CDMTCS
Research
Report
Series

Probabilistic cardinality constraints

Tania K. Roblot
University of Auckland,
Auckland, New Zealand

Sebastian Link
University of Auckland,
Auckland, New Zealand

CDMTCS-481
March 2015

Centre for Discrete Mathematics and
Theoretical Computer Science

Probabilistic cardinality constraints

Tania K. Roblot
University of Auckland, New Zealand
tkr@auckland.ac.nz

Sebastian Link
The University of Auckland, Private Bag 92019, New Zealand
s.link@auckland.ac.nz

March 18, 2015

Abstract

Probabilistic databases address well the requirements of an increasing number of modern applications that produce large collections of uncertain data. We propose probabilistic cardinality constraints as a principled tool to control the occurrences of data patterns in probabilistic databases. Our constraints help balance the consistency and completeness targets for the quality of an organization's data, and can be used to predict with which probability a given number of query answers will be returned without actually querying the data. These target applications are unlocked by developing algorithms to reason efficiently about probabilistic cardinality constraints, and to help analysts acquire the marginal probability by which cardinality constraints should hold in a given application domain. For this purpose, we overcome technical challenges to compute Armstrong PC-sketches as succinct data samples that perfectly visualize any given perceptions about these marginal probabilities.

Keywords: Cardinality constraints; Data and knowledge visualization; Data models; Database semantics; Management of integrity constraints; Probabilistic database; Requirements engineering

1 Introduction

Background. The notion of cardinality constraints is fundamental for understanding the structure and semantics of data. In traditional conceptual modeling, cardinality constraints were already introduced in Chen's seminal paper [2]. They have attracted significant interest and tool support ever since. Intuitively, a cardinality constraint consists of a set of attributes and a positive integer b, and holds in a relation if it does not contain $b + 1$ distinct tuples that all have matching values on all the attributes of the

Table 1: Probabilistic relation

w_1 ($p_1 = 0.75$)			w_2 ($p_2 = 0.15$)			w_3 ($p_3 = 0.1$)		
rfid	*time*	*zone*	*rfid*	*time*	*zone*	*rfid*	*time*	*zone*
w2	06	z1	w1	08	z4	w1	08	z4
w2	07	z1	w1	08	z5	w1	08	z5
w3	15	z7	w1	08	z6	w1	08	z6
w3	16	z8	w2	05	z1	w2	05	z1
w3	17	z9	w2	06	z1	w2	06	z1
w10	10	z11	w2	07	z1	w2	07	z1
w11	10	z12	w4	11	z3	w4	11	z3
w12	10	z13	w5	12	z3	w5	12	z3
w4	11	z3	w6	13	z3	w6	13	z3
w5	12	z3	w7	14	z3	w7	14	z3
w6	13	z3	w8	09	z2	w8	09	z2
w7	14	z3	w9	09	z2	w9	09	z2
						w0	09	z2

constraint. For example, bank customers with no more than 5 withdrawals from their bank account per month may qualify for a special interest rate. Traditionally, cardinality constraints empower applications to control the occurrences of certain data, and have applications in data cleaning, integration, modeling, processing, and retrieval among others.

Example. Relational databases target applications with certain data, such as accounting, inventory and payroll. Modern applications, such as data integration, information extraction, scientific data management, and financial risk assessment produce large amounts uncertain data. For instance, RFID (radio frequency identification) is used to track movements of endangered species of animals, such as wolverines. Here it is sensible to apply probabilistic databases. Table 1 shows a probabilistic relation (p-relation) over TRACKING=$\{$*rfid*,*time*,*zone*$\}$, which is a probability distribution over a finite set of possible worlds, each being a relation.

Data patterns occur with different frequency in different worlds. That is, different worlds satisfy different cardinality constraints. For example, the cardinality constraint $c_1 = card(time, zone) \leq 1$ holds in the world w_1 and $\{time,zone\}$ is therefore a key in this world, and $c_2 = card(time, zone) \leq 2$ holds in the world w_1 and w_2. Typically, the likelihood of a cardinality constraint to hold in a given application domain, i.e. the constraint's degree of meaningfulness, should be reflected by its marginal probability. In the example above, c_1 and c_2 have marginal probability 0.75 and 0.9, respectively, and we may write $(card(time, zone) \leq 1, \geq 0.75)$ and $(card(time, zone) \leq 2, \geq 0.9)$ to denote the *probabilistic cardinality constraints* (pCCs) that c_1 holds at least with probability 0.75 and c_2 holds at least with probability 0.9.

Applications. PCCs have important applications. *Data quality.* Foremost, they can express desirable properties of modern application domains that must accommodate uncertain data. This raises the ability of database systems to enforce higher levels of consistency in probabilistic databases, as updates to data are questioned when they re-

Table 2: PC-sketch of Table 1

card	*rfid*	*time*	*zone*	ι
3	w1	08	$*$	2,3
2	w2	$*$	z1	1,2,3
1	w2	$*$	z1	2,3
2	$*$	09	z2	2,3
1	$*$	09	z2	3
3	w3	$*$	$*$	1
3	$*$	10	$*$	1
4	$*$	$*$	z3	1,2,3

ι	$\Pi(\iota)$
1	.75
2	.15
3	.1

sult in violations of some pCC. Enforcing hard constraints, holding with probability 1, may remove plausible worlds and lead to an incomplete representation. The marginal probability of cardinality constraints can balance the consistency and completeness targets for the quality of an organization's data [19, 24]. *Query estimation.* PCCs can be used to obtain lower bounds on the probability by which a given maximum number of answers to a given query will be returned, without having to evaluate the query on any portion of the given, potentially big, database. For example, the query

SELECT *rfid* FROM `Tracking` WHERE *zone*='z2' AND *time*='09'

asks for the rfid of wolverines recorded in zone z2 at 09am. Reasoning about our pCCs tells us that at most 3 answers will be returned with probability 1, at most 2 answers will be returned with minimum probability 0.9, and at most 1 answer will be returned with minimum probability 0.75. A service provider may return these numbers, or approximate costs derived from them, to a customer, who can make a more informed decision whether to pay for the service. The provider, on the other hand, does not need to utilize unpaid resources for querying the potentially big data source to return the feedback.

Contributions. The applications motivate us to stipulate lower bounds on the marginal probability of cardinality constraints. The main inhibitor for the uptake of pCCs is the identification of the right lower bounds on their marginal probabilities. While it is already challenging to identify traditional cardinality constraints which are semantically meaningful in a given application domain, identifying the right probabilities is an even harder problem. Lower bounds appear to be a realistic compromise here. Our contributions can be summarized as follows. *1) Modeling.* We propose pCCs as a natural class of semantic integrity constraints over uncertain data. Their main target is to help organizations derive more value from data by ensuring higher levels of data quality and assist with data processing. *2) Reasoning.* We characterize the implication problem of pCCs by a simple finite set of Horn rules, as well as a linear time decision algorithm. This enables organizations to reduce the overhead of data quality management by pCCs to a minimal level necessary. For example, enforcing $(\mathit{card}(\mathit{rfid}) \leq 3, \geq 0.9)$, $(\mathit{card}(\mathit{zone}) \leq 4, \geq 0.9)$ and $(\mathit{card}(\mathit{rfid}, \mathit{zone}) \leq 3, \geq 0.75)$ would be redundant as the enforcement of $(\mathit{card}(\mathit{rfid}, \mathit{zone}) \leq 3, \geq 0.75)$ is already implicitly done by enforcing $(\mathit{card}(\mathit{rfid}) \leq 3, \geq 0.9)$.

3) Acquisition. For acquiring the right marginal probabilities by which pCCs hold, we show how to visualize concisely any given system of pCCs in the form of an Armstrong PC-sketch. Recall that every p-relation can be represented by some PC-table. Here, we introduce Armstrong PC-sketches as finite semantic representations of some possibly infinite p-relation which satisfies every cardinality constraint with the exact marginal probability by which it is currently perceived to hold. Problems with such perceptions are explicitly pointed out by the PC-sketch. For example, Figure 2 shows a PC-sketch for the p-relation from Table 1, which is Armstrong for the pCCs satisfied by the p-relation. The sketch shows which patterns of data must occur in how many rows (represented in column *card*) in which possible worlds (represented by the world identifiers in column ι). The symbol $*$ represents some data value that is unique within each world of the p-relations derived from the sketch. Π defines the probability distribution over the resulting possible worlds. Even when they represent finite p-relations, PC-sketches are still more concise since they only show patterns that matter and how often these occur.

Organization. We discuss related work in Section 2. PCCs are introduced in Section 3, and reasoning tools for them are established in Section 4. These form the foundation for computational support to acquire the correct marginal probabilities in Section 5. We conclude and outline future work in Section 6.

2 Related Work

Cardinality constraints are one of the most influential contributions conceptual modeling has made to the study of database constraints. They were already present in Chen's seminal paper [2] on conceptual database design. It is no surprise that today they are part of all major languages for data and knowledge modeling, including UML, EER, ORM, XSD, and OWL. Cardinality constraints have been extensively studied in database design [3, 4, 5, 6, 9, 12, 10, 11, 13, 17, 18, 22, 23, 26]. For a recent survey, see [27].

Closest to our approach is the work on possibilistic cardinality constraints [14], where tuples are attributed some degree of possibility and cardinality constraints some degree of certainty saying to which tuples they apply. In general, possibility theory is a qualitative approach, while probability theory is a quantitative approach to uncertainty. This research thereby complements the qualitative approach to cardinality constraints in [14] by a quantitative approach.

Our contribution extends results on cardinality constraints from traditional relations, which are covered by our framework as the special case where the p-relation consists of only one possible world. Extensions include work on the classical implication problem [7] and Armstrong relations [1, 8, 20]. As pCCs form a new class of integrity constraints, their associated implication problem and properties of Armstrong p-relations have not been investigated before.

There is also a large body of work on the discovery of "approximate" business rules, such as keys, functional and inclusion dependencies [21]. Here, approximate means that almost all tuples satisfy the given rule; hence allowing for very few exceptions. Our constraints are not approximate since they are either satisfied or violated by the given p-relation or the PC-sketch that represents it.

3 Cardinality Constraints on Probabilistic Databases

Next we introduce some preliminary concepts from probabilistic databases and the central notion of a probabilistic cardinality constraint. We use the symbol $\mathbb{N}_1^\infty$ to denote the positive integers together with the symbol ∞ for infinity.

A *relation schema* is a finite set R of attributes A. Each attribute A is associated with a domain $dom(A)$ of values. A tuple t over R is a function that assigns to each attribute A of R an element $t(A)$ from the domain $dom(A)$. A *relation* over R is a finite set of tuples over R. Relations over R are also called *possible worlds* of R here. An expression $card(X) \leq b$ over R with some non-empty subset $X \subseteq R$ and $b \in \mathbb{N}_1^\infty$ is called a *cardinality constraint over* R. In what follows, we will always assume that a subset of R is non-empty without mentioning it explicitly. A cardinality constraint $card(X) \leq b$ over R is said to *hold* in a possible world w of R, denoted by $w \models card(X) \leq b$, if and only if there are not $b+1$ different tuples $t_1, \cdots, t_{b+1} \in W$ such that for all $1 \leq i < j \leq b+1$, $t_i \neq t_j$ and $t_i(X) = t_j(X)$.

A *probabilistic relation* (p-relation) over R is a pair $r = (W, P)$ of a finite non-empty set w of possible worlds over R and a probability distribution $P : W \to (0, 1]$ such that $\sum_{w \in W} P(w) = 1$ holds.
Table 1 shows a p-relation over relation schema WOLVERINE=$\{$*rfid*,*time*,*zone*$\}$. World w_2 satisfies the CCs $card(rfid) \leq 3$, $card(time) \leq 3$, $card(zone) \leq 4$, $card(rfid, time) \leq 3$, $card(rfid, zone) \leq 3$, and $card(time, zone) \leq 2$ but violates the CC $card(time, zone) \leq 1$.

A cardinality constraint $card(X) \leq b$ over R is said to *hold with probability* $p \in [0, 1]$ in the p-relation $r = (W, P)$ if and only if $\sum_{w \in W, w \models card(X) \leq b} P(w) = p$. In other words, the probability of a cardinality constraint in a p-relation is the marginal probability with which it holds in the p-relation. We will now introduce the central notion of a cardinality constraint on probabilistic databases.

Definition 1 *A* probabilistic cardinality constraint*, or* pCC *for short, over relation schema R is an expression $(card(X) \leq b, \geq p)$ where $X \subseteq R$, $b \in \mathbb{N}_1^\infty$ and $p \in [0, 1]$. The pCC $(card(X) \leq b, \geq p)$ over R is said to* hold *in the p-relation r over R if and only if the probability with which the cardinality constraint $card(X) \leq b$ holds in r is at least p.*

Example 1 *In our running example over relation schema* WOLVERINE*, the p-relation from Table 1 satisfies the set Σ of the following pCCs $(card(rfid) \leq 3, \geq 1)$, $(card(time) \leq 3, \geq 1)$, $(card(zone) \leq 4, \geq 1)$, $(card(time, zone) \leq 2, \geq 0.9)$, $(card(rfid, time) \leq 1, \geq 0.75)$, $(card(rfid, zone) \leq 2, \geq 0.75)$, as well as $(card(time, zone) \leq 1, \geq 0.75)$. It violates the pCC $(card(rfid, time) \leq 1, \geq 0.9)$.*

4 Reasoning Tools

When enforcing sets of pCCs to improve data quality, the overhead they cause must be reduced to a minimal level necessary. In practice, this requires us to reason about pCCs efficiently. We will now establish basic tools for this purpose. The tools compute efficiently the largest probability for which a cardinality constraint is implied by a given set of pCCs, to discover any pCCs that are enforced redundantly, and to obtain the probability that some given query has some given number of answers. The results also help us in subsequent sections.

4.1 Finite and Unrestricted Implication Problems

Let $\Sigma \cup \{\varphi\}$ denote a set of constraints over relation schema R. We say Σ *(finitely) implies* φ, denoted by $\Sigma \models_{(f)} \varphi$, if every (finite) p-relation r over R that satisfies Σ, also satisfies φ. We use $\Sigma^*_{(f)} = \{\varphi \mid \Sigma \models_{(f)} \varphi\}$ to denote the *(finite) semantic closure* of Σ. For a class $\mathcal{C}$ of constraints, the (finite) $\mathcal{C}$-implication problem is to decide for a given relation schema R and a given set $\Sigma \cup \{\varphi\}$ of constraints in $\mathcal{C}$ over R, whether Σ (finitely) implies φ. We will now characterize the (finite) $\mathcal{C}$-implication problem for the class of pCCs axiomatically by a simple finite set of Horn rules, and algorithmically by a linear time algorithm. Our first result is that the implication problem and the finite implication problem coincide for the class of probabilistic cardinality constraints.

Theorem 1 *For the class of probabilistic cardinality constraints, the finite implication problem and the implication problem coincide.*

Proof Clearly, if $\Sigma \models \varphi$ holds, then $\Sigma \models_{(f)} \varphi$ holds too, since every finite relation is also a relation. It remains to show that if $\Sigma \models \varphi$ does not hold, then $\Sigma \models_{(f)} \varphi$ does also not hold. So, suppose that $\Sigma \models \varphi = (\mathit{card}(X) \leq b, \geq p)$ does not hold. Then there is some p-relation $r = (W, P)$ and some subset $W' \subseteq W$ such that for all $w \in W'$, w does not satisfy $\mathit{card}(X) \leq b$ and $\sum_{w \in W'} P(w) > 1 - p$. That is, for all $w \in W'$ there must exist $b + 1$ distinct tuples $t^w_0, \ldots, t^w_b \in W$ such that $t^w_i(X) = t^w_j(X)$ holds for all $0 \leq i \leq j \leq b$. Let $r' = (W_{\text{new}}, P)$ be the finite p-relation that results from r by replacing every world $w \in W'$ by the finite world $\{t^w_0, \ldots, t^w_b\}$, and by replacing every world $w \in W - \mathcal{W}'$ by the finite world $\{t^w\}$ where t is some arbitrary tuple in w. By construction, r' violates φ, and since every world in W_{new} is a subset of the world in W it results from, r' satisfies all pCCs in Σ. Consequently, r' is a witness that $\Sigma \models_{(f)} \varphi$ does also not hold. ∎

Theorem 1 allows us to speak about *the* implication problem of pCCs. For each $X \subseteq R$ we can potentially specify an infinite number of cardinality constraints of the form $\mathit{card}(X) \leq b$, say, for all odd $b \in \mathbb{N}_1$. However, this infinite set of different cardinality constraints is equivalent to the single cardinality constraint $\mathit{card}(X) \leq b$ where b is the smallest bound that has been specified. For the same reason, any potentially infinite set Σ of cardinality constraints is equivalent to a finite set of cardinality constraints, in the sense that they have the same semantic closure. In particular, the infinite semantic closure of a given set of cardinality constraints can be represented by an equivalent finite set.

These arguments do not carry over to the probabilistic case. For example, the set $\{(\mathit{card}(A) \leq n, \geq 1 - 1/n) \mid n \geq 2\}$ is infinite and none of its elements is implied by the remaining elements. We therefore assume from now on that sets of probabilistic cardinality constraints over a given relation schema are finite.

Table 3: Axiomatization $\mathfrak{P} = \{\mathcal{D}, \mathcal{Z}, \mathcal{U}, \mathcal{S}, \mathcal{B}, \mathcal{P}\}$

$\dfrac{}{(\mathit{card}(R) \leq 1, \geq 1)}$ (Duplicate-free, $\mathcal{D}$)	$\dfrac{}{(\mathit{card}(X) \leq b, \geq 0)}$ (Zero, $\mathcal{Z}$)	$\dfrac{}{(\mathit{card}(X) \leq \infty, \geq 1)}$ (Unbounded, $\mathcal{U}$)
$\dfrac{(\mathit{card}(X) \leq b, \geq p)}{(\mathit{card}(XY) \leq b, \geq p)}$ (Superset, $\mathcal{S}$)	$\dfrac{(\mathit{card}(X) \leq b, \geq p)}{(\mathit{card}(X) \leq b + b', \geq p)}$ (Bound, $\mathcal{B}$)	$\dfrac{(\mathit{card}(X) \leq b, \geq p + q)}{(\mathit{card}(X) \leq b, \geq p)}$ (Probability, $\mathcal{P}$)

4.2 Axiomatic Characterization

We determine the semantic closure by applying *inference rules* of the form $\frac{\text{premise}}{\text{conclusion}}$. For a set $\mathfrak{R}$ of inference rules let $\Sigma \vdash_{\mathfrak{R}} \varphi$ denote the *inference* of φ from Σ by $\mathfrak{R}$. That is, there is some sequence $\sigma_1, \ldots, \sigma_n$ such that $\sigma_n = \varphi$ and every σ_i is an element of Σ or is the conclusion that results from an application of an inference rule in $\mathfrak{R}$ to some premises in $\{\sigma_1, \ldots, \sigma_{i-1}\}$. Let $\Sigma^+_{\mathfrak{R}} = \{\varphi \mid \Sigma \vdash_{\mathfrak{R}} \varphi\}$ be the *syntactic closure* of Σ under inferences by $\mathfrak{R}$. $\mathfrak{R}$ is *sound* (*complete*) if for every set Σ over every $(R, \mathcal{S})$ we have $\Sigma^+_{\mathfrak{R}} \subseteq \Sigma^*$ ($\Sigma^* \subseteq \Sigma^+_{\mathfrak{R}}$). The (finite) set $\mathfrak{R}$ is a (finite) *axiomatization* if $\mathfrak{R}$ is both sound and complete. In the set $\mathfrak{P}$ of inference rules from Table 3, R denotes the underlying relation schema, X and Y form attribute subsets of R, $b, b' \in \mathbb{N}^{\infty}_1$, and p, q as well as $p + q$ are probabilities.

Theorem 2 *$\mathfrak{P}$ forms a finite axiomatization for the implication of probabilistic cardinality constraints.*

Proof The soundness is a straightforward consequence of the definitions. Indeed, $\mathcal{D}$ is sound since every possible world over R is a relation that cannot contain two different tuples with matching values on all the attributes of R. The soundness of $\mathcal{Z}$ and $\mathcal{U}$ are both satisfied trivially. The soundness of $\mathcal{S}$ follows from the fact that every possible world that satisfies $\mathit{card}(X) \leq b$ also satisfies $\mathit{card}(XY) \leq b$. The soundness of $\mathcal{B}$ follows from the fact that every possible world that satisfies $\mathit{card}(X) \leq b$ also satisfies $\mathit{card}(X) \leq b + b'$. The soundness of $\mathcal{P}$ follows immediately from the definition of a pCC.

For the completeness of $\mathfrak{P}$ let R be some relation schema and $\Sigma \cup \{(\mathit{card}(X) \leq b, \geq p)\}$ be a set of pCCs over R such that $(\mathit{card}(X) \leq b, \geq p) \notin \Sigma^+_{\mathfrak{P}}$. We need to show that $(\mathit{card}(X) \leq b, \geq p) \notin \Sigma^*$. From $(\mathit{card}(X) \leq b, \geq p) \notin \Sigma^+_{\mathfrak{P}}$ we conclude that $p > 0$, $b < \infty$ and $R - X \neq \emptyset$, due to $\mathcal{Z}$; $\mathcal{U}, \mathcal{P}$; and $\mathcal{D}, \mathcal{P}$, respectively. Let $p' := \sup\{q \mid \exists Z \subseteq X, b' \leq b((\mathit{card}(Z) \leq b', \geq q) \in \Sigma)\}$. In particular, $p' = 0$, if there is no $(\mathit{card}(Z) \leq b', \geq q) \in \Sigma$ where $Z \subseteq X$ and $b' \leq b$. We conclude that $p' < p$ for the following reason. If $p' \geq p > 0$, then there is some $(\mathit{card}(Z) \leq b', \geq q) \in \Sigma$ such that $Z \subseteq X$, $b' \leq b$ and $q \geq p$. Applications of $\mathcal{S}$, $\mathcal{B}$ and $\mathcal{P}$ would result in

Table 4: P-relation $r = (W, P)$ from the proof of Theorem 2

w_1, $p_1 = 1 - p'$	
X	$R - X$
$0 \cdots 0$	$0 \cdots 0$
$0 \cdots 0$	$1 \cdots 1$
$\vdots$	$\vdots$
$0 \cdots 0$	$b \cdots b$

w_2, $p_2 = p'$	
X	$R - X$
$0 \cdots 0$	$0 \cdots 0$

$(\mathit{card}(X) \leq b, \geq p) \in \Sigma$, which is a contradiction to our assumption. Hence, $p' < p$ and, in particular, $p' < 1$.

We now define the p-relation $r = (W, P)$ in Table 4 over R and distinguish between two cases. In the first case, $p' = 0$. Then $W = \{w_1\}$ with $p_1 = P(w_1) = 1$. Since w_1 violates $\mathit{card}(X) \leq b$, r violates $(\mathit{card}(X) \leq b, \geq p)$ since $\sum_{w \in W, w \models \mathit{card}(X) \leq b} P(W) = 0 < p$. Moreover, r satisfies every $(\mathit{card}(Z) \leq b', \geq q) \in \Sigma$, since either i) $q = 0$ or ii) $Z \not\subseteq X$ or $b' > b$.

In the second case, $p' > 0$. Then $W = \{w_1, w_2\}$. As $\mathit{card}(X) \leq b$ does not hold in world w_1, it follows that $\mathit{card}(X) \leq b$ holds with probability p' on r. Since $p' < p$, we conclude that $(\mathit{card}(X) \leq b, \geq p)$ does not hold on r. It remains to show that every pCC $\sigma = (\mathit{card}(Z) \leq b', \geq q) \in \Sigma$ holds on r. If $Z \not\subseteq X$ or $b' > b$, then both w_1 and w_2 satisfy $\mathit{card}(Z) \leq b'$ and thus $\sum_{w \in W, w \models \mathit{card}(Z) \leq b'} = 1 \geq q$, that is, r satisfies σ. Otherwise, $Z \subseteq X$ and $b' \leq b$. Consequently, w_1 violates $\mathit{card}(Z) \leq b'$ and $\sum_{w \in W, w \models \mathit{card}(Z) \leq b'} = p'$. However, $p' \geq q$ by definition. This means, r satisfies σ in this case, too. We have shown that Σ does not imply φ. ■

Example 2 *The set Σ of pCCs from Example 1 implies $\varphi = (\mathit{card}(\mathit{rfid}, \mathit{time}) \leq 4, \geq 0.8)$, but not $\varphi' = (\mathit{card}(\mathit{rfid}, \mathit{time}) \leq 1, \geq 0.8)$. In fact, φ can be inferred from Σ by applying $\mathcal{S}$ to $(\mathit{card}(\mathit{rfid}) \leq 3, \geq 1)$ to infer $(\mathit{card}(\mathit{rfid}, \mathit{time}) \leq 3, \geq 1)$, applying $\mathcal{B}$ to this pCC to infer $(\mathit{card}(\mathit{rfid}, \mathit{time}) \leq 4, \geq 1)$, and then applying $\mathcal{P}$.*

If a data set is validated against a set Σ of pCCs, then the data set does not need to be validated against any pCC φ implied by Σ. The larger the data set, the more time is saved by avoiding redundant validation checks.

4.3 Algorithmic Characterisation

In practice it is often unnecessary to determine all implied pCCs. In fact, the implication problem for pCCs has as input $\Sigma \cup \{\varphi\}$ and the question is whether Σ implies φ. Computing Σ^* and checking whether $\varphi \in \Sigma^*$ is hardly efficient. Indeed, we will now establish a linear-time algorithm for computing the maximum probability p, such that $\varphi = (\mathit{card}(X) \leq b, \geq p)$ is implied by Σ.

Theorem 3 *Let $\Sigma \cup \{(\mathit{card}(X) \leq b, \geq p)\}$ denote a set of pCCs over relation schema R. Then Σ implies $(\mathit{card}(X) \leq b, \geq p)$ if and only if i) $X = R$ or ii) $p = 0$ or iii) $b = \infty$ or iv) there is some $(\mathit{card}(Z) \leq b', \geq q) \in \Sigma$ such that $Z \subseteq X$, $b' \leq b$, and $q \geq p$.*

Algorithm 1 Inference

```
Require: $R, \Sigma, card(X) \leq b$
Ensure: $\max\{p : \Sigma \models (card(X) \leq b, \geq p)\}$
if $X = R$ or $b = \infty$ then
    $p \leftarrow 1$;
else
    $p \leftarrow 0$;
    for all $(card(Z) \leq b', \geq q) \in \Sigma$ do
        if $Z \subseteq X$ and $b' \leq b$ and $q > p$ then
            $p \leftarrow q$;
return $p$;
```

Proof We show the sufficiency first. If $X = R$, then the soundness of $\mathcal{D}$, $\mathcal{B}$, and $\mathcal{P}$ imply that $\Sigma \models (card(X) \leq b, \geq p)$. If $p = 0$, then the soundness of $\mathcal{Z}$ ensures that $\Sigma \models (card(X) \leq b, \geq p)$. If $b = \infty$, then the soundness of $\mathcal{U}$ and $\mathcal{P}$ ensure that $\Sigma \models (card(X) \leq b, \geq p)$. Finally, if there is some $(card(Z) \leq b', \geq q) \in \Sigma$ such that $Z \subseteq X$, $b' \leq b$, and $q \geq p$, then the soundness of $\mathcal{S}$, $\mathcal{B}$ and $\mathcal{P}$ imply that $\Sigma \models (card(X) \leq b, \geq p)$.

It remains to show the necessity. Let $R - X \neq \emptyset$, $p > 0$, $b < \infty$ and let Σ be such that for all $(card(Z) \leq b', \geq q) \in \Sigma$ where $Z \leq X$ and $b' \leq b$ we have $q < p$. Using the terminology from the completeness proof of Theorem 2 it follows that $p' := \sup\{q \mid \exists Z \subseteq X, b' \leq b((card(Z) \leq b', \geq q) \in \Sigma)\} < p$. Consequently, the p-relation r from the completeness proof of Theorem 2 shows that Σ does not imply $(card(X) \leq b, \geq p)$. ■

Example 3 *Continuing Example 2, we can apply Theorem 3 directly to see that* Σ *implies* $\varphi = (card(rfid, time) \leq 4, \geq 0.8)$. *Indeed, the pCC* $(card(rfid) \leq 3, \geq 1) \in \Sigma$ *satisfies the sufficient conditions of Theorem 3 to imply* φ, *since* $\{rfid\} \subseteq \{rfid, time\}$, $3 \leq 4$, *and* $1 \geq 0.8$.

Theorem 3 enables us to design Algorithm 1, which returns for a given cardinality constraint $card(X) \leq b$ the maximum probability p by which $(card(X) \leq b, \geq p)$ is implied by a given set Σ of pCCs over R. If $X = R$ or $b = \infty$, then we return probability 1. Otherwise, starting with $p = 0$ the algorithm scans all input pCCs $(card(Z) \leq b', \geq q)$ and sets p to q whenever q is larger than the current p, X contains Z and $b' \leq b$.

Theorem 4 states the correctness of Algorithm 1, which follows from Theorem 3, as well as the time complexity. Note that $||\Sigma||$ denotes the sum of the total number of attributes and the logarithm of the integer bounds that occur in the pCCs of Σ. Here, we assume without loss of generality that ∞ does not occur.

Theorem 4 *On input* $(R, \Sigma, card(X) \leq b)$ *our algorithm returns in* $\mathcal{O}(||\Sigma \cup \{(card(X) \leq b, \geq p)\}||)$ *time the maximum probability* p *with which* $(card(X) \leq b, \geq p)$ *is implied by* Σ. ■

Example 4 *Continuing Example 1, we can apply Algorithm 1 to the schema* Wolverine, *pCC set* Σ, *and the cardinality constraint* $card(rfid, time) \leq 4$, *which gives us the maximum probability* 1 *for which it is implied by* Σ.

Theorem 4 allows us to decide the associated implication problem efficiently, too. Given $R, \Sigma, (\mathit{card}(X) \leq b, \geq p)$ as an input to the implication problem we can use Algorithm 1 to compute $p' := \max\{q : \Sigma \models \mathit{card}(X) \leq b, \geq q\}$ and return an affirmative answer if and only if $p' \geq p$.

Corollary 5 *The implication problem of probabilistic cardinality constraints can be decided in linear time.* ▮

Example 5 *Continuing Example 4 we can see directly that* Σ *implies the pCC* $\varphi = (\mathit{card}(\mathit{rfid}, \mathit{time}) \leq 4, \geq 0.8)$ *since our algorithm returned 1 as the maximum probability for which* $\mathit{card}(\mathit{rfid}, \mathit{time}) \leq 4$ *is implied by* Σ*. Since the given probability of* 0.8 *does not exceed* $p = 1$*,* φ *is indeed implied.*

5 Acquiring Probabilistic Cardinality Constraints

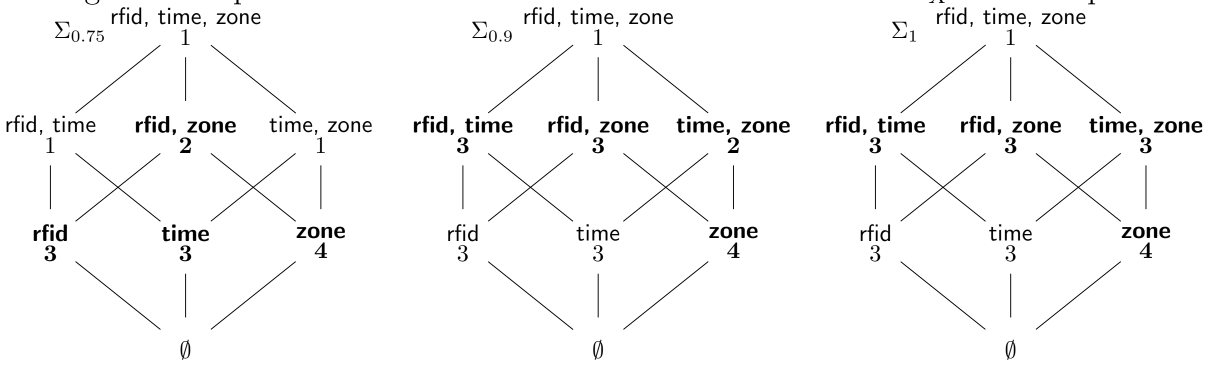
Data quality, and therefore largely the success of data-driven organizations, depend on the ability of analysts to identify the rules that govern the data. For cardinality constraints $(\mathit{card}(X) \leq b, \geq p)$ this means that the "right" marginal probability p and the "right" upper bound b must be chosen for a given set X of attributes. Choosing p too big or b too small prevents the entry of clean data, and choosing p too small or b too high enables the entry of dirty data. Analysts benefit from computational support to improve upon their ad-hoc perceptions on an appropriate probability p and bound b.

5.1 Goal

Armstrong relations constitute a useful tool in consolidating the perception of business analysts about the ordinary cardinality constraints that hold in the given application domain. Starting off with a set Σ, the tool creates a small relation that satisfies Σ and violates all ordinary cardinality constraints not implied by Σ. Hence, the relation is a perfect sample for the semantics encoded by Σ, since an arbitrary CC is satisfied by the relation if and only if it is implied by Σ. Therefore, having an Armstrong relation for Σ means that for *every* CC φ, the implication problem $\Sigma \models \varphi$ reduces to the problem of checking whether the Armstrong relation satisfies φ.

Our goal is to develop the tool of Armstrong p-relations for probabilistic cardinality constraints. Here, the situation becomes even more intriguing. In fact, an Armstrong p-relation for a given set Σ of pCCs has the following property: the marginal probability by which an arbitrary ordinary CC $\mathit{card}(X) \leq b$ holds on the Armstrong p-relation is the maximum probability p by which the pCC $(\mathit{card}(X) \leq b, \geq p)$ is implied by Σ. So, if a business analyst wants to check for an arbitrary pCC $(\mathit{card}(X) \leq b, \geq p)$ whether it is already captured by Σ, she can compute the marginal probability p' by which $\mathit{card}(X) \leq b$ holds on the Armstrong p-relation and verify that $p \geq p'$. In other words, solving an infinite number of implication problems $\Sigma \models (\mathit{card}(X) \leq b, \geq p)$ reduces to determining the marginal probabilities of $\mathit{card}(X) \leq b$ on the Armstrong p-relation.

Figure 1: Duplicate sets X in bold font and their cardinalities b_X for Example 6



In what follows, we will repeat some fundamental results from the theory of Armstrong relations for ordinary CCs and add new results to these. Based on this, we will then devise our construction of Armstrong p-relations and concise representations of these.

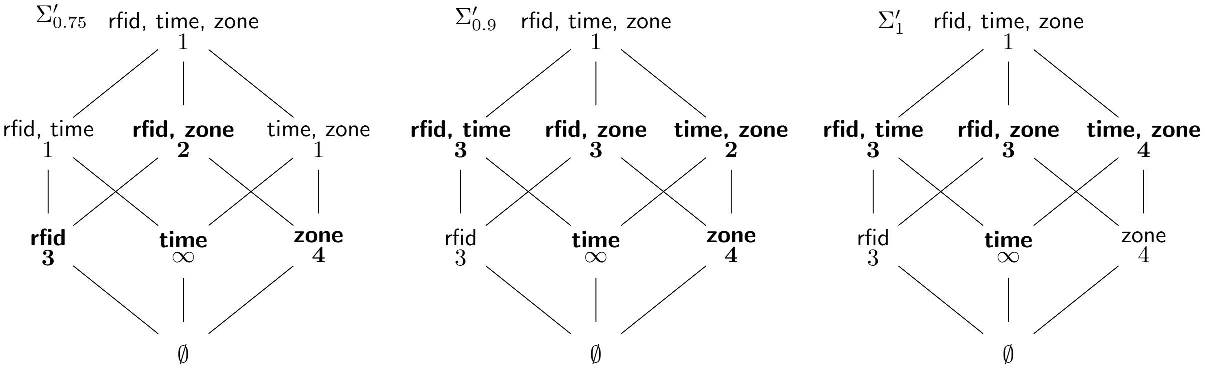
5.2 Armstrong relations and Armstrong sketches

An Armstrong relation w for a given set Σ of CCs over relation schema R violates all CCs $card(X) \leq b$ over R which are not implied by Σ. However, $\Sigma \models card(X) \leq b$ if and only if $X = R$ or $b = \infty$ or there is some $card(Z) \leq b' \in \Sigma$ where $Z \subseteq X$ and $b' \leq b$. Hence, if $\Sigma \not\models card(X) \leq b$, then $X \neq R$, $b < \infty$ and for all $card(Z) \leq b' \in \Sigma$ where $Z \subseteq X$ we have $b' > b$. Our strategy is therefore to find for all subsets X, the smallest upper bound b_X that applies to the set X. In other words, $b_X = \inf\{b \mid \Sigma \models card(X) \leq b\}$. Moreover, if $b_{XY} = b_X$ for some attribute sets X, Y, then it suffices to violate $card(XY) \leq b_{XY} - 1$. For this reason, the set $dup_\Sigma(R)$ of *duplicate sets* is defined as $dup_\Sigma(R) = \{\emptyset \subset X \subset R \mid b_X > 1 \wedge (\forall A \in R - X(b_{XA} < b_X))\}$. For each duplicate set $X \in dup_\Sigma(R)$, we introduce b_X new tuples $t_1^X, \ldots, t_{b_X}^X$ that all have matching values on all the attributes in X and all have unique values on all the attributes in $R - X$. An Armstrong relation for Σ is obtained by taking the disjoint union of $\{t_1^X, \ldots, t_{b_X}^X\}$ for all duplicate sets X.

Example 6 *For a probability p and a given set Σ of pCCs let $\Sigma_p = \{card(X) \leq b \mid \exists p' \in (0,1](card(X) \leq b, \geq p') \in \Sigma\}$. Continuing Example 1 consider the sets $\Sigma_{0.75}$, $\Sigma_{0.9}$ and Σ_1 of traditional cardinality constraints on* WOLVERINE. *The attribute subsets which are duplicate with respect to these sets are illustrated in Figure 1, together with their associated cardinalities. The worlds w_1, w_2 and w_3 in Table 1 are Armstrong relations for $\Sigma_{0.75}$, $\Sigma_{0.9}$ and Σ_1, respectively.*

While this construction works well in theory, a problem occurs with the actual use of these Armstrong relations in practice. In some cases, the Armstrong relation will be infinite and therefore of no use. These cases occur exactly if there is some attribute $A \in R$ for which $b_A = \infty$, in other words, if there is some attribute for which no finite upper bound has been specified.

Figure 2: Duplicate sets X in bold font and their cardinalities b_X for Example 7



Example 7 *Let Σ' denote the set of pCCs resulting from the set Σ from Example 1 by removing $(card(time) \leq 3, \geq 1)$. Figure 2 shows the attribute subsets which are duplicate with respect to $\Sigma'_{0.75}$, $\Sigma'_{0.9}$, and Σ'_{1}, respectively, together with their associated cardinalities.*

Fortunately, there is a simple solution to circumvent this problem and take full advantage of Armstrong relations in practice. We now present this simple solution. We introduce Armstrong sketches, which are finite representations of possibly infinite Armstrong relations.

Let R_* denote a relation schema resulting from R by extending the domain of each attribute of R by the distinguished symbol $*$. A *sketch* $\varsigma = (card, \omega)$ over R consists of a finite relation $\omega = \{\tau_1, \ldots, \tau_n\}$ over R_*, and a function *card* that maps each tuple $\tau_i \in \omega$ to a value $b_i = card(\tau_i) \in \mathbb{N}_1^\infty$. An *expansion* of ς is a relation w over R such that

- $w = \bigcup_{i=1}^n \{t_i^1, \ldots, t_i^{b_i}\}$,
- (preservation of domain values) for all $i = 1, \ldots, n$, for all $k = 1, \ldots, b_i$, for all $A \in R$, if $\tau_i(A) \neq *$, then $t_i^k(A) = \tau_i(A)$,
- (uniqueness of values substituted for $*$) for all $i = 1, \ldots, n$, for all $A \in R$, if $\tau_i(A) = *$, then for all $k = 1, \ldots, b_i$, for all $j = 1, \ldots, n$, and for all $l = 1, \ldots, b_j$ (where $l \neq k$, if $j = i$), $t_i^k(A) \neq t_j^l(A)$.

We call ς an *Armstrong sketch* for Σ, if every expansion of ς is an Armstrong relation for Σ. The following simple algorithm can be used to construct an Armstrong sketch $\varsigma = (card, \omega)$ for Σ: for each duplicate set $X \in dup_\Sigma(R)$ we introduce a tuple τ_X into ω such that, for all $A \in X$, $\tau_X(A)$ has some unique domain value from $dom(A) - \{*\}$, and for all $A \in R - X$, $\tau_X(A) = *$, and $card(\tau_X) = b_X$. The main advantage of Armstrong sketches over Armstrong relations is their smaller number of tuples. In fact, this number coincides with the number of duplicate sets which is guaranteed to be finite. In contrast, if some $b_X = \infty$, then every Armstrong relation must be infinite.

Example 8 *Continuing Example 6 the following tables show Armstrong sketches (A-sketches) for the sets $\Sigma_{0.75}$, $\Sigma_{0.9}$, and Σ_1, which have expansions w_1, w_2, and w_3 as shown in Table 1, respectively.*

A-sketch for $\Sigma_{0.75}$

card	*rfid*	*time*	*zone*
2	w2	$*$	z1
3	w3	$*$	$*$
3	$*$	10	$*$
4	$*$	$*$	z3

A-sketch for $\Sigma_{0.9}$

card	*rfid*	*time*	*zone*
3	w1	08	$*$
3	w2	$*$	z1
2	$*$	09	z2
4	$*$	$*$	z3

A-sketch for Σ_{1}

card	*rfid*	*time*	*zone*
3	w1	08	$*$
3	w2	$*$	z1
4	$*$	$*$	z3
3	$*$	09	z2

Example 9 *Continuing Example 7, the following table shows Armstrong sketches (A-sketches) for the sets* $\Sigma'_{0.75}$*,* $\Sigma'_{0.9}$*, and* Σ'_{1}*, respectively.*

A-sketch for $\Sigma'_{0.75}$

card	*rfid*	*time*	*zone*
2	w2	$*$	z1
3	w3	$*$	$*$
∞	$*$	10	$*$
4	$*$	$*$	z3

A-sketch for $\Sigma'_{0.9}$

card	*rfid*	*time*	*zone*
3	w1	08	$*$
3	w2	$*$	z1
2	$*$	09	z2
∞	$*$	10	$*$
4	$*$	$*$	z3

A-sketch for Σ'_{1}

card	*rfid*	*time*	*zone*
3	w1	08	$*$
3	w2	$*$	z1
3	$*$	09	z2
∞	$*$	10	$*$

5.3 Armstrong p-sketches and their construction

An *Armstrong p-relation* for a set Σ of pCCs over R is a p-relation r over R such that for all pCCs φ over R the following holds: $\Sigma \models \varphi$ if and only if r satisfies φ. As relations are the idealized special case of p-relations in which the relation forms the only possible world of the p-relation, there are sets of pCCs for which no finite Armstrong p-relation exists, i.e., the Armstrong p-relation contains some possible world that is infinite. For this reason we introduce probabilistic sketches and their expansions, as well as Armstrong p-sketches which are guaranteed to be finite p-relations.

A *probabilistic sketch* (p-sketch) over R is a probabilistic relation $s = (\mathcal{W}, \mathcal{P})$ over R_* where the possible worlds in $\mathcal{W}$ are sketches over R. A *probabilistic expansion* (p-expansion) of s is a p-relation $r = (W, P)$ where W contains for every sketch $\varsigma \in \mathcal{W}$ a single expansion w over R of ς, and $P(w) = \mathcal{P}(\varsigma)$.

An *Armstrong p-sketch* for a set Σ of pCCs over R is a p-sketch over R such that each of its p-expansions is an Armstrong p-relation for Σ.

Example 10 *Continuing Example 1 the following table shows an Armstrong p-sketch* s *for the given set* Σ *of pCCs.*

$\varsigma_1(p_1 = 0.75)$

card$_1$	*rfid*	*time*	*zone*
2	w2	$*$	z1
3	w3	$*$	$*$
3	$*$	10	$*$
4	$*$	$*$	z3

$\varsigma_2(p_2 = 0.15)$

card$_2$	*rfid*	*time*	*zone*
3	w1	08	$*$
3	w2	$*$	z1
4	$*$	$*$	z3
2	$*$	09	z2

$\varsigma_3(p_3 = 0.1)$

card$_3$	*rfid*	*time*	*zone*
3	w1	08	$*$
3	w2	$*$	z1
4	$*$	$*$	z3
3	$*$	09	z2

A p-expansion of s *is the finite Armstrong p-relation of Table 1.*

Example 11 *Continuing Example 7 the following table shows an Armstrong p-sketch for the set Σ' of pCCs.*

$\varsigma'_1(p'_1 = 0.75)$

card$'_1$	*rfid*	*time*	*zone*
2	*w2*	$*$	*z1*
3	*w3*	$*$	$*$
∞	$*$	*10*	$*$
4	$*$	$*$	*z3*

$\varsigma'_2(p'_2 = 0.15)$

card$'_2$	*rfid*	*time*	*zone*
3	*w1*	*08*	$*$
3	*w2*	$*$	*z1*
2	$*$	*09*	*z2*
∞	$*$	*10*	$*$
4	$*$	$*$	*z3*

$\varsigma'_3(p'_3 = 0.1)$

card$'_3$	*rfid*	*time*	*zone*
3	*w1*	*08*	$*$
3	*w2*	$*$	*z1*
3	$*$	*09*	*z2*
∞	$*$	*10*	$*$

Naturally the question arises whether Armstrong p-sketches exist for any given set of pCCs over any given relation schema. The next theorem shows that every distribution of probabilities to a finite set of cardinality constraints, that follows the inference rules from Table 3, can be represented by a single p-relation which exhibits this distribution in the form of marginal probabilities.

Theorem 6 *Let* $l : 2^R \times \mathbb{N}_1^\infty \to [0,1]$ *be a function such that the image of* l *is a finite subset of* $[0,1]$, $l(R,1) = 1$ *and for all* $X \subseteq R$, $l(X,\infty) = 1$, *and for all* $X, Y \subseteq R$ *and* $b, b' \in \mathbb{N}_1$, $l(X,b) \leq l(XY, b+b')$ *holds. Then there is some p-sketch* s *over* R *such that every p-expansion* r *of* s *satisfies* $(\mathit{card}(X) \leq b, \geq l(X,b))$, *and for all* $X \subseteq R$, $b \in \mathbb{N}_1^\infty$ *and* $p \in [0,1]$ *such that* $p > l(X,b)$, r *violates* $(\mathit{card}(X) \leq b, \geq p)$.

Proof Let $\{l_1, \ldots, l_n\}$ be the finite image of l, such that $l_1 < l_2 < \ldots < l_n$, and let $l_0 = 0$. Define a p-sketch $s = (\{\varsigma_1, \ldots, \varsigma_n\}, \mathcal{P})$ as follows. For all $i = 1, \ldots, n$, the sketch ς_i is an Armstrong sketch for the set $\Sigma_i = \{\mathit{card}(Z) \leq b' : l(Z,b') \geq l_i\}$ of ordinary cardinality constraints, and $\mathcal{P}(\varsigma_i) = l_i - l_{i-1}$. For all $X \subseteq R$ and $b \in \mathbb{N}_1^\infty$, let $l(X,b) = l_j$ for $j \in \{1, \ldots, n\}$. Then for every p-expansion $r = (\{w_1, \ldots, w_n\}, P)$ of s, $\mathit{card}(X) \leq b$ holds on w_i if and only if $i \leq j$. Consequently, $\mathit{card}(X) \leq b$ has marginal probability $l(X,b) = l_j$ with respect to r, and r satisfies $(\mathit{card}(X) \leq b, \geq l(X,b))$. However, r violates $(\mathit{card}(X) \leq b, \geq p)$ for every $p > l(X,b)$. ▮

We say that pCCs *enjoy* Armstrong p-sketches, if for every relation schema R and for every finite set Σ of pCCs over R there is some p-sketch over R that is Armstrong for Σ.

Theorem 7 *Probabilistic cardinality constraints enjoy Armstrong p-sketches.*

Proof Let Σ be a set of pCCs over R. For all $X \subseteq R$ and $b \in \mathbb{N}_1^\infty$, let $p_{X,b} := \sup\{p : \exists Z \subseteq X, b' \leq b((\mathit{card}(Z) \leq b', \geq p) \in \Sigma \cup \{(\mathit{card}(X) \leq 1, \geq 1), (\mathit{card}(X) \leq \infty, \geq 1)\})\}$. Then for all $X \subseteq R$, $b \in \mathbb{N}_1^\infty$, Σ implies $(\mathit{card}(X) \leq b, \geq p)$ if and only if $p \leq p_{X,b}$. Now, let $l(X,b) := p_{X,b}$. Then the image of l is finite as we assume that Σ is finite. Furthermore, $l(R,1) = p_{R,1} = 1$, $l(X,\infty) = 1$ and $l(XY, b+b') = p_{XY,b+b'} \geq p_{X,b} = l(X,b)$. By Theorem 6 it follows that there is some Armstrong p-sketch s over R, since for every p-expansion r of s, for all $X \subseteq R$, for all $b \in \mathbb{N}_1^\infty$ and for all $p \in [0,1]$, Σ implies $(\mathit{card}(X) \leq b, \geq p)$ if and only if r satisfies $(\mathit{card}(X) \leq b, \geq p)$. ▮

5.4 Armstrong PC-sketches

Probabilistic databases can have huge numbers of possible worlds. It is therefore important to represent and process probabilistic data concisely. Probabilistic conditional databases, or short PC-tables [25] are a popular system that can represent any given probabilistic database. Considering our aim of finding concise data samples of pCCs, we would like to compute Armstrong p-sketches in the form of Armstrong PC-sketches.

For this purpose, we first adapt the standard definition of PC-tables [25] to that of PC-sketches. A *conditional sketch* or *c-sketch*, is a tuple $\Gamma = \langle \varsigma, \iota \rangle$, where $\varsigma = (card, \omega)$ is a sketch (where ω may contain duplicate tuples), and ι assigns to each tuple τ in ω a finite set ι_τ of positive integers. The set of *world identifiers* of Γ is the union of the sets ι_τ for all tuples τ of ω. Given a world identifier i of Γ, the possible world sketch $\varsigma_i = (card_i, \omega_i)$ associated with i is $\omega_i = \{\tau | \tau \in \omega \text{ and } i \in \iota_\tau\}$ and $card_i$ is the restriction of $card$ to ω_i. The *representation* of a c-sketch $\Gamma = \langle \varsigma, \iota \rangle$ is the set $\mathcal{W}$ of possible world sketches ς_i where i denotes some world identifier of Γ. A *probabilistic conditional sketch* or *PC-sketch*, is a pair $\langle \Gamma, \Pi \rangle$ where Γ is a c-sketch, and Π is a probability distribution over the set of world identifiers of Γ. The *representation* of a PC-sketch $\langle \Gamma, \Pi \rangle$ is the p-sketch $s = (\mathcal{W}, \mathcal{P})$ where $\mathcal{W}$ is the set of possible world sketches associated with Γ and the probability $\mathcal{P}$ of each possible world sketch $\varsigma_i \in \mathcal{W}$ is defined as the probability $\Pi(i)$ of its world identifier i. It is simple to see that every p-sketch can be represented as a PC-sketch.

Theorem 8 *Every p-sketch can be represented as a PC-sketch.*

Proof Let $s = (\mathcal{W}, \mathcal{P})$ be a p-sketch. We define a PC-sketch $\langle \Gamma, \Pi \rangle$ with $\Gamma = \langle (card, \omega), \iota \rangle$ as follows. The only elements of ω are tuples τ with cardinality $card(\tau) = b$ that occur as tuples with the same cardinality b in some possible world sketch ς_i of $\mathcal{W}$. For such tuples τ we define $\iota(\tau)$ as the set of world identifiers i such that τ occurs with cardinality b in ς_i. Moreover, $\Pi(i) = \mathcal{P}(\varsigma_i)$. It is immediate that s is the representation of this PC-sketch. ∎

A PC-sketch is called an *Armstrong PC-sketch* for Σ if and only if its representation is an Armstrong p-sketch for Σ.

Example 12 *The following table shows an Armstrong PC-sketch for the set Σ' of pCCs from Example 7.*

Γ

card	*rfid*	*time*	*zone*	ι
3	*w1*	*08*	$*$	*2,3*
2	*w2*	$*$	*z1*	*1,2,3*
1	*w2*	$*$	*z1*	*2,3*
2	$*$	*09*	*z2*	*2,3*
2	$*$	*09*	*z2*	*3*
3	*w3*	$*$	$*$	*1*
∞	$*$	*10*	$*$	*1,2,3*
4	$*$	$*$	*z3*	*1,2*

Π

ι	$\Pi(\iota)$
1	*.75*
2	*.15*
3	*.1*

Algorithm 2 Armstrong PC-sketch

```
Require: $R, \Sigma$
Ensure: Armstrong PC-sketch $\langle\langle(card, \omega), \iota\rangle, \Pi\rangle$ for $\Sigma$
 1: Let $p_1 < \cdots < p_n$ be the probabilities in $\Sigma$;   ▷ If $p_n < 1$, $n \leftarrow n+1$ and $p_n \leftarrow 1$
 2: $p_0 \leftarrow 0$; $\Pi \leftarrow \emptyset$;
 3: for $i = 1, \dots, n$ do   ▷ Process one possible world sketch at a time
 4:     $\Pi \leftarrow \Pi \cup \{(i, p_i - p_{i-1})\}$;   ▷ World $i$ has probability $p_i - p_{i-1}$
 5:     Compute $\{b_X^i \mid X \subseteq R\}$;   ▷ Smallest upper bound for each $X$ in world $i$
 6:     $dup_i \leftarrow$ Set of duplicate sets for $\Sigma_{p_i}$;   ▷ Duplicate sets to realize in world $i$
 7: $\omega \leftarrow \emptyset$; $k \leftarrow 0$;
 8: $dup \leftarrow \{(X, \{i \mid X \in dup_i\}) \mid X \in dup_i$ for some $i\}$;
 9: for all $(X, W) \in dup$ do   ▷ For each $X$ that is a duplicate set in every world in $W$
10:     $b \leftarrow 0$; $j \leftarrow k + 1$;
11:     for $i = 1, \dots, n$ do   ▷ Add some $\tau_k$ that realizes $X$ in every world in $W$
12:         if $X \in dup_i$ and $b_X^i > b$ then   ▷ if there are any remaining cardinalities
13:             $k \leftarrow k + 1$;
14:             for all $A \in R$ do   ▷ Define $\tau_k$ with...
15:                 if $A \in X$ then
16:                     $\tau_k(A) \leftarrow j$;   ▷ ...fixed values on $X$
17:                 else
18:                     $\tau_k(A) \leftarrow *$;   ▷ ...and unique values outside of $X$
19:             $\omega \leftarrow \omega \cup \{\tau_k\}$;   ▷ Add new tuple
20:             $card(\tau_k) \leftarrow b_X^i - b$;   ▷ Stipulate remaining cardinality
21:             $\iota(\tau_k) \leftarrow W - \{1, \dots, i-1\}$;   ▷ Worlds that require this cardinality
22:             $b \leftarrow b_X^i$;   ▷ Mark cardinalities as already realized
23: return $\langle\langle(card, \omega), \iota\rangle, \Pi\rangle$;
```

Example 13 *Table 2 shows a PC-sketch $\langle\Gamma, \Pi\rangle$ that is Armstrong for the set Σ of pCCs from Example 1.*

Algorithm 2 computes an Armstrong PC-sketch for every given set Σ of pCCs over every given relation schema R. In our construction, the number of possible worlds is determined by the number of distinct probabilities that occur in Σ. For that purpose, for every given set Σ of pCCs over R and every probability $p \in [0,1]$, let $\Sigma_p = \{card(X) \leq b \mid \exists q \in [0,1](card(X) \leq b, \geq q) \in \Sigma \wedge q \geq p\}$ denote the *p-cut* of Σ, i.e., the set of cardinality constraints over R which hold with a probability at least p. It is possible that Σ does not contain any pCC $(card(X) \leq b, \geq p)$ where $p = 1$. In this case, Algorithm 2 computes an Armstrong PC-sketch for Σ that contains one more possible world than the number of distinct probabilities occurring in Σ. Lines (3-6) compute for all possible worlds $i = 1, \dots, n$, their probabilities as $p_i - p_{i-1}$ where $p_0 = 0$, the cardinalities b_X^i for each attribute set, and the duplicate sets dup_i for each p_i-cut Σ_{p_i} of Σ. For each duplicate set X we record the pair of X and the set W of possible worlds i for which X is a duplicate set (line 8). For each pair (X, W), Algorithm 2 adds in lines (9-22) for $i = 1, \dots, n$ a new tuple τ_k to ω that represents $b_X^i - b$ actual tuples that all agree on X

in all worlds that occur in $W - \{1, \ldots, i-1\}$. Here, b quantifies the number of tuples that have already been introduced to these worlds in previous rounds, so $b_X^i - b$ is the number of remaining tuples for these worlds.

Theorem 9 *For every set Σ of pCCs over relation schema R, Algorithm 2 computes an Armstrong PC-sketch for Σ.* ▌

Finally, we derive some bounds on the time complexity of finding Armstrong PC-sketches. Since the relational model is subsumed there are cases, where the number of tuples in every Armstrong PC-sketch for Σ over R is exponential in $||\Sigma||$. Such a case is given by $R_n = \{A_1, \ldots, A_{2n}\}$ and $\Sigma_n = \{(card(A_{2i-1}, A_{2i}) \leq 1, \geq 1) \mid i = 1, \ldots, n\}$ with $||\Sigma_n|| = 2 \cdot n$. Indeed, every Armstrong PC-sketch for Σ_n must feature 2^n different tuples to accommodate the 2^n different duplicate sets X with associated cardinality $b_X^1 = \infty$, and there is only one possible world. Algorithm 2 was designed with the goal that the worst-case time bound from the traditional relational case does not deteriorate in our more general setting. This is indeed achieved, as the computationally most demanding part of Algorithm 2 is the computation of the cardinalities in line (5) which is achieved in time exponential in $max(||\Sigma||, |R|)$, where $|R|$ denotes the number of attributes in R.

Theorem 10 *The time complexity to find an Armstrong PC-sketch for a given set Σ of pCCs over schema R is precisely exponential in $max(||\Sigma||, |R|)$.*

Proof Given R and Σ as input, Algorithm 2 computes an Armstrong PC-sketch for Σ in time at most exponential in $max(||\Sigma||, |R|)$. Indeed, the main computational effort goes into the computation of the numbers b_X^i for each attribute set $X \subseteq R$ and every p_i-cut of Σ. Clearly, the number of attribute subsets is exponential in the size of the relation schema R, and for each attribute subset X, b_X^i can be determined by scanning all the elements of Σ_{p_i} once where $||\Sigma_{p_i}|| \leq ||\Sigma||$.

There are also cases, where the number of tuples in every Armstrong PC-sketch for Σ over R is exponential in $||\Sigma||$. Such a case is given by $R_n = \{A_1, \ldots, A_{2n}\}$ and $\Sigma_n = \{(card(A_{2i-1}, A_{2i}) \leq 1, \geq 1) \mid i = 1, \ldots, n\}$ with $||\Sigma_n|| = 2 \cdot n$. Every Armstrong PC-sketch for Σ_n must feature 2^n different tuples to accommodate the 2^n different duplicate sets X with associated cardinality $b_X^1 = \infty$, and there is only one possible world. ▌

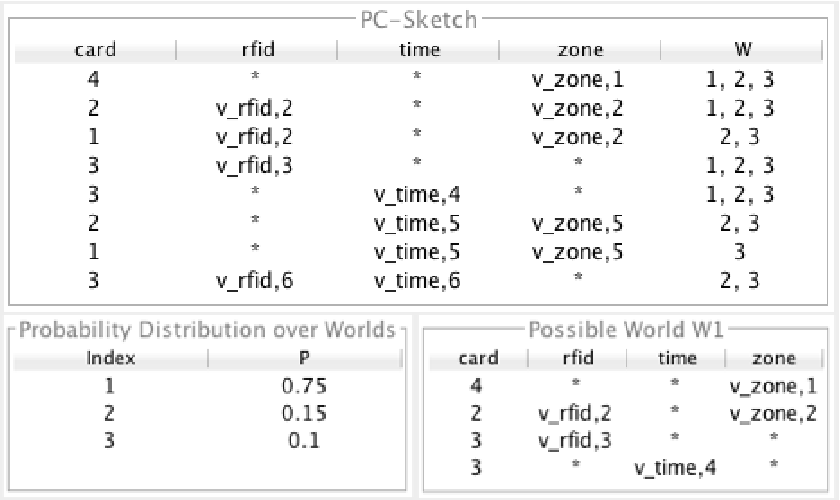
There are also cases where the number of tuples in some Armstrong PC-sketch for Σ over R is logarithmic in $||\Sigma||$. Such a case is given by $R_n = \{A_1, \ldots, A_{2n}\}$ and $\Sigma_n = \{(card(X_1 \cdots X_n) \leq 1, \geq 1) \mid X_i \in \{A_{2i-1}, A_{2i}\}$ for $i = 1, \ldots, n\}$ with $||\Sigma_n|| = n \cdot 2^n$. There is some Armstrong PC-sketch for Σ that contains one tuple for each of the n duplicate sets $X = R - \{A_{2i-1}, A_{2i}\}$ with associated cardinality $b_X^1 = \infty$.

In practice we recommend to use both representations of business rules: one in the form of the set Σ of pCCs itself and one in the form of an Armstrong PC-sketch for Σ. Our results confirm that this is always possible. We conjecture that Armstrong PC-sketches help identify bounds b that are too low and/or probabilities p that are too high, while the set Σ helps identify bounds b that are too high and/or probabilities p that are too low.

5.5 Graphical User Interface

We have implemented Algorithm 2 in the form of a graphical user interface (GUI) called *Fortuna*[1]. A user can enter some attributes and specify probabilistic cardinality constraints using any combination of these. The GUI shows an Armstrong PC-sketch for the specified input, sketches of the possible worlds can be brought up, and their individual tuples can be expanded at will. Figure 3 shows a partial screenshot of our GUI *Fortuna* with some outputs for our running example.

Figure 3: Screenshot of the GUI *Fortuna*



PC-Sketch

card	rfid	time	zone	W
4	*	*	v_zone,1	1, 2, 3
2	v_rfid,2	*	v_zone,2	1, 2, 3
1	v_rfid,2	*	v_zone,2	2, 3
3	v_rfid,3	*	*	1, 2, 3
3	*	v_time,4	*	1, 2, 3
2	*	v_time,5	v_zone,5	2, 3
1	*	v_time,5	v_zone,5	3
3	v_rfid,6	v_time,6	*	2, 3

Probability Distribution over Worlds

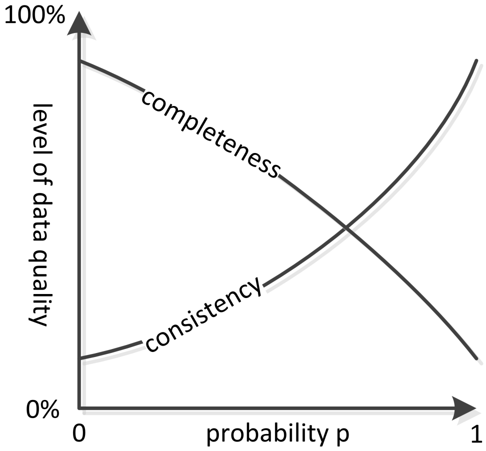
Index	P
1	0.75
2	0.15
3	0.1

Possible World W1

card	rfid	time	zone
4	*	*	v_zone,1
2	v_rfid,2	*	v_zone,2
3	v_rfid,3	*	*
3	*	v_time,4	*

6 Conclusion and Future Work

Figure 4: Control mechanism p



Probabilistic cardinality constraints were introduced to stipulate lower bounds on the marginal probability by which a maximum number of the same data pattern can occur in sets of uncertain data. As shown in Figure 4 the marginal probability can be used to balance the consistency and completeness targets for the quality of data, enabling organizations to derive more value from it. Axiomatic and algorithmic tools were developed to reason efficiently about probabilistic cardinality constraints. This can help minimize the overhead in using them for data quality purposes or deriving probabilities on the maximum number of query answers without querying any data. These applications are effectively unlocked by developing computational support in the form of probabilistic Armstrong samples for identifying the right marginal probabilities by which cardinality constraints should hold in a given application domain. Analysts and domain experts can jointly inspect Armstrong samples which point out any flaws in the current perception of the marginal probabilities. Our tool *Fortuna* can

[1] Available for download at `https://www.cs.auckland.ac.nz/~tkr/fortuna.html`

be used to generate Armstrong samples for any input, and to explore the possible worlds it represents.

Our results constitute the core foundation for probabilistic cardinality constraints, which can be extended into various directions in future work. It will be interesting to raise the expressivity of probabilistic cardinality constraints by allowing the stipulation of lower bounds on the number of the same data patterns, and/or upper bounds on the marginal probabilities, for examples. For a given PC-table it would be interesting to develop efficient algorithms that compute the marginal probability by which cardinality constraints hold on the data the table represents. Experiments with our implementation are expected to provide further insight into the average case performance of Algorithm 2 in relationship to the worst- and best-cases discussed. Finally, it would be interesting to conduct an empirical investigation into the usefulness of our framework for acquiring the right marginal probabilities of cardinality constraints in a given application domain. This will also require us to extend empirical measures from certain [16, 15] to probabilistic data sets. Particularly intriguing will be the question which of Armstrong PC-sketches and Armstrong p-sketches are actually more useful. While Armstrong PC-sketches are more concise, they may prove to be too concise to draw the attention of analysts and domain experts to critical constraint violations.

Acknowledgement. This research is supported by the Marsden fund council from Government funding, administered by the Royal Society of New Zealand.

References

[1] Beeri, C., Dowd, M., Fagin, R., Statman, R.: On the structure of Armstrong relations for functional dependencies. J. ACM 31(1), 30–46 (1984)

[2] Chen, P.P.: The Entity-Relationship model - toward a unified view of data. ACM Trans. Database Syst. 1(1), 9–36 (1976)

[3] Currim, F., Neidig, N., Kampoowale, A., Mhatre, G.: The CARD system. In: Parsons, J., Saeki, M., Shoval, P., Woo, C.C., Wand, Y. (eds.) Conceptual Modeling - ER 2010, 29th International Conference on Conceptual Modeling, Vancouver, BC, Canada, November 1-4, 2010. Proceedings. Lecture Notes in Computer Science, vol. 6412, pp. 433–437. Springer (2010)

[4] Ferrarotti, F., Hartmann, S., Link, S.: Efficiency frontiers of XML cardinality constraints. Data Knowl. Eng. 87, 297–319 (2013)

[5] Ferrarotti, F., Hartmann, S., Link, S., Marín, M., Muñoz, E.: Soft cardinality constraints on XML data - how exceptions prove the business rule. In: Lin, X., Manolopoulos, Y., Srivastava, D., Huang, G. (eds.) Web Information Systems Engineering - WISE 2013 - 14th International Conference, Nanjing, China, October 13-15, 2013, Proceedings, Part I. Lecture Notes in Computer Science, vol. 8180, pp. 382–395. Springer (2013)

[6] Hartmann, S.: Decomposition by pivoting and path cardinality constraints. In: ER. pp. 126–139 (2000)

[7] Hartmann, S.: On the implication problem for cardinality constraints and functional dependencies. Ann. Math. Artif. Intell. 33(2-4), 253–307 (2001)

[8] Hartmann, S., Kirchberg, M., Link, S.: Design by example for SQL table definitions with functional dependencies. VLDB J. 21(1), 121–144 (2012)

[9] Hartmann, S., Köhler, H., Leck, U., Link, S., Thalheim, B., Wang, J.: Constructing Armstrong tables for general cardinality constraints and not-null constraints. Ann. Math. Artif. Intell. 73(1-2), 139–165 (2015)

[10] Hartmann, S., Link, S.: Efficient reasoning about a robust XML key fragment. ACM Trans. Database Syst. 34(2) (2009)

[11] Hartmann, S., Link, S.: Numerical constraints on XML data. Inf. Comput. 208(5), 521–544 (2010)

[12] Hartmann, S., Link, S.: The implication problem of data dependencies over SQL table definitions: Axiomatic, algorithmic and logical characterizations. ACM Trans. Database Syst. 37(2), 13 (2012)

[13] Jones, T.H., Song, I.Y.: Analysis of binary/ternary cardinality combinations in Entity-Relationship modeling. Data Knowl. Eng. 19(1), 39–64 (1996)

[14] Köhler, H., Link, S., Prade, H., Zhou, X.: Cardinality constraints for uncertain data. In: Yu, E., Dobbie, G., Jarke, M., Purao, S. (eds.) Conceptual Modeling - 33rd International Conference, ER 2014, Atlanta, GA, USA, October 27-29, 2014. Proceedings. Lecture Notes in Computer Science, vol. 8824, pp. 108–121. Springer (2014)

[15] Langeveldt, W.D., Link, S.: Empirical evidence for the usefulness of Armstrong relations in the acquisition of meaningful functional dependencies. Inf. Syst. 35(3), 352–374 (2010)

[16] Le, V.B.T., Link, S., Ferrarotti, F.: Effective recognition and visualization of semantic requirements by perfect SQL samples. In: Ng, W., Storey, V.C., Trujillo, J. (eds.) Conceptual Modeling - 32th International Conference, ER 2013, Hong-Kong, China, November 11-13, 2013. Proceedings. Lecture Notes in Computer Science, vol. 8217, pp. 227–240. Springer (2013)

[17] Lenzerini, M., Nobili, P.: On the satisfiability of dependency constraints in entity-relationship schemata. Inf. Syst. 15(4), 453–461 (1990)

[18] Liddle, S.W., Embley, D.W., Woodfield, S.N.: Cardinality constraints in semantic data models. Data Knowl. Eng. 11(3), 235–270 (1993)

[19] Link, S.: Consistency enforcement in databases. In: Bertossi, L.E., Katona, G.O.H., Schewe, K.D., Thalheim, B. (eds.) Semantics in Databases, Second International Workshop, Dagstuhl Castle, Germany, January 7-12, 2001, Revised Papers. Lecture Notes in Computer Science, vol. 2582, pp. 139–159. Springer (2003)

[20] Link, S.: Armstrong databases: Validation, communication and consolidation of conceptual models with perfect test data. In: Ghose, A., Ferrarotti, F. (eds.) Eighth Asia-Pacic Conference on Conceptual Modelling, APCCM 2012, Melbourne, Australia, January 2012. CRPIT, vol. 130, pp. 3–20. Australian Computer Society (2012)

[21] Liu, J., Li, J., Liu, C., Chen, Y.: Discover dependencies from data - A review. IEEE Trans. Knowl. Data Eng. 24(2), 251–264 (2012)

[22] McAllister, A.J.: Complete rules for n-ary relationship cardinality constraints. Data Knowl. Eng. 27(3), 255–288 (1998)

[23] Queralt, A., Artale, A., Calvanese, D., Teniente, E.: OCL-Lite: Finite reasoning on UML/OCL conceptual schemas. Data Knowl. Eng. 73, 1–22 (2012)

[24] Sadiq, S.: Handbook of Data Quality. Springer (2013)

[25] Suciu, D., Olteanu, D., Ré, C., Koch, C.: Probabilistic Databases. Synthesis Lectures on Data Management, Morgan & Claypool Publishers (2011)

[26] Thalheim, B.: Entity-relationship modeling. Springer (2000)

[27] Thalheim, B.: Integrity constraints in (conceptual) database models. In: The Evolution of Conceptual Modeling. pp. 42–67 (2008)