Differential constraints

Conference Paper · January 2005

DOI: 10.1145/1065167.1065213 · Source: DBLP

2 authors:

Bassem Sayrafi
Birzeit University
5 PUBLICATIONS **20** CITATIONS
SEE PROFILE

Dirk Van Gucht
Indiana University Bloomington
174 PUBLICATIONS **3,307** CITATIONS
SEE PROFILE

Differential Constraints

Bassem Sayrafi and Dirk Van Gucht *
Computer Science Department
Indiana University
Lindley Hall 215
Bloomington, IN 47405-4101, USA
{bsayrafi,vgucht}@cs.indiana.edu

ABSTRACT

Differential constraints are a class of finite difference equations specified over functions from the powerset of a finite set into the reals. We characterize the implication problem for such constraints in terms of lattice decompositions, and give a sound and complete set of inference rules. We relate differential constraints to a subclass of propositional logic formulas, allowing us to show that the implication problem is coNP-complete. Furthermore, we apply the theory of differential constraints to the problem of concise representations in the frequent itemset problem by linking differential constraints to *disjunctive rules*. We also establish a connection to relational databases by associating differential constraints to *positive boolean dependencies*.

1. INTRODUCTION

Constraints (dependencies) have long been studied in the area of relational databases. In the late 1970's and early 1980's, the theory of functional, multivalued, join, inclusion, equality and tuple generating dependencies, etc was worked out [1, 16]. However, recently some of these dependencies have been re-visited, as in the context of XML databases [3, 15].

The study of constraints in data mining is more recent. Of particular interest for this paper are constraints that surface in the theory and the applications of the *frequent itemset problem* (FIS) [2]; the objective in this problem is to find itemsets that are frequently contained inside other sets given in a list (list of *baskets*). To illustrate where constraints in the FIS problem arise, consider a list of baskets $\mathcal{B}$ over some set of items S. The *support function* f associated with $\mathcal{B}$ gives for each itemset $X \subseteq S$, the value $f(X)$ which is the number of times that X is contained in sets occurring in $\mathcal{B}$. Given f, we can reason about $\mathcal{B}$ satisfying certain types of constraints. For example, the constraint $f(X) = 0$ states that $\mathcal{B}$ does not have baskets containing X. More generally, if k and l are integers, then the constraint $k \leq f(X) \leq l$ states that $\mathcal{B}$ has at least k, but not more than l, baskets containing X. These types of support (frequency) constraints were studied by Calders and Paredaens [7, 9]. Here we study constraints of a different type. Consider the constraint $f(X) = f(X \cup Y)$ which states that whenever $\mathcal{B}$ has a basket containing X, then that basket must also contain Y. A more subtle example is the constraint $f(X) - f(X \cup Y) = f(X \cup Z) - f(X \cup Y \cup Z)$. This constraint can be shown to mean that if $\mathcal{B}$ has a basket containing X then that basket must also contain Y or Z, or both. Observe that we can write these two last constraints as finite difference equations, and from there the notion *differential constraints*:

$$f(X) - f(X \cup Y) = 0$$
$$(f(X) - f(X \cup Y)) - (f(X \cup Z) - f(X \cup Y \cup Z)) = 0$$

A useful way to relate the work of Calders and Paredaens with ours is that we consider constraints on the differentials of support functions rather than on the support functions themselves.

The main purpose of this paper is the study of differential constraints and their associated *implication problem*, i.e., given a set of differential constraints $\mathcal{C}$, what other differential constraints are implied by $\mathcal{C}$? We develop this problem in the context of the class of all functions from the powerset of a finite set S into the reals. We also specialize the problem to certain subclasses of these functions, including the class of all support functions.

We show that the differential constraints implication problem can be syntactically characterized using certain *lattice decompositions*, a concept earlier considered by Demetrovics, Libkin, and Muchnik for functional dependencies [11]. We provide another syntactic characterization for the implication problem by providing a *sound and complete system of inference rules*. (Such rules already appeared in [24, 25, 26], but in that work neither the connection to lattice decompositions, nor their completeness was shown.) We relate the class of differential constraints with a class of propositional logic formulas, a technique pioneered in the study of relational constraints for functional, multivalued and other dependencies [12, 22]. The link to propositional logic enables us to prove that the implication problem for differential constraints is complete for coNP. We then apply our

*The authors were supported by NSF Grant IIS-0082407.

PODS 2005 June 13-15, 2005, Baltimore, Maryland.

work to constraints that surface in the FIS problem and in relational databases. For the FIS problem, we show that differential constraints generalize the class of disjunctive rules [6]. These constraints are used in algorithms for the FIS problem because they permit *concise representations* of certain classes of frequent or infrequent itemsets [6, 8, 17]. In particular, we show that the implication problems for differential and disjunctive constraints are equivalent. Finally, for relational constraints, we show how coupling certain real-valued functions to the attribute space of relations, a technique introduced by Lee and Malvestuto [18, 19], and later rediscovered by Dalkilic and Robertson [10], leads to a connection between differential constraints and the class of *positive boolean dependencies*. Here too, we can show that the implication problems are equivalent.

2. PRELIMINARIES

In this paper, S and all other sets, except the set of real numbers, will be assumed to be finite sets. We will use the notation $\mathcal{F}(S)$ to denote the set of all functions from 2^S into the reals. We will also use the notation $A_1A_2\cdots A_n$ to denote the set $\{A_1,\ldots,A_n\}$. If $\mathcal{Y}$ is a set of subsets of S then we use the notation $\bigcup\mathcal{Y}$ for the set $\bigcup_{Y\in\mathcal{Y}} Y$.

2.1 Differentials and density functions

Reconsider the following constraints discussed in the introduction:

$$f(X) = 0 \quad (1)$$
$$f(X) - f(X \cup Y) = 0 \quad (2)$$
$$(f(X) - f(X \cup Y)) - (f(X \cup Z) - f(X \cup Y \cup Z)) = 0 \quad (3)$$

We can write these constraints in a single format as follows:

$$\sum_{\mathcal{Z}\subseteq\mathcal{Y}} (-1)^{|\mathcal{Z}|} f(X \cup \bigcup \mathcal{Z}) = 0,$$

where for constraint (1), $\mathcal{Y} = \emptyset$, for constraint (2), $\mathcal{Y} = \{Y\}$, and for constraint (3), $\mathcal{Y} = \{Y, Z\}$. This leads to the definition of differentials and their associated density functions. (Differentials were considered by the authors and Gyssens [24, 25], however in a setting less general than here.)

DEFINITION 2.1. *Let $\mathcal{Y}$ be a set of subsets of S and let $f \in \mathcal{F}(S)$. The* $\mathcal{Y}$-differential *of f, denoted $D_f^{\mathcal{Y}}$, is the function in $\mathcal{F}(S)$ such that for each $X \subseteq S$,*

$$D_f^{\mathcal{Y}}(X) = \sum_{\mathcal{Z}\subseteq\mathcal{Y}} (-1)^{|\mathcal{Z}|} f(X \cup \bigcup \mathcal{Z}).$$

The density function *of f, denoted d_f, is the function in $\mathcal{F}(S)$ such that for each $X \subseteq S$,*

$$d_f(X) = D_f^{\{\{y\}\mid y\in\overline{X}\}}(X).$$

EXAMPLE 2.2. Let $S = \{A, B, C, D\}$ and let $f \in \mathcal{F}(S)$. Then,

$$\begin{aligned} D_f^{\{\{B\},\{C,D\}\}}(\{A\}) &= f(\{A\}) - f(\{A,B\}) \\ &\quad -f(\{A,C,D\}) + f(\{A,B,C,D\}), \end{aligned}$$

or, more succinctly,

$$D_f^{\{B,CD\}}(A) = f(A) - f(AB) - f(ACD) + f(ABCD).$$

For d_f, at sets $\{A\}$, $\{A, C\}$ and $\{A, D\}$, we have

$$\begin{aligned} d_f(A) &= D_f^{\{B,C,D\}}(A) \\ d_f(AC) &= D_f^{\{B,D\}}(AC) \\ d_f(AD) &= D_f^{\{B,C\}}(AD) \end{aligned}$$

REMARK 2.3. There is a relationship between a function $f \in \mathcal{F}(S)$ and its density function d_f. This relationship is also exhibited in the work of Calders for support functions in the frequent itemset problem [7]. Let $f \in \mathcal{F}(S)$. Then, d_f is the only function $d \in \mathcal{F}(S)$, such that for each $X \subseteq S$,

$$d(X) = \sum_{X\subseteq U\subseteq S} (-1)^{|U|-|X|} f(U) \quad (4)$$
$$f(X) = \sum_{X\subseteq U\subseteq S} d(U). \quad (5)$$

The function d_f is known as the *Möbius inverse* of f.

EXAMPLE 2.4. Continuing with Example 2.2, we have

$$\begin{aligned} d_f(A) &= f(A) - f(AB) - f(AC) - f(AD) + \\ &\quad f(ABC) + f(ABD) + f(ACD) - \\ &\quad f(ABCD) \\ d_f(AC) &= f(AC) - f(ABC) - f(ACD) + f(ABCD) \\ d_f(AD) &= f(AD) - f(ABD) - f(ACD) + f(ABCD) \end{aligned}$$

$$\begin{aligned} f(A) &= d_f(A) + d_f(AB) + d_f(AC) + d_f(AD) + \\ &\quad d_f(ABC) + d_f(ABD) + \\ &\quad d_f(ACD) + d_f(ABCD) \\ f(AC) &= d_f(AC) + d_f(ABC) + d_f(ACD) + \\ &\quad d_f(ABCD) \\ f(AD) &= d_f(AD) + d_f(ABD) + d_f(ACD) + \\ &\quad d_f(ABCD) \end{aligned}$$

2.2 Lattice decompositions

In order theory, a *semilattice* is defined as a partially ordered set wherein each pair of elements either have a meet (infimum) or a join (supremum). Let R be an nonempty partially ordered set, then R is called a *meet-semilattice* (*join-semilattice*), if each pair of elements of R has a meet (a join, respectively) in R [13]. In our context, R is a set of subsets of S, and the meet (join) of any two sets in R is their intersection (union, respectively). In this subsection, we will introduce the notions of *witness sets* and *lattice decompositions* which are formed as the union of semilattices associated with certain witness sets.

We begin with the definition of witness sets and lattice decompositions. We use the following notation: for $X, Z \subseteq S$, $[X, Z]$ denotes the *interval* $\{U \mid X \subseteq U \subseteq Z\}$.

DEFINITION 2.5. *Let $\mathcal{Y}$ be a set of subsets of S. A subset W of S is called a* witness set *of $\mathcal{Y}$ if $W \subseteq \bigcup\mathcal{Y}$ and for each $Y \in \mathcal{Y}$, $Y \cap W \neq \emptyset$. $\mathcal{W}(\mathcal{Y})$ denotes the set of all witness sets of $\mathcal{Y}$. (Observe that $\mathcal{W}(\emptyset) = \{\emptyset\}$.)*

DEFINITION 2.6. *Let $X \subseteq S$, and let $\mathcal{Y}$ be a set of subsets of S. The* lattice decomposition *of $\mathcal{Y}$ relative to X, denoted $L(X, \mathcal{Y})$, is defined such that,*

$$L(X, \mathcal{Y}) = \bigcup_{W \in \mathcal{W}(\mathcal{Y})} [X, \overline{W}].$$

EXAMPLE 2.7. Continuing with Example 2.2,

$$\mathcal{W}(\{B, CD\}) = \{BC, BD, BCD\},$$

and therefore,

$$\begin{aligned} L(A, \{B, CD\}) &= [A, \overline{BC}] \cup [A, \overline{BD}] \cup [A, \overline{BCD}] \\ &= \{A, AC, AD\}. \end{aligned}$$

An example that highlights *overlap* is as follows:

$$\mathcal{W}(\{BC, BD\}) = \{B, BC, BD, CD, BCD\},$$

and therefore,

$$\begin{aligned} L(A, \{BC, BD\}) &= [A, \overline{B}] \cup [A, \overline{BC}] \cup [A, \overline{BD}] \cup \\ &\quad [A, \overline{CD}] \cup [A, \overline{BCD}] \\ &= \{A, AB, AC, AD, ACD\}. \end{aligned}$$

The use of lattice decompositions to study data dependencies was first considered by Demetrovics, Libkin, and Muchnik in their study of a lattice-theoretic formulation for the implication problem for functional dependencies [11]. (If $X \rightarrow Y$ is a functional dependency over S then the lattice decomposition given by these authors is, in our notation, $L(X, \{Y\})$.) More recently, Baixeries and Balcázar, used *concept lattices* (occurring in the theory of *formal concept analysis*) to study the implication problems for functional dependencies [4] and that for degenerate multivalued dependencies [5].

The following proposition relates various lattice decompositions.

PROPOSITION 2.8. Let S be a finite set, let X and Z be subsets of S, and let $\mathcal{Y}$ be a set of subsets of S. Then,

$$L(X, \mathcal{Y}) = L(X, \mathcal{Y} \cup \{Z\}) \cup L(X \cup Z, \mathcal{Y}).$$

PROOF. $L(X, \mathcal{Y} \cup \{Z\}) \subseteq L(X, \mathcal{Y})$. Let $U \in L(X, \mathcal{Y} \cup \{Z\})$. Then there exists a witness set W of $\mathcal{Y} \cup \{Z\}$ such that $X \subseteq U \subseteq \overline{W}$. Let $W' = W \cap \bigcup \mathcal{Y}$. Then $W' \subseteq \bigcup \mathcal{Y}$, $\bigwedge_{Y \in \mathcal{Y}} Y \cap W' \neq \emptyset$, and since $W' \subseteq W$ and $U \subseteq \overline{W}$, $U \subseteq \overline{W'}$. Thus W' is a witness set of $\mathcal{Y}$. Consequently, $U \in L(X, \mathcal{Y})$.

$L(X \cup Z, \mathcal{Y}) \subseteq L(X, \mathcal{Y})$. Let $U \in L(X \cup Z, \mathcal{Y})$. Then there exists a witness set W of $\mathcal{Y}$ such that $X \cup Z \subseteq U \subseteq \overline{W}$. Clearly, U is in $L(X, \mathcal{Y})$.

$L(X, \mathcal{Y}) \subseteq L(X, \mathcal{Y} \cup \{Z\}) \cup L(X \cup Z, \mathcal{Y})$. Let $U \in L(X, \mathcal{Y})$. Then there exists a witness set W of $\mathcal{Y}$ such that $X \subseteq U \subseteq \overline{W}$. We consider 3 cases. Case 1: $Z \subseteq U$. Then $U \in L(X \cup Z, \mathcal{Y})$. Case 2: $Z \cap W \neq \emptyset$. Then W is a witness set for $\mathcal{Y} \cup \{Z\}$, and therefore $U \in L(X, \mathcal{Y} \cup \{Z\})$. Case 3: $Z \not\subseteq U$ and $Z \cap W = \emptyset$. Let $W' = W \cup (Z - U)$. Then $W' \subseteq \bigcup \mathcal{Y} \cup Z$, $\bigwedge_{Y \in \mathcal{Y}} Y \cap W' \neq \emptyset$, $Z \cap W' \neq \emptyset$, and $U \subseteq \overline{W'}$. Therefore, W' is a witness set of $\mathcal{Y} \cup \{Z\}$. Thus, $U \in L(X, \mathcal{Y} \cup \{Z\})$. □

The following proposition shows how differentials are related to their associated density functions via lattice decompositions.

PROPOSITION 2.9. Let X be a subset of S, let $\mathcal{Y}$ be a set of subsets of S, and let $f \in \mathcal{F}(S)$. Then,

$$D_f^{\mathcal{Y}}(X) = \textstyle\sum_{U \in L(X, \mathcal{Y})} d_f(U).$$

PROOF. By Definition 2.1, we have

$$\begin{aligned} D_f^{\mathcal{Y}}(X) &= \sum_{\mathcal{Z} \subseteq \mathcal{Y}} (-1)^{|\mathcal{Z}|} f(X \cup \bigcup \mathcal{Z}) \\ &= \sum_{\mathcal{Z} \subseteq \mathcal{Y}} (-1)^{|\mathcal{Z}|} \sum_{X \cup \bigcup \mathcal{Z} \subseteq U \subseteq S} d_f(U) \text{ (by (5))} \\ &= \sum_{X \subseteq U \subseteq S} d_f(U) \sum_{\mathcal{Z} \subseteq \{Y \in \mathcal{Y} | Y \subseteq U\}} (-1)^{|\mathcal{Z}|} \\ &\quad \text{(by commuting sums)} \\ &= \sum_{X \subseteq U \subseteq S \,\&\, \{Y \in \mathcal{Y} | Y \subseteq U\} = \emptyset} d_f(U). \end{aligned}$$

(Note that $\sum_{\mathcal{Z} \subseteq \{Y \in \mathcal{Y} | Y \subseteq U\}} (-1)^{|\mathcal{Z}|}$ equals 1 when $\{Y \in \mathcal{Y} \mid Y \subseteq U\} = \emptyset$, and equals 0, otherwise.) It suffices to show that $L(X, \mathcal{Y}) = \{U \mid X \subseteq U \subseteq S \,\&\, \{Y \in \mathcal{Y} \mid Y \subseteq U\} = \emptyset\}$.

Let $U \in L(X, \mathcal{Y})$. Thus $X \subseteq U \subseteq \overline{W}$, for some witness set W of $\mathcal{Y}$. Assume that $\{Y \in \mathcal{Y} \mid Y \subseteq U\} \neq \emptyset$. Then there exists a $Y \in \mathcal{Y}$ such that $Y \subseteq U$ and, therefore, $Y \subseteq \overline{W}$. But this is impossible since $W \cap Y \neq \emptyset$.

Let U be such that $X \subseteq U \subseteq S$ and $\{Y \in \mathcal{Y} \mid Y \subseteq U\} = \emptyset$. Let $W = \bigcup \mathcal{Y} - U$. Clearly, $W \subseteq \bigcup \mathcal{Y}$, and $\bigwedge_{Y \in \mathcal{Y}} W \cap Y \neq \emptyset$ is true. Since $U \cap W = \emptyset$, it follows that $U \subseteq \overline{W}$, and therefore, W is a witness set of $\mathcal{Y}$. Thus, $U \in L(X, \mathcal{Y})$. □

EXAMPLE 2.10. Examples 2.2 and 2.7 can be connected as follows:

$$\begin{aligned} D_f^{\{B, CD\}}(A) &= \textstyle\sum_{U \in L(A, \{B, CD\})} d_f(U) \\ &= d_f(A) + d_f(AC) + d_f(AD). \end{aligned}$$

3. DIFFERENTIAL CONSTRAINTS

In this section, we introduce differential constraints. The satisfaction of these constraints are defined in terms of density functions, in particular where these functions are 0. In addition, we give a lattice-theoretic characterization for the implication problem for differential constraints.

DEFINITION 3.1. *Let X be a subset of S, and let $\mathcal{Y}$ be a set of subsets of S. Then $X \rightarrow \mathcal{Y}$ is called a* differential constraint *over S. If $X \subseteq Y$ for some $Y \in \mathcal{Y}$, then $X \rightarrow \mathcal{Y}$ is called a* trivial *differential constraint.*

For a function $f \in \mathcal{F}(S)$, f satisfies *$X \rightarrow \mathcal{Y}$ if for each $U \in L(X, \mathcal{Y})$, $d_f(U) = 0$.*

EXAMPLE 3.2. Let $S = \{A, B, C\}$, and let $f \in \mathcal{F}(S)$ such that $f(\emptyset) = f(C) = 2$, and $f = 1$ elsewhere. Hence, $d_f(C) = d_f(ABC) = 1$, and $d_f = 0$ elsewhere. f satisfies the differential constraints $A \to \{B\}$ and $B \to \{C\}$ since $L(A, \{B\}) = \{A, AC\}$ and $L(B, \{C\}) = \{A, AB\}$. However, f does not satisfy the differential constraint $C \to \{A\}$ since $L(C, \{A\}) = \{C, BC\}$, and $d_f(C) = 1$.

We now define the implication problem for differential constraints.

DEFINITION 3.3. *Let $\mathcal{C}$ be a set of differential constraints over S, let $X \to \mathcal{Y}$ be a differential constraint, and let $\mathcal{G} \subseteq \mathcal{F}(S)$. We say that $\mathcal{C}$* implies *$X \to \mathcal{Y}$ in $\mathcal{G}$, denoted $\mathcal{C} \models_{\mathcal{G}} X \to \mathcal{Y}$, if for each $f \in \mathcal{G}$ that satisfies all the constraints in $\mathcal{C}$, f also satisfies $X \to \mathcal{Y}$. We denote by $C^*_{\mathcal{G}}$ the set of all differential constraints over S that are implied by $\mathcal{C}$ in $\mathcal{G}$. When $\mathcal{G} = \mathcal{F}(S)$, we will write $\mathcal{C} \models X \to \mathcal{Y}$ instead of $\mathcal{C} \models_{\mathcal{F}(S)} X \to \mathcal{Y}$, and C^* instead of $C^*_{\mathcal{F}(S)}$.*

EXAMPLE 3.4. Continuing with Example 3.2, let

$$\mathcal{C} \quad = \quad \{A \to \{B\}, B \to \{C\}\}.$$

Then, $\mathcal{C} \models A \to \{C\}$. Indeed, if $f \in \mathcal{F}(S)$ that satisfies all the constraints in $\mathcal{C}$, then $d_f(A) = d_f(AC) = d_f(B) = d_f(AB) = 0$. Thus, f also satisfies $A \to \{C\}$.

The following theorem gives a syntactic characterization for the implication problem for differential constraints. (Given a set $\mathcal{C}$ of differential constraints, we will use the notation $L(\mathcal{C})$ to denote the set $\bigcup_{X' \to \mathcal{Y}' \in \mathcal{C}} L(X', \mathcal{Y}')$.)

THEOREM 3.5. Let $\mathcal{C}$ be a set of differential constraints over S, and let $X \to \mathcal{Y}$ be a differential constraint over S. Then, $\mathcal{C} \models X \to \mathcal{Y}$ if and only if $L(\mathcal{C}) \supseteq L(X, \mathcal{Y})$.

PROOF. If $\mathcal{C} \models X \to \mathcal{Y}$ then for each $U \in L(X, \mathcal{Y})$, $d_f(U) = 0$ for each function $f \in \mathcal{F}(S)$ that satisfies all the constraints in $\mathcal{C}$. Assume that $L(X, \mathcal{Y}) - L(\mathcal{C}) \neq \emptyset$, and let U be an element of this set. Let c be a *nonzero* real number. Define the function f^U such that $f^U(W) = c$ if $W \subseteq U$, and $f^U(W) = 0$, otherwise. Therefore, $d_{f^U}(U) = c$ and $d_{f^U} = 0$, everywhere else. Clearly, f^U satisfies all the constraints in $\mathcal{C}$, but f^U violates the constraint $X \to \mathcal{Y}$, a contradiction.

If f satisfies the constraints in $\mathcal{C}$ then $d_f(U) = 0$ for each $U \in L(\mathcal{C})$. Since $L(\mathcal{C}) \supseteq L(X, \mathcal{Y})$, it follows that f satisfies $X \to \mathcal{Y}$. □

Demetrovics, Libkin, and Muchnik proved a result similar to Theorem 3.5 for the implication problem for functional dependencies in the relational model [11]: if F is a set of functional dependencies over S, then F implies the functional dependency $X \to Y$ if and only if (in our notation) $L(F) \supseteq L(X, \{Y\})$.

Triviality	**Addition**
$\dfrac{Y \in \mathcal{Y} \;\&\; Y \subseteq X}{X \to \mathcal{Y}}$	$\dfrac{X \to \mathcal{Y}}{X \to \mathcal{Y} \cup \{Z\}}$
Augmentation	**Elimination**
$\dfrac{X \to \mathcal{Y}}{X \cup Z \to \mathcal{Y}}$	$\dfrac{X \to \mathcal{Y} \cup \{Z\} \quad X \cup Z \to \mathcal{Y}}{X \to \mathcal{Y}}$

Figure 1: A sound and complete inference system for differential constraints.

REMARK 3.6. Observe that if $f \in \mathcal{F}(S)$ satisfies the differential constraint $X \to \mathcal{Y}$ then $D^{\mathcal{Y}}_f(X) = 0$ by Proposition 2.9. However, the converse does not hold. Indeed, let $S = \{A\}$, and let $f(\emptyset) = 0$, and $f(A) = 1$. Then $d_f(\emptyset) = -1$ and $d_f(A) = 1$ and therefore, $D^{\emptyset}_f(\emptyset) = 0$. But, f does not satisfy $\emptyset \to \emptyset$ since $L(\emptyset, \emptyset) = \{\emptyset, A\}$.

The authors and Gyssens also studied differential constraints [24, 25, 26]. However, instead of using the *density-based* semantics for such constraints defined in this paper, they considered a *differential-based* semantics. In particular, under the differential-based semantics, a function f satisfies the differential constraint $X \to \mathcal{Y}$ if $D^{\mathcal{Y}}_f(X) = 0$. Thus, if f satisfies $X \to \mathcal{Y}$ under the density-based semantics, then f satisfies $X \to \mathcal{Y}$ under the differential-based semantics. However, as shown in Remark 3.6, the converse does not hold. Consequently, the implication problem for differential constraints under the density-based semantics is *not* equivalent to the implication problem for such constraints under the differential-based semantics. In fact, the relationship between these two implication problems is not yet well-understood. However, in the special case where we only consider functions whose corresponding density functions are nonnegative or nonpositive, the two logical implication problems are equivalent. This follows from Proposition 2.9.

4. SOUND AND COMPLETE INFERENCE SYSTEM

Consider the inference rules shown in Fig. 1. We will show that the inference system consisting of these rules is sound and complete for the implication problem for differential constraints.

The inference rules in Fig. 1 were first considered by the authors for certain subclasses of $\mathcal{F}(S)$, and the soundness of the rules was proved for the differential-based semantics (see Remark 3.6) [24]. But, that paper does not address the issue of the completeness of the rules under this semantics, nor the complexity of the implication problem. However, under differential-based semantics, the authors and Gyssens gave sound and complete inference systems for certain subclasses of differential constraints whose righthand sides have at most two subsets of S [25, 26].

DEFINITION 4.1. *Let $\mathcal{C}$ be a set of differential constraints over S, and let $X \to \mathcal{Y}$ be a differential constraint. We write $\mathcal{C} \vdash X \to \mathcal{Y}$ to denote that $X \to \mathcal{Y}$ can be derived*

Chain rule

$$\frac{X \to \mathcal{Y} \cup \{Y\} \qquad X \cup Y \to \mathcal{Y} \cup \{Z\}}{X \to \mathcal{Y} \cup \{Y \cup Z\}}$$

Projection

$$\frac{X \to \mathcal{Y} \cup \{Y \cup Z\}}{X \to \mathcal{Y} \cup \{Y\}}$$

Transitivity

$$\frac{X \to \mathcal{Y} \cup \{Y\} \qquad Y \to \mathcal{Y} \cup \{Z\}}{X \to \mathcal{Y} \cup \{Z\}}$$

Separation

$$\frac{X \to \mathcal{Y} \cup \{Y \cup Z\}}{X \to \mathcal{Y} \cup \{Y\} \cup \{Z\}}$$

Union

$$\frac{X \to \mathcal{Y} \cup \{Y\} \qquad X \to \mathcal{Y} \cup \{Z\}}{X \to \mathcal{Y} \cup \{Y \cup Z\}}$$

Figure 2: Derivable inference rules for differential constraints.

from $\mathcal{C}$ using zero or more applications of the triviality, addition, augmentation, and elimination rules. Furthermore, $\mathcal{C}^+$ denotes the set $\{X' \to \mathcal{Y}' \mid \mathcal{C} \vdash X' \to \mathcal{Y}'\}$.

4.1 Soundness

PROPOSITION 4.2. Let $\mathcal{C}$ be set of differential constraints over S. If $\mathcal{C} \vdash X \to \mathcal{Y}$ then $\mathcal{C} \models X \to \mathcal{Y}$.

PROOF. The soundness of the **triviality** rule follows since, in that case $L(X, \mathcal{Y}) = \emptyset$, and thus, by Theorem. 3.5, $\mathcal{C} \models X \to \mathcal{Y}$. The soundness of the **addition**, the **augmentation**, and the **elimination** rules follow from Theorem 3.5 since, by Proposition 2.8, $L(X, \mathcal{Y} \cup \{Z\}) \subseteq L(X, \mathcal{Y})$, $L(X \cup Z, \mathcal{Y}) \subseteq L(X, \mathcal{Y})$, and $L(X, \mathcal{Y}) \subseteq L(X, \mathcal{Y} \cup \{Z\}) \cup L(X \cup Z, \mathcal{Y})$, respectively. □

In Fig. 2, we show additional inference rules for differential constraints that can be derived from the triviality, addition, augmentation, and elimination rules. These rules are useful to show the completeness of the inference system.

EXAMPLE 4.3. Let $S = \{A, B, C, D\}$, and let

$$\mathcal{C} \;=\; \{A \to \{BC, CD\}, C \to \{D\}\}.$$

We can derive the constraint $AB \to \{D\}$ as follows:

$C \to \{D\}$	(a)	given
$A \to \{BC, CD\}$	(b)	given
$A \to \{BC, C\}$	(c)	**projection** on (b)
$A \to \{C\}$	(d)	**projection** on (c)
$AB \to \{C\}$	(e)	**augmentation** on (d)
$AB \to \{D\}$		**transitivity** on (e) *and* (a).

4.2 Completeness

To prove the completeness of the inference system, we first show how a differential constraint can be decomposed into that of simpler constraints.

For a subset U of a set S, we will use the notation $\mathbf{U}$ for the set $\{\{u\} \mid u \in U\}$, and the notation $\texttt{atom}(U)$ for the differential constraint $U \to \{\{z\} \mid z \in \overline{U}\}$. We refer to $\texttt{atom}(U)$ as an *atomic* differential constraint.

DEFINITION 4.4. *Let $X \to \mathcal{Y}$ be a differential constraint over S. The* decomposition *of $X \to \mathcal{Y}$, $\texttt{decomp}(X \to \mathcal{Y})$, and the* atomic decomposition *of $X \to \mathcal{Y}$, $\texttt{atoms}(X \to \mathcal{Y})$, are defined as follows:*

$$\begin{aligned} \texttt{decomp}(X \to \mathcal{Y}) &= \{X \to \mathbf{W} \mid W \in \mathcal{W}(\mathcal{Y})\} \\ \texttt{atoms}(X \to \mathcal{Y}) &= \{\texttt{atom}(U) \mid U \in L(X, \mathcal{Y})\}. \end{aligned}$$

For example, let $S = \{A, B, C, D\}$. Then,

$$\begin{aligned} \texttt{decomp}(A \to \{B, CD\}) \;=\; & \{A \to \{B, C\}, \\ & A \to \{B, D\}, \\ & A \to \{B, C, D\}\}, \end{aligned}$$

and for atomic decompositions,

$$\begin{aligned} \texttt{atoms}(A \to \{B, CD\}) \;=\; & \{A \to \{B, C, D\}, \\ & AC \to \{B, D\}, \\ & AD \to \{B, C\}\}. \end{aligned}$$

REMARK 4.5. For each witness set $W \in \mathcal{W}(\mathcal{Y})$, $\mathcal{W}(\mathbf{W}) = \{W\}$, and for each $U \in L(X, \mathcal{Y})$, $L(U, \{\{z\} \mid z \in \overline{U}\}) = \{U\}$. Therefore, by Theorem 3.5, $\{X \to \mathcal{Y}\}^* = \texttt{decomp}(X \to \mathcal{Y})^* = \texttt{atoms}(X \to \mathcal{Y})^*$.

PROPOSITION 4.6. Let $X \to \mathcal{Y}$ be a differential constraint over S. Then, $\{X \to \mathcal{Y}\}^+ = \texttt{decomp}(X \to \mathcal{Y})^+$.

PROOF. If $X \to \mathcal{Y}$ is a trivial differential constraint, then $\mathcal{W}(\mathcal{Y}) = \emptyset$. The statement of the proposition then follows from the **triviality** rule. In the rest of the proof we will assume that $X \to \mathcal{Y}$ is a nontrivial differential constraint.

$\{X \to \mathcal{Y}\}^+ \supseteq \texttt{decomp}(X \to \mathcal{Y})^+$. Consider a constraint $X \to \mathbf{W} \in \texttt{decomp}(X \to \mathcal{Y})$. By repeated application of the **projection** rule, $X \to \mathcal{Y} \vdash X \to \{Y \cap W \mid Y \in \mathcal{Y}\}$. Then, by repeated application of the **separation** rule, $X \to \{Y \cap W \mid Y \in \mathcal{Y}\} \vdash X \to \mathbf{W}$. (Recall $W \subseteq \bigcup \mathcal{Y}$, and therefore $\bigcup_{Y \in \mathcal{Y}} Y \cap W = W$.)

$\{X \to \mathcal{Y}\}^+ \subseteq \texttt{decomp}(X \to \mathcal{Y})^+$. When $\mathcal{Y} = \emptyset$, or when $\mathcal{Y}$ consists of only singletons, then $\mathcal{W}(\mathcal{Y}) = \{\bigcup \mathcal{Y}\}$. Thus $\texttt{decomp}(X \to \mathcal{Y}) = \{X \to \mathcal{Y}\}$ and, again, the statement of the proposition holds. Otherwise, $\mathcal{Y}$ contains a set Y with two different, nonempty subsets Y_1 and Y_2 such that $Y = Y_1 \cup Y_2$. Let $\mathcal{Y}' = \mathcal{Y} - \{Y\}$ and consider the following constraints:

$$\begin{aligned} X \to \{Y_1\} \cup \mathcal{Y}' \quad &\text{(a)} \\ X \to \{Y_2\} \cup \mathcal{Y}' \quad &\text{(b)} \end{aligned}$$

$X \to \mathcal{Y}$ can be inferred from (a) and (b) using the **union** rule. By a structural induction argument, we can assume that $X \to \{Y_1\} \cup \mathcal{Y}'^+ = \texttt{decomp}(X \to \{Y_1\} \cup \mathcal{Y}')^+$ and

$X \to \{Y_2\} \cup \mathcal{Y}'^{+} = \texttt{decomp}(X \to \{Y_2\} \cup \mathcal{Y}')^{+}$. Consequently, $\{X \to \mathcal{Y}\}^{+} \subseteq \texttt{decomp}(X \to \{Y_1\} \cup \mathcal{Y}') \cup \texttt{decomp}(X \to \{Y_2\} \cup \mathcal{Y}') \vdash X \to \mathcal{Y}$. All that remains to show is $\texttt{decomp}(X \to \{Y_1\} \cup \mathcal{Y}') \cup \texttt{decomp}(X \to \{Y_2\} \cup \mathcal{Y}') \subseteq \texttt{decomp}(X \to \mathcal{Y})$. But this follows from the fact that $\mathcal{W}(\{Y_1\} \cup \mathcal{Y}') \cup \mathcal{W}(\{Y_2\} \cup \mathcal{Y}') \subseteq \mathcal{W}(Y)$. □

PROPOSITION 4.7. Let $X \to \mathcal{Y}$ be a differential constraint over S. Then, $\{X \to \mathcal{Y}\}^{+} = \texttt{atoms}(X \to \mathcal{Y})^{+}$.

PROOF. If $X \to \mathcal{Y}$ is a trivial differential constraint, then $L(X, \mathcal{Y}) = \emptyset$. The statement of the proposition then follows from the **triviality** rule. In the rest of the proof we will assume that $X \to \mathcal{Y}$ is a nontrivial differential constraint.

$\{X \to \mathcal{Y}\}^{+} \supseteq \texttt{atoms}(X \to \mathcal{Y})^{+}$. A constraint in $\texttt{atoms}(X \to \mathcal{Y})$ is of the form $U \to \{\{z\} \mid z \in \overline{U}\}$, where $U \in [X, \overline{W}]$ for some witness set $W \in \mathcal{W}(\mathcal{Y})$. By Proposition 4.6, $\{X \to \mathcal{Y}\}^{+} \supseteq \{X \to \mathbf{W}\}^{+}$. Applying the **augmentation** rule to $X \to \mathbf{W}$ results in the constraint $U \to \mathbf{W}$. Repeatedly applying the **addition rule** to this rule yields the atomic constraint $\texttt{atom}(U)$.

$\{X \to \mathcal{Y}\}^{+} \subseteq \texttt{atoms}(X \to \mathcal{Y})^{+}$. By Proposition 4.6, all that remains to show is that for each $W \in \mathcal{W}(\mathcal{Y})$, $X \to \mathbf{W} \in \texttt{atoms}(X \to \mathcal{Y})^{+}$. This can be done by considering the set of atomic constraints of the form $\texttt{atom}(U)$ for $U \in [X, \overline{W}]$. Each of these constraints is of the form $U \to \mathbf{W} \cup \{\{v\} \mid v \in S - (U \cup W)\}$. What we need to do is successively eliminate each element in $S - (X \cup W)$. This is done using the **elimination** rule as follows. For $v' \in S - (U \cup W)$, consider each pair of rules of the form $U \cup \{v'\} \to \mathbf{W} \cup \{\{v\} \mid v \in S - (U \cup W) - \{v'\}\}$ and $U \to \mathbf{W} \cup \{\{v\} \mid v \in S - (U \cup W) - \{v'\}\} \cup \{v'\}$. By the **elimination** rule, we can infer the differential constraint $U \to \mathbf{W} \cup \{\{v\} \mid v \in S - (U \cup W) - \{v'\}\}$. We have now obtained a set of rules where from v' has disappeared and from which we can, in the same manner, using the **elimination** rule, further eliminate elements from $S - (X \cup W)$. After all these eliminations, we will have derived $X \to \mathbf{W}$. □

We are now ready to prove the completeness of the inference system for differential constraints.

THEOREM 4.8. Let $\mathcal{C}$ be set of differential constraints over S, and let $X \to \mathcal{Y}$ be a differential constraint over S. If $\mathcal{C} \models X \to \mathcal{Y}$, then $\mathcal{C} \vdash X \to \mathcal{Y}$.

PROOF. Since $\mathcal{C} \models X \to \mathcal{Y}$, Theorem 3.5 implies that $L(X, \mathcal{Y}) \subseteq \bigcup_{X' \to \mathcal{Y}' \in \mathcal{C}} L(X', \mathcal{Y}')$. Therefore, $\texttt{atoms}(X \to \mathcal{Y}) \subseteq \bigcup_{X' \to \mathcal{Y}' \in \mathcal{C}} \texttt{atoms}(X' \to \mathcal{Y}')$, and, thus $\texttt{atoms}(X \to \mathcal{Y})^{+} \subseteq \left(\bigcup_{X' \to \mathcal{Y}' \in \mathcal{C}} \texttt{atoms}(X' \to \mathcal{Y}')\right)^{+}$. By Proposition 4.7, it follows that $\{X \to \mathcal{Y}\}^{+} \subseteq (\bigcup_{X' \to \mathcal{Y}' \in \mathcal{C}} \texttt{atoms}(X' \to \mathcal{Y}'))^{+}$. Since

$$\left(\bigcup_{X' \to \mathcal{Y}' \in \mathcal{C}} \texttt{atoms}(X' \to \mathcal{Y}')\right)^{+} = \left(\bigcup_{X' \to \mathcal{Y}' \in \mathcal{C}} \texttt{atoms}(X' \to \mathcal{Y}')^{+}\right)^{+},$$

it follows from Proposition 4.7 that

$$\{X \to \mathcal{Y}\}^{+} \subseteq \left(\bigcup_{X' \to \mathcal{Y}' \in \mathcal{C}} \{X' \to \mathcal{Y}'\}^{+}\right)^{+} = \mathcal{C}^{+}.$$

Thus, $\mathcal{C} \vdash X \to \mathcal{Y}$. □

5. PROPOSITIONAL LOGIC

In this section, we establish a characterization of the implication problem for differential constraints in terms of the logical implication problem for a fragment of propositional logic.

DEFINITION 5.1. *Let S be a set of propositional variables and let X be a subset of S. The* minterm *associated with X, denoted $\mathbf{X}$ is the formula $\bigwedge_{A \in X} A \wedge \bigwedge_{B \in \overline{X}} \neg B$.*

Let ϕ be a propositional formula over S. The minset *of ϕ, denoted $\texttt{minset}(\phi)$, is the set $\{X \mid \mathbf{X} \models \phi\}$, where $\models$ is the logical implication relation of the propositional logic. The* negative minset *of ϕ, denoted $\texttt{negminset}(\phi)$, is the set $\texttt{minset}(\neg\phi)$.*

The importance of minterms stems from the fact that formulas ϕ and $\bigvee_{X \in \texttt{minset}(\phi)} \mathbf{X}$ are logically equivalent. Furthermore, if Φ is a set of propositional formulas over S, and ϕ is propositional formula over S, then it is well-known that $\Phi \models \phi$ if and only if $\texttt{negminset}(\phi) \subseteq \bigcup_{\phi' \in \Phi} \texttt{negminset}(\phi')$. (Notice the resemblance with Theorem 3.5.)

DEFINITION 5.2. *Let S be a set of propositional variables, let X be a subset of S, and let $\mathcal{Y}$ be a set of subsets of S. Then $X \Rightarrow_{\texttt{prop}} \mathcal{Y}$ is called an* implication constraint *and it denotes the formula $\bigwedge X \Rightarrow \bigvee_{Y \in \mathcal{Y}} \bigwedge Y$.*

PROPOSITION 5.3. Let $X \to \mathcal{Y}$ be an implication constraint over S. Then $\texttt{negminset}(X \Rightarrow_{\texttt{prop}} \mathcal{Y}) = L(X, \mathcal{Y})$.

For example, let α be $A \to B \vee (C \wedge D)$, i.e $\neg A \vee B \vee (C \wedge D)$. Since $\texttt{negminset}(\alpha) = \texttt{minset}(\neg\alpha)$, and $\neg\alpha = A \wedge \neg B \wedge (\neg C \vee \neg D)$, then $\texttt{negminset}(\alpha) = \{A, AC, AD\}$. This corresponds to $L(A, \{B, CD\})$ (Example 2.7).

Now, by Theorem 3.5 and Proposition 5.3,

PROPOSITION 5.4. Let $\mathcal{C}$ be a set of differential constraints over S, and let $X \to \mathcal{Y}$ be a differential constraint over S. Let $\mathcal{C}^{\texttt{prop}} = \{X' \Rightarrow_{\texttt{prop}} \mathcal{Y}' \mid X' \to \mathcal{Y}' \in \mathcal{C}\}$. Then, $\mathcal{C} \models X \to \mathcal{Y}$ if and only if $\mathcal{C}^{\texttt{prop}} \models X \Rightarrow_{\texttt{prop}} \mathcal{Y}$.

Proposition 5.4 allows us to characterize the complexity of the implication problem for differential constraints. The proof is by reduction from the tautology problem for propositional logic [21].

PROPOSITION 5.5. The implication problem for differential constraints is complete for coNP.

PROOF. Since the logical implication problem for implication constraints is in coNP, we have that the implication problem for differential constraints is in coNP.

We now show that the implication problem for differential constraints is coNP-complete. Consider the class of propositional formulas in disjunctive normal form. The problem of

deciding whether a formula ϕ in this class is a tautology is known to be coNP-complete. The formula ϕ is of the form $\bigvee \Psi$ where each formula $\psi \in \Psi$ is a conjunction of literals, i.e., ψ is of the form

$$\bigwedge \mathcal{P}^\psi \wedge \bigwedge_{q \in \mathcal{Q}^\psi} \neg q,$$

where $\mathcal{P}^\psi$ and $\mathcal{Q}^\psi$ are sets propositional variables. Thus ϕ is of the form

$$\bigvee_{\psi \in \Psi} (\bigwedge \mathcal{P}^\psi \wedge \bigwedge_{q \in \mathcal{Q}^\psi} \neg q).$$

The formula ϕ is a tautology if and only if $\neg\phi$ is a contradiction. The formula $\neg\phi$ can be written as $\bigwedge_{\psi \in \Psi} ((\bigvee_{p \in \mathcal{P}^\psi} \neg p) \vee (\bigvee_{q \in \mathcal{Q}^\psi} q))$, or, equivalently as

$$\bigwedge_{\psi \in \Psi} \mathcal{P}^\psi \rightarrow \{\{q\} \mid q \in \mathcal{Q}^\psi\}.$$

Since $\neg\phi$ is a contradiction if and only if $\mathcal{C}^\phi \models \emptyset \rightarrow \emptyset$, where $\mathcal{C}^\phi = \{\mathcal{P}^\psi \rightarrow \{\{q\} \mid q \in \mathcal{Q}^\psi\} \mid \psi \in \Psi\}$, it follows that the implication problem for differential constraints is coNP-complete. □

6. FIS CONSTRAINTS

In this section, we introduce the notion of *frequency functions* and show that the implication problem for differential constraints restricted to this class of functions is equivalent to the implication problem for differential constraints over the entire class of functions. In addition, we establish a connection between differential constraints and *disjunctive rules*, a class of constraints introduced in [6, 17, 20] in the context of the *frequent itemset problem* [2]. We also indicate how this connection allows reasoning about *concise representations* of frequent itemsets [6, 8, 17].

A function f from 2^S into the reals is called a *frequency function* if for each set $\mathcal{Y}$ of subsets of S, $D_f^{\mathcal{Y}}$ is a nonnegative function. If $X \rightarrow \mathcal{Y}$ is a differential constraint over S, and f is a frequency function, then one can show that f satisfies $X \rightarrow \mathcal{Y}$ if and only if $D_f^{\mathcal{Y}}(X) = 0$. We denote the class of all frequency functions over S by $\texttt{positive}(S)$.

The importance of these functions stems from the fact that it is possible to induce a basket space from each of these functions, and vice versa.

6.1 The frequent itemset problem

The frequent itemset (FIS) problem is the following: given a set S of items, a list $\mathcal{B}$ of subsets (*baskets*) of S, and a nonnegative threshold κ, determine for each $X \subseteq S$ whether its *support*, $s^{\mathcal{B}}(X)$, is at least κ, where $s^{\mathcal{B}}(X) = |\mathcal{B}(X)|$ and $\mathcal{B}(X) = \{i \mid X \subseteq \mathcal{B}[i]\}$. If $s^{\mathcal{B}}(X) \geq \kappa$ then X is said to be a *frequent itemset*. Otherwise, X is an *infrequent itemset*.

The support function $s^{\mathcal{B}}$ is a frequency function. Indeed, let $d^{\mathcal{B}}$ be the function from 2^S into the nonnegative reals such that $d^{\mathcal{B}}(X) = |\{i \mid \mathcal{B}[i] = X\}|$, for $X \subseteq S$. In other words, $d^{\mathcal{B}}(X)$ is the number of times X appears as an element in the list of baskets $\mathcal{B}$. Clearly, for each $X \subseteq S$, $s^{\mathcal{B}}(X) = \sum_{X \subseteq U \subseteq S} d^{\mathcal{B}}(U)$. Consequently, by (Remark 2.3), $d_{s^{\mathcal{B}}} = d^{\mathcal{B}}$. Hence $d_{s^{\mathcal{B}}}$ is a nonnegative function. Therefore, by Proposition 2.9, $s^{\mathcal{B}}$ is a frequency function.

6.1.1 Concise representations

The frequency status of an itemset, in the context of a list of baskets, can be determined by explicitly *counting* the number of baskets that contain the itemset, or by *deducing* it from the frequency statuses of other itemsets. Counting is typically more expensive than deduction since it requires visiting a potentially large database of baskets. Thus, algorithms for the FIS problem try to optimize the number of deductions.

The most commonly used deduction principle is the *monotonicity rule* which states that if an itemset is infrequent then so are all of its supersets. This rule is commonly known as the Apriori rule and is at the core of the Apriori Algorithm [2]. In fact, the Apriori Algorithm computes the *negative border* associated with a list of baskets. This border consists of minimal infrequent itemsets. Since each infrequent itemset contains a subset that is minimally infrequent, the negative border is a *concise representation* of the set of all infrequent itemsets. (The notion of concise representations was first introduced by Mannila and Toivonen [20]).

Besides the Apriori rule, other techniques to avoid explicit counting were introduced in [6, 17, 20]. The main observation made in these papers is that if certain relationships are known between the frequencies of particular itemsets, then the frequencies of other itemsets can be derived. For example, assume that $\mathcal{B}(\{a\}) = \mathcal{B}(\{a, b\})$. Then, clearly, $\mathcal{B}(\{a, c\}) = \mathcal{B}(\{a, b, c\})$, and therefore, the support for itemset $\{a, b, c\}$ does not need to be counted if the support for itemset $\{a, c\}$ is known (note this is an example of applying augmentation to *pure association rules* introduced in [25]).

Building on this idea, Bykowski and Rigotti introduced the concept of *disjunctive-free* itemsets [6]. (Kryszkiewicz and Gajek [17] considered a generalization of such itemsets.) Actually, for this purpose, it is easier to consider itemsets that are *not* disjunctive-free. An itemset X is not disjunctive-free if it contains a subset X' and two attributes y_1 and y_2 in $X - X'$ (y_1 and y_2 can be equal) such that $\mathcal{B}(X') = \mathcal{B}(X' \cup \{y_1\}) \cup \mathcal{B}(X' \cup \{y_2\})$. Therefore, by inclusion-exclusion reasoning on $\mathcal{B}(X - \{y_1\})$ and $\mathcal{B}(X - \{y_2\})$, $s^{\mathcal{B}}(X) = s^{\mathcal{B}}(X - \{y_1\}) + s^{\mathcal{B}}(X - \{y_2\}) - s^{\mathcal{B}}(X - \{y_1, y_2\})$. Hence, the support of an itemset X that is not disjunctive-free can be derived, without counting, from the supports of the set $X - \{y_1\}$, $X - \{y_2\}$, and $X - \{y_1, y_2\}$. These insights lead us to introduce the concepts of *disjunctive constraints* and *disjunctive itemsets*.

DEFINITION 6.1. *Let X be a subset of S, and let $\mathcal{Y}$ be a set of subsets of S. Then $X \Rightarrow_{\texttt{disj}} \mathcal{Y}$ denotes a* disjunctive constraint. *We say that a list of baskets $\mathcal{B}$ satisfies $X \Rightarrow_{\texttt{disj}} \mathcal{Y}$ if $\mathcal{B}(X) = \bigcup_{Y \in \mathcal{Y}} \mathcal{B}(X \cup Y)$, i.e., if X is contained in a basket $B \in \mathcal{B}$ then, for some $Y \in \mathcal{Y}$, $X \cup Y$ is also contained in B.*

Disjunctive rules [6] and *generalized-disjunctive rules* [17] are special cases of disjunctive constraints introduced here. In [6, 17], only nonempty sets containing only *singleton* itemsets can occur in the righthand side of the constraints. For us, the righthand side of a constraint can be empty, or contain nonsingleton itemsets.

DEFINITION 6.2. *Let $\mathcal{B}$ be a list of baskets of S and let $X \subseteq S$. We say that X is a* disjunctive itemset *of $\mathcal{B}$ if $\mathcal{B}$ satisfies a nontrivial disjunctive constraint $X' \Rightarrow_{\texttt{disj}} \mathcal{Y}'$ such that $X \supseteq X' \cup \bigcup \mathcal{Y}'$. An itemset is* disjunctive-free *if it is not disjunctive.*

Disjunctive-free itemsets [6] and generalized disjunctive-free itemsets [17] are special cases of disjunctive-free itemsets.

We now relate the concept of disjunctive constraints to differential constraints.

PROPOSITION 6.3. Let X be a subset of S, and let $\mathcal{Y}$ be a set of subsets of S. For each list of baskets $\mathcal{B}$ over S, $\mathcal{B}$ satisfies the disjunctive constraint $X \Rightarrow_{\texttt{disj}} \mathcal{Y}$ if and only if the support function of $\mathcal{B}$, $s^{\mathcal{B}}$, satisfies the differential constraint $X \rightarrow \mathcal{Y}$.

The following proposition links the implication problem for differential constraints to the implication problem for disjunctive constraints. In this proposition, $\texttt{support}(S)$ denotes the set of all support functions of lists of baskets over S, i.e, $\texttt{support}(S) = \{s^{\mathcal{B}} \mid \mathcal{B} \text{ is a list of baskets over } S\}$.

PROPOSITION 6.4. Let $\mathcal{C}$ be a set of differential constraints over S, and let $X \rightarrow \mathcal{Y}$ be a differential constraint over S, and $\mathcal{C}^{\texttt{disj}}$ and $X \Rightarrow_{\texttt{disj}} \mathcal{Y}$ be the corresponding disjunctive constraints. Then, the following statements are equivalent:

$$\begin{array}{llll} (1) & \mathcal{C} & \models & X \rightarrow \mathcal{Y} \\ (2) & \mathcal{C} & \models_{\texttt{positive}(S)} & X \rightarrow \mathcal{Y} \\ (3) & \mathcal{C} & \models_{\texttt{support}(S)} & X \rightarrow \mathcal{Y} \\ (4) & \mathcal{C}^{\texttt{disj}} & \models & X \Rightarrow_{\texttt{disj}} \mathcal{Y} \end{array}$$

PROOF. Since, $\mathcal{F}(S) \supset \texttt{positive}(S) \supset \texttt{support}(S)$, (1) implies (2) implies (3). By Proposition 6.3, (3) and (4) are equivalent. All that remains to show is that (3) implies (1). By Theorem 3.5, it suffices to prove that if $\mathcal{C} \models_{\texttt{support}(S)} X \rightarrow \mathcal{Y}$ then $L(X, \mathcal{Y}) \subseteq \bigcup_{X' \rightarrow \mathcal{Y}' \in \mathcal{C}} L(X', \mathcal{Y}')$. In analogy with the proof of Theorem 3.5, assume that $L(X, \mathcal{Y}) - L(\mathcal{C}) \neq \emptyset$, and let U be an element of this set. Denote by f^U the support function of the list of baskets (U). Therefore, $d_{f^U}(U) = 1$ and $d_{f^U} = 0$, everywhere else. Clearly, f^U satisfies all the constraints in $\mathcal{C}$, but f^U violates the constraint $X \rightarrow \mathcal{Y}$, a contradiction. □

We conclude this section with a discussion on disjunctive sets. Bykowski and Rigotti [6] introduced, for a list of baskets $\mathcal{B}$ and some threshold κ, the sets $\texttt{FDFree}(\mathcal{B}, \kappa)$ and $\texttt{Bd}^-(\mathcal{B}, \kappa)$ which together constitute a concise representation for determining the frequency status of all itemsets, and in the case of frequent itemsets also their frequencies. $\texttt{FDFree}(\mathcal{B}, \kappa)$ consists of all frequent, disjunctive-free itemsets, and $\texttt{Bd}^-(\mathcal{B}, \kappa)$ is the set of all *minimal* itemsets of $\overline{\texttt{FDFree}(\mathcal{B}, \kappa)}$. Since

$$\overline{\texttt{FDFree}(\mathcal{B}, \kappa)} = \texttt{Infreq}(\mathcal{B}, \kappa) \cup \texttt{Disjunctive}(\mathcal{B}),$$

disjunctive itemsets feature prominently in this representation.

If $s^{\mathcal{B}}$ satisfies the nontrivial differential constraint $X \rightarrow \mathcal{Y}$, then itemset $X \cup \bigcup \mathcal{Y}$ is a disjunctive itemset. Furthermore, by the **augmentation** rule, $s^{\mathcal{B}}$ satisfies $X \cup Z \rightarrow \mathcal{Y}$, and therefore $(X \cup \bigcup \mathcal{Y}) \cup Z$ is also a disjunctive set. From this we can deduce that if a set W is disjunctive, then all of its supersets are also disjunctive. This idea is already present in the [6]. However, our results concerning the inference system for differential constraints permit additional inferencing for disjunctive itemsets. For example, assume that $\{A, B, D\}$ and $\{B, C, D\}$ are disjunctive sets on account of the constraints $A \rightarrow \{B, D\}$ and $B \rightarrow \{C, D\}$, respectively. Then, by the **transitivity** rule, it follows that $\{A, C, D\}$ is disjunctive. Thus, it it not necessary to retain $\{A, C, D\}$ as a disjunctive set since it can already be derived.

This suggests that the theory of concise representations for the frequent itemset problem can be advanced by considering how disjunctive constraints can be incorporated into the representations themselves and how inferencing can be utilized to reason about them. Practically this may be difficult since we can show that, given a set $\mathcal{C}$ of disjunctive constraints over S, and a subset X of S, the problem of deciding whether X is a disjunctive itemset, according to $\mathcal{C}$, is in the Σ_2 complexity class of the polynomial hierarchy.

7. RELATIONAL CONSTRAINTS

In this section, we will show how differential constraints feature in the theory of relational database constraints. In particular, we show how differential constraints are related to the class of *positive boolean dependencies* introduced by Sagiv, Delobel, Parker, and Fagin [22, 23]. To accomplish this, we need to introduce the notion of the Simpson function of a nonempty probabilistic relation.

DEFINITION 7.1. *Let r be a nonempty finite relation over the relation schema S (with tuple-component values in some finite set), and let p be a probability distribution associated with r such that $p(t) \neq 0$ for all tuples $t \in r$, and $p(t) = 0$ for all tuples $t \notin r$. For $X \subseteq S$, define p_X to be the marginal probability distribution of p on X. Thus, for $x \in \pi_X(r)$, $p_X(x) = \sum_{\{t \in r \mid t[X] = x\}} p(t)$. The* Simpson function *over r with distribution p is defined as follows:*

$$\texttt{simpson}_{r,p}(X) = \sum_{x \in \pi_X(r)} p_X^2(x)$$

The Simpson function $\texttt{simpson}_{r,p}$ can be interpreted as measuring, for attribute set $X \subseteq S$, *the degree of uniformity*, according to p, among the X-components of the tuples in r [24, 25, 26, 28]. The technique of associating a probability distribution with a relation r, and then introducing a function defined over the space of attributes sets was introduced by Malvestuto and Lee [18, 19] to provide *information theoretic* characterizations of functional and multivalued dependencies. Instead of using the Simpson function, these authors introduced a measure related to Shannon's entropy [27]. Later, but independently, Dalkilic and Robertson rediscovered this technique [10]. (It remains an open problem whether results in this section apply to Shannon functions.)

The class of all Simpson functions over S will be denoted by $\texttt{simpson}(S)$. It follows from the next proposition and

Proposition 2.9 that each Simpson function is a frequency function.

PROPOSITION 7.2. Let r be a nonempty relation and let p be a probability distribution associated with p. Then $d_{\texttt{simpson}_{r,p}}$, the density function of the Simpson function $\texttt{simpson}_{r,p}$, is a nonnegative function such that, for $X \subseteq S$,

$$d_{\texttt{simpson}_{r,p}}(X) = \sum_{t,t' \in r \ \&\ c(X,t,t')} p(t)p(t'),$$

where $c(X,t,t')$ is the condition,

$$(t[X] = t'[X]) \ \& \ \bigwedge_{y \in \overline{X}} t(y) \neq t'(y).$$

It is instructive to consider when $d_{\texttt{simpson}_{r,p}}(X) = 0$, for $X \subseteq S$. Then $\sum_{t,t' \in r \ \&\ c(X,t,t')} p(t)p(t') = 0$, or equivalently, r satisfies the statement

$$\forall t, t' \in r : t[X] = t'[X] \Rightarrow \bigvee_{y \in \overline{X}} t(y) = t'(y).$$

In fact, we are able to show the following:

PROPOSITION 7.3. Let $X \rightarrow \mathcal{Y}$ be a differential constraint over S, let r be a nonempty relation over S, and let p be a probability distribution over r such that $p(t) \neq 0$ for each $t \in r$. Then, $\texttt{simpson}_{r,p}$ satisfies $X \rightarrow \mathcal{Y}$ if and only if r satisfies the statement:

$$\forall t, t' \in r : t[X] = t'[X] \Rightarrow \bigvee_{Y \in \mathcal{Y}} t[Y] = t'[Y] \qquad (6)$$

Formula (6) is an instance of the class of *positive boolean dependencies* introduced in [22, 23] in the context of relational databases. (Observe that boolean dependencies generalize functional dependencies: set $\mathcal{Y} = \{Y\}$ for some subset Y of S.) We will use the notation $X \Rightarrow_{\texttt{boolean}} \mathcal{Y}$ to denote formula (6). Sagiv, Delobel, Parker, and Fagin established an equivalence between the implication problem for positive boolean dependencies and the corresponding propositional formulas [22, 23]. An immediate corollary of this and Propositions 7.3 and 5.4 follows next.

COROLLARY 7.4. Let $\mathcal{C}$ be a set of differential constraints over S and let $X \rightarrow \mathcal{Y}$ be a differential constraint over S. Then $\mathcal{C} \models_{\texttt{simpson}(S)} X \rightarrow \mathcal{Y}$ if and only if $\mathcal{C}^{\texttt{boolean}} \models X \Rightarrow_{\texttt{boolean}} \mathcal{Y}$.

8. CONCLUSION

The results of our paper are summarized in the following theorem.

THEOREM 8.1. Let $\mathcal{C}$ be a set of differential constraints over S and let $X \rightarrow \mathcal{Y}$ be a differential constraint over S. Then the following statements are equivalent:

$$\begin{array}{rcl}
\mathcal{C} & \models & X \rightarrow \mathcal{Y} \\
\mathcal{C} & \models_{\texttt{positive}(S)} & X \rightarrow \mathcal{Y} \\
\mathcal{C} & \models_{\texttt{support}(S)} & X \rightarrow \mathcal{Y} \\
\mathcal{C} & \models_{\texttt{simpson}(S)} & X \rightarrow \mathcal{Y} \\
\mathcal{C}^{\texttt{prop}} & \models & X \Rightarrow_{\texttt{prop}} \mathcal{Y} \\
\mathcal{C}^{\texttt{disj}} & \models & X \Rightarrow_{\texttt{disj}} \mathcal{Y} \\
\mathcal{C}^{\texttt{boolean}} & \models & X \Rightarrow_{\texttt{boolean}} \mathcal{Y} \\
\mathcal{C} & \vdash & X \rightarrow \mathcal{Y} \\
L(\mathcal{C}) & \supseteq & L(X, \mathcal{Y})
\end{array}$$

Furthermore, each of these implication problems is complete for coNP.

This theorem has two aspects to it, one positive, the other negative. On the positive side, the theorem relates classes of constraints that occur in different domains and shows that their respective implication problems are equivalent. In addition, two syntactic characterizations are provided for these problems, one with lattices, the other with inference rules. On the negative side is the complexity of these implication problems. However, here too progress can be made. In particular, one could consider subclasses of differential constraints. For example, in the case where the righthand sides of constraints only contain one element, it is possible to show that the implication problem for such constraints is equivalent to the implication problem for functional dependencies, a problem in P.

In general, we think that by relating constraints from different domains, cross-fertilization between these domains can occur. In this regard, the authors and Gyssens [26] are doing research on *measure-based constraints*. A subtopic in this research is to discover the relationship between the density-based semantics and the differential-based semantics for differential constraints. The authors of [26] show that measure-based constraints occur naturally in the FIS problem, relational databases, and also in the Dempster-Shafer theory of reasoning about uncertainty. (A good exposition of the Dempster-Shafer theory can be found in [14].)

It also our plan to consider more general differential constraints. In such constraints, it would be possible to specify additional constraints on the density functions (for example, requiring that the density functions obtain certain values (not necessarily equal to 0) at certain subsets of S. This would permit a study of the relationship between such constraints and the frequency constraints considered by Calders and Paredaens [7, 9].

9. ACKNOWLEDGMENTS

First and foremost, we like to thank Marc Gyssens for all his input related to this paper. He suggested the term *witness set*, which helped with the presentation and offered better insights. We also like to thank Leonid Libkin for pointing out recent work by Baixeries and Balcázar. Additionally, the authors extend their thanks to Toon Calders, Paul Purdom and Ed Robertson for helpful comments.

10. REFERENCES

[1] Serge Abiteboul, Richard Hull, and Victor Vianu. *Foundations of Database Systems*. Addison-Wesley, Reading, Mass., 1995.

[2] Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules. In *Proc. 20th Int. Conf. Very Large Data Bases, VLDB*, pages 487–499, 1994.

[3] Marcelo Arenas and Leonid Libkin. A normal form for xml documents. In *Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 85–96. ACM Press, 2002.

[4] Jaume Baixeries. A formal concept analysis framework to mine functional dependencies. *Workshop on mathematical mehtods for learning, Villa Geno, Como, Italy*, 2004.

[5] Jaume Baixeries and Jos Lus Balcázar. Characterization and armstrong relations for degenerate multivalued dependencies using formal concept analysis. In *Proceedings of third international conference on formal concept analysis*, LNCS. To appear Springer-Verlag, 2005.

[6] Artur Bykowski and Christophe Rigotti. A condensed representation to find frequent patterns. In *Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 267–273. ACM Press, 2001.

[7] Toon Calders. Computational complexity of itemset frequency satisfiability. In *Proceedings of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 143–154, 2004.

[8] Toon Calders and Bart Goethals. Mining all non-derivable frequent itemsets. In *Proceedings European Conference on Principles of Data Mining and Knowledge Discovery*, volume 2431 of *LNCS*, pages 74–85. Springer-Verlag, 2002.

[9] Toon Calders and Jan Paredaens. Axiomatization of frequent sets. In *Proceedings of the international conference on database theory*, pages 204–218, 2001.

[10] Mehmet M. Dalkilic and Edward L. Robertson. Information dependencies. In *Proceedings of the nineteenth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 245–253. ACM Press, 2000.

[11] János Demetrovics, Leonid Libkin, and Ilya B. Muchnik. Functional dependencies in relational databases: A lattice point of view. *Discrete Applied Mathematics*, 40(2):155–185, 1992.

[12] Ronald Fagin. Functional dependencies in a relational data base and propositional logic. *IBM Journal of Research and Development*, 21(6):543–544, 1977.

[13] George A. Grätzer. *General Lattice Theory*. Birkhäuser-Verlag, 2nd edition, 1998. XIX + 663 pages.

[14] Joseph Y. Halpern. *Reasoning about Uncertainty*. The MIT Press, 2003.

[15] Sven Hartmann and Sebastian Link. Multi-valued dependencies in the presence of lists. In *Proceedings of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 330–341, 2004.

[16] Paris C. Kanellakis. Elements of relational database theory. *Handbook of theoretical computer science (vol. B): formal models and semantics*, pages 1073–1156, 1990.

[17] Marzena Kryszkiewicz and Marcin Gajek. Concise representation of frequent patterns based on generalized disjunction-free generators. In *Proceedings of the 6th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, pages 159–171. Springer-Verlag, 2002.

[18] Tony T. Lee. An information-theoretic analysis of relational databases – part i: Data dependencies and information metric. *IEEE Transactions on Software Engineering*, SE-13:1049–1061, 1987.

[19] Francesco M. Malvestuto. Statistical treatment of the information content of a database. *Information Systems*, 11:211–223, 1986.

[20] Heikki Mannila and Hannu Toivonen. Multiple uses of frequent sets and condensed representations (extended abstract). In *Knowledge Discovery and Data Mining*, pages 189–194, 1996.

[21] Christos H. Papadimitriou. *Computational Complexity*. Addison-Wesley, 1994.

[22] Yehoshua Sagiv, Claude Delobel, Douglas Stott Parker Jr., and Ronald Fagin. An equivalence between relational database dependencies and a fragment of propositional logic. *J. ACM*, 28(3):435–453, 1981.

[23] Yehoshua Sagiv, Claude Delobel, Douglas Stott Parker Jr., and Ronald Fagin. Correction to "an equivalence between relational database dependencies and a fragment of propositional logic". *J. ACM*, 34(4):1016–1018, 1987.

[24] Bassem Sayrafi and Dirk Van Gucht. Inference systems derived from additive measures. *Workshop on Causality and Causal Discovery, London, Canada*, 2004.

[25] Bassem Sayrafi, Dirk Van Gucht, and Marc Gyssens. Measures in databases and datamining. Tech. Report TR602, Indiana University Computer Science, 2004.

[26] Bassem Sayrafi, Dirk Van Gucht, and Marc Gyssens. The implication problem for measure constraints. *Submitted*, 2005.

[27] Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423 and 623–656, July and October 1948.

[28] E. H. Simpson. Measurement of diversity. In *Nature*, volume 163, page 688, 1949.