frac{|r|}{c(x_i)})$.

3. If $|\Pi_X(r)| = 1$, then $p = 1$ and $c(x_1) = |r|$, so, $\mathcal{H}_X(r) = \frac{|r|}{|r|} lg(\frac{|r|}{|r|}) = 0$. On the other hand, assume $\mathcal{H}_X(r) = 0$. Since $c(x_i) \geq 1$ for all $1 \leq i \leq p$, then it follows that $\frac{c(x_i)}{|r|} lg(\frac{|r|}{c(x_i)}) = 0$. Hence $c(x_i) = |r|$ for all i, therefore $|\Pi_X(r)| = 1$ (because $\sum_{i=1}^{p} c(x_i) = |r|$).

4. Since $c(x_i) \geq 1$ for all $1 \leq i \leq p$, it is straight-forward to show that $\frac{c(x_i)}{|r|} lg(\frac{|r|}{c(x_i)}) = \frac{lg(|r|)}{|r|}$ for all i if and only if $|\Pi_X(r)| = |r|$. Hence $\mathcal{H}_X(r) = \sum_{i=1}^{|r|} \frac{c(x_i)}{|r|} lg(\frac{|r|}{c(x_i)}) = lg(|r|)$ if and only if $|\Pi_X(r)| = |r|$.

5. Assume $|\Pi_X(r)| = 1$. Then $p = 1$ and $c(x_1, y_j) = c(y_j)$ for all $1 \leq j \leq q$. Hence, $\mathcal{H}_{X \cup Y}(r) = \sum_{i=1}^{p} \sum_{j=1}^{q} \frac{c(x_i, y_j)}{|r|} lg(\frac{|r|}{c(x_i, y_j)}) = \sum_{j=1}^{q} \frac{c(x_1, y_j)}{|r|} lg(\frac{|r|}{c(x_1, y_j)}) = \sum_{j=1}^{q} \frac{c(y_j)}{|r|} lg(\frac{|r|}{c(y_j)}) = \mathcal{H}_Y(r)$. [3] □

Part 1. of the following theorem forms the basis of our information dependency based AFD approximation measure. Part 1. also completely characterizes the class of relation instances which minimize the InD measure. Part 2., completely characterizes the class which maximize the InD measure.

Theorem 1 *1.* $0 \leq \mathcal{H}_{X \rightarrow Y}(r) \leq lg(|r|)$ *and the lower bound obtains if and only if* $X \rightarrow Y$ *holds over* r.
2. $|\Pi_X(r)| = 1$ *and* $|\Pi_Y(r)| = |r|$ *if and only if* $\mathcal{H}_{X \rightarrow Y}(r) = lg(|r|)$.

Proof: 1. See [2]. Now consider 2.. By Fact 1 parts 1. and 2. it follows that, $\mathcal{H}_{X \rightarrow Y}(r) = lg(|r|)$ if and only if $\mathcal{H}_{X \cup Y}(r) = lg(|r|)$ and $\mathcal{H}_X(r) = 0$. By parts 3. and 4. it follows that $\mathcal{H}_{X \cup Y}(r) = lg(|r|)$ and $\mathcal{H}_X(r) = 0$ if and only if $|\Pi_{X \cup Y}(r)| = |r|$ and $|\Pi_X(r)| = 1$. But this can easily be seen to hold if and only if $|\Pi_Y(r)| = |r|$ and $|\Pi_X(r)| = 1$. □

We define an approximation measure for AFDs as follows.

Definition 3 *Information theoretic AFD approximation measure.*

$$IA_{X \rightarrow Y}(r) = \frac{\mathcal{H}_{X \cup Y}(r) - \mathcal{H}_X(r)}{lg(|r|)}.$$

$IA_{X \rightarrow Y}$ is normalized to lie between zero and one (inclusive) and equals zero if and only if $X \rightarrow Y$ holds. Note that $IA_{X \rightarrow Y}(r)$ can be computed in time $O(|r| lg(|r|))$.

[3] If $c(x_i, y_j)$ equals zero, then we define $\frac{c(x_i, y_j)}{|r|} lg(\frac{|r|}{c(x_i, y_j)})$ to be zero.

4 Comparison of the Approximation Measures

In this section we compare the approximation measure, $\hat{g_3}$, and the InD based approximation measure, IA. At their endpoints, these measures correspond completely.

Theorem 2 *$\hat{g_3}_{X\to Y}(r) = 0$ or 1 if and only if $IA_{X\to Y}(r) = 0$ or 1, respectively.*

Proof: By definition, $\hat{g_3}_{X\to Y}(r) = 0$ if and only if the $X \to Y$ holds over r. So, the desired result for zero follows from Theorem 1 part 1.. Also, by definition, $\hat{g_3}_{X\to Y}(r) = 1$ if and only if $|\Pi_Y(r)| = |r|$ and $|\Pi_X(r)| = 1$. So, the desired result for one follows from Theorem 1 part 2.. □

In-between the endpoints the classical and InD measures do not correspond very well. The next theorem demonstrates just how badly they correspond.

Theorem 3 *For any rational number $\frac{a}{b} \in [0,1)$ and any positive multiple, m, of b, there exists s with m tuples such that $\hat{g_3}_{X\to Y}(s) = \frac{\frac{a}{b}m}{m-1}$ and $IA_{X\to Y}(s) = \frac{\frac{a}{b}lg(b)-(1-\frac{a}{b})lg(1-\frac{a}{b})}{lg(m)}$.*

Proof: Let s be as seen in Fig. 2 (schema A, B, C and m rows). Note that if $\frac{a}{b} = 0$, then s consists entirely of only those rows in the above table with "1" in their B column. Let $X = \{A\}$ and $Y = \{B\}$. Consider $\hat{g_3}_{A\to B}(s)$. Since $\frac{a}{b}$ is a non-negative, rational number less than one, then $a + 1 \leq b$. Hence $(1 - \frac{a}{b}) \geq \frac{1}{b}$. So, the B value which occurs in a maximal number of tuples is 1 which occurs in $(1 - \frac{a}{b})m$ tuples. Thus, $\hat{g_3}_{A\to B}(s) = \frac{m-(1-\frac{a}{b})m}{m-1} = \frac{\frac{a}{b}m}{m-1}$, as needed.

Now consider $IA_{A\to B}(s)$. Since $|\Pi_A(s)| = 1$, then by Fact 1 parts 3. and 5., it follows that $IA_{A\to B}(s) = \frac{\mathcal{H}_Y(s)}{lg(m)}$. By observation it can be shown that $\mathcal{H}_Y(s) = -(1 - \frac{a}{b})lg(1 - \frac{a}{b}) + \frac{a}{b}lg(b)$. Hence $IA_{A\to B}(s) = \frac{\frac{a}{b}lg(b)-(1-\frac{a}{b})lg(1-\frac{a}{b})}{lg(m)}$, as needed. □

To better understand Theorem 3, let $\frac{a}{b} = \frac{1}{2}$. s, in this case, is of the form seen in Fig. 3. We have that $\hat{g_3}_{A\to B}(s) = \frac{\frac{1}{2}m}{m-1}$ and $IA_{A\to B}(s) = \frac{1}{lg(m)}$. In this example, for all but small m, the measures are quite different. In fact, it will be shown later, that, for large m, the instance s witnesses the maximum difference between the measures and this difference is $\hat{g_3}_{A\to B}(s)$. We examine the limiting behavior of the measures in the next subsection.

A	B	C
1	1	1
1	1	2
⋮	⋮	⋮
1	1	$(1-\frac{a}{b})m$
1	2	1
1	2	2
⋮	⋮	⋮
1	2	$\frac{m}{b}$
	⋮	
	⋮	
1	$a+1$	1
1	$a+1$	2
⋮	⋮	⋮
1	$a+1$	$\frac{m}{b}$

Fig. 2.

4.1 Limiting AFD Measure Behavior

Let (r_n) be a sequence of relation instances over schema $\mathcal{S}$ where for all n, $|r_n| < |r_{n+1}|$. Let us call sequences of this form, *growing sequences*. We compare $L_{\hat{g}_3} = lim_{n\to\infty}\hat{g}_{3\,X\to Y}(r_n)$ and $L_{IA} = lim_{n\to\infty}\ IA_{X\to Y}(r_n)$ under the assumption that both these limits exist. Clearly $-1 \leq L_{\hat{g}_3} - L_{IA} \leq 1$. We address the question: how large or small can $L_{\hat{g}_3} - L_{IA}$ be? Also, we address the following questions for $k = 0, 1$. If $L_{\hat{g}_3} = k$ does $L_{IA} = k$? If $L_{IA} = k$ does $L_{\hat{g}_3} = k$?

To address the question of how large or small $L_{\hat{g}_3} - L_{IA}$ can be, we show that: $L_{\hat{g}_3} - L_{IA} \geq 0$, and for any rational number $\frac{a}{b} \in [0, 1)$, $L_{\hat{g}_3} - L_{IA}$ can be made as large as $L_{\hat{g}_3}$ where $L_{\hat{g}_3} = \frac{a}{b}$. Since $L_{IA} \geq 0$, then $L_{\hat{g}_3} - L_{IA}$ is bounded above by $L_{\hat{g}_3}$. Our result shows that this bound obtains and the difference can be made as arbitrarily large as possible (i.e. arbitrarily close to one).

A	B	C
1	1	1
1	1	2
⋮	⋮	⋮
1	1	$(1-\frac{1}{2})m$
1	2	1
1	2	2
⋮	⋮	⋮
1	2	$\frac{m}{2}$

Fig. 3.

Theorem 4 *Assume both $L_{\hat{g}_3}$ and L_{IA} exist. Then $L_{\hat{g}_3} \geq L_{IA}$, and, for any rational number $\frac{a}{b} \in [0, 1)$, there exists a growing sequence (s_n), such that $L_{\hat{g}_3} - L_{IA} = L_{\hat{g}_3} = \frac{a}{b}$.*

Proof: From the proof of Theorem 3, we see how to construct a growing sequence (s_n) such that $L_{\hat{g}_3} = lim_{n\to\infty}\frac{\frac{a}{b}|s_n|}{|s_n|-1} = \frac{a}{b}$ and $L_{IA} = lim_{n\to\infty}\frac{\frac{a}{b}lg(b)-(1-\frac{a}{b})lg(1-\frac{a}{b})}{lg(|s_n|)} = 0$. To complete the proof it suffices to show that $L_{\hat{g}_3} \geq L_{IA}$ (for (r_n)).

Let r_n be any relation in the sequence (r_n) where $|r_n| \geq 4$. Let D_n denote $min\{\hat{g}_{3\,X\to Y}(r_n) - IA_{X\to Y}(r_n), 0\}$. It suffices to show that $lim_{n\to\infty}(D_n) = 0$.

Let $\Pi_X(r_n) = \{x_1, \ldots, x_p\}$ and $\Pi_Y(r_n) = \{y_1, \ldots, y_q\}$. For each i, let $m_i = max\{c(x_i, y_1), \ldots, c(x_i, y_q)\}$ and let $q_i = |\{c(x_i, y_j) \geq 1 : 1 \leq j \leq q\}|$. Assume without loss of generality that for all $1 \leq i \leq p$, we have that $m_i = c(x_i, y_1) \geq c(x_i, y_2) \geq \cdots \geq c(x_i, y_q)$ (this implies that: $c(x_i, y_{q_i+1}) = \cdots = c(x_1, y_q) = 0$). By definition and Fact 1 part 1., we have that

$$
\begin{aligned}
IA_{X\to Y}(r_n) &= \frac{\sum_{i=1}^{p}\sum_{j=1}^{q}\frac{c(x_i,y_j)}{|r_n|}lg(\frac{|r_n|}{c(x_i,y_j)}) - \sum_{i=1}^{p}\frac{c(x_i)}{|r_n|}lg(\frac{|r_n|}{c(x_i)})}{lg(|r_n|)} \\
&= \frac{\sum_{i=1}^{p}[c(x_i)lg(c(x_i)) - \sum_{j=1}^{q_i}c(x_i,y_j)lg(c(x_i,y_j))]}{|r_n|lg(|r_n|)} \\
&\leq \frac{\sum_{i=1}^{p}[c(x_i)lg(c(x_i)) - m_i lg(m_i)]}{|r_n|lg(|r_n|)}.
\end{aligned}
$$

It can be seen that $\hat{g}_{3\,X\to Y}(r_n) = \sum_{i=1}^{p}\frac{c(x_i)-m_i}{|r_n|-1}$. So, we get $\hat{g}_{3\,X\to Y}(r_n) - IA_{X\to Y}(r_n)$ is bounded below by

$$
\frac{\sum_{i=1}^{p}[c(x_i)(lg(|r_n|) - lg(c(x_i))) - m_i(lg(|r_n|) - lg(m_i))]}{|r_n|lg(|r_n|)}. \qquad (1)
$$

Let N_i denote the i^{th} summation term in the numerator of (1). We assert that: (a) if $c(x_i) \leq \frac{|r_n|}{2}$, then $N_i \geq 0$; (b) if $c(x_i) > \frac{|r_n|}{2}$, then $N_i \geq -\frac{|r_n|}{2}$.

It will follow that (1) is bounded below by $-\frac{p_>}{2lg(|r_n|)}$ where $p_>$ denotes $|\{1 \leq i \leq p | c(x_i) > \frac{|r_n|}{2}\}|$. Hence it will follow that $0 \geq D_n \geq -\frac{p_>}{2lg(|r_n|)}$. However, since $\sum_{i=1}^{p} c(x_i) = |r_n|$, then $p_> \leq 1$. Thus, it will follow that $lim_{n\to\infty} D_n = 0$, as desired. All that remains is to prove the assertion.

(a) The derivative of N_i with respect to m_i equals zero if and only if $m_i = \frac{|r_n|}{2}$ (under the constraint that $1 \leq m_i \leq c(x_i)$). Assume that $c(x_i) \leq \frac{|r_n|}{2}$. Then N_i is monotonic in m_i. Thus, N_i is bounded below by the minimum of N_i at $m_i = 1, c(x_i)$. At $m_i = c(x_i)$, $N_i = 0$, and at $m_i = 1$, $N_i = c(x_i)[lg(|r_n|) - lg(c(x_i))] - lg(|r_n|)$. The latter expression is monotonic for $1 \leq c(x_i) \leq \frac{|r_n|}{2}$, and at $c(x_i) = 1$, it equals zero while at $c(x_i) = \frac{|r_n|}{2}$, it equals $\frac{|r_n|}{2} - lg(|r_n|)$. Since $|r_n| \geq 4$, $\frac{|r_n|}{2} - lg(|r_n|) \geq 0$, so, $c(x_i)[lg(|r_n|) - lg(c(x_i))] - lg(|r_n|) \geq 0$ for $1 \leq c(x_i) \leq \frac{|r_n|}{2}$. Hence, $N_i \geq 0$ for $1 \leq m_i \leq c(x_i)$.

(b) Assume $c(x_i) > \frac{|r_n|}{2}$. Since the derivative of N_i with respect to m_i is zero if and only if $m_i = \frac{|r_n|}{2}$, then N_i is bounded below by the minimum of N_i at $m_i = 1, \frac{|r_n|}{2}$, and $c(x_i)$. At $m_i = 1$, $N_i = c(x_i)[lg(|r_n|) - lg(c(x_i))] - lg(|r_n|)$; at $m_i = \frac{|r_n|}{2}, N_i = c(x_i)[lg(|r_n|) - lg(c(x_i))] - \frac{|r_n|}{2}$; and at $m_i = c(x_i), N_i = 0$. Since $|r_n| \geq 4$, then $\frac{|r_n|}{2} \geq lg(|r_n|)$, so, it suffices to show that $c(x_i)[lg(|r_n|) - c(x_i)] - \frac{|r_n|}{2} \geq -\frac{|r_n|}{2}$. But, this clearly holds since $c(x_i) \geq 1$. □

A	B	C
1	1	1
1	1	2
⋮	⋮	⋮
1	1	n
	⋮	
	⋮	
1	n	1
1	n	2
⋮	⋮	⋮
1	n	n.

Fig. 4.

The following Corollary to Theorem 4 addresses the questions (for $k = 0, 1$): does $L_{\hat{g}_3} = k$ imply $L_{IA} = k$ and does $L_{IA} = k$ imply $L_{\hat{g}_3} = k$?

Corollary 1 *Assume both $L_{\hat{g}_3}$ and L_{IA} exist.*

1. *If $L_{\hat{g}_3} = 0$, then $L_{IA} = 0$, but, the converse does not hold.*
2. *If $L_{IA} = 1$, then $L_{\hat{g}_3} = 1$, but, the converse does not hold.*

Proof: Part 1. and the "If ... then" statement in part 2. follow directly from Theorem 4. To see that the converse in part 2. does not hold, consider the relation instance, s_n, over schema A, B, C (with n^2 rows) in Fig. 4. Careful inspection shows that $\hat{g}_{3A\to B}(s_n) = \frac{n^2-n}{n^2-1}$ which goes to 1 as $n \to \infty$. Also, by Fact 1 parts 3. and 5., it follows that $IA_{A\to B}(s_n) = \frac{\mathcal{H}_B(s_n)}{lg(n^2)}$ which equals $\frac{lg(n)}{2lg(n)} = \frac{1}{2}$. Hence, $L_{\hat{g}_3} = 1$, but, $L_{IA} = \frac{1}{2}$. □

Conclusion: $\hat{g}_3$ and IA agree at their endpoints, but, behave quite differently in-between. In fact, in several cases, the measures behave quite differently even as one approaches an endpoint. We conclude that $\hat{g}_3$ and IA are quite different measures.

5 Other AFD Measures

In this section we describe two other approximation measures for AFDs from the literature (the two AFD measures mentioned in the first two paragraphs of section 2.1).

5.1 Measure of Cavallo and Pittarelli

The first approximation measure from the literature is an information theoretic measure proposed by Cavallo and Pittarelli in [1]. Their measure is defined to lie between zero and one (inclusive) and equal *one* if and only if the FD holds. To make their measure comparable with IA and g_3 we subtract it from one (so that it equals *zero* if and only if the FD holds). The result is the following.

$$CP_{X \to Y}(r) = \begin{cases} 0 & \text{if } \mathcal{H}_Y(r) = 0 \\ \frac{\mathcal{H}_{X \to Y}(r)}{\mathcal{H}_Y(r)} & \text{otherwise.} \end{cases}$$

The only difference between IA and CP is the normalization; both are the information dependency measure normalized to lie between zero and one (inclusive). IA is normalized by $lg(|r|)$ while CP is normalized by $\mathcal{H}_Y(r)$. Because CP, IA, and $\hat{g_3}$ are all zero exactly when $X \to Y$ holds over r, then these three measures agree at the endpoint, zero.

A fundamental difference in the behavior of CP and IA (also $\hat{g_3}$) lies in what happens when these measures are maximized. We saw from Theorem 1 part 2. that IA and $\hat{g_3}$ are maximized exactly when $|\Pi_X(r)| = 1$ and $|\Pi_Y(r)| = |r|$. However, CP is maximized exactly when X and Y are independent in the sense that knowing the X value of a tuple gives no information whatsoever as to the Y value in the tuple. Formally stated, X and Y are independent if $|\Pi_Y(r)| \neq 1$ and for all $1 \leq i \leq p, 1 \leq j \leq q$ it is the case that $c(x_i, y_j) = \frac{c(x_i)c(y_j)}{|r|}$. The definition of independence can be understood as follows. The probability that a tuple contains x_i is $\frac{c(x_i)}{|r|}$, contains y_j is $\frac{c(y_j)}{|r|}$, and contains both x_i and y_j is $\frac{c(x_i,y_j)}{|r|}$. Independence implies $\frac{c(x_i,y_j)}{|r|} = \frac{c(x_i)}{|r|}\frac{c(y_j)}{|r|}$, thus, $c(x_i, y_j) = \frac{c(x_i)c(y_j)}{|r|}$.

If $IA_{X \to Y}(r) = 1$ (equivalently, $\hat{g_3}$), then $|\Pi_X(r)| = 1$, so, $p = 1$ and $c(x_1) = |r|$. Thus, $\frac{c(x_1,y_j)}{c(x_1)} = \frac{c(y_j)}{|r|}$ for all $1 \leq j \leq q$. So, X and Y are independent, therefore, $CP_{X \to Y}(r) = 1$. On the other hand, there are instances, s, in which X and Y are independent, but, $IA_{X \to Y}(s) < 1$. In fact, $IA_{X \to Y}(s) = 1$ only if $|\Pi_Y(s)| = |s|$, but, independnece does not require such a strong condition. So, the conditions under which CP is maximized are much weaker than that of IA and $\hat{g_3}$. CP does not agree with IA and $\hat{g_3}$ at the endpoint, one.

Moreover, it can be shown that $\hat{g_3}$ and CP do not agree at either endpoint in the limit. In other words, there are growing sequences for which: $\hat{g_3} \to 1$, but $CP \nrightarrow 1$; $\hat{g_3} \to 0$, but $CP \nrightarrow 0$; $CP \to 1$, but, $\hat{g_3} \nrightarrow 1$; $CP \to 0$, but $\nrightarrow 0$. As for IA, the situation is a bit less constrasting. Since IA is normalized by a larger factor than CP, then $IA_{X \to Y}(r) \leq CP_{X \to Y}(r)$. Hence, $IA \to 1$ implies $CP \to 1$

and $CP \rightarrow 0$ implies $IA \rightarrow 0$. But, in the other two endpoint limit cases, IA and CP do not agree. Namely, there exist growing sequences for which: $IA \rightarrow 0$, but, $CP \nrightarrow 0$; $CP \rightarrow 1$, but $IA \nrightarrow 1$.

In summary, CP behaves quite differently than $\hat{g_3}$ and IA with a only few exceptions. One of the fundamental differences is that the condition for maximizing IA and $\hat{g_3}$ is much weaker than that for maximizing CP.

5.2 τ Measure

The second approximation measure in the literature was originally defined by Goodman and Kruskal [3] and is called τ. Later Piatetski-Shapiro [13] defines probabilisitic data dependencies and based on these he goes on to define a normalized measure for AFDs which corresponds to τ. Just as the measure of Cavallo and Pittarelli, τ is defined to lie between zero and one (inclusive) and equal *one* if and only if the FD holds. To make their measure comparible with IA and $\hat{g_3}$ we subtract it from one (so that it equals *zero* if and only if the FD holds). The result is the following.

$$\hat{\tau}_{X \rightarrow Y}(r) = \begin{cases} 0 & \text{if } |\Pi_Y(r)| = 1 \\ 1 - \frac{(\sum_{i=1}^{p} \sum_{j=1}^{q} \frac{c(x_i, y_j)^2}{|r| c(x_i)}) - \sum_{j=1}^{q} \frac{c(y_j)^2}{|r|^2}}{1 - \sum_{j=1}^{q} \frac{c(y_j)^2}{|r|^2}} & \text{otherwise.} \end{cases}$$

As with CP, a fundamental difference in the behavior of $\hat{\tau}$ and IA (also $\hat{g_3}$) lies in what happens when these measures are maximized. $\hat{\tau}$ reaches one exactly when CP does, namely, when X and Y are independent. So, the conditions under which $\hat{\tau}$ is maximized are much weaker than that of IA and $\hat{g_3}$. $\hat{\tau}$ does not agree with IA and $\hat{g_3}$ at the endpoint, one.

Moreover, it can be shown that there are limiting cases where $\hat{\tau}$ and g_3, IA do not agree at the endpoints. Namely, there exist growing sequences such that: $\hat{\tau} \rightarrow 1$ but $g_3, IA \nrightarrow 1$ and $g_3, IA \rightarrow 0$ but $\hat{\tau} \nrightarrow 0$, respectively. We did not investigate the two other limiting cases, because it is clear already that $\hat{\tau}$ is quite different from both of g_3 and IA.

We know that $\hat{\tau}$ and CP agree on their endpoints, but, we did not compare these measures further. This analysis is left as future work.

6 Further Directions

In [6], three different measures for AFDs are defined (one of which is g_3) and the authors state that it is not clear which, if any, of these measures is the most natural measure of the degree to which an FD is approximate. Adding IA, CP, and $\hat{\tau}$ to the mix, the situation is made further unclear. The following question is raised. Is there a rigorous way to determine when one measure better captures the meaning of the degree to which an FD is "approximate"?

We intend to address this question by hypothesizing that the degree to which an FD is approximate is the degree to which a the relation instance determines

a function between $\Pi_X(r)$ and $\Pi_Y(r)$. We feel that an effort an axiomatizing the degree to which a function is determined will shed light on our hypothesis. We feel that particular attention should be paid to conditions under which a function is the "furthest" from being determined.

Acknowledgments

The authors thank the following individuals for contributions to this paper: Dennis Groth, Memo Dalkilic, Dirk Van Gucht, and Jan Paredaens.

References

1. Cavallo R. and Pittarelli M. The theory of probabilistic databases. In *Proceedings of the 13th International Conference on Very Large Databases (VLDB)*, pages 71–81, 1987.
2. Dalkilic M. and Robertson E. Information dependencies. In *Proceedings of the 19th ACM SIGMOD-SIGACT-SIGART Symposium on Principals of Database Systems (PODS)*, pages 245–253, 2000.
3. Goodman L. and Kruskal W. Measures of associations for cross classifications. *Journal of the American Statistical Association*, 49:732–764, 1954.
4. Huhtala Y., Kärkkäinen J., Porkka P., and Toivonen H. Tane: An efficient algorithm for discovering functional and approximate dependencies. *The Computer Journal*, 42(2):100–111, 1999.
5. Kantola M., Mannila H., Räihä K., and Siirtola H. Discovering functional and inclusion dependencies in relational databases. *International Journal of Intelligent Systems*, 7:591–607, 1992.
6. Kivinen J., Mannila H. Approximate inference of functional dependencies from relations. *Theoretical Computer Science*, 149:129–149, 1995.
7. Lopes S., Petit J., and Lakhal L. Efficient discovery of functional dependencies and armstrong relations. In *Lecture Notes in Computer Science 1777 (first appeared in the Proceedings of the Seventh International Conference on Extending Database Technology (EDBT))*, pages 350–364, 2000.
8. Malvestuto F. Theory of random observables in relational databases. *Information Systems*, 8(4):281–289, 1983.
9. Mannila H. and Räihä K. Dependency inference. In *Proceedings of the 13th International Conference on Very Large Databases (VLDB)*, pages 155–158, 1987.
10. Mannila H. and Räihä K. Algorithms for inferring functional dependencies. *Data & Knowledge Engineering*, 12:83–99, 1994.
11. Nambiar K. K. Some analytic tools for the design of relational database systems. In *Proceedings of the 6th International Conference on Very Large Databases (VLDB)*, pages 417–428, 1980.
12. Novelli N., Cicchetti R. Fun: an efficient algorithm for mining functional and embedded dependencies. In *Lecture Notes in Computer Science 1973 (Proceedings of the 8th International Conference on Database Theory (ICDT))*, pages 189–203, 2001.
13. Piatatsky-Shapiro G. Probabilistic data dependencies. In *Proceedings of the ML-92 Workshop on Machine Discovery, Aberdeen, UK*, pages 11–17, 1992.

14. Ramakrishnan R., Gehrke J. *Database Management Systems Second Edition.* McGraw Hill Co., New York, 2000.
15. Wyss C., Giannella C., and Robertson E. Fastfds: A heuristic-driven, depth-first algorithm for mining functional dependencies from relation instances. In *To appear in Lecture Notes in Computer Science (Proceedings of the Third International Conference on Data Warehousing and Knowledge Discovery)*, pages ??–??, 2001.