Probabilistic Cardinality Constraints

Tania Roblot and Sebastian Link(✉)

Department of Computer Science, University of Auckland, Auckland, New Zealand
{tkr,s.link}@auckland.ac.nz

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 for controlling the occurrences of data patterns in probabilistic databases. Our constraints help organizations balance their targets for different data quality dimensions, and infer probabilities on the number of query answers. These 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 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: Data and knowledge visualization · Data models · Database semantics · Management of integrity constraints · 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 [3]. They have attracted significant interest and tool support ever since. Intuitively, a cardinality constraint $card(X) \leq b$ stipulates for an attribute set X and a positive integer b that a relation must not contain more than b different tuples that all have matching values on all the attributes in X. 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

P. Johannesson et al. (Eds.): ER 2015, LNCS 9381, pp. 214–228, 2015.
DOI: 10.1007/978-3-319-25264-3_16

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.

Table 1. Probabilistic relation

w_1 ($p_1 = 0.75$)		
rfid	*time*	*zone*
w2	06	z1
w2	07	z1
w3	15	z7
w3	16	z8
w3	17	z9
w10	10	z11
w11	10	z12
w12	10	z13
w4	11	z3
w5	12	z3
w6	13	z3
w7	14	z3

w_2 ($p_2 = 0.15$)		
rfid	*time*	*zone*
w1	08	z4
w1	08	z5
w1	08	z6
w2	05	z1
w2	06	z1
w2	07	z1
w4	11	z3
w5	12	z3
w6	13	z3
w7	14	z3
w8	09	z2
w9	09	z2

w_3 ($p_3 = 0.1$)		
rfid	*time*	*zone*
w1	08	z4
w1	08	z5
w1	08	z6
w2	05	z1
w2	06	z1
w2	07	z1
w4	11	z3
w5	12	z3
w6	13	z3
w7	14	z3
w8	09	z2
w9	09	z2
w0	09	z2

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 result 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. *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 $(card(rfid) \leq 3, \geq 0.9)$, $(card(zone) \leq 4, \geq 0.9)$ and $(card(rfid, zone) \leq 3, \geq 0.75)$ would be redundant as the enforcement of $(card(rfid, zone) \leq 3, \geq 0.75)$ is already implicitly done by enforcing $(card(rfid) \leq 3, \geq 0.9)$.

(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, Fig. 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.

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

Organization. We discuss related work in Sect. 2. PCCs are introduced in Sect. 3, and reasoning tools for them are established in Sect. 4. These form the

foundation for computational support to acquire the correct marginal probabilities in Sect. 5. We conclude and outline future work in Sect. 6. Due to lack of space, all proofs have been made available in the technical report [16].

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 [3]. 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 [4–9,12,14,15,18]. For a recent survey, see [19].

Probabilistic cardinality constraints $card(X) \leq b$, introduced in this paper, subsume the class of probabilistic keys [2] as the special case where $b = 1$.

For possibilistic cardinality constraints [10], 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. Our research thereby complements the qualitative approach to cardinality constraints in [10] 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 [1,6]. 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 [13]. 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 $card(X) \leq b$ 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.

Implication. Let $\Sigma \cup \{\varphi\}$ denote a finite set of constraints over relation schema R, in particular Σ is always finite. 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. Finite implication problem and implication problem coincide for the class of pCCs, and we thus speak of *the* implication problem.

Axioms. 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. Due to lack of space we omit the soundness and completeness proof of the following theorem, see [16].

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

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

Theorem 1. *$\mathfrak{P}$ forms a finite axiomatization for the implication of probabilistic cardinality constraints.*

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

Algorithms. 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 $\varphi = (card(X) \leq b, \geq p)$ is implied by Σ. The following theorem provides the foundation for the algorithm [16].

Theorem 2. *Let $\Sigma \cup \{(card(X) \leq b, \geq p)\}$ denote a set of pCCs over relation schema R. Then Σ implies $(card(X) \leq b, \geq p)$ if and only if (i) $X = R$ or (ii) $p = 0$ or (iii) $b = \infty$ or (iv) there is some $(card(Z) \leq b', \geq q) \in \Sigma$ such that $Z \subseteq X$, $b' \leq b$, and $q \geq p$.*

Example 3. Continuing Example 2, we can apply Theorem 2 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 2 to imply φ, since $\{rfid\} \subseteq \{rfid, time\}$, $3 \leq 4$, and $1 \geq 0.8$.

Theorem 2 motivates the following algorithm that 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 $(card(Z) \leq b', \geq q) \in \Sigma$ and sets p to q whenever q is larger than the current p, X contains Z and $b' \leq b$. $||\Sigma||$ denotes the total number of attributes together with the logarithm of the integer bounds in Σ. Here, we assume without loss of generality that ∞ does not occur.

Theorem 3. *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 our algorithm 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 3 allows us to decide the associated implication problem efficiently, too. Given $R, \Sigma, (card(X) \leq b, \geq p)$ as an input to the implication problem, we use our algorithm to compute $p' := \max\{q : \Sigma \models card(X) \leq b, \geq q\}$ and return an affirmative answer if and only if $p' \geq p$.

Corollary 1. *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 $\varphi = (card(rfid, time) \leq 4, \geq 0.8)$ since our algorithm returned 1 as the maximum probability for which $card(rfid, time) \leq 4$ 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 semantic integrity constraints that govern the data. For cardinality constraints $(card(X) \leq b, \geq p)$ the "right" marginal probability p and the "right" upper bound b must be identified for a given set X of attributes. Choosing p too big or b too small prevents the entry of clean data, resulting in a lower level of data completeness. Choosing p too small or b too high can lead to the entry of dirty data, resulting in a lower level of data consistency. Analysts benefit from computational support to improve upon their ad-hoc perceptions on an appropriate probability p and bound b.

Goal. Armstrong relations are a useful tool for consolidating the perception of analysts about the cardinality constraints (CCs) of a given application domain. Starting with a set Σ, the tool creates a small relation r that satisfies Σ and violates all CCs not implied by Σ. This property makes r a perfect sample for Σ: any CC is satisfied by the relation if and only if it is implied by Σ.

Our goal is to develop the tool of Armstrong p-relations for a given set Σ of pCCs: the marginal probability by which a traditional constraint $card(X) \leq b$ holds on the Armstrong p-relation is the maximum probability p by which the pCC $(card(X) \leq b, \geq p)$ is implied by Σ. So, if an analyst wants to check for an arbitrary pCC $(card(X) \leq b, \geq p)$ whether it is implied by Σ, she can compute the marginal probability p' by which the CC $card(X) \leq b$ holds on the Armstrong p-relation and verify that $p \geq p'$. For the remainder of this section, we will review Armstrong relations, add new results, and then devise our construction of Armstrong p-relations and more concise representations thereof.

Armstrong Relations. 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 Fig. 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.

Armstrong Sketches. 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. For a practical 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

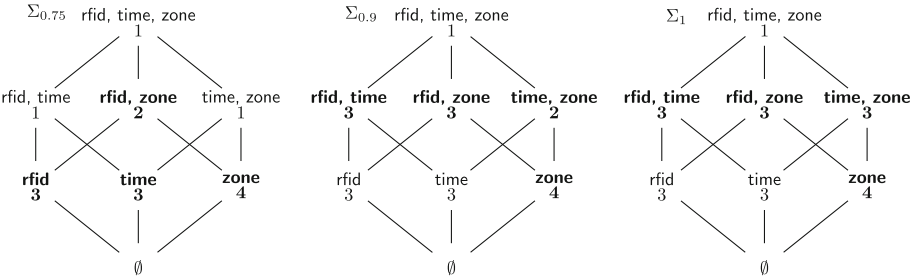

Fig. 1. Duplicate sets X in bold font and their cardinalities b_X for Example 6

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 7. 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

Armstrong p-sketches. 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 8. 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.

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 [16].

Theorem 4. *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 $(card(X) \leq b, \geq l(X,b))$, 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 $(card(X) \leq b, \geq p)$.*

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 [16].

Theorem 5. *Prob. cardinality constraints enjoy Armstrong p-sketches.*

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 [17] 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 [17] 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 [16].

Theorem 6. *Every p-sketch can be represented as a 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 9. Table 2 shows a PC-sketch $\langle \Gamma, \Pi \rangle$ that is Armstrong for the set Σ of pCCs from Example 1.

Algorithm 1 contains the pseudo-code and comments how to compute an Armstrong PC-sketch for any given set Σ of pCCs over any given relation schema R. In particular, line (5) uses the definition of the cardinality $b^i_X := \inf\{b \mid card(Y) \leq b \in \Sigma_{p_i} \wedge Y \subseteq X\}$ to compute them.

Theorem 7. *For every set Σ of pCCs over relation schema R, Algorithm 1 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 $b^1_X = \infty$, and there is only one possible world. Algorithm 1 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 1

Algorithm 1. Armstrong PC-sketch

Require: R, Σ
Ensure: Armstrong PC-sketch $\langle\langle(card, \omega), \iota\rangle, \Pi\rangle$ for Σ
1: Let $p_1 < \cdots < p_n$ be the probabilities in Σ; ▷ If $p_n < 1$, $n \leftarrow n+1$ and $p_n \leftarrow 1$
2: $p_0 \leftarrow 0$; $\Pi \leftarrow \emptyset$;
3: **for** $i = 1, \ldots, n$ **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 $\{b^i_X \mid X \subseteq R\}$; ▷ Smallest upper bound for each X in world i
6: $dup_i \leftarrow$ Set of duplicate sets for Σ_{p_i}; ▷ Duplicate sets to realize in world i
7: $\omega \leftarrow \emptyset$; $k \leftarrow 0$;
8: $dup \leftarrow \{(X, \{i \mid X \in dup_i\}) \mid X \in dup_i \text{ for some } i\}$;
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** $i = 1, \ldots, n$ **do** ▷ Add some τ_k that realizes X in every world in W
12: **if** $X \in dup_i$ **and** $b^i_X > b$ **then** ▷ if there are any remaining cardinalities
13: $k \leftarrow k+1$;
14: **for all** $A \in R$ **do** ▷ Define τ_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: $card(\tau_k) \leftarrow b^i_X - b$; ▷ Stipulate remaining cardinality
21: $\iota(\tau_k) \leftarrow W - \{1, \ldots, i-1\}$; ▷ Worlds that require this cardinality
22: $b \leftarrow b^i_X$; ▷ Mark cardinalities as already realized
23: **return** $\langle\langle(card, \omega), \iota\rangle, \Pi\rangle$;

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

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 $\Sigma_n = \{(card(X_1 \cdots X_n) \leq 1, \geq 1) \mid X_i \in \{A_{2i-1}, A_{2i}\} \text{ for } i = 1, \ldots, n\}$ with $||\Sigma_n|| = n \cdot 2^n$. There is an Armstrong PC-sketch for Σ that contains only one tuple for each of the n duplicate sets $X = R - \{A_{2i-1}, A_{2i}\}$ with associated cardinality $b^1_X = \infty$.

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. This is always possible by our results. We think Armstrong PC-sketches help identify bounds b that are too low or probabilities p that are too high, while the set Σ helps identify bounds b that are too high or probabilities p that are too low.

Graphical User Interface. We have implemented Algorithm 1 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 2 shows a partial screenshot of our GUI *Fortuna* with some outputs for our running example.

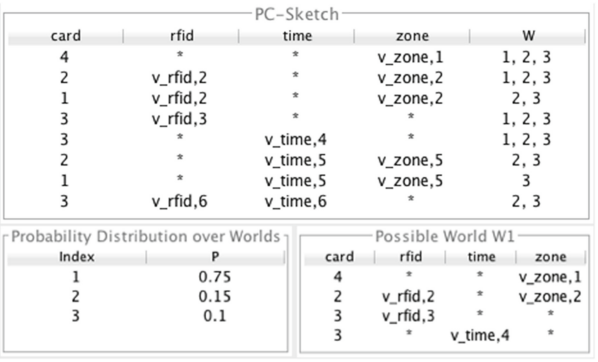

Fig. 2. Screenshot of the GUI *Fortuna*

6 Conclusion and Future Work

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 Fig. 3 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 be used to generate Armstrong samples for any input, and to explore the possible worlds it represents.

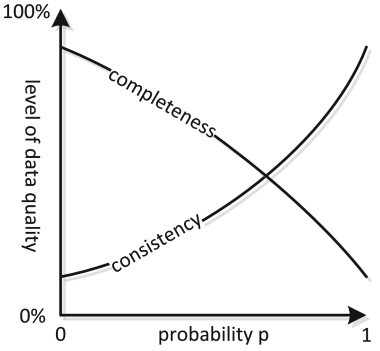

Fig. 3. Control mechanism p

Our results constitute the core foundation for probabilistic cardinality constraints, which can be extended into various directions in future work. It will

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

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 1 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 [11] 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. Brown, P., Link, S.: Probabilistic keys for data quality management. In: Zdravkovic, J., Kirikova, M., Johannesson, P. (eds.) CAiSE 2015. LNCS, vol. 9097, pp. 118–132. Springer, Heidelberg (2015)
3. Chen, P.P.: The Entity-Relationship model - toward a unified view of data. ACM Trans. Database Syst. **1**(1), 9–36 (1976)
4. Currim, F., Neidig, N., Kampoowale, A., Mhatre, G.: The CARD system. In: Parsons, J., Saeki, M., Shoval, P., Woo, C., Wand, Y. (eds.) ER 2010. LNCS, vol. 6412, pp. 433–437. Springer, Heidelberg (2010)
5. Ferrarotti, F., Hartmann, S., Link, S.: Efficiency frontiers of XML cardinality constraints. Data Knowl. Eng. **87**, 297–319 (2013)
6. 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)
7. Hartmann, S., Link, S.: Efficient reasoning about a robust XML key fragment. ACM Trans. Database Syst. 34(2) (2009)
8. Hartmann, S., Link, S.: Numerical constraints on XML data. Inf. Comput. **208**(5), 521–544 (2010)
9. Jones, T.H., Song, I.Y.: Analysis of binary/ternary cardinality combinations in entity-relationship modeling. Data Knowl. Eng. **19**(1), 39–64 (1996)
10. Koehler, H., Link, S., Prade, H., Zhou, X.: Cardinality constraints for uncertain data. In: Yu, E., Dobbie, G., Jarke, M., Purao, S. (eds.) ER 2014. LNCS, vol. 8824, pp. 108–121. Springer, Heidelberg (2014)

11. 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)
12. Liddle, S.W., Embley, D.W., Woodfield, S.N.: Cardinality constraints in semantic data models. Data Knowl. Eng. **11**(3), 235–270 (1993)
13. Liu, J., Li, J., Liu, C., Chen, Y.: Discover dependencies from data - a review. IEEE Trans. Knowl. Data Eng. **24**(2), 251–264 (2012)
14. McAllister, A.J.: Complete rules for n-ary relationship cardinality constraints. Data Knowl. Eng. **27**(3), 255–288 (1998)
15. Queralt, A., Artale, A., Calvanese, D., Teniente, E.: OCL-Lite: Finite reasoning on UML/OCL conceptual schemas. Data Knowl. Eng. **73**, 1–22 (2012)
16. Roblot, T., Link, S.: Probabilistic cardinality constraints. Tech. Rep. 481. https://www.cs.auckland.ac.nz/research/groups/CDMTCS/researchreports/ (2015)
17. Suciu, D., Olteanu, D., Ré, C., Koch, C.: Probabilistic Databases, Synthesis Lectures on Data Management. Morgan and Claypool Publishers, San Rafael (2011)
18. Thalheim, B.: Entity-Relationship Modeling. Springer, Heidelberg (2000)
19. Thalheim, B.: Integrity constraints in (conceptual) database models. In: Kaschek, R., Delcambre, L. (eds.) The Evolution of Conceptual Modeling. LNCS, vol. 6520, pp. 42–67. Springer, Heidelberg (2011)