



**HAL**  
open science

## Cardinality constraints on qualitatively uncertain data

Neil Hall, Hemming Köhler, Sebastian Link, Henri Prade, Xiaofang Zhou

► **To cite this version:**

Neil Hall, Hemming Köhler, Sebastian Link, Henri Prade, Xiaofang Zhou. Cardinality constraints on qualitatively uncertain data. *Data and Knowledge Engineering*, 2015, 99 (Special issue: Selected Papers from the 33rd International Conference on Conceptual Modeling (ER 2014)), pp.126-150. 10.1016/j.datak.2015.06.002 . hal-03516777

**HAL Id: hal-03516777**

**<https://hal.science/hal-03516777>**

Submitted on 7 Jan 2022

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



## Open Archive Toulouse Archive Ouverte


OATAO is an open access repository that collects the work of Toulouse researchers and makes it freely available over the web where possible

This is an author's version published in:

<http://oatao.univ-toulouse.fr/24807>

### Official URL

DOI : <https://doi.org/10.1016/j.datak.2015.06.002>

**To cite this version:** Hall, Neil and Köhler, Hemming and Link, Sebastian and Prade, Henri  and Zhou, Xiaofang *Cardinality constraints on qualitatively uncertain data*. (2015) *Data and Knowledge Engineering*, 99. 126-150. ISSN 0169-023X

Any correspondence concerning this service should be sent to the repository administrator: [tech-oatao@listes-diff.inp-toulouse.fr](mailto:tech-oatao@listes-diff.inp-toulouse.fr)

# Cardinality constraints on qualitatively uncertain data

Neil Hall<sup>a</sup>, Henning Koehler<sup>b</sup>, Sebastian Link<sup>a,\*</sup>, Henri Prade<sup>c</sup>, Xiaofang Zhou<sup>d,e</sup>

<sup>a</sup> Department of Computer Science, The University of Auckland, New Zealand

<sup>b</sup> School of Engineering & Advanced Technology, Massey University, New Zealand

<sup>c</sup> IRIT, CNRS Université de Toulouse III, France

<sup>d</sup> School of Information Technology and Electrical Engineering, The University of Queensland, Australia

<sup>e</sup> Soochow University, Suzhou, China

## A B S T R A C T

Modern applications require advanced techniques and tools to process large volumes of uncertain data. For that purpose we introduce cardinality constraints as a principled tool to control the occurrences of uncertain data. Uncertainty is modeled qualitatively by assigning to each object a degree of possibility by which the object occurs in an uncertain instance. Cardinality constraints are assigned a degree of certainty that stipulates on which objects they hold. Our framework empowers users to model uncertainty in an intuitive way, without the requirement to put a precise value on it. Our class of cardinality constraints enjoys a natural possible world semantics, which is exploited to establish several tools to reason about them. We characterize the associated implication problem axiomatically and algorithmically in linear input time. Furthermore, we show how to visualize any given set of our cardinality constraints in the form of an Armstrong sketch. Even though the problem of finding an Armstrong sketch is precisely exponential, our algorithm computes a sketch with conservative use of time and space. Data engineers may therefore compute Armstrong sketches that they can jointly inspect with domain experts in order to consolidate the set of cardinality constraints meaningful for a given application domain.

### Keywords:

Data and knowledge visualization

Data models

Database semantics

Management of integrity constraints

Requirements engineering

## 1. Introduction

### 1.1. Background

The notion of cardinality constraints is fundamental for understanding the structure and semantics of data. In traditional conceptual modeling, cardinality constraints were introduced in Chen's seminal paper [7]. 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 an instance if there are no  $b + 1$  distinct objects in the instance that have matching values on all the attributes of the 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 therefore have significant applications in data cleaning, integration, modeling, processing, and retrieval.

### 1.2. Motivation

Traditional conceptual modeling was targeted at certain data for applications such as accounting, inventory and payroll. Modern applications, such as information extraction, radio-frequency identification (RFID), scientific data management, data cleaning, and

\* Corresponding author.

E-mail addresses: hall.neil@gmail.com (N. Hall), h.koehler@massey.ac.nz (H. Koehler), s.link@auckland.ac.nz (S. Link), prade@irit.fr (H. Prade), zxf@itee.uq.edu.au (X. Zhou).

financial risk assessment produce large volumes of uncertain data. For example, RFID can track movements of endangered species of animals, such as the Indiana bat in Georgia, USA. For such an application, data comes in the form of objects associated with some discrete level of confidence in the signal reading; for example based on the quality of the signal received. More generally, uncertainty can be modeled qualitatively by associating objects with the degree of possibility (p-degree) that the object is perceived to occur in the instance. Fig. 1 shows such a possibilistic instance (p-instance), where each object is associated with an element from a finite scale of p-degrees:  $\alpha_1 > \dots > \alpha_{k+1}$ . The top degree  $\alpha_1$  is reserved for objects that are 'fully possible', the bottom degree  $\alpha_{k+1}$  for objects that are 'impossible' to occur. Intermediate degrees are used as required and linguistic interpretations attached as preferred, such as 'quite possible' ( $\alpha_2$ ) and 'somewhat possible' ( $\alpha_3$ ).

As this scenario is typical for a broad range of applications, we investigate in this article how cardinality constraints can benefit from the p-degrees assigned to objects. More specifically, we investigate cardinality constraints on uncertain data, where uncertainty is modeled qualitatively in the form of p-degrees.

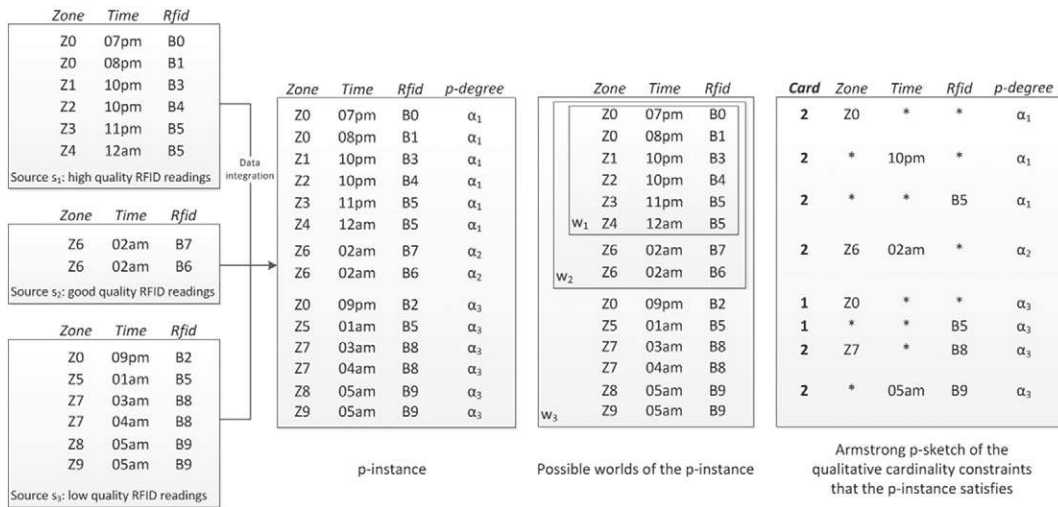
The degrees of possibility are a natural source for extending the expressivity of traditional cardinality constraints. In fact, our use of p-degrees enjoys a natural possible world semantics, as illustrated on the running example in Fig. 1. Here, the world  $w_1$  contains the RFID readings of high quality only, that is, all the objects with p-degree  $\alpha_1$ . The world  $w_2$  contains RFID readings of high or good quality, that is, all the objects with p-degree  $\alpha_1$  or  $\alpha_2$ . Finally, world  $w_3$  contains RFID readings of high, good, or low quality, that is, all the objects with p-degree  $\alpha_1, \alpha_2$  or  $\alpha_3$ . This possible world semantics enables us to express traditional cardinality constraints with different degrees of certainty. The certainty by which a traditional cardinality constraint holds is derived from the possible worlds in which it holds.

For example, we can express that for all low, good, and high quality readings, there are at most three readings recorded in the same zone, by declaring the cardinality constraint  $card(Zone) \leq 3$  to be 'fully certain'. That is,  $card(Zone) \leq 3$  must hold in the largest possible world  $w_3$ , and therefore also in all the worlds it contains. Similarly, we can express that for all good and high quality readings, at most two bats are recorded in the same zone at the same time, by declaring the cardinality constraint  $card(Zone, Time) \leq 2$  to be 'quite certain'. That is,  $card(Zone, Time) \leq 2$  must hold in the second largest possible world  $w_2$ , but not necessarily in the largest world  $w_3$ . Finally, we can express that for all high quality readings, the zone and time together identify the bat, by declaring the cardinality constraint  $card(Zone, Time) \leq 1$  to be 'somewhat certain'. That is,  $card(Zone, Time) \leq 1$  must hold in the smallest possible world  $w_1$ , but not necessarily in the worlds  $w_2$  or  $w_3$ .

### 1.3. Contributions

Our objective is to apply possibility theory from artificial intelligence to establish qualitative cardinality constraints (QCs) as a fundamental tool to control the occurrences of uncertain data. Our contributions can be summarized as follows:

- **Modeling.** We introduce qualitative cardinality constraints as a class of integrity constraints on uncertain data. Here, uncertainty is modeled qualitatively by assigning to each object a degree of possibility with which it occurs in the instance. The p-degrees bring forward a nested chain of possible worlds, with each world being a classic instance that has some possibility. Hence, the higher the possibility of a world the fewer objects it contains. This empowers us to assign degrees of certainty to cardinality constraints, stipulating to which possible worlds they apply. The degrees of certainty (c-degree) are usually denoted by  $\beta_1 > \dots > \beta_k > \beta_{k+1}$ , where  $\beta_{k+1}$  denotes the bottom c-degree reserved for constraints that are satisfied by any p-instance. Cardinality constraints that apply to the largest possible world hold with 'full certainty', denoted by the top c-degree  $\beta_1$ , while cardinality constraints that apply to the smallest possible world are only 'somewhat certain' to hold, denoted by the c-degree  $\beta_k$ . Fig. 1 shows the possible



**Fig. 1.** P-instance and its possible worlds as the result of integrating RFID readings of different qualities; Armstrong p-sketch of the qualitative cardinality constraints that the p-instance satisfies.

worlds  $w_1$ ,  $w_2$  and  $w_3$  of our running example. Here,  $w_2$  satisfies  $\text{card}(\text{Zone}, \text{Time}) \leq 2$  but violates  $\text{card}(\text{Zone}, \text{Time}) \leq 1$ , which holds only on world  $w_1$ .

- Reasoning.** We establish axiomatic and algorithmic solutions to the implication problem associated with qualitative cardinality constraints. The implication problem is to decide, for any given qualitative cardinality constraint and any set of such constraints over a given object type, whether the constraint is implied by the set, that is, whether every instance over the object type that satisfies every element of the set also satisfies the constraint. Technically, the algorithmic solution to the implication problem is derived from a linear-time characterization of the inference problem, where one must compute for any given traditional cardinality constraint and any given set of qualitative cardinality constraints on any given object type, the highest degree of certainty with which the constraint is implied by the given set. Our algorithmic solution allows us to detect and remove any redundant constraints from a given set, thereby reducing the number of cardinality constraints that must actively be enforced on given sets of objects to a minimal level necessary. This ability results in time savings proportional to the size of the data sets. That means our solutions empower us to efficiently enforce many desirable properties of uncertain data arising from modern application domains. For example, Fig. 2 shows a cover for the qualitative cardinality constraints that the p-instance from Fig. 1 satisfies. In the figure a qualitative cardinality constraint ( $\text{card}(X) \leq b, \beta$ ) is represented as follows: the cube under the c-degree  $\beta$  features all attribute sets  $X$  with the minimal upper bound  $b$  that applies to them. The constraints that form a minimal cover  $\Sigma$  for the set of all qualitative cardinality constraints satisfied by the p-instance are shown in bold font. For example, the constraint ( $\text{card}(\text{Time}) \leq 2, \beta_1$ ) is implied by the constraint ( $\text{card}(\text{Time}) \leq 2, \beta_1$ ). We also show that our findings cannot only be used to control the integrity of uncertain data, but also have interesting applications to query processing. For example, our solution of the implication problem can be used to compute upper bounds on the number of query answers without actually having to query the potentially big data set.
- Acquisition.** The benefits of applying qualitative cardinality constraints effectively to an application are inhibited by the difficulty of identifying those constraints that actually hold on the domain of the application. In the idealized special case where only fully certain objects occur in the data, it is already difficult to identify the correct upper bound. In the context of uncertain data, the problem becomes even more intricate as the correct degree of certainty has to be identified for any given upper bound. It is therefore important to provide computational support to business analysts who need to discover meaningful qualitative cardinality constraints. For this purpose, we investigate Armstrong samples for the class of qualitative cardinality constraints. A p-instance is said to be Armstrong for a given set of qualitative cardinality constraints if and only if for every qualitative cardinality constraint, it holds that it is implied by the given set if and only if it is satisfied by the p-instance. An Armstrong p-instance therefore tells us the highest degree of certainty by which a qualitative cardinality constraint is implied by the given set. While there are sets of qualitative cardinality constraints which require every Armstrong p-instance to be infinite, we show that every set of qualitative cardinality constraints enjoys Armstrong p-sketches, which are finite representations of potentially infinite Armstrong p-instances. Even though the problem of finding an Armstrong p-sketch is precisely exponential, we establish an algorithm that computes an Armstrong p-sketch with conservative use of time and space. Business analysts may therefore compute Armstrong p-sketches that they can jointly inspect with domain experts in order to consolidate the correct degree of certainty with which cardinality constraints should hold in the given application domain. For example, the p-instance from Fig. 1 is a finite Armstrong p-instance for the set  $\Sigma$  of qualitative cardinality constraints above. An Armstrong p-sketch for  $\Sigma$  is shown on the right of Fig. 1. Although p-sketches are mostly useful to finitely represent infinite Armstrong p-instances, they are also more concise representations of finite Armstrong p-instances. They are more concise as they require fewer objects and focus the attention of the people who inspect them on only the relevant patterns of data. For instance, the row (**Card**:2, Zone:Z7, Time:\*, RFID:R8,  $\alpha_3$ ) summarizes the fact that any p-instance that the p-sketch represents must feature two different objects that both have the value Z7 on Zone, each have unique values on Time, and both have the value R8 on RFID, and both have associated p-degree  $\alpha_3$ .
- Tool support and experiments.** We implemented our algorithm for computing Armstrong p-sketches in a prototype system, and conducted several experiments regarding the size of the output and the time to compute the sketches. Our prototype successfully transfers the concept of Armstrong p-sketches from theory into practice. Our results suggest that Armstrong p-sketches, as computed by our prototype, are small enough for effective use during the requirements acquisition phase, and can be computed very quickly.

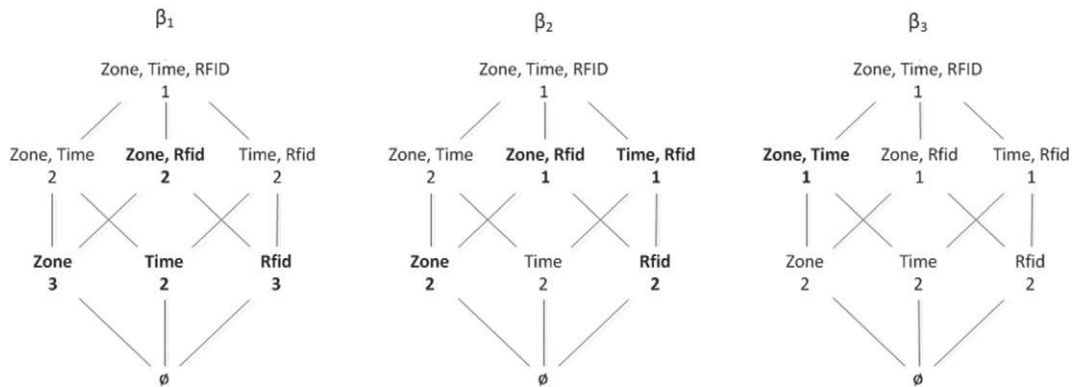


Fig. 2. Minimal cover for the set of qualitative cardinality constraints that the p-instance from Fig. 1 satisfies.

In summary, we introduce a new class of cardinality constraint that is useful in terms of i) expressing the semantics of uncertain data within a given application domain, ii) the small effort required to reason about them and process updates and queries more efficiently, and iii) the computational support available to acquire them.

#### 1.4. Organization

In Section 2 we summarize briefly the vast research on cardinality constraints from the community. This points out the lack of qualitative approaches to constraints on uncertain data. We propose a semantics for qualitative cardinality constraints on instances of uncertain data in Section 3. In Section 4 we establish axiomatic and linear-time algorithmic characterizations for the associated implication problem of qualitative cardinality constraints, and linear-time algorithmic characterizations for the associated inference problem. This section also features two applications of our results. The first application is an algorithm for computing a minimal cover for a given set of qualitative cardinality constraints, which can be used to determine a minimal set of constraints that must be enforced when updates are processed. The second application illustrates the usefulness of our results for processing queries. Section 5 details how to visualize arbitrary sets of qualitative cardinality constraints in the form of Armstrong p-sketches. While the problem of finding an Armstrong p-sketch is shown to be precisely exponential, our computed Armstrong p-sketch is always at most quadratic in the size of a minimum-sized Armstrong p-sketch and the given set of constraints. We briefly present our prototype system in Section 6, and the results of our experiments in Section 7. In Section 8 we conclude and discuss future work.

## 2. Related work

Cardinality constraints are one of the most influential contributions conceptual modeling has made to the study of database constraints. They were present in Chen's seminal paper [7] 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, or OWL. Cardinality constraints have been extensively studied in database design [1,6,8,14,15,18,19,23–25,27,31,32,35,37,42,43]. For a recent survey, see [44].

There are many quantitative approaches to uncertain data, foremost probability theory [41]. Research about constraints on probabilistic data is still in its infancy [5,26,38]. Qualitative approaches to uncertain data deal with either query languages or extensions of functional dependencies [4]. In [28] we introduced the class of possibilistic keys on qualitatively uncertain data, established axiomatic and algorithmic characterizations of their associated implication problem, and showed how to construct finite Armstrong p-instances for them. Possibilistic keys can be expressed by qualitative cardinality constraints of the form  $(card(X) \leq 1, \beta)$ . In contrast to qualitative cardinality constraints, Armstrong p-instances for any set of possibilistic keys are guaranteed to be finite. *Qualitative approaches to cardinality constraints on uncertain data have not been studied yet to the best of our knowledge.* Our contributions extend results on cardinality constraints from traditional conceptual modeling, covered by the special case of two degrees of possibility. These include findings on the implication problem and Armstrong databases [20]. The definition of Armstrong p-sketches as finite representations of potentially infinite Armstrong p-instances is original.

Possibilistic logic is a well-established tool for reasoning about uncertainty [9,12] with numerous applications in artificial intelligence [11], including approximate reasoning [45], non-monotonic reasoning [16], qualitative reasoning [40], belief revision [10,17,36], soft constraint satisfaction problems [3], decision-making under uncertainty [39], and pattern classification and preferences [2]. Our results show that possibilistic logic is suitable to extend the classical notion of cardinality constraints from certain to qualitatively uncertain data.

The current article is an extended version of the conference paper [29]. The extensions are manifold. 1) We introduce the new concept of Armstrong p-sketches. This concept is highly useful for the discovery of meaningful qualitative cardinality constraints because finite Armstrong p-sketches exist for any given set of these constraints. In contrast, there are sets of qualitative cardinality constraints that require infinite Armstrong p-instances. The conference paper [29] was restricted to the study of Armstrong p-instances only. As these are only finite in cases where for each underlying attribute some finite upper bound has been specified with full certainty, the use of Armstrong p-instances is limited in practice. Even in the case where a finite Armstrong p-instance does exist, an Armstrong p-sketch still provides a more concise summary. 2) We study the inference problem of qualitative cardinality constraints, which has not been considered in previous research. Our linear-time solution to this problem can also be used to decide the associated implication problem in linear time in the input. 3) We have included some applications of our results, notably an algorithm to compute some minimal cover for a given set of qualitative cardinality constraints and can be applied to enforce cardinality constraints on uncertain data without redundancy; as well as an example that illustrates the applications of our findings to query processing and cardinality estimation. 4) While the conference paper [29] did not include any proofs, we include all proofs in the current article. This not only makes it possible to understand the validity of our results and algorithms, but also provides all details for the techniques and constructions we establish. 5) A detailed running example is provided to illustrate the concepts and findings throughout the article. Our proofs and findings may become more accessible for the reader, or at least provide a showcase to which they are applied. 6) We have transferred the concept of Armstrong p-sketches from theory into practice by implementing our algorithm in a prototype system. 7) We conducted several experiments with our prototype, showing that Armstrong p-sketches are small enough to use them successfully during the requirements acquisition phase and can be computed very quickly with our prototype.

### 3. Qualitative cardinality constraints

In this section we extend object types that model certain objects in traditional conceptual modeling to model uncertain objects qualitatively. Based on our model to attribute to each object a degree of possibility with which it occurs, we can attribute degrees of certainty to traditional cardinality constraints that say to which objects they apply.

We start by recalling some basic definitions of attributes, object types, objects and their projections, and instances. An object type, denoted by  $O$ , is a finite non-empty set of *attributes*. Each attribute  $A \in O$  has a *domain*  $dom(A)$  of values. An *object*  $o$  over  $O$  is an element of the Cartesian product  $\prod_{A \in O} dom(A)$ . For  $X \subseteq O$  we denote by  $o(X)$  the *projection* of  $o$  on  $X$ . An *instance* over  $O$  is a set  $\iota$  of objects over  $O$ . Note that an instance may be infinite.

As our running example we use the object type `TRACKING` with attributes `Zone`, `Time`, and `Rfid`. Objects either belong or do not belong to an instance. For example, we cannot express that we have less confidence for the bat identified by `Rfid` value `B5` to be in `Zone Z5` at `01 am` than for the same bat to be in `Z4` at `12 am`.

We model uncertain instances by assigning to each object some degree of possibility with which the object occurs in an instance. Formally, we have a *possibility scale*, that is, a finite strict linear order  $\mathcal{S} = (S, <)$  with  $k + 1$  elements, denoted by  $\alpha_1 > \dots > \alpha_k > \alpha_{k+1}$ . The elements  $\alpha_i \in S$  are called *possibility degrees*, or p-degrees for short. Here,  $\alpha_1$  is reserved for objects that are ‘fully possible’ to occur, while  $\alpha_{k+1}$  is reserved for objects that are ‘impossible’ to occur in an instance, and any intermediate p-degree might linguistically be interpreted by some graded version of possibility such as ‘somewhat possible’ or ‘rather possible’. Of course, a linguistic interpretation is not necessary at all. The use of a specific possibility scale should simply reflect the requirements of an organization to distinguish between different degrees of possibility with which it perceives its data to occur. In our running example, we choose  $k = 3$  and interpret  $\alpha_1$  as ‘fully possible’,  $\alpha_2$  as ‘quite possible’,  $\alpha_3$  as ‘somewhat possible’, and  $\alpha_4$  as ‘impossible’, reflecting the perceived quality of the RFID readings. We point out that humans like to use simple scales in everyday life to communicate, compare, or rank. Here, simple means to classify items qualitatively rather than quantitatively by putting precise values on them. Finally, we point out that classical instances are subsumed by the special case where  $k = 1$ . Here, objects that are assigned p-degree  $\alpha_1$  are the objects of the instance, while all objects that do not occur in the instance are assumed to be assigned p-degree  $\alpha_2$ . As we demonstrate below, objects that are assigned the top p-degree  $\alpha_1$  are not just ‘fully possible’ to occur, but in fact, ‘fully certain’ to occur as well. Therefore, classical instances are a special case of uncertain instances.

A *possibilistic object type*  $(O, S)$ , or p-object type, consists of an object type  $O$  and a possibility scale  $S$ . A *possibilistic instance*, or p-instance, over  $(O, S)$  consists of an instance  $\iota$  over  $O$ , and a function  $Poss_\iota$  that assigns to each object  $o \in \iota$  a p-degree  $Poss_\iota(o) \in S - \{\alpha_{k+1}\}$ . We sometimes omit  $Poss$  when denoting a p-instance. Fig. 1 shows a p-instance over  $(\text{TRACKING}, S = \{\alpha_1, \dots, \alpha_4\})$ .

P-instances enjoy a possible world semantics. For  $i = 1, \dots, k$  let  $w_i$  consist of all objects in  $\iota$  that have p-degree at least  $\alpha_i$ , that is,  $w_i = \{o \in \iota \mid Poss_\iota(o) \geq \alpha_i\}$ . Indeed, we have  $w_1 \subseteq w_2 \subseteq \dots \subseteq w_k$ . The possibility distribution  $\pi_\iota$  for this linear chain of possible worlds is defined by  $\pi_\iota(w_i) = \alpha_i$ . Note that  $w_{k+1}$  is not a possible world, since its p-degree  $\pi_\iota(w_{k+1}) = \alpha_{k+1}$  means ‘impossible’. Vice versa,  $Poss_\iota(o)$  for an object  $o \in \iota$  is the maximum p-degree  $\max\{\alpha_i \mid o \in w_i\}$  of a world to which  $o$  belongs. If  $o \notin w_k$ , then  $Poss_\iota(o) = \alpha_{k+1}$ . Every object that is ‘fully possible’ occurs in every possible world, and is therefore also ‘fully certain’. Hence, instances are a special case of uncertain instances. Fig. 1 shows the possible worlds  $w_1 \subseteq w_2 \subseteq w_3$  of the p-instance in the same figure.

We introduce qualitative cardinality constraints, or QCs, as cardinality constraints that have some associated degree of certainty. As cardinality constraints are fundamental to applications with certain data, QCs will serve a similar role for applications with uncertain data. A *cardinality constraint* over object type  $O$  is an expression  $card(X) \leq b$  where  $\emptyset \neq X \subseteq O$  and  $b$  is a positive integer. The cardinality constraint  $card(X) \leq b$  over  $O$  is satisfied by an instance  $w$  over  $O$ , denoted by  $\models_w card(X) \leq b$ , if there are no  $b + 1$  distinct objects  $o_1, \dots, o_{b+1} \in w$  with matching values on all the attributes in  $X$ . For example, Fig. 1 shows that  $card(\text{Zone}, \text{Time}) \leq 1$  is not satisfied by any instance  $w_1, w_2$  or  $w_3$ ;  $card(\text{Zone}, \text{Time}) \leq 1$  is satisfied by  $w_1$ , but not by  $w_2$  nor  $w_3$ ;  $card(\text{Rfid}) \leq 2$  is satisfied by  $w_1$  and  $w_2$ , but not by  $w_3$ ; and  $card(\text{Rfid}) \leq 3$  is satisfied by  $w_1, w_2$  and  $w_3$ .

The p-degrees of objects result in degrees of certainty by which QCs hold. As  $card(\text{Rfid}) \leq 3$  holds in every possible world, it is ‘fully certain’ to hold on  $\iota$ . As  $card(\text{Rfid}) \leq 2$  is only violated in a ‘somewhat possible’ world  $w_3$ , it is ‘quite certain’ to hold on  $\iota$ . As the smallest world that violates  $card(\text{Zone}, \text{Time}) \leq 1$  is the ‘quite possible’ world  $w_2$ , it is ‘somewhat certain’ to hold on  $\iota$ . As  $card(\text{Zone}) \leq 1$  is violated in the ‘fully possible’ world  $w_1$ , it is ‘not certain at all’ to hold on  $\iota$ .

Similar to the scale  $S$  of p-degrees  $\alpha_i$  for objects we use a scale  $S^T$  of certainty degrees  $\beta_j$ , or c-degrees, for cardinality constraints. As indicated in the last paragraph, we use ‘fully certain’, ‘quite certain’, ‘somewhat certain’, and ‘not certain at all’ in our running example. Formally, the correspondence between p-degrees in  $S$  and the c-degrees in  $S^T$  is defined by the mapping  $\alpha_i \mapsto \beta_{k+2-i}$  for  $i = 1, \dots, k + 1$ . Hence, the certainty  $C_\iota(card(X) \leq b)$  by which the cardinality constraint  $card(X) \leq b$  holds on the uncertain instance  $\iota$  is either the top degree  $\beta_1$  if  $card(X) \leq b$  is satisfied by  $w_k$ , or the minimum among the c-degrees  $\beta_{k+2-i}$  that correspond to possible worlds  $w_i$  in which  $card(X) \leq b$  is violated, that is,

$$C_\iota(card(X) \leq b) = \begin{cases} \beta_1 & , \text{if } \models_{w_k} card(X) \leq b \\ \min\{\beta_{k+2-i} \mid \not\models_{w_i} card(X) \leq b\} & , \text{otherwise} \end{cases}$$

When  $\iota$  denotes the p-instance from Fig. 1, then the c-degree  $C_\iota(card(\text{Rfid}) \leq 3)$  is  $\beta_1$  as  $card(\text{Rfid}) \leq 3$  is even satisfied in the world  $w_3$ . Similarly, the c-degree  $C_\iota(card(\text{Rfid}) \leq 2)$  is  $\beta_2$  as the smallest possible world that violates  $card(\text{Rfid}) \leq 2$  is  $w_3$ . The c-degree  $C_\iota(card(\text{Zone}, \text{Time}) \leq 1)$  is  $\beta_3$  as the smallest possible world that violates  $card(\text{Zone}, \text{Time}) \leq 1$  is  $w_2$ . Finally, the c-degree  $C_\iota(card(\text{Zone}) \leq 1)$  is  $\beta_4$  as the smallest possible world that violates  $card(\text{Zone}) \leq 1$  is  $w_1$ .

We can now define the syntax and semantics of qualitative cardinality constraints.

**Definition 1.** Let  $(O, S)$  denote a  $p$ -object type. A qualitative cardinality constraint (QC) over  $(O, S)$  is an expression  $(\text{card}(X) \leq b, \beta)$  where  $\text{card}(X) \leq b$  denotes a cardinality constraint over  $O$  and  $\beta \in S^T$ . A  $p$ -instance  $(\iota, \text{Poss}_\iota)$  over  $(O, S)$  satisfies the QC  $(\text{card}(X) \leq b, \beta)$  if and only if  $C_i(\text{card}(X) \leq b) \geq \beta$ .

Qualitative cardinality constraints form a class of integrity constraints tailored to uncertain data. Indeed, a QC  $(\text{card}(X) \leq b, \beta_i)$  separates semantically meaningful from meaningless  $p$ -relations by allowing violations of the cardinality constraint  $\text{card}(X) \leq b$  only by objects with a  $p$ -degree  $\alpha_j$  where  $j \leq k + 1 - i$ . For  $i = 1, \dots, k$ , the  $c$ -degree  $\beta_i$  of  $(\text{card}(X) \leq b, \beta_i)$  means that the cardinality constraint  $\text{card}(X) \leq b$  must hold in the possible world  $w_{k+1-i}$ . This constitutes a conveniently flexible mechanism to enforce the targeted level of integrity effectively.

**Example 1.** Let  $\Sigma$  denote the set consisting of the following qualitative cardinality constraints:  $(\text{card}(\text{Zone}) \leq 3, \beta_1)$ ,  $(\text{card}(\text{Time}) \leq 2, \beta_1)$ ,  $(\text{card}(\text{Rfid}) \leq 3, \beta_1)$ ,  $(\text{card}(\text{Zone}, \text{Rfid}) \leq 2, \beta_1)$ ,  $(\text{card}(\text{Zone}) \leq 2, \beta_2)$ ,  $(\text{card}(\text{Rfid}) \leq 2, \beta_2)$ ,  $(\text{card}(\text{Zone}, \text{Rfid}) \leq 1, \beta_2)$ ,  $(\text{card}(\text{Time}, \text{Rfid}) \leq 1, \beta_2)$ ,  $(\text{card}(\text{Zone}, \text{Time}) \leq 1, \beta_3)$ . The  $p$ -instance  $\iota$  from Table 3 satisfies all of these QCs. However,  $\iota$  violates  $(\text{card}(\text{Rfid}) \leq 2, \beta_1)$ ,  $(\text{card}(\text{Rfid}) \leq 1, \beta_2)$ , and  $(\text{card}(\text{Zone}, \text{Time}) \leq 1, \beta_2)$ .

#### 4. Reasoning about qualitative cardinality constraints

In this section we will establish tools to reason about qualitative cardinality constraints. These subsume existing tools for the reasoning about traditional cardinality constraints as the special case where only two  $p$ -degrees are used. We will introduce implication and inference problems as core problems associated with the reasoning about qualitative cardinality constraints. We will then establish a theorem that allows us to reduce any instance of the implication problem for qualitative cardinality constraints to an instance of the implication problem for traditional cardinality constraints. This result will be used to establish a finite axiomatization for the implication of qualitative cardinality constraints by a simple set of Horn axioms. These axioms will allow us to establish a linear-time algorithm for solving the inference problem, which can also be used to decide the implication problem in linear time. Finally, we will show that efficient integrity enforcement and query processing are two major areas in which our results can be applied.

##### 4.1. Implication and inference problems

We first define two core problems associated with the reasoning about qualitative cardinality constraints. For this purpose, let  $\Sigma \cup \{\varphi\}$  denote a set of QCs over  $(O, S)$ . As we will show later, we can always assume without loss of generality that this set is finite. We say that  $\Sigma$  (finitely) implies  $\varphi$ , denoted by  $\Sigma \models_{(f)} \varphi$ , if and only if every (finite)  $p$ -instance  $(\iota, \text{Poss}_\iota)$  over  $(O, S)$  that satisfies every QC in  $\Sigma$  also satisfies  $\varphi$ . In other words, there is no (finite)  $p$ -instance  $(\iota, \text{Poss}_\iota)$  over  $(O, S)$  that satisfies every QC in  $\Sigma$  but violates  $\varphi$ . We use  $\Sigma_{(f)}^* = \{\varphi \mid \Sigma \models_{(f)} \varphi\}$  to denote the (finite) semantic closure of  $\Sigma$ . The (finite) implication problem for QCs is to decide, given any  $p$ -object type, and any set  $\Sigma \cup \{\varphi\}$  of QCs over the  $p$ -object type, whether  $\Sigma \models_{(f)} \varphi$  holds.

PROBLEM:	(Finite) Implication problem for qualitative cardinality constraints
INPUT:	Object type $(O, S)$ , Finite set $\Sigma \cup \{\varphi\}$ of QCs over $(O, S)$
OUTPUT:	Yes, if $\Sigma \models_{(f)} \varphi$ ; No, otherwise

Our first observation is that the finite implication problem and the implication problem coincide for the class of qualitative cardinality constraints. That is, for every object type and every set  $\Sigma \cup \{\varphi\}$  of QCs over that object type, it is true that  $\Sigma \models \varphi$  if and only if it is true that  $\Sigma \models_{(f)} \varphi$ .

**Theorem 1.** Finite and unrestricted implication problem coincide for the class of qualitative cardinality constraints.

**Proof.** Let  $\Sigma \cup \{\varphi\}$  denote a finite set of QCs over object type  $(O, S)$ .

If  $\Sigma$  implies  $\varphi$ , then it follows immediately that  $\Sigma$  finitely implies  $\varphi$  since every finite  $p$ -instance is also a  $p$ -instance.

It remains to show the following: if  $\Sigma$  does not imply  $\varphi$ , then  $\Sigma$  does not finitely imply  $\varphi$ . Let  $\varphi = (\text{card}(X) \leq b, \geq \beta_i)$  and suppose that  $\Sigma \not\models \varphi$ . Hence, there must be some (possibly infinite)  $p$ -instance  $(\iota, \text{Poss}_\iota)$  over  $(O, S)$  that satisfies all QCs in  $\Sigma$  and violates  $\varphi$ . Consequently, there must be  $b + 1$  distinct objects  $o_1, \dots, o_{b+1} \in \iota$  such that  $\text{Poss}_\iota(o_j) \geq \alpha_{k+1-i}$  holds for all  $j = 1, \dots, b + 1$ , and  $o_i(X) = o_j(X)$  holds for all  $1 \leq i \leq j \leq b + 1$ . Let  $(\iota_f, \text{Poss}_{\iota_f})$  denote the finite  $p$ -instance over  $(O, S)$  where  $\iota_f = \{o_1, \dots, o_{b+1}\}$  and  $\text{Poss}_{\iota_f}(o_j) = \text{Poss}_\iota(o_j)$  for all  $j = 1, \dots, b + 1$ . By construction,  $(\iota_f, \text{Poss}_{\iota_f})$  is finite and violates  $\varphi$ . In addition,  $(\iota_f, \text{Poss}_{\iota_f})$  also satisfies every QC in  $\Sigma$  since  $\iota_f \subseteq \iota$  holds and  $(\iota, \text{Poss}_\iota)$  satisfies every QC in  $\Sigma$ . We have just shown that  $\Sigma$  does not finitely imply  $\varphi$ , which completes the proof.  $\square$



Theorem 1 allows us to speak of the implication problem of qualitative cardinality constraints.

**Example 2.** Let  $\Sigma$  be as in Example 1. Further, let  $\varphi$  denote the QC  $(\text{card}(\text{Rfid}) \leq 2, \beta_1)$ . Then  $\Sigma$  does not imply  $\varphi$  as the following p-instance witnesses:

Zone	Time	Rfid	Poss. degree
Z3	11 pm	B5	$\alpha_1$
Z4	12 am	B5	$\alpha_1$
Z5	01 am	B5	$\alpha_3$

We now return to our previous claim that we can assume without loss of generality that a given set  $\Sigma$  of qualitative cardinality constraints over a given object type is finite. In a nutshell, for each fixed attribute set  $X$  and each c-degree  $\beta_i$  it only matters which smallest upper bound  $b_X^i$  is given to us. If  $\Sigma$  is infinite, then there must be some infinite subset of  $\Sigma$  of the form  $\Sigma_{X,i} = \{(\text{card}(X) \leq b_j, \beta_j) \in \Sigma\}$ . In this case, however, we can replace  $\Sigma_{X,i}$  in  $\Sigma$  by the singleton  $(\text{card}(X) \leq b_X^i, \beta_i)$  where  $b_X^i = \min\{b_j \mid (\text{card}(X) \leq b_j, \beta_j) \in \Sigma_{X,i}\}$ . If  $\Sigma_f$  denotes the result of replacing for every non-empty  $X \subseteq O$  and every  $i = 1, \dots, k$ ,  $\Sigma_{X,i}$  by the singleton  $(\text{card}(X) \leq b_X^i, \beta_i)$ , then  $\Sigma$  implies every elements of  $\Sigma_f$  and  $\Sigma_f$  implies every element of  $\Sigma$ . That is,  $\Sigma_f$  is a cover of  $\Sigma$ , which is finite. In particular, the semantic closure  $\Sigma^*$  of a given  $\Sigma$  of QCs is always infinite, but has a finite cover by the construction above. We may therefore assume without loss of generality that a set of qualitative cardinality constraints is given in the form of a finite cover.

While the implication problem is a decision problem, qualitative cardinality constraints also have an interesting computational problem associated with them. Given a set  $\Sigma$  of QCs and a traditional cardinality constraint  $\text{card}(X) \leq b$ , we may ask what the maximum c-degree  $\beta$  is, with which  $(\text{card}(X) \leq b, \beta)$  is implied by  $\Sigma$ . This is the *inference problem* of qualitative cardinality constraints.

PROBLEM:	Inference problem for qualitative cardinality constraints
INPUT:	Object type $(O, S)$ , Finite set $\Sigma$ of QCs over $(O, S)$ , and Cardinality constraint $\text{card}(X) \leq b$ over $O$
OUTPUT:	$\max\{\beta \in S^T \mid \Sigma \models (\text{card}(X) \leq b, \beta)\}$

Note that the inference problem also has a finite and unrestricted version, which both coincide due to Theorem 1.

**Example 3.** Let  $\Sigma$  be as in Example 1. Then the maximum c-degree with which  $\text{card}(\text{Rfid}) \leq 2$  is implied by  $\Sigma$  is  $\beta_2$ . In fact, Example 2 has shown that  $(\text{card}(\text{Rfid}) \leq 2, \beta_1)$  is not implied by  $\Sigma$ , and  $(\text{card}(\text{Rfid}) \leq 2, \beta_2) \in \Sigma$ .

In what follows we will establish an axiomatic characterization of the implication problem, from which we will derive algorithmic characterizations of the inference and implication problems.

#### 4.2. The magic of $\beta$ -cuts

We will now establish a strong correspondence between instances of the implication problem for qualitative cardinality constraints and instances of the implication problem for cardinality constraints.

**Definition 2.** Let  $\Sigma$  denote a set of qualitative cardinality constraints over the possibilistic object type  $(O, S)$ . For each c-degree  $\beta \in S^T$  where  $\beta > \beta_{k+1}$ , let  $\Sigma_\beta$  denote those cardinality constraints  $\text{card}(X) \leq b$  over object type  $O$  for which there is some  $(\text{card}(X) \leq b, \beta') \in \Sigma$  where  $\beta' \geq \beta$ , that is,

$$\Sigma_\beta = \{\text{card}(X) \leq b \mid (\text{card}(X) \leq b, \beta') \in \Sigma \text{ and } \beta' \geq \beta\}.$$

We call  $\Sigma_\beta$  the  $\beta$ -cut of  $\Sigma$ .

For the set  $\Sigma$  of QCs from Example 1, the following cardinality constraints form the  $\beta_2$ -cut of  $\Sigma$ :  $\text{card}(\text{Zone}) \leq 3$ ,  $\text{card}(\text{Time}) \leq 2$ ,  $\text{card}(\text{Rfid}) \leq 3$ ,  $\text{card}(\text{Zone}, \text{Rfid}) \leq 2$ ,  $\text{card}(\text{Zone}) \leq 2$ ,  $\text{card}(\text{Rfid}) \leq 2$ ,  $\text{card}(\text{Zone}, \text{Rfid}) \leq 1$ , and  $\text{card}(\text{Time}, \text{Rfid}) \leq 1$ .

It turns out that  $\beta$ -cuts suffice to decide the implication problem for qualitative cardinality constraints, as the following theorem establishes.

**Theorem 2.** Let  $\Sigma \cup \{(\text{card}(X) \leq b, \beta)\}$  be a QC set over  $(O, S)$  where  $\beta > \beta_{k+1}$ . Then  $\Sigma \models (\text{card}(X) \leq b, \beta)$  if and only if  $\Sigma_\beta \models \text{card}(X) \leq b$ .

**Proof.** Suppose  $(\iota, \text{Poss}_\iota)$  is some p-instance over  $(O, S)$  that satisfies  $\Sigma$ , but violates  $(\text{card}(X) \leq b, \beta)$ . In particular,  $C_\iota(\text{card}(X) \leq b) < \beta$  implies that there is some world  $w_i$  that violates  $\text{card}(X) \leq b$  and where  $\beta_{k+2-i} < \beta$ .

Let  $\text{card}(Y) \leq b' \in \Sigma_\beta$ , where  $(\text{card}(Y) \leq b', \beta') \in \Sigma$ . Since  $\iota$  satisfies  $(\text{card}(Y) \leq b', \beta') \in \Sigma$  we have  $C_\iota(\text{card}(Y) \leq b') \geq \beta' \geq \beta$ . If  $w_i$  violated  $\text{card}(Y) \leq b'$ , then  $\beta > \beta_{k+2-i} \geq C_\iota(\text{card}(Y) \leq b') \geq \beta$ , a contradiction. Hence,  $w_i$  satisfies  $\Sigma_\beta$  and violates  $\text{card}(X) \leq b$ .

Let  $\iota$  denote some instance that satisfies  $\Sigma_\beta$  and violates  $\text{card}(X) \leq b$ , without loss of generality  $\iota' = \{o_1, \dots, o_{b+1}\}$ . Let  $\iota$  be the p-instance over  $(O, S)$  that consists of  $\iota'$  and where  $\text{Poss}_\iota(o_1) = \dots = \text{Poss}_\iota(o_b) = \alpha_1$  and  $\text{Poss}_\iota(o_{b+1}) = \alpha_i$ , such that  $\beta_{k+1-i} = \beta$ . Then  $\iota$  violates  $(\text{card}(X) \leq b, \beta)$  since  $C_\iota(\text{card}(X) \leq b) = \beta_{k+2-i}$ , as  $w_i = \iota'$  is the smallest world that violates  $\text{card}(X) \leq b$ , and  $\beta_{k+2-i} < \beta_{k+1-i} = \beta$ . For  $(\text{card}(Y) \leq b', \beta') \in \Sigma$  we distinguish two cases. If  $w_i$  satisfies  $\text{card}(Y) \leq b'$ , then  $C_\iota(\text{card}(Y) \leq b') = \beta_1 \geq \beta$ . If  $w_i$  violates  $\text{card}(Y) \leq b'$ , then  $\text{card}(Y) \leq b' \notin \Sigma_\beta$ , i.e.,  $\beta' < \beta = \beta_{k+1-i}$ . Therefore,  $\beta' \leq \beta_{k+2-i} = C_\iota(\text{card}(Y) \leq b')$  as  $w_i = \iota'$  is the smallest world that violates  $\text{card}(Y) \leq b'$ . We conclude that  $C_\iota(\text{card}(Y) \leq b') \geq \beta'$ . Consequently,  $(\iota, \text{Poss}_\iota)$  is a p-instance that satisfies  $\Sigma$  and violates  $(\text{card}(X) \leq b, \beta)$ .  $\square$

**Theorem 2** allows us to apply achievements from cardinality constraints for certain data to qualitative cardinality constraints. It is a major tool to establish the remaining results in this article.

**Example 4.** Let  $\Sigma$  be as in **Example 1**. Then  $\Sigma_{\beta_1}$  consists of the cardinality constraints  $\text{card}(\text{Zone}) \leq 3$ ,  $\text{card}(\text{Time}) \leq 2$ ,  $\text{card}(\text{Rfid}) \leq 3$  and  $\text{card}(\text{Zone}, \text{Rfid}) \leq 2$ . **Theorem 2** says that  $\Sigma_{\beta_1}$  does not imply  $\text{card}(\text{Rfid}) \leq 2$ . The possible world  $w_3$  of the p-instance from **Example 2**:

Zone	Time	Rfid
Z3	11 pm	B5
Z4	12 am	B5
Z5	01 am	B5

satisfies  $\Sigma$ , and violates  $\text{card}(\text{Rfid}) \leq 2$ .

### 4.3. Axiomatic characterization

In this section we will establish an axiomatic characterization for the implication problem of qualitative cardinality constraints by a finite set of Horn axioms. For this purpose, we first recall some basic definitions regarding axiomatizations.

In fact, we determine the semantic closure  $\Sigma^*$  of a set  $\Sigma$  of QCs by applying *inference rules* or *axioms* of the form  $\frac{\text{premise}}{\text{conclusion}}$ . In logic, inference rules of this form are known as Horn axioms. For a set  $\mathfrak{R}$  of inference rules let  $\Sigma \vdash_{\mathfrak{R}} \varphi$  denote the *inference* of  $\varphi$  from  $\Sigma$  by  $\mathfrak{R}$ . That is, there is some sequence  $\sigma_1, \dots, \sigma_n$  such that  $\sigma_n = \varphi$  and every  $\sigma_i$  is an element of  $\Sigma$  or is the conclusion that results from an application of an inference rule in  $\mathfrak{R}$  to some premises in  $\{\sigma_1, \dots, \sigma_{i-1}\}$ . Let  $\Sigma_{\mathfrak{R}}^+ = \{\varphi \mid \Sigma \vdash_{\mathfrak{R}} \varphi\}$  be the *syntactic closure* of  $\Sigma$  under inferences by  $\mathfrak{R}$ .  $\mathfrak{R}$  is *sound* (*complete*) if for every set  $\Sigma$  over every p-object type  $(O, S)$  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. **Table 1** shows an axiomatization  $\mathfrak{C}'$  for the implication problem of traditional cardinality constraints [21]. In these rules, it is assumed that  $O$  is an arbitrarily given object type,  $X, Y \subseteq O$  are non-empty and  $b$  a positive integer. **Theorem 2** and the fact that  $\mathfrak{C}'$  forms a finite axiomatization for the implication of cardinality constraints can be exploited to show directly that the set  $\mathfrak{C}$  from **Table 2** forms an axiomatization for the implication of QCs. Here, it is assumed that  $(O, S)$  is an arbitrarily given p-object type,  $X, Y \subseteq O$  are non-empty,  $b$  a positive integer, and  $\beta, \beta' \in S^T$  some c-degrees. In particular,  $\beta_{k+1}$  denotes the bottom certainty degree in  $S^T$ .

**Theorem 3.** *The set  $\mathfrak{C}$  forms a finite axiomatization for the implication of qualitative cardinality constraints.*

**Proof.** The soundness proof is straightforward. Let  $\Sigma$  denote a set of QCs over p-object type  $(O, S)$ . Let  $(\iota, \text{Poss}_\iota)$  denote a p-instance over  $(O, S)$ . The soundness of the top axiom  $\mathcal{T}$  follows from the fact that  $\iota$  is a set of objects over  $O$  and can therefore not contain any duplicate objects. Consequently,  $\text{card}(O) \leq 1$  holds with c-degree  $\beta_1$  (and therefore with any other c-degree). For the soundness of the relax axiom  $\mathcal{R}$  suppose that  $(\iota, \text{Poss}_\iota)$  satisfies  $(\text{card}(X) \leq b, \beta_i)$ . That is the instance  $w_{k+1-i}$  satisfies  $\text{card}(X) \leq b$ . Since  $\mathcal{R}$  is sound for the implication of cardinality constraints,  $w_{k+1-i}$  also satisfies  $\text{card}(X) \leq b+1$ , which means that  $(\iota, \text{Poss}_\iota)$  satisfies  $(\text{card}(X) \leq b+1, \beta_i)$ . For the soundness of the superset axiom  $\mathcal{S}$  suppose that  $(\iota, \text{Poss}_\iota)$  satisfies  $(\text{card}(X) \leq b, \beta_i)$ . That is the instance  $w_{k+1-i}$  satisfies  $\text{card}(X) \leq b$ . Since  $\mathcal{S}$  is sound for the implication of cardinality constraints,  $w_{k+1-i}$  also satisfies  $\text{card}(XY) \leq b$ , which means that  $(\iota, \text{Poss}_\iota)$  satisfies  $(\text{card}(XY) \leq b, \beta_i)$ . Since for  $(\text{card}(X) \leq b, \beta_{k+1})$  to be satisfied by some p-instance there is no requirement that any possible world satisfies  $\text{card}(X) \leq b$ , the bottom axiom  $\mathcal{B}$  is sound, too. Finally, for the soundness of the weakening axiom  $\mathcal{W}$  assume that  $(\iota, \text{Poss}_\iota)$  satisfies  $(\text{card}(X) \leq b, \beta_i)$ . Consequently,  $w_{k+1-i}$  satisfies  $\text{card}(X) \leq b$  and every world that  $w_{k+1-i}$  contains must also satisfy  $\text{card}(X) \leq b$ , including  $w_{k+1-j}$  for every  $k \geq j \geq i$ . Hence,  $(\iota, \text{Poss}_\iota)$  satisfies  $(\text{card}(X) \leq b, \beta_j)$  for every  $\beta_j \leq \beta_i$ . This establishes the soundness of  $\mathfrak{C}$ .

**Table 1**  
Axiomatization  $\mathfrak{C}' = \{\mathcal{T}', \mathcal{R}', \mathcal{S}'\}$  of traditional cardinality constraints.

	$\text{card}(X) \leq b$	$\text{card}(X) \leq b$
$\text{card}(O) \leq 1$ (top, $\mathcal{T}'$ )	$\text{card}(X) \leq b+1$ (relax, $\mathcal{R}'$ )	$\text{card}(XY) \leq b$ (superset, $\mathcal{S}'$ )

**Table 2**  
Axiomatization  $\mathcal{C} = \{\mathcal{T}, \mathcal{R}, \mathcal{S}, \mathcal{B}, \mathcal{W}\}$  of qualitative cardinality constraints.

$\frac{}{(card(O) \leq \beta)}$ (top, $\mathcal{T}$ )	$\frac{(card(X) \leq b, \beta)}{(card(X) \leq b + 1, \beta)}$ (relax, $\mathcal{R}$ )	$\frac{(card(X) \leq b, \beta)}{(card(XY) \leq b, \beta)}$ (superset, $\mathcal{S}$ )
$\frac{(card(X) \leq b, \beta)}{(card(X) \leq b, \beta)}$ (bottom, $\mathcal{B}$ )	$\frac{(card(X) \leq b, \beta)}{(card(X) \leq b, \beta')} \beta' \leq \beta$ (weakening, $\mathcal{W}$ )	

For completeness, we apply [Theorem 2](#) and the fact that  $\mathcal{C}'$  axiomatizes the implication of cardinality constraints. Let  $(O, \mathcal{S})$  be a p-object type with  $|S| = k + 1$ , and  $\Sigma \cup \{(card(X) \leq b, \beta)\}$  a QC set such that  $\Sigma \models (card(X) \leq b, \beta)$ . We need to show that  $\Sigma \vdash_{\mathcal{C}'} (card(X) \leq b, \beta)$  holds.

For  $\Sigma \models (card(X) \leq b, \beta_{k+1})$  we have  $\Sigma \vdash_{\mathcal{C}'} (card(X) \leq b, \beta_{k+1})$  by applying  $\mathcal{B}$ . Let now  $\beta < \beta_{k+1}$ . From  $\Sigma \models (card(X) \leq b, \beta)$  we conclude  $\Sigma_{\beta} \models card(X) \leq b$  by [Theorem 2](#). Since  $\mathcal{C}'$  is complete for the implication of cardinality constraints,  $\Sigma_{\beta} \vdash_{\mathcal{C}'} card(X) \leq b$  holds. Let  $\Sigma_{\beta}^{\beta} = \{(card(Y) \leq b', \beta) \mid card(Y) \leq b' \in \Sigma_{\beta}\}$ . Thus, the inference of  $card(X) \leq b$  from  $\Sigma_{\beta}$  using  $\mathcal{C}'$  can be turned into an inference of  $(card(X) \leq b, \beta)$  from  $\Sigma_{\beta}^{\beta}$  by  $\mathcal{C}'$ , simply by adding  $\beta$  to each QC in the inference. Hence, whenever  $\mathcal{T}'$  or  $\mathcal{S}'$  is applied, one applies instead  $\mathcal{T}$  or  $\mathcal{S}$ , respectively. Consequently,  $\Sigma_{\beta}^{\beta} \vdash_{\mathcal{C}'} (card(X) \leq b, \beta)$ . The definition of  $\Sigma_{\beta}^{\beta}$  ensures that every QC in  $\Sigma_{\beta}^{\beta}$  can be inferred from  $\Sigma$  by applying  $\mathcal{W}$ . Hence,  $\Sigma_{\beta}^{\beta} \vdash_{\mathcal{C}'} (card(X) \leq b, \beta)$  means that  $\Sigma \vdash_{\mathcal{C}'} (card(X) \leq b, \beta)$ .  $\square$

#### Algorithm 1. Inference

**Require:**  $(O, \mathcal{S}), \Sigma, card(X) \leq b$

**Ensure:**  $\max\{\beta \in \mathcal{S}^T : \Sigma \models (card(X) \leq b, \beta)\}$

```

1: if  $X = O$  then                                     ▶ The cardinality constraint holds with full certainty
2:   return  $\beta_1$ ;
3: else
4:    $\beta \leftarrow \beta_{k+1}$ ;                                   ▶ Starting from the bottom degree
5:   while  $\Sigma \neq \emptyset$  do                             ▶ As long as some input has not been considered
6:     Select  $(card(Y) \leq b', \beta') \in \Sigma$ ;             ▶ Consider one of these next
7:      $\Sigma \leftarrow \Sigma - \{(card(Y) \leq b', \beta')\}$ ;   ▶ and remove it from  $\Sigma$ 
8:     if  $Y \subseteq X$  and  $b' \leq b$  and  $\beta' > \beta$  then       ▶ If the given constraint holds with some stronger c-degree
9:        $\beta \leftarrow \beta'$ ;                                   ▶ Increase the current output to this new degree
10:    end if
11:  end while
12: end if
13: return  $\beta$ ;

```

**Example 5.** Let  $\Sigma$  be as in [Example 1](#). The QC  $(card(\text{Zone}, \text{Rfid}) \leq 4, \beta_2)$  is implied by  $\Sigma$ . Indeed, applying the superset rule  $\mathcal{S}$  to  $(card(\text{Zone}) \leq 3, \beta_1) \in \Sigma$  results in  $(card(\text{Zone}, \text{Rfid}) \leq 3, \beta_1) \in \Sigma_{\beta_1}^+$ . Applying the relax rule  $\mathcal{R}$  to this QC results in  $(card(\text{Zone}, \text{Rfid}) \leq 4, \beta_1) \in \Sigma_{\beta_1}^+$ . Finally, an application of the weakening rule  $\mathcal{W}$  to the last QC results in  $(card(\text{Zone}, \text{Rfid}) \leq 4, \beta_2) \in \Sigma_{\beta_1}^+$ .

#### 4.4. Algorithmic characterization

In practice, the semantic closure  $\Sigma^*$  of a finite set  $\Sigma$  of QCs is infinite and even though there always is some finite cover, it is often unnecessary to determine all implied QCs. In fact, the implication problem for QCs has as input  $\Sigma \cup \{\varphi\}$  and the question is whether  $\Sigma$  implies  $\varphi$ . Computing  $\Sigma^*$  and checking whether  $\varphi \in \Sigma^*$  is hardly efficient. In fact, we will now establish a linear-time algorithm for computing the maximum c-degree  $\beta$  such that  $(card(X) \leq b, \beta)$  is implied by  $\Sigma$ . The following theorem allows us to reduce the implication problem for QCs to a single scan of the input.

**Theorem 4.** Let  $\Sigma \cup \{(card(X) \leq b, \beta)\}$  denote a set of QCs over  $(O, \mathcal{S})$  with  $|S| = k + 1$ . Then  $\Sigma$  implies  $(card(X) \leq b, \beta)$  if and only if  $\beta = \beta_{k+1}$ , or  $X = O$ , or there is some  $(card(Y) \leq b', \beta') \in \Sigma$  such that  $Y \subseteq X$ ,  $b' \leq b$  and  $\beta' \geq \beta$ .

**Proof.** [Theorem 2](#) shows for  $i = 1, \dots, k$  that  $\Sigma$  implies  $(card(X) \leq b, \beta_i)$  if and only if  $\Sigma_{\beta_i}$  implies  $card(X) \leq b$ . It is easy to observe from the axiomatization  $\mathcal{C}'$  of cardinality constraints that  $\Sigma_{\beta_i}$  implies  $card(X) \leq b$  if and only if  $O = X$ , or there is some  $card(Y) \leq b' \in \Sigma_{\beta_i}$  such that  $Y \subseteq X$  and  $b' \leq b$  hold. As  $\Sigma$  implies  $(card(X) \leq b, \beta_{k+1})$ , the theorem follows.  $\square$

[Theorem 4](#) enables us to design [Algorithm 1](#), which returns for a given cardinality constraint  $card(X) \leq b$  the maximum c-degree  $\beta$  for which  $(card(X) \leq b, \beta)$  is implied by a given set  $\Sigma$  of QCs over p-object type  $(O, \mathcal{S})$ . If  $X = O$ , then we return  $\beta_1$  due to the soundness

of the top axiom  $\mathcal{T}$ . Otherwise, starting with  $\beta = \beta_{k+1}$  the algorithm scans all input QCs ( $\text{card}(Y) \leq b', \beta'$ ) and sets  $\beta$  to  $\beta'$  whenever  $\beta'$  is larger than the current  $\beta$ ,  $X$  contains  $Y$  and  $b' \leq b$ .

**Theorem 5** states the correctness of **Algorithm 1**, which follows from **Theorem 4**, as well as the time complexity. Note that  $\|\Sigma\|$  denotes the sum of the total number of attributes and the logarithm of the associated  $c$ -degree's index that occur in the QCs of  $\Sigma$ .

**Theorem 5.** On input  $((O, \mathcal{S}), \Sigma, \text{card}(X) \leq b)$ , **Algorithm 1** returns in  $\mathcal{O}(\|\Sigma\| \cup \{\text{card}(X) \leq b, \beta_{k+1}\})$  time the maximum  $c$ -degree  $\beta$  for which  $(\text{card}(X) \leq b, \beta)$  is implied by  $\Sigma$ .

#### Algorithm 2. Minimal Cover

**Require:** Set  $\Sigma$  of Qualitative Cardinality Constraints over  $(O, \mathcal{S})$

**Ensure:** Subset  $\Sigma_c \subseteq \Sigma$  that forms a minimal cover of  $\Sigma$

```

1:  $\Sigma_c \leftarrow \Sigma;$  ▷ Start with  $\Sigma$  as a cover of itself
2: for all  $\sigma \in \Sigma_c$  do ▷ Check all elements of the current cover
3:   if  $\Sigma_c - \{\sigma\}$  implies  $\sigma$  then ▷ whether they are redundant
4:      $\Sigma_c \leftarrow \Sigma_c - \{\sigma\};$  ▷ and remove them from the cover if they are
5:   end if
6: end for
7: return  $\Sigma_c;$  ▷ The result is a cover that contains no redundant elements

```

**Example 6.** Let  $\Sigma$  be as in **Example 1**, and use **Algorithm 1** to determine the maximum  $c$ -degree  $\beta$  for which QC  $(\text{card}(\text{Rfid}) \leq 2, \beta)$  is implied by  $\Sigma$ . In fact,  $\beta$  becomes  $\beta_2$  as soon as the QC  $(\text{card}(\text{Rfid}) \leq 2, \beta_2)$  is selected as part of the input  $\Sigma$ . This  $c$ -degree cannot be increased and is therefore the output.

**Theorem 5** allows us to decide the associated implication problem efficiently, too. Given  $((O, \mathcal{S}), \Sigma, (\text{card}(X) \leq b, \beta'))$  as an input to the implication problem we can use **Algorithm 1** to compute  $\beta := \max\{\beta'' \mid \Sigma \models (\text{card}(X) \leq b, \beta'')\}$  and return an affirmative answer if and only if  $\beta' \leq \beta$ .

**Corollary 1.** The implication problem of qualitative cardinality constraints can be decided in linear time in the input. □

**Example 7.** Following on from **Example 6**, let  $\Sigma$  be as in **Example 1**. Then the QC  $(\text{card}(\text{Rfid}) \leq 2, \beta_3)$  is implied by  $\Sigma$ . Indeed, the maximum  $c$ -degree  $\beta$  for which  $(\text{card}(\text{Rfid}) \leq 2, \beta)$  is implied by  $\Sigma$  was determined as  $\beta_2$  in **Example 6**. Since  $\beta_3 \leq \beta_2$ , the given QC is indeed implied.

#### 4.5. Applications to integrity enforcement and query processing

We conclude this section with some applications of our results. Our first application is the computation of a minimal cover for a given set  $\Sigma$  of qualitative cardinality constraints.

Recall from before that a *cover* of the given set  $\Sigma$  is a set  $\Sigma'$  of QCs such that every  $\sigma' \in \Sigma'$  is implied by  $\Sigma$  and every  $\sigma \in \Sigma$  is implied by  $\Sigma'$ . In other words,  $\Sigma'$  is a faithful representation of  $\Sigma$ . A cover  $\Sigma'$  of  $\Sigma$  is said to be *minimal* if and only if there is no proper subset  $\Sigma''$  of  $\Sigma'$  that is also a cover of  $\Sigma$ . In other words, a minimal cover  $\Sigma'$  does not feature any QCs  $\sigma'$  that are *redundant* with respect to  $\Sigma'$ , i.e., redundant in the sense that  $\sigma'$  is implied by  $\Sigma' - \{\sigma'\}$ . Our algorithmic solution to the implication problem suggests the following strategy for computing a minimal cover of a given QC set  $\Sigma$ : scan, one by one, each element  $\sigma \in \Sigma$  whether it is implied by  $\Sigma - \{\sigma\}$ , and remove  $\sigma$  from  $\Sigma$  if that is the case. The resulting subset of  $\Sigma$  is a minimal cover. This strategy is manifested in **Algorithm 2**.

The upper time bound of the following theorem follows from the linear time complexity of the associated implication problem, as established in **Corollary 1**.

**Theorem 6.** **Algorithm 2** computes a minimal cover in time  $\mathcal{O}(\|\Sigma\|^2)$  in the size of the input  $\Sigma$  of qualitative cardinality constraints. □

The importance of minimal covers results from their application in integrity enforcement. To guarantee that the objects resulting from updates against the given instance conform to the rules of the application, the resulting instance must be validated against the given set of business rules. The overhead for this enforcement of business rules can be optimized by removing redundant business rules. The optimization is indirectly proportional to the number of objects in the instance. That is, the more objects are in the instance the more time savings can be achieved by enforcing only non-redundant rules. For this purpose, a minimal cover is desirable.

**Example 8.** Let  $\Sigma'$  be the set of qualitative cardinality constraints represented in **Fig. 2**. Recall that these form a cover of the qualitative cardinality constraints satisfied by the  $p$ -instance in **Fig. 1**. Given  $\Sigma'$ , **Algorithm 2** may compute a minimal cover  $\Sigma$  of  $\Sigma'$  as given in **Example 1**, or in other words, the qualitative cardinality constraints highlighted in bold font in **Fig. 2**.

As our second application we demonstrate the benefit of qualitative cardinality constraints on query processing. Therefore, we simply add the attribute  $P$ -degree to the object type `TRACKING` with attributes `Zone`, `Time`, and `Rfid`. Suppose we are interested in finding

out which bats have been tracked in which zone, but we are only interested in answers that come from ‘certain’ or ‘quite possible’ objects in the instance. A user might enter the following SQL query:

Zone	Rfid	P-degree
Z0	B0	$\alpha_1$
Z0	B1	$\alpha_1$
Z1	B3	$\alpha_1$
Z2	B4	$\alpha_1$
Z3	B5	$\alpha_1$
Z4	B5	$\alpha_1$
Z6	B6	$\alpha_2$
Z6	B7	$\alpha_2$

```
SELECT DISTINCT Zone, Rfid, P-degree
FROM TRACKING
WHERE P-degree =  $\alpha_1$  or P-degree =  $\alpha_2$ 
ORDER BY P-degree ASC
```

which removes duplicate answers. When applied to the p-instance from Fig. 1, the query returns the answers on the right.

Our framework allows users to ask such queries having available the p-degrees of objects. Answers can be ordered according to the p-degree, which makes it possible for users to appreciate their significance. The example shows how our framework can be embedded with standard technology, here SQL. Finally, the QC  $card(Zone, Rfid) \leq 1$  holds with maximum c-degree  $\beta_2$ . That is,  $\{Zone, Rfid\}$  forms a key on the world  $w_2$  that contains fully certain and quite possible objects. Consequently, the `DISTINCT` clause becomes superfluous in the query above. A query optimizer, capable of reasoning about QCs, can remove the `DISTINCT` clause from the input query without affecting its output. This optimization saves response time when answering queries, as duplicate elimination is an expensive operation and therefore not executed by default in SQL databases. Note that we do not view the enforcement of meaningful constraints as an overhead, but a requirement that is necessary in data processing. Qualitative cardinality constraints, and the ability to reason about them, have therefore direct applications to query processing. We illustrate this point by a further example. Suppose we have a provider answering queries on large data sets as a service to customers. The customer will only pay for the service when the price is not too high, but the provider will only want to invoke the service for a paying customer. Our reasoning abilities can be used to determine a “quote” for the price of some queries in the form of an upper bound on the number of query answers without the need to evaluate the query at all. The query

```
SELECT Rfid, P-degree
FROM TRACKING
WHERE (P-degree =  $\alpha_1$  or P-degree =  $\alpha_2$ ) and Zone = Z6 and Time = 02am
ORDER BY P-degree ASC
```

will return at most two answers when evaluated on any data set conforming to the set  $\Sigma$  of rules from Example 1. The reason is that the qualitative cardinality constraint ( $card(Zone, Time) \leq 2, \beta_2$ ) is implied by  $\Sigma$ . Being able to decide implication in linear time of the constraints not only makes this feedback to the customer very efficient in practice, but also very affordable to the provider.

### 5. Armstrong samples for qualitative cardinality constraints

In this section we develop a theory of Armstrong samples for sets of qualitative cardinality constraints. The concept of an Armstrong database is well established in database research [13]. They are widely regarded as an effective tool to visualize abstract sets of constraints in a user-friendly way [13,33,34]. As such data engineers exploit Armstrong databases as a communication tool in their interaction with domain experts in order to determine the set of constraints that are meaningful to the application at hand [22,30,33,34]. As Fig. 3 illustrates, a design team generates an Armstrong sample  $db_\Sigma$  that perfectly represents their current perceptions of the set  $\Sigma$  of constraints that should be enforced. The team then jointly inspects the sample with domain experts in order to discover flaws or shortcomings in the perception of the design team. Evidently, the Armstrong sample helps designers and domain experts communicate more effectively, thereby overcoming the mismatch in expertise [30]. This process repeats until all parties are happy.

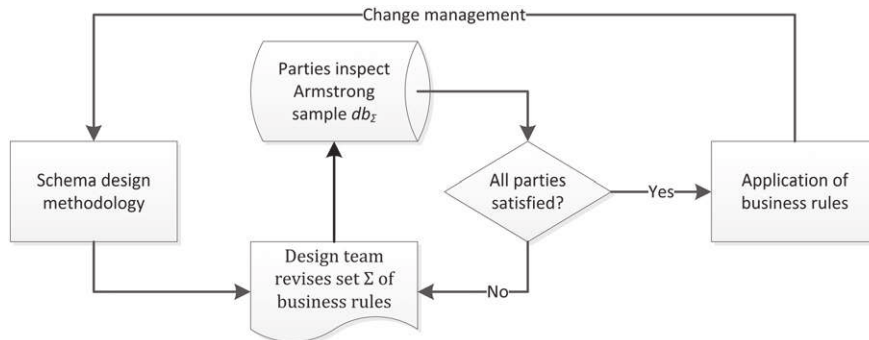


Fig. 3. The use of Armstrong samples in requirements engineering.

We now introduce the concept of an Armstrong p-instance for qualitative cardinality constraints. While we show that Armstrong p-instances do exist for arbitrary sets of QCs, some of these sets require every Armstrong p-instance to be infinite. Obviously, this is a strong inhibitor to the usefulness of Armstrong samples in practice, as illustrated in Fig. 3. We overcome this challenge by introducing the new concept of an Armstrong p-sketch, which is a finite representation of some potentially infinite Armstrong p-instances. We establish sufficient and necessary conditions for when a given p-instance is Armstrong for a given set of qualitative cardinality constraints. Based on these conditions, we show how to compute Armstrong p-sketches for an arbitrary set of such constraints. While the problem of finding an Armstrong p-sketch is precisely exponential in the size of the input, our algorithm computes an Armstrong p-sketch whose size is always guaranteed to be bounded by the product of the size of a minimum-sized Armstrong p-sketch and the cardinality of the input. Finally, we characterize the situation when finite Armstrong p-instances exist and that their existence can be decided in linear time in the input.

### 5.1. Armstrong instances and sketches

We first restate the original definition of Armstrong databases [13] in our context.

**Definition 3.** A p-instance  $\iota$  is said to be Armstrong for a given set  $\Sigma$  of qualitative cardinality constraints on a given p-object type  $(O, S)$  if and only if for all qualitative cardinality constraints  $\varphi$  over  $(O, S)$  it is true that  $\iota$  satisfies  $\varphi$  if and only if  $\Sigma$  implies  $\varphi$ .

Armstrong p-instances for  $\Sigma$  are exact visual representations of  $\Sigma$ . Fig. 1 shows an Armstrong p-instance for the set  $\Sigma$  of QCs from Example 1. While Armstrong p-instances always exist there are QC sets that require infinite Armstrong p-instances. As infinite samples cannot be used directly to communicate with domain experts, we introduce the new concept of Armstrong p-sketches.

**Definition 4.** Let  $\Sigma$  be a set of qualitative cardinality constraints over p-object type  $(O, S)$ . Let  $O_*$  denote the object type that results from  $O$  by adding to the domain of each attribute the distinguished symbol  $*$ . A p-sketch over  $(O, S)$  consists of a finite p-instance  $(\varsigma = \{\omega_1, \dots, \omega_n\}, \text{Poss}_\varsigma)$  over  $(O_*, S)$ , and a function  $\text{Card}_\varsigma$  that maps each  $\omega_i \in \varsigma$  to a value  $c_i = \text{Card}_\varsigma(\omega_i) \in \mathbb{N} \cup \{\infty\}$ . A p-expansion of  $(\varsigma, \text{Poss}_\varsigma, \text{Card}_\varsigma)$  is a p-instance  $(\iota, \text{Poss}_\iota)$  over  $(O, S)$  such that

- $\iota = \bigcup_{i=1}^n \{o_i^1, \dots, o_i^{c_i}\}$ ,
- (preservation of p-degrees) for all  $i = 1, \dots, n$ , for all  $j = 1, \dots, c_i$ ,  $\text{Poss}_\iota(o_i^j) = \text{Poss}_\varsigma(\omega_i)$ , and all other objects receive the bottom p-degree,
- (preservation of domain values) for all  $i = 1, \dots, n$ , for all  $j = 1, \dots, c_i$ , for all  $A \in O_*$ , if  $\omega_i(A) \neq *$ , then  $o_i^j(A) = \omega_i(A)$ ,
- (uniqueness of values substituted for  $*$ ) for all  $i = 1, \dots, n$ , for all  $A \in O_*$ , if  $\omega_i(A) = *$ , then for all  $j = 1, \dots, c_i$ , for all  $l = 1, \dots, n$ , and for all  $m = 1, \dots, c_l$  (where  $j \neq m$ , if  $l = i$ ),  $o_i^j(A) \neq o_l^m(A)$ .

Sometimes we omit  $\text{Poss}_\varsigma$  and  $\text{Card}_\varsigma$  and simply refer to the p-sketch  $(\varsigma, \text{Poss}_\varsigma, \text{Card}_\varsigma)$  by  $\varsigma$ . We call  $\varsigma$  an Armstrong p-sketch for  $\Sigma$  if and only if every p-expansion of  $\varsigma$  is an Armstrong p-instance for  $\Sigma$ .

**Example 9.** Fig. 1 shows an Armstrong p-sketch  $\varsigma$  for the QC set  $\Sigma$  from Example 1. The Armstrong p-instance  $\iota$  in Table 3 is a p-expansion of  $\varsigma$ , which yields the p-instance from Fig. 1 after suitable substitutions.

Armstrong p-sketches are most beneficial when they have only infinite expansions, as characterized in Theorem 11 later. The following example illustrates how Armstrong p-sketches can still finitely represent sets of qualitative cardinality constraints for which only infinite Armstrong p-instances exist.

**Example 10.** Table 4 shows an Armstrong p-sketch for the following set of QCs:  $(\text{card}(\text{Zone}) \leq 3, \beta_1)$ ,  $(\text{card}(\text{Time}) \leq 3, \beta_1)$ ,  $(\text{card}(\text{Time}, \text{Rfid}) \leq 2, \beta_1)$ ,  $(\text{card}(\text{Rfid}) \leq 1, \beta_2)$  and  $(\text{card}(\text{Zone}, \text{Time}) \leq 2, \beta_3)$ . Notably, every p-expansion of this p-sketch requires infinitely many objects. Designers and domain expert who jointly inspect this p-sketch are immediately alerted to the fact that no 'fully certain' finite upper bound has been specified on Rfid.

Our ultimate aim in this section is to compute Armstrong p-sketches for any given set of qualitative cardinality constraints. For this purpose it is useful to find conditions that allow us to say when a given p-instance is Armstrong for a given set of qualitative cardinality constraints. Section 5.2 addresses this subject.

### 5.2. Structural characterization

For characterizing the structure of Armstrong p-instances we define notions of agreement between objects of an instance. Cardinality constraints require us to compare any number of distinct objects. Intuitively, the agree set of two objects consists of all attributes on which the two objects have the same value; and the b-agree set of an instance consists of the intersection of all agree sets for all pairs of any  $b$  distinct objects that are part of the instance.

**Definition 5.** Let  $O$  be an object type,  $w$  an instance, and  $o_1, o_2$  two objects of  $O$ . The agree set of  $o_1$  and  $o_2$  is defined as  $\text{ag}(o_1, o_2) = \{A \in O \mid o_1(A) = o_2(A)\}$ . For every  $b \in \mathbb{N} \cup \{\infty\}$ ,  $b > 1$  we define the b-agree set of  $w$  as  $\text{ag}_b(w) = \{\bigcap_{1 \leq i < j \leq b} \text{ag}(o_i, o_j) \mid \exists o_1, \dots, o_b \in w (\forall 1 \leq i < j \leq b (o_i \neq o_j))\}$ , and  $\text{ag}_1(w) = \{O\}$ .

**Table 3**

Armstrong p-sketch for set  $\Sigma$  from Example 1 and one of its p-expansions.

p-Sketch $\varsigma$				
Card	Zone	Time	Rfid	p-Degree
2	$C_{Z,1}$	*	*	$\alpha_1$
1	$C_{Z,1}$	*	*	$\alpha_3$
2	*	$C_{T,2}$	*	$\alpha_1$
2	*	*	$C_{R,3}$	$\alpha_1$
1	*	*	$C_{R,3}$	$\alpha_3$
2	$C_{Z,4}$	$C_{T,4}$	*	$\alpha_2$
2	$C_{Z,5}$	*	$C_{R,5}$	$\alpha_3$
2	*	$C_{T,6}$	$C_{R,6}$	$\alpha_3$

p-Expansion $\iota$ of $\varsigma$			
Zone	Time	Rfid	p-Degree
$C_{Z,1}$	$C_{T,1}^1$	$C_{R,1}^1$	$\alpha_1$
$C_{Z,1}$	$C_{T,1}^2$	$C_{R,1}^2$	$\alpha_1$
$C_{Z,1}$	$C_{T,1}^3$	$C_{R,1}^3$	$\alpha_3$
$C_{Z,2}^1$	$C_{T,2}$	$C_{R,2}^1$	$\alpha_1$
$C_{Z,2}^2$	$C_{T,2}$	$C_{R,2}^2$	$\alpha_1$
$C_{Z,3}^1$	$C_{T,3}$	$C_{R,3}$	$\alpha_1$
$C_{Z,3}^2$	$C_{T,3}$	$C_{R,3}$	$\alpha_1$
$C_{Z,3}^3$	$C_{T,3}$	$C_{R,3}$	$\alpha_3$
$C_{Z,4}$	$C_{T,4}$	$C_{R,4}^1$	$\alpha_2$
$C_{Z,4}$	$C_{T,4}$	$C_{R,4}^2$	$\alpha_2$
$C_{Z,5}$	$C_{T,5}$	$C_{R,5}$	$\alpha_3$
$C_{Z,5}^1$	$C_{T,5}$	$C_{R,5}$	$\alpha_3$
$C_{Z,6}^1$	$C_{T,6}$	$C_{R,6}$	$\alpha_3$
$C_{Z,6}^2$	$C_{T,6}$	$C_{R,6}$	$\alpha_3$

Note that  $ag_\infty(w) = \emptyset$  whenever  $w$  is finite.

**Example 11.** For the worlds  $w_1, w_2$  and  $w_3$  from our running example in Fig. 1 we obtain  $ag_2(w_1) = \{\{Zone\}, \{Time\}, \{Rfid\}\}$  and  $ag_b(w_1) = \emptyset$  for all  $b > 2$ ,  $ag_2(w_2) = \{\{Zone\}, \{Time\}, \{Rfid\}, \{Zone, Time\}\}$  and  $ag_b(w_2) = \emptyset$  for all  $b > 2$ ,  $ag_2(w_3) = \{\{Zone\}, \{Time\}, \{Rfid\}, \{Zone, Time\}, \{Zone, Rfid\}, \{Time, Rfid\}\}$ , and  $ag_3(w_3) = \{\{Zone\}, \{Rfid\}\}$  and  $ag_b(w_3) = \emptyset$  for all  $b > 3$ .

An Armstrong p-instance  $\iota$  violates all QCs not implied by the given QC set  $\Sigma$ . It suffices for any non-empty set  $X$  to have  $card(X) \leq b_X^i - 1$  violated by the world  $w_{k+1-i}$  of  $\iota$  where  $b_X^i$  denotes the minimum positive integer for which  $card(X) \leq b_X^i$  is implied by  $\Sigma_{\beta_i}$ . If there are implied  $card(X) \leq b_X^i$  and  $card(Y) \leq b_Y^j$  such that  $b_X^i = b_Y^j$  and  $Y \subseteq X$ , then it suffices to have  $card(X) \leq b_X^i - 1$  violated by  $w_{k+1-i}$ . Finally, if there are implied  $card(X) \leq b_X^i$  and  $card(X) \leq b_X^j$  such that  $b_X^i = b_X^j$  and  $i < j$ , then it suffices to have  $card(X) \leq b_X^i - 1$  violated by  $w_{k+1-i}$ . This motivates the following definition of duplicate sets  $X$  with certainty  $\beta_i$  and their associated cardinalities  $b_X^i$ . Intuitively, if we know these sets and cardinalities, we can construct an Armstrong p-sketch by generating objects with associated p-degrees that violate the qualitative cardinality constraints ( $card(X) \leq b_X^i - 1, \beta_i$ ).

**Definition 6.** Let  $\Sigma$  be a set of qualitative cardinality constraints over p-object type  $(O, S)$  with  $|S| = k + 1$ . For  $\emptyset \neq X \subseteq O$  and  $i = 1, \dots, k$ , let

$$b_X^i = \begin{cases} \min\{b \in \mathbb{N} \mid \Sigma_{\beta_i} \models card(X) \leq b\} & , \text{ if } \{b \in \mathbb{N} \mid \Sigma_{\beta_i} \models card(X) \leq b\} \neq \emptyset \\ \infty & , \text{ otherwise} \end{cases}$$

The set  $dup_{\Sigma_{\beta_i}}(O)$  of duplicate sets of c-degree  $\beta_i$  is defined as  $dup_{\Sigma_{\beta_i}}(O) = \{X \subseteq O \mid b_X^i > 1 \wedge (\forall A \in O - X (b_{XA}^i < b_X^i)) \wedge \forall j > i (b_X^j < b_X^i)\}$ .

**Table 4**

An Armstrong p-sketch for the QC set from Example 10 without finite p-expansion.

Card	Zone	Time	Rfid	p-Degree
2	$C_{Z,1}$	$C_{T,1}$	*	$\alpha_1$
1	$C_{Z,1}$	$C_{T,1}$	*	$\alpha_2$
3	$C_{Z,2}$	*	*	$\alpha_1$
3	*	$C_{T,2}$	*	$\alpha_1$
2	*	$C_{T,3}$	$C_{R,1}$	$\alpha_3$
3	$C_{Z,3}$	*	$C_{R,2}$	$\alpha_3$
$\infty$	*	*	$C_{R,3}$	$\alpha_3$

**Example 12.** Consider the set  $\Sigma$  over  $p$ -object type  $(O, S)$  from Example 1. Fig. 4 shows the non-trivial subsets  $X$  of  $O$  associated with their  $b_X^i$  values for  $i = 1, 2, 3$  from left to right. Among these, the duplicate sets of  $c$ -degree  $\beta_1$  (left figure),  $\beta_2$  (middle figure) and  $\beta_3$  (right figure) are indicated in bold font. Here,  $Y = \{\text{Zone}, \text{Time}\}$  is not a duplicate set of  $c$ -degree  $\beta_1$  as  $b_Y^1 = 2 = b_Y^1$ . Similarly,  $Z = \{\text{Rfid}\}$  is not a duplicate set of  $c$ -degree  $\beta_2$  since  $b_Z^2 = 2 = b_Z^2$ .

Next we characterize the structure of Armstrong  $p$ -instances. A given  $p$ -instance satisfies a given QC  $\text{card}(X) \leq b \in \Sigma_{\beta_i}$  if there are not  $b + 1$  distinct objects in world  $w_{k+1-i}$  that have matching values on  $X$ . Also, a given  $p$ -instance violates all non-implies QCs if every duplicate set  $X$  of  $c$ -degree  $\beta_i$  is contained by some attribute set on which  $b_X^i$  distinct objects in  $w_{k+1-i}$  agree.

**Theorem 7.** Let  $\Sigma$  denote a set of qualitative cardinality constraints, and let  $(\iota, \text{Poss}_\iota)$  denote a  $p$ -instance over  $(O, S)$  with  $|S| = k + 1$ . Then  $(\iota, \text{Poss}_\iota)$  is an Armstrong  $p$ -instance for  $\Sigma$  if and only if for all  $i = 1, \dots, k$ , the world  $w_{k+1-i}$  is Armstrong for  $\Sigma_{\beta_i}$ . That is, for all  $i = 1, \dots, k$ , for all  $X \in \text{dup}_{\Sigma_{\beta_i}}(O)$  there is some  $Z \in \text{ag}_{b_X^i}(w_{k+1-i})$  such that  $X \subseteq Z$ , and for all  $\text{card}(X) \leq b \in \Sigma_{\beta_i}$  and for all  $Z \in \text{ag}_{b+1}(w_{k+1-i})$ ,  $X \not\subseteq Z$ .

**Proof.** The  $p$ -instance  $(\iota, \text{Poss}_\iota)$  is Armstrong for  $\Sigma$  if and only if for all  $i = 1, \dots, k$ , for all QCs  $(\varphi, \beta_i)$ , it holds that  $\models_{(\iota, \text{Poss}_\iota)} (\varphi, \beta_i)$  iff  $\Sigma \models (\varphi, \beta_i)$ . However,  $\models_{(\iota, \text{Poss}_\iota)} (\varphi, \beta_i)$  iff  $\models_{w_{k+1-i}} \varphi$ , and  $\Sigma \models (\varphi, \beta_i)$  iff  $\Sigma_{\beta_i} \models \varphi$ . Therefore, the  $p$ -instance  $(\iota, \text{Poss}_\iota)$  is Armstrong for  $\Sigma$  if and only if for all  $i = 1, \dots, k$ ,  $w_{k+1-i}$  is Armstrong for  $\Sigma_{\beta_i}$ . The second statement follows from the known result that a world  $w$  is Armstrong for a set  $\Sigma$  of cardinality constraints if and only if for all  $X \in \text{dup}_\Sigma(O)$  there is some  $Z \in \text{ag}_{b_X}(w)$  such that  $X \subseteq Z$ , and for all  $\text{card}(X) \leq b \in \Sigma$  and for all  $Z \in \text{ag}_{b+1}(w_{k+1-i})$ ,  $X \not\subseteq Z$  [20,33].  $\square$

**Example 13.** Consider the  $p$ -instance  $\iota$  from Fig. 1 and the set  $\Sigma$  of QCs from Example 1. Examples 11 and 12 show that  $\iota$  satisfies the conditions of Theorem 7, and is therefore a finite Armstrong  $p$ -instance for  $\Sigma$ .

### 5.3. Computational characterization

We now apply Theorem 7 to compute Armstrong  $p$ -sketches for any given QC set over any given  $p$ -object type. It follows that Armstrong  $p$ -sketches always exist, even though finite Armstrong  $p$ -instances may not. While the problem of finding an Armstrong  $p$ -sketch is precisely exponential in the size of the given constraints we show that the size of our output Armstrong  $p$ -sketch is always bounded by the product of the number of the given constraints and the size of a minimum-sized Armstrong  $p$ -sketch. Finally, we show that there are Armstrong  $p$ -sketches whose size is logarithmic in the size of the given constraints. We recommend using both representations: i) the set  $\Sigma$  which explicitly lists the qualitative cardinality constraints, and ii) an Armstrong  $p$ -sketch for  $\Sigma$ .

For a given QC set  $\Sigma$  over a given  $p$ -object type  $(O, S)$  and  $|S| = k + 1$ , we visualize  $\Sigma$  by computing an Armstrong  $p$ -sketch  $\varsigma$  for  $\Sigma$ . If finite Armstrong  $p$ -instances exist for  $\Sigma$ , then we may compute one in the form of a  $p$ -expansion of  $\varsigma$ . Theorem 7 provides us with a strategy to compute an Armstrong  $p$ -sketch for  $\Sigma$ . The main complexity of this strategy goes into the computation of duplicate sets and their associated cardinalities. Conceptually, we could proceed in three stages. First, we compute for all  $i = 1, \dots, k$  and for all non-trivial  $X \subset O$ ,  $b_X^i$  by starting with  $\infty$  and setting  $b_X^i$  to  $b$  whenever there is some  $\text{card}(Y) \leq b \in \Sigma_{\beta_i}$  such that  $Y \subseteq X$  and  $b < b_X^i$ . Secondly, for all  $i = 1, \dots, k$  and starting with all non-trivial subsets  $X$  as the set of duplicate sets of  $c$ -degree  $\beta_i$ , we remove  $X$  whenever  $b_X^i = 1$  or there is some  $A \in O - X$  such that  $b_{XA}^i = b_X^i$ . Algorithm 3 calls this procedure in the loop at lines 1–3, which computes an Armstrong  $p$ -sketch for a given QC set  $\Sigma$  over some given  $p$ -object type  $(O, S)$ . We now outline the remaining steps of Algorithm 3.

Algorithm 3 computes objects over  $O_*$  for each duplicate set  $X$  of  $c$ -degree  $\beta_i$ , starting from  $i = k$  down to 1. Before moving on to another duplicate set of  $c$ -degree  $\beta_i$ , the algorithm processes all occurrences of  $X$  as a duplicate set of  $c$ -degree  $\beta_i \geq \beta_i$  (lines 8–9), introducing an object  $\omega_r$  (lines 10–18) with  $p$ -degree  $\alpha_{k+1-i}$  (line 16) and cardinality  $b_X^i - b$  (line 17) where  $b$  is the cardinality of the duplicate set  $X$  already processed in the previous steps (line 19). Line 23 marks  $X$  as processed to exclude it from repeated computations in the future (line 6).

For a set  $S$  let  $|S|$  denote the number of elements in  $S$ . An Armstrong  $p$ -sketch for  $\Sigma$  is said to be *minimum-sized* if there is no Armstrong  $p$ -sketch for  $\Sigma$  with fewer objects.

**Theorem 8.** Let  $\varsigma^{\min}$  denote a minimum-sized Armstrong  $p$ -sketch for  $\Sigma$ . Algorithm 3 computes an Armstrong  $p$ -sketch  $\varsigma_c$  for  $\Sigma$  such that  $|\varsigma_c| \leq |\varsigma^{\min}| \times |\Sigma|$ .

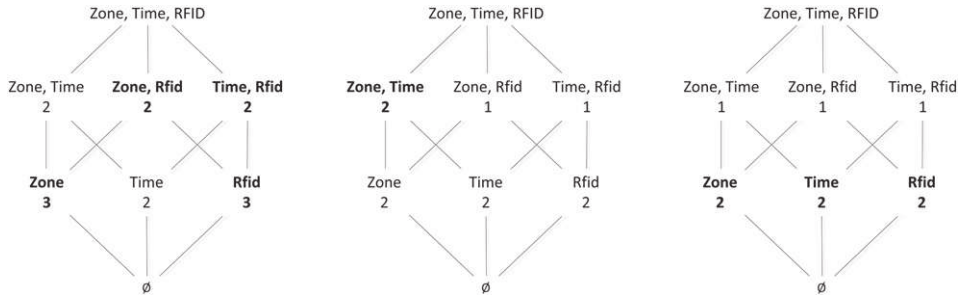


Fig. 4. Attribute sets  $X$ , duplicate sets in bold font and their cardinalities  $b_X^i$  for  $i = 1, 2, 3$  from left to right.



**Proof.** The soundness of [Algorithm 3](#) follows from [Theorem 7](#): every p-expansion of the Armstrong p-sketch computed by [Algorithm 3](#) meets the conditions in [Theorem 7](#) by construction. The upper bound on the size of the Armstrong p-sketch  $\varsigma_c$  computed by [Algorithm 3](#) follows from a series of arguments. Firstly, the number of objects in  $\varsigma_c$  equals the number of duplicate sets of c-degree  $\beta_i$  for  $i = 1, \dots, k$ . Secondly, for each duplicate set of c-degree  $\beta_i$  there is some  $(\text{card}(Y) \leq b, \beta_j) \in \Sigma$  such that  $Y \subseteq X$  and  $j \leq i$ . Thirdly, different duplicate sets  $X, Z \in \text{dup}_{\Sigma\beta_i}(O)$  that both derive their cardinalities  $b_X^i = b_Z^i = b$  from the same  $(\text{card}(Y) \leq b, \beta_j) \in \Sigma$  must have different objects with p-degree  $\alpha_{k+1-i}$  in every Armstrong p-sketch. Finally, the number of objects in any Armstrong p-sketch  $\varsigma$  equals the number of objects in its largest “world”  $w_k$ . We therefore get the following:

$$\begin{aligned}
|\varsigma_c| &= \sum_{i=1}^k \sum_{X \in \text{dup}_{\Sigma\beta_i}(O)} 1 \\
&\leq \sum_{i=1}^k \left( \sum_{(\sigma, \beta_j) \in \Sigma, j \leq i} \left( \sum_{X \in \text{dup}_{\Sigma\beta_i}(\sigma, \beta_j)} 1 \right) \right) \\
&\leq \sum_{i=1}^k \left( \sum_{(\sigma, \beta_j) \in \Sigma, j \leq i} \left( |w_{k+1-i}^{\min}| - |w_{k+1-i-1}^{\min}| \right) \right) \\
&\leq |\Sigma| \cdot |w_k^{\min}| \\
&= |\Sigma| \cdot |\varsigma^{\min}|
\end{aligned}$$

which shows the upper bound. □

**Algorithm 3.** Armstrong p-sketch

---

**Require:** Set  $\Sigma$  of qualitative cardinality constraints over p-object type  $(O, \{\beta_1, \dots, \beta_k, \beta_{k+1}\})$

**Ensure:** Armstrong p-sketch  $(\varsigma, \text{Poss}_{\varsigma}, \text{Card}_{\varsigma})$  for  $\Sigma$

```

1: for  $i = 1, \dots, k$  do ▷ Compute duplicate sets  $X$  of c-degree  $\beta_i$  and  $b_X^i$ 
2:   Compute  $(\text{dup}_{\Sigma\beta_i}(O), \{b_X^i \mid X \in \text{dup}_{\Sigma\beta_i}(O)\})$ 
3: end for
4:  $j \leftarrow 0; r \leftarrow 0; \varsigma \leftarrow \emptyset; \text{dup}_{\Sigma}(O) \leftarrow \emptyset;$ 
5: for  $i = k$  downto 1 do
6:   for all  $X \in \text{dup}_{\Sigma\beta_i}(O) - \text{dup}_{\Sigma}(O)$  do ▷ Duplicate sets not processed yet
7:      $b \leftarrow 0; j \leftarrow j + 1;$ 
8:     for  $l = i$  downto 1 do ▷ Represent  $X$  in all possible worlds required
9:       if  $X \in \text{dup}_{\Sigma\beta_l}(O)$  then ▷  $X$  requires more objects in world  $w_l$ 
10:         $r \leftarrow r + 1;$ 
11:        for all  $A \in O$  do
12:          if  $A \in X$  then  $\omega_r(A) \leftarrow c_{A,j};$  ▷  $\{c_{A,j} : A \in X\}$  represent  $X$ 
13:          else  $\omega_r(A) \leftarrow *;$  ▷  $*$  represents unique value in expansion
14:          end if
15:        end for
16:         $\varsigma \leftarrow \varsigma \cup \{\omega_r\};$  ▷ Add new object to sketch
17:         $\text{Poss}_{\varsigma}(\omega_r) \leftarrow \alpha_{k+1-l};$  ▷ Add p-degree to new object
18:         $\text{Card}_{\varsigma}(\omega_r) \leftarrow b_X^i - b;$  ▷ Add remaining cardinality to new object
19:         $b \leftarrow b_X^i;$  ▷ Book-keeping for cardinalities already represented
20:      end if
21:    end for
22:  end for
23:   $\text{dup}_{\Sigma}(O) \leftarrow \text{dup}_{\Sigma}(O) \cup \text{dup}_{\Sigma\beta_i}(O);$  ▷ Mark duplicate set  $X$  as processed
24: end for
25: return  $(\varsigma, \text{Poss}_{\varsigma}, \text{Card}_{\varsigma});$ 

```

---

Note that [Theorem 8](#) shows that qualitative cardinality constraints enjoy Armstrong sketches, that is, for every given p-object type and every given QC set  $\Sigma$  over this p-object type there is an Armstrong p-sketch for  $\Sigma$ .

**Corollary 2.** Qualitative cardinality constraints enjoy Armstrong p-sketches.

We show next that the computational problem of finding an Armstrong p-sketch for  $\Sigma$  is precisely exponential in the size of  $\Sigma$ . That is, an Armstrong p-sketch for  $\Sigma$  can be found in time at most exponential in the size of  $\Sigma$ , and there are QC sets  $\Sigma$  such that every Armstrong p-sketch for  $\Sigma$  requires a number of objects that is exponential in the size of  $\Sigma$ .

**Theorem 9.** *Finding an Armstrong p-sketch is precisely exponential in the size of the given set  $\Sigma$  of qualitative cardinality constraints.*

**Proof.** Algorithm 3 computes an Armstrong p-sketch for  $\Sigma$  in time at most exponential in its size. Some QC sets  $\Sigma$  have only Armstrong p-sketches with exponentially many objects in the size of  $\Sigma$ . For  $O = \{A_1, \dots, A_{2n}\}$ ,  $S = \{\alpha_1, \alpha_2\}$  and  $\Sigma = \{(card(A_1, A_2) \leq 1, \beta_1), \dots, (card(A_{2n-1}, A_{2n}) \leq 1, \beta_1)\}$  with size  $2 \cdot n$ ,  $dup_{\Sigma, \beta_1}(O)$  consists of the  $2^n$  duplicate sets  $\cup_{j=1}^n X_j$  where  $X_j \in \{A_{2j-1}, A_{2j}\}$ .

We also show that there are other extreme cases where there are Armstrong p-sketches for QC sets  $\Sigma'$  that only require a size logarithmic in that of  $\Sigma$ .

**Theorem 10.** *There are sets  $\Sigma'$  of qualitative cardinality constraints for which there are Armstrong p-sketches whose size is logarithmic in that of  $\Sigma$ .*

**Proof.** Such a set  $\Sigma'$  is given by the following  $2^n$  QCs: for all  $i = 1, \dots, n$ , for all  $X = \cup_{i=1}^n X_i$  where  $X_i \in \{A_{2i-1}, A_{2i}\}$ ,  $(card(X) \leq 1, \beta_1)$ . Then the size of  $\Sigma'$  is  $n \cdot 2^n \in \mathcal{O}(2^n)$  and there is no equivalent set for  $\Sigma'$  of smaller size. Furthermore,  $dup_{\Sigma, \beta_1}(O)$  consists of the  $n$  sets  $O - \{A_{2i-1}, A_{2i}\}$  for  $i = 1, \dots, n$ . Thus, Algorithm 3 computes an Armstrong p-sketch for  $\Sigma'$  whose number of objects is in  $\mathcal{O}(n)$ .  $\square$

Due to Theorems 9 and 10 we recommend the use of both abstract constraint sets and their Armstrong p-sketches. Indeed, the constraint sets enable design teams to identify constraints that they currently incorrectly perceive as semantically meaningful; and the Armstrong p-sketches enable design teams to identify constraints that they currently incorrectly perceive as semantically meaningless.

Our final result characterizes the situations in which finite Armstrong p-instances exist, and that the problem of deciding whether there is a finite Armstrong p-instance for a given QC set can be decided efficiently.

**Theorem 11.** *Let  $\Sigma$  be a set of QCs over some given p-object type  $(O, S)$ . Then there is a finite Armstrong p-instance for  $\Sigma$  if and only if for all  $A \in O$  there is some  $b \in \mathbb{N}$  such that  $\Sigma$  implies  $(card(A) \leq b, \beta_1)$ . It can therefore be decided in time  $\mathcal{O}(|O| \times \|\Sigma\|)$  whether there is a finite Armstrong p-instance for  $\Sigma$ .*

**Proof.** If it is true that for all  $A \in O$  there is some  $b \in \mathbb{N}$  such that  $\Sigma$  implies  $(card(A) \leq b, \beta_1)$ , then  $b_X^i < \infty$  for all non-empty  $X \subset O$ . Hence, every p-expansion of an Armstrong p-sketch that Algorithm 3 computes is finite. Vice versa, suppose there is some  $A \in O$  such that  $\Sigma$  does not imply  $(card(A) \leq b, \beta_1)$  for any  $b \in \mathbb{N}$ . Consequently, there is some duplicate set  $X$  of c-degree  $\beta_1$  and  $A \in X$  such that  $b_X^i = \infty$ . Theorem 7 shows that every Armstrong p-instance for  $\Sigma$  must contain infinitely many objects that agree on  $X$ . The condition that for all  $A \in O$  there is some  $b \in \mathbb{N}$  such that  $\Sigma$  implies  $(card(A) \leq b, \beta_1)$  can be verified in time  $\mathcal{O}(|O| \times \|\Sigma\|)$ .  $\square$

## 6. Implementation

We have implemented Algorithm 3 within a prototype system. The system enables users to enter a p-object type and a finite set of p-cardinality constraints over this type, and computes an Armstrong p-sketch for the constraint set. We illustrate the basic functionality of our system by some examples.

### 6.1. Running example

We begin with some screenshots that show how the instance from Example 10 is processed by our system. The left of Fig. 5 shows the input interface where the p-object type and input constraint set of Example 10 have been entered. The figure also shows that our system can randomly generate input that complies with any user specification of the following parameters: number of p-degrees, number of attributes, and number of cardinality constraints.

The left of Fig. 6 shows the Armstrong p-sketch for the input constraint set of Example 10 as computed by our prototype system. Note that only integer values are used as domain values and that these can be interpreted as indices for actual domain values, where different indices represent different domain values. Similarly, the integer  $i$  in the column *Poss* represents the p-degree  $\alpha_i$ . Observe that the p-sketch shown in the left of Fig. 6 is isomorphic to the p-sketch shown in Table 4.

As an explanation of the p-sketch, our prototype system can also show the duplicate sets it computes as part of Algorithm 3. The right of Fig. 6 shows the duplicate sets for the input constraint set of Example 10. The system shows for each attribute subset  $X$  and for each c-degree  $\beta_i$ , the minimum upper bound  $b_X^i$  that can be inferred from the input constraints. The system indicates by  $T$  that a given attribute subset is a duplicate set of c-degree  $\beta_i$ .

### 6.2. Extreme cases

Fig. 5 shows that our system can automatically generate instances of the extreme cases of p-cardinality constraints reported in Theorems 9 and 10, for any given even number of attributes. The left of Fig. 7 shows the output for the case  $n = 3$  of Theorem 9,

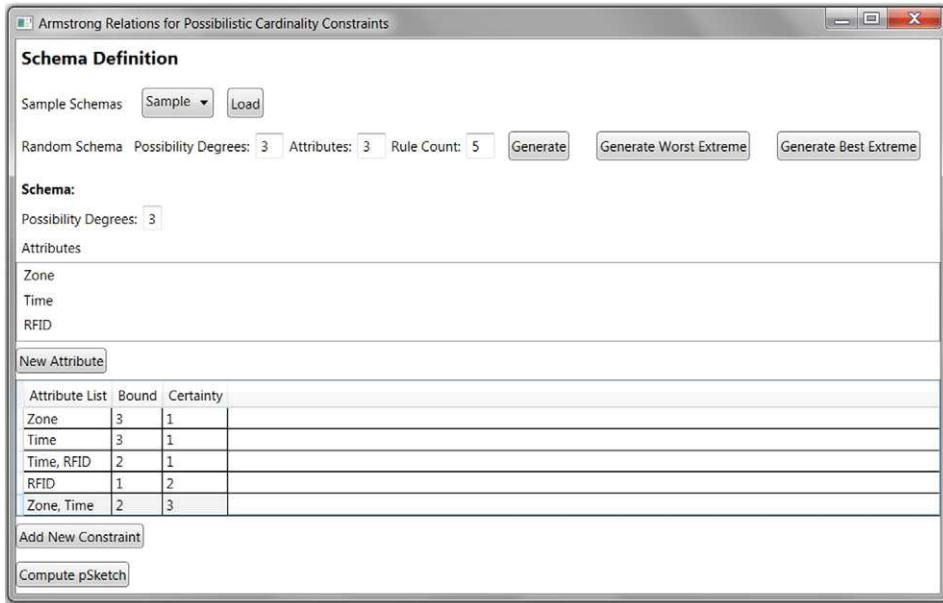


Fig. 5. User interface of our prototype system to enter input for the computation of Armstrong p-sketches.

where the attributes  $A, \dots, F$  are used instead of  $A_1, \dots, A_6$ , respectively. In this case, the three input constraints result in a p-sketch with eight objects. The right of Fig. 7 shows the output for the case  $n = 3$  of Theorem 10, with the same use of attribute names. In this case, the eight input constraints result in a p-sketch with three objects.

## 7. Computational experiments

We conducted a series of experiments with our prototype system. These provide some insight into the actual feasibility for the usefulness of Armstrong p-sketches in the acquisition of meaningful possibilistic cardinality constraints. Section 7.1 presents our results for the case of p-cardinality constraints from Theorem 9. Among other results, the experiments show that the computation of Armstrong p-sketches with 512 objects for an input constraint set with 18 attributes can be done in less than 3 s. Section 7.2 presents our results for the case of p-cardinality constraints from Theorem 10. Among other results, the experiments show that the computation of Armstrong p-sketches with 9 objects for an input constraint set with 4608 attributes can be done in less than 3 s, too. Finally, Section 7.3 presents our results for the average case in which sets of p-cardinality constraints were generated randomly. The results show that, on average, the growth in size of p-Armstrong sketches is low-degree polynomial in both the size of the input constraint set, and the number of different p-degrees available for a fixed input constraint set. Furthermore, the results show that, on average, the growth in time required to compute p-Armstrong sketches is low-degree polynomial in the size of the input constraint set, and linear in the number of different p-degrees available on a fixed input constraint set. We conclude that on problem instances

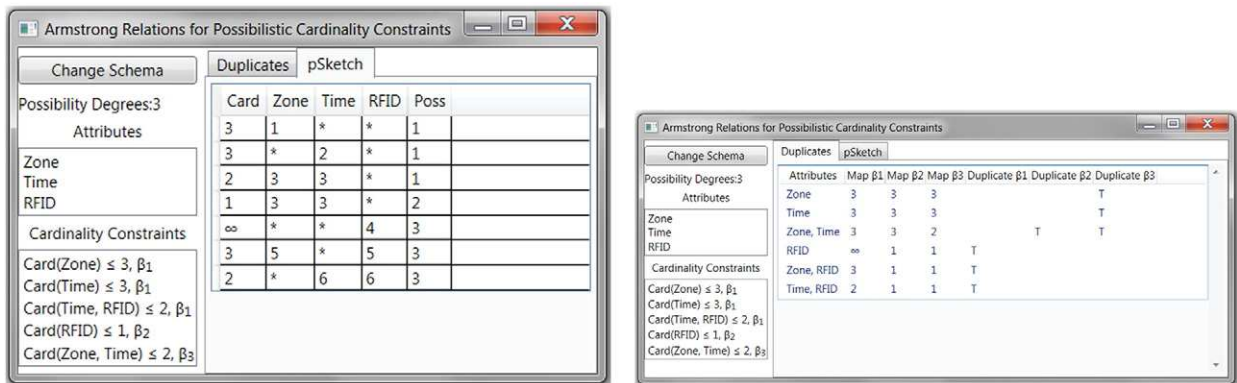


Fig. 6. Armstrong p-sketch and duplicate sets for the input from Example 10.

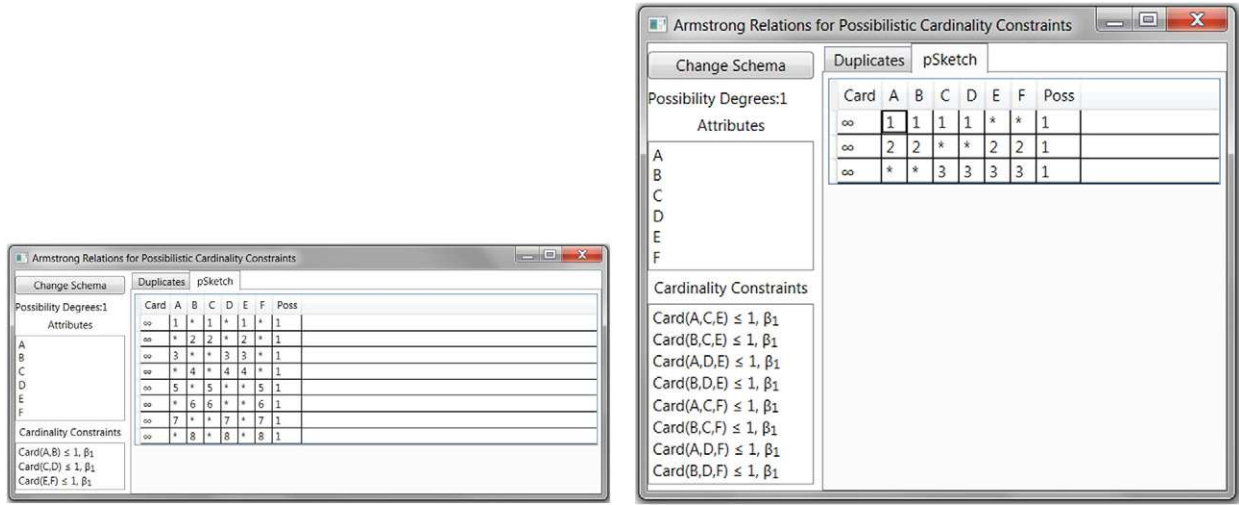


Fig. 7. Armstrong p-sketches for the input constraints of case  $n = 3$  from Theorems 9 and 10.

of practically relevant sizes, Armstrong p-sketches can be computed very quickly and consist of a number of objects that makes it possible to use the sketches in the requirements acquisition phase. We will now discuss our experiments in more detail.

### 7.1. Exponential case

Here, we consider the case  $\Sigma$  from Theorem 9 in which only Armstrong p-sketches with exponentially many objects in the size of  $\Sigma$  exist. For  $O_n = \{A_1, \dots, A_{2n}\}$ ,  $S = \{\alpha_1, \alpha_2\}$  and  $\Sigma_n = \{(card(A_1, A_2) \leq 1, \beta_1), \dots, (card(A_{2n-1}, A_{2n}) \leq 1, \beta_1)\}$  with size  $2 \cdot n$ ,  $dup_{\Sigma_n, \beta_1}(O)$  consists of the  $2^n$  duplicate sets  $U_{j=1}^n X_j$  where  $X_j \in \{A_{2j-1}, A_{2j}\}$ . The left of Fig. 8 shows the exponential growth of the output size in the input size for  $n = 1, \dots, 9$ . The right of Fig. 8 shows the time required by our tool to compute the sketches. In fact, the following table shows these times.

$n$	1	2	3	4	5	6	7	8	9
$\ \Sigma_n\ $	2	4	6	8	10	12	14	16	18
time (ms)	0.07	0.12	0.18	0.76	3.39	18.93	105.37	557.32	2683.24

### 7.2. Logarithmic case

Here, we consider the case  $\Sigma'$  from Theorem 10 in which Armstrong p-sketches exist that have a number of objects that is logarithmic in the size of  $\Sigma'$ . For all  $i = 1, \dots, n$ , for all  $X = \cup_{i=1}^n X_i$  where  $X_i \in \{A_{2i-1}, A_{2i}\}$ ,  $(card(X) \leq 1, \beta_1)$ . Then the size of  $\Sigma'$  is  $n \cdot 2^n$  and  $dup_{\Sigma', \beta_1}(O)$  consists of the  $n$  sets  $O - \{A_{2i-1}, A_{2i}\}$  for  $i = 1, \dots, n$ . The left of Fig. 9 shows the logarithmic growth of the output size in the input size for  $n = 1, \dots, 9$ . The right of Fig. 9 shows the time required by our tool to compute the sketches. In fact, the following table shows these times.

$n$	1	2	3	4	5	6	7	8	9
$\ \Sigma'_n\ $	4	8	24	64	160	384	896	2048	4608
time (ms)	0.01	0.03	0.18	0.68	3.21	17.6	104.53	561.46	2706.84

### 7.3. Average case

Experiments were conducted over a set of randomly generated input constraints in order to determine the impact of increasing numbers of p-degrees and increasing sizes of input on the computational effort required, and the size of the resulting Armstrong p-sketch.

#### 7.3.1. Experiment design

We experiment over different input constraints. For each number  $n = 3, \dots, 12$  of attributes, the object type  $O_n$  contains  $n$  attributes. We randomly generate 500 constraint sets for every  $O_n$ . The classical constraint sets, denoted  $\Sigma_{n,j}$  for  $j = 1, \dots, 500$ , are generated to contain between  $n$  and  $n^2/2$  cardinality constraints. Each cardinality constraint is constructed by selecting a random attribute

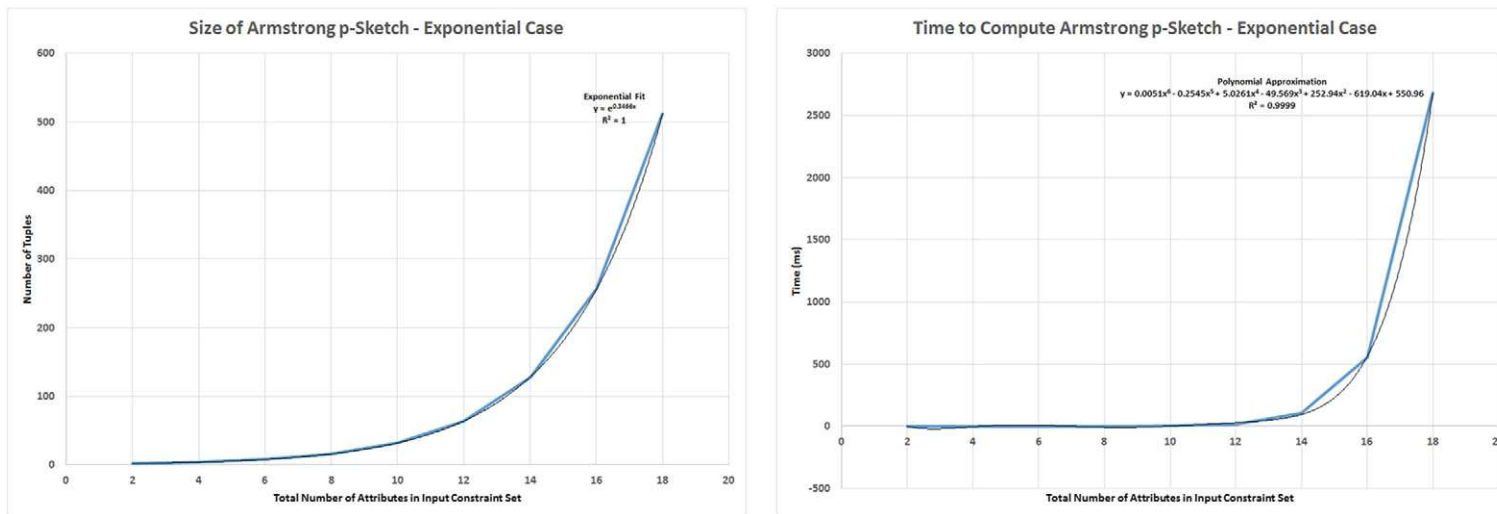
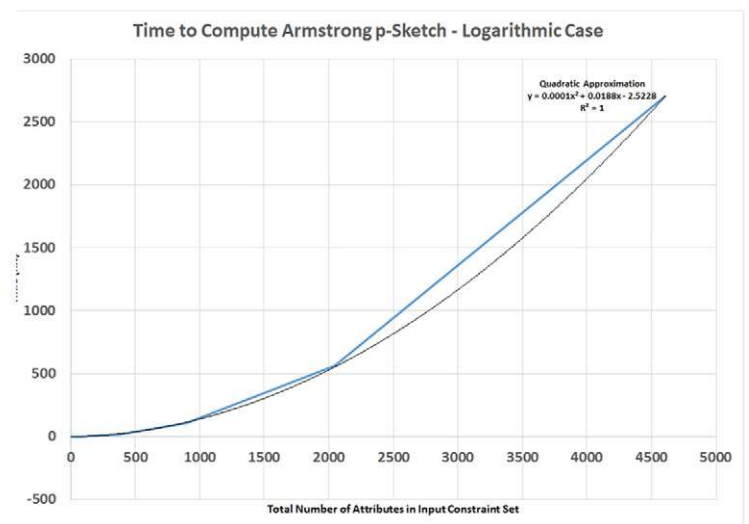
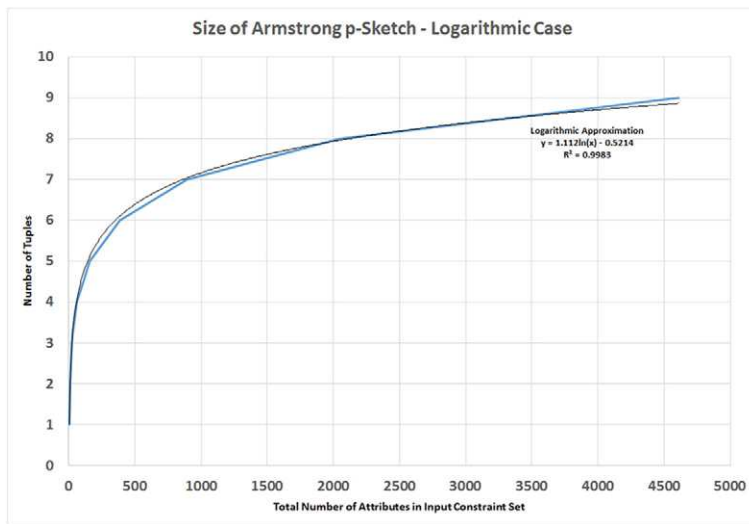


Fig. 8. Size of Armstrong p-sketches and time to compute them for the exponential case.



**Fig. 9.** Size of Armstrong p-sketches and time to compute them for the logarithmic case.

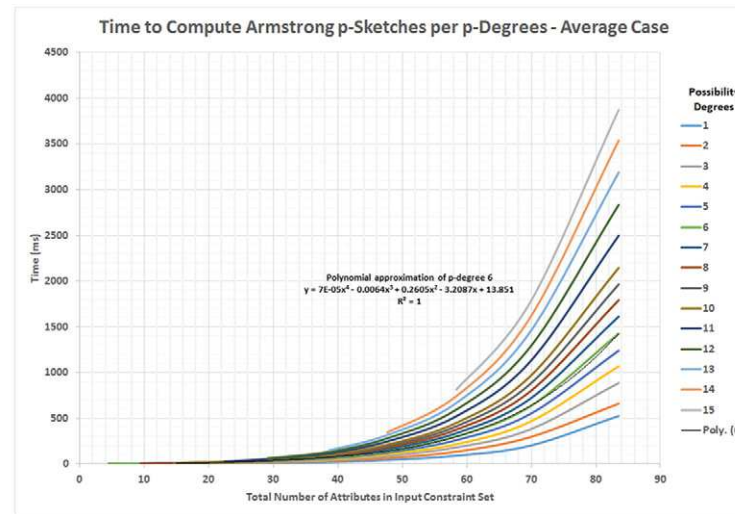
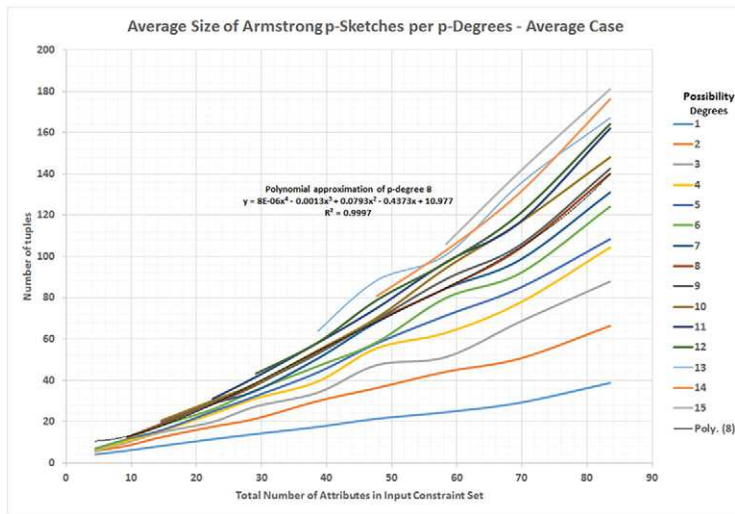


Fig. 10. Size of Armstrong p-sketches and time to compute them in the size of the input for the average case.

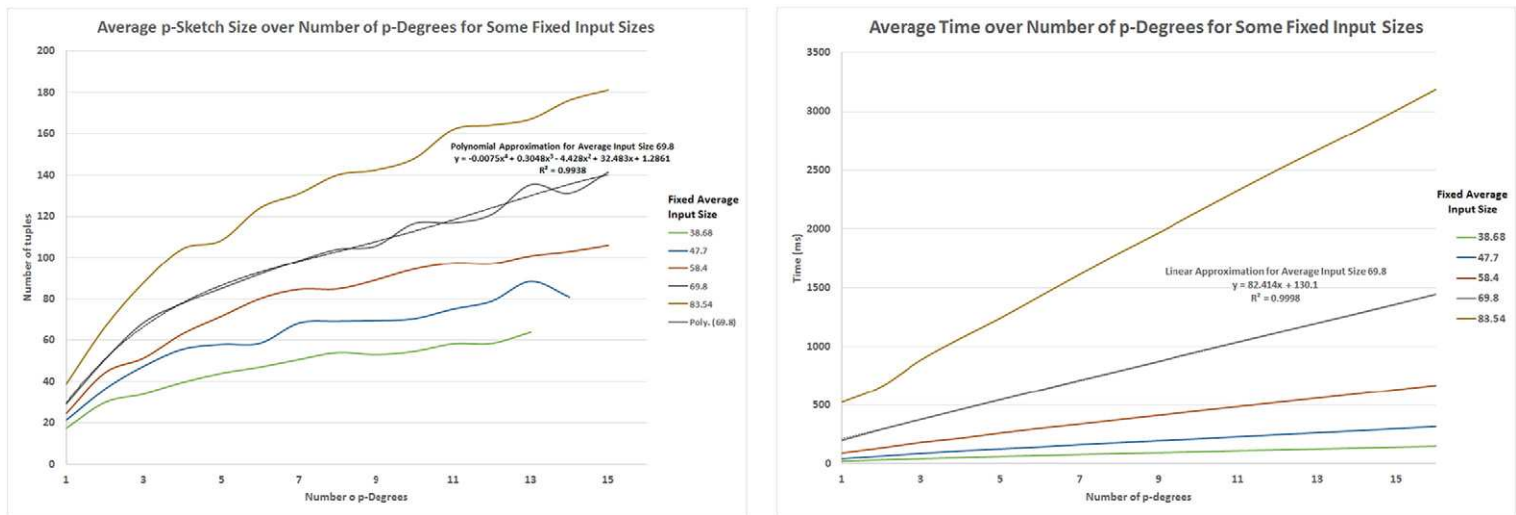


Fig. 11. Size of Armstrong p-sketches and time to compute them in the number of different p-degrees for the average case.



from  $O_n$  and adding an additional attribute while a coin toss is heads. From each of these classical constraint sets, a range of possibilistic cardinality constraint sets are generated by assigning a maximum number  $k$  of  $p$ -degrees and assigning to each cardinality constraint in  $\Sigma_{n,j}$  a  $c$ -degree from  $\beta_1$  to  $\beta_k$ . The maximum  $k$  is iterated from 1 to 15. This gives us for each  $n$  and  $k$ , 500 possibilistic constraint sets denoted as  $\Sigma_{n,j,k}$ , whose average size is determined by dividing the sum  $\sum_j |\Sigma_{n,j,k}|$  by 500. The average input size for each  $n = 3, \dots, 12$  forms the x-axis, which is the same for all  $k = 1, \dots, 15$  and each fixed  $n$ . The y-axes are, for each  $k$ , made up of the average sizes of the computed Armstrong  $p$ -sketches and the average computation time required, respectively.

Fig. 10 shows for each number  $k$  of available  $p$ -degrees the average size of the computed Armstrong  $p$ -sketches and the average time to compute them, respectively, in the average input constraint size for each number  $n = 3, \dots, 12$ . The results suggest that both size and time exhibit a low-degree polynomial growth in the input size. The average sizes of the computed Armstrong  $p$ -sketches suggest that it is possible in practice to effectively use the sketches in the requirements analysis phase, and the average execution times suggest that the sketches can be computed quickly.

Fig. 11 shows for each of the fixed input constraint sizes for  $n = 8, \dots, 12$ , the growth of the computed Armstrong  $p$ -sketches and the growth in time to compute them, respectively, in the growing number  $k = 1, \dots, 15$  of available  $p$ -degrees. The results suggest that for each fixed input constraint size, the size of the computed Armstrong  $p$ -sketches exhibits a low-degree polynomial growth and the time to compute them exhibits a linear growth in the number of available  $p$ -degrees. This provides some insight for selecting an appropriate number of  $p$ -degrees in practice.

## 8. Conclusion and future work

Cardinality constraints occur naturally in most aspects of life. Consequently, they have received invested interest from the data and knowledge engineering community over the last three decades. We have introduced cardinality constraints to control the occurrences of uncertain data patterns in modern applications, including big data. Uncertainty has been modeled qualitatively by applying the framework of possibility theory. Our cardinality constraints stipulate upper bounds on the number of occurrences of uncertain data patterns, an ability that can capture many real-world requirements. Our results show that cardinality constraints can be reasoned about efficiently. We have illustrated how reasoning about cardinality constraints is useful for the efficient processing of updates and queries. Despite several challenges we have shown that every set of cardinality constraints can be visualized perfectly in the form of an Armstrong  $p$ -sketch. The concept of an Armstrong  $p$ -sketch overcomes the problem of visualizing infinite Armstrong  $p$ -instances. We also implemented a tool that computes Armstrong  $p$ -sketches for any given set of cardinality constraints. Our experiments with the tool suggest that, on average, the times for computing Armstrong  $p$ -sketches show low-degree polynomial growth in the size of the input constraint set, and linear growth in the number of available  $p$ -degrees on a fixed input size. In practice, execution times were fast, ranging from an average of about half a second to an average of under four seconds on the largest fixed input size with one to fifteen available  $p$ -degrees, respectively. Our experiments suggest further that, on average, the sizes of the computed Armstrong  $p$ -sketches show low-degree polynomial growth in both the size of the input constraint set, and in the number of available  $p$ -degrees per fixed input size. The sizes were ranging from an average of 40 objects to an average of 180 objects on the largest fixed input size ranging from one to fifteen available  $p$ -degrees, respectively. Business analysts can therefore show our small sketches to domain experts in order to jointly consolidate the cardinality constraints that are meaningful for a given application domain.

Our framework opens up several questions for future investigation, including the benefits of processing data with the help of cardinality constraints, more expressive cardinality constraints and their interaction with other constraints, and empirical evaluations for the use of Armstrong  $p$ -instances and  $p$ -sketches. It is interesting to apply the concept of Armstrong sketches to other classes of constraints, including constraints in standard relational databases. Finally, constraints have not received much attention yet in probabilistic databases.

## Acknowledgment

This research is partially supported by the Marsden Fund Council from Government funding administered by the Royal Society of New Zealand, by the Natural Science Foundation of China (Grant No. 61472263) and by the Australian Research Council (Grant No. DP140103171).

## References

- [1] A. Artale, D. Calvanese, R. Kontchakov, V. Ryzhikov, M. Zakharyashev, Reasoning over extended ER models, in: C. Parent, K. Schewe, V.C. Storey, B. Thalheim (Eds.), *Conceptual Modeling – ER 2007, 26th International Conference on Conceptual Modeling*, Auckland, New Zealand, November 5–9, 2007, Proceedings, Lecture Notes in Computer Science, vol. 4801, Springer 2007, pp. 277–292.
- [2] S. Benferhat, D. Dubois, H. Prade, Towards a possibilistic logic handling of preferences, *Appl. Intell.* 14 (2001) 303–317.
- [3] S. Bistarelli, P. Codognot, F. Rossi, Abstracting soft constraints: framework, properties, examples, *Artif. Intell.* 139 (2002) 175–211.
- [4] P. Bosc, O. Pivert, On the impact of regular FDs when moving to a possibilistic database framework, *Fuzzy Sets Syst.* 140 (2003) 207–227.
- [5] P. Brown, S. Link, Probabilistic keys for data quality management, in: J. Zdravkovic, M. Kirikova, P. Johannesson (Eds.), *Advanced Information Systems Engineering – 27th International Conference, CAiSE 2015, Stockholm, Sweden, June 8–12, 2015, Proceedings, Lecture Notes in Computer Science*, vol. 9097, Springer 2015, pp. 118–132.
- [6] D. Calvanese, M. Lenzerini, On the interaction between ISA and cardinality constraints, *Proceedings of the Tenth International Conference on Data Engineering*, February 14–18, 1994, IEEE Computer Society, Houston, Texas, USA 1994, pp. 204–213.
- [7] P.P. Chen, The entity-relationship model – toward a unified view of data, *ACM Trans. Database Syst.* 1 (1976) 9–36.

- [8] F. Currim, N. Neidig, A. Kampooowale, G. Mhatre, The CARD system, in: J. Parsons, M. Saeki, P. Shoval, C.C. Woo, Y. Wand (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, Springer 2010, pp. 433–437.
- [9] D. Dubois, J. Lang, H. Prade, Automated reasoning using possibilistic logic: semantics, belief revision, and variable certainty weights, *IEEE Trans. Knowl. Data Eng.* 6 (1994) 64–71.
- [10] D. Dubois, H. Prade, Epistemic entrenchment and possibilistic logic, *Artif. Intell.* 50 (1991) 223–239.
- [11] D. Dubois, H. Prade, Fuzzy set and possibility theory-based methods in artificial intelligence, *Artif. Intell.* 148 (2003) 1–9.
- [12] D. Dubois, H. Prade, S. Schockaert, Stable models in generalized possibilistic logic, in: G. Brewka, T. Eiter, S.A. McIlraith (Eds.), *Principles of Knowledge Representation and Reasoning: Proceedings of the Thirteenth International Conference, KR 2012, Rome, Italy, June 10–14, 2012*, AAAI Press, 2012.
- [13] R. Fagin, Horn clauses and database dependencies, *J. ACM* 29 (1982) 952–985.
- [14] F. Ferrarotti, S. Hartmann, S. Link, A precious class of cardinality constraints for flexible XML data processing, in: M.A. Jeusfeld, L.M.L. Delcambre, T.W. Ling (Eds.), *Conceptual Modeling – ER 2011, 30th International Conference, ER 2011, Brussels, Belgium, October 31–November 3, 2011*. Proceedings, Lecture Notes in Computer Science, vol. 6998, Springer 2011, pp. 175–188.
- [15] F. Ferrarotti, S. Hartmann, S. Link, Efficiency frontiers of XML cardinality constraints, *Data Knowl. Eng.* 87 (2013) 297–319.
- [16] P. Gärdenfors, D. Makinson, Nonmonotonic inference based on expectations, *Artif. Intell.* 65 (1994) 197–245.
- [17] A. Grove, Two modellings for theory change, *J. Philos. Log.* 17 (1988) 157–170.
- [18] S. Hartmann, On the consistency of int-cardinality constraints, in: T.W. Ling, S. Ram, M. Lee (Eds.), *Conceptual Modeling – ER '98, 17th International Conference on Conceptual Modeling*, Singapore, November 16–19, 1998. Proceedings, Lecture Notes in Computer Science, vol. 1507, Springer 1998, pp. 150–163.
- [19] S. Hartmann, Decomposition by pivoting and path cardinality constraints, in: A.H.F. Laender, S.W. Liddle, V.C. Storey (Eds.), *Conceptual Modeling – ER 2000, 19th International Conference on Conceptual Modeling*, Salt Lake City, Utah, USA, October 9–12, 2000. Proceedings, Lecture Notes in Computer Science, vol. 1920, Springer 2000, pp. 126–139.
- [20] S. Hartmann, On the implication problem for cardinality constraints and functional dependencies, *Ann. Math. Artif. Intell.* 33 (2001) 253–307.
- [21] S. Hartmann, Reasoning about participation constraints and Chen's constraints, in: K. Schewe, X. Zhou (Eds.), *Database Technologies 2003, Proceedings of the 14th Australasian Database Conference, ADC 2003, Adelaide, South Australia, February 2003*, CRPIT, vol. 17, Australian Computer Society 2003, pp. 105–113.
- [22] S. Hartmann, M. Kirchberg, S. Link, Design by example for SQL table definitions with functional dependencies, *VLDB J.* 21 (2012) 121–144.
- [23] S. Hartmann, H. Köhler, U. Leck, S. Link, B. Thalheim, J. Wang, Constructing Armstrong tables for general cardinality constraints and not-null constraints, *Ann. Math. Artif. Intell.* 73 (2015) 139–165.
- [24] S. Hartmann, S. Link, Efficient reasoning about a robust XML key fragment, *ACM Trans. Database Syst.* 34 (2009).
- [25] S. Hartmann, S. Link, Numerical constraints on XML data, *Inf. Comput.* 208 (2010) 521–544.
- [26] A.K. Jha, V. Rastogi, D. Suciu, Query evaluation with soft-key constraints, in: M. Lenzerini, D. Lembo (Eds.), *Proceedings of the Twenty-Seventh ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, PODS 2008, June 9–11, 2008*, ACM, Vancouver, BC, Canada 2008, pp. 119–128.
- [27] T.H. Jones, I.Y. Song, Analysis of binary/ternary cardinality combinations in entity-relationship modeling, *Data Knowl. Eng.* 19 (1996) 39–64.
- [28] H. Köhler, U. Leck, S. Link, H. Prade, Logical foundations of possibilistic keys, in: E. Fermé, J. Leite (Eds.), *Logics in Artificial Intelligence – 14th European Conference, JELIA 2014, Funchal, Madeira, Portugal, September 24–26, 2014*. Proceedings, Lecture Notes in Computer Science, vol. 8761, Springer 2014, pp. 181–195.
- [29] H. Köhler, S. Link, H. Prade, X. Zhou, Cardinality constraints for uncertain data, in: E.S.K. Yu, G. Dobbie, M. Jarke, S. Purao (Eds.), *Conceptual Modeling – 33rd International Conference, ER 2014, Atlanta, GA, USA, October 27–29, 2014*. Proceedings, Lecture Notes in Computer Science, vol. 8824, Springer 2014, pp. 108–121.
- [30] W.D. Langeveldt, S. Link, Empirical evidence for the usefulness of Armstrong relations in the acquisition of meaningful functional dependencies, *Inf. Syst.* 35 (2010) 352–374.
- [31] M. Lenzerini, P. Nobili, On the satisfiability of dependency constraints in entity-relationship schemata, *Inf. Syst.* 15 (1990) 453–461.
- [32] S.W. Liddle, D.W. Embley, S.N. Woodfield, Cardinality constraints in semantic data models, *Data Knowl. Eng.* 11 (1993) 235–270.
- [33] S. Link, Armstrong databases: validation, communication and consolidation of conceptual models with perfect test data, in: A. Ghose, F. Ferrarotti (Eds.), *Eighth Asia-Pacific Conference on Conceptual Modelling, APCCM 2012, Melbourne, Australia, January 2012*, Conferences in Research and Practice in Information Technology, vol. 130, Australian Computer Society 2012, pp. 3–20.
- [34] H. Mannila, K.J. Räihä, Design by example: an application of Armstrong relations, *J. Comput. Syst. Sci.* 33 (1986) 126–141.
- [35] A.J. McAllister, Complete rules for  $n$ -ary relationship cardinality constraints, *Data Knowl. Eng.* 27 (1998) 255–288.
- [36] G. Qi, K. Wang, Conflict-based belief revision operators in possibilistic logic, in: J. Hoffmann, B. Selman (Eds.), *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*, July 22–26, 2012, Toronto, Ontario, Canada, AAAI Press, 2012.
- [37] A. Queral, A. Artale, D. Calvanese, E. Teniente, OCL-Lite: finite reasoning on UML/OCL conceptual schemas, *Data Knowl. Eng.* 73 (2012) 1–22.
- [38] T. Roblot, S. Link, Probabilistic cardinality constraints, in: P. Johannesson, M.L. Lee, S. Liddle, O. Pastor (Eds.), *Conceptual Modeling – ER 2015, 34th International Conference on Conceptual Modeling*, Stockholm, Sweden, October 19–22, 2015. Proceedings, Lecture Notes in Computer Science, Springer, 2015.
- [39] R. Sabbadin, H. Fargier, J. Lang, Towards qualitative approaches to multi-stage decision making, *Int. J. Approx. Reason.* 19 (1998) 441–471.
- [40] Q. Shen, R. Leitch, Fuzzy qualitative simulation, *IEEE Trans. Syst. Man Cybern.* 23 (1993) 1038–1061.
- [41] D. Suciu, D. Olteanu, C. Ré, C. Koch, *Probabilistic Databases, Synthesis Lectures on Data Management*, Morgan & Claypool Publishers, 2011.
- [42] B. Thalheim, Fundamentals of cardinality constraints, in: G. Pernul, A.M. Tjoa (Eds.), *Entity-Relationship Approach – ER'92, 11th International Conference on the Entity-Relationship Approach*, Karlsruhe, Germany, October 7–9, 1992. Proceedings, Lecture Notes in Computer Science, vol. 645, Springer 1992, pp. 7–23.
- [43] B. Thalheim, *Entity-relationship modeling*, Springer, 2000.
- [44] B. Thalheim, Integrity constraints in (conceptual) database models, in: R.H. Kaschek, L.M.L. Delcambre (Eds.), *The Evolution of Conceptual Modeling – From a Historical Perspective towards the Future of Conceptual Modeling [outcome of a Dagstuhl seminar held 2008]*, Lecture Notes in Computer Science, vol. 6520, Springer 2008, pp. 42–67.
- [45] L.A. Zadeh, Approximate reasoning based on fuzzy logic, in: B.G. Buchanan (Ed.), *Proceedings of the Sixth International Joint Conference on Artificial Intelligence, IJCAI 79, Tokyo, Japan, August 20–23, 1979*, 2 vols., William Kaufmann 1979, pp. 1004–1010.



**Neil Hall** received a Bachelor of Science with Honours degree from the University of Auckland in 1992. Over twenty years later, he has returned to the University where he is currently studying the Masters Programme of Professional Studies in Data Science.



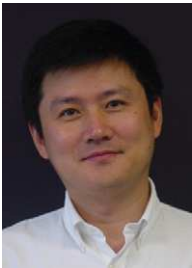
**Henning Koehler** received his PhD in Information Systems from Massey University, New Zealand, in 2008. He is currently a Senior Lecturer in the Department of Computer Science and Information Technology at Massey University. His research interest includes database design and semantics, data cleaning and integration, and computational graph theory. He has served as a reviewer or program chair, and published dozens of research papers, in various conferences and journals.



**Sebastian Link** received a DSc from the University of Auckland in 2015, and a PhD in Information Systems from Massey University in 2005. Currently, he is an Associate Professor at the Department of Computer Science in the University of Auckland. His research interests include conceptual data modeling, semantics in databases, foundations of mark-up languages, and applications of discrete mathematics to computer science. Sebastian received the Chris Wallace Award for Outstanding Research Contributions from the Computing Research and Education Association of Australasia in recognition of his work on the semantics of SQL and XML data. Sebastian has published more than 125 research papers, and served as a reviewer for numerous conferences and journals. He is a member of the editorial board of the journal Information Systems.



**Henri Prade** is “Directeur de Recherche” at the National Center for Scientific Research (C.N.R.S.), France, and works at IRIT in Toulouse. He received a Doctor-Engineer degree from Ecole Nationale Supérieure de l’Aéronautique et de l’Espace, in Toulouse (1977), his “Doctorat d’Etat” (1982) and the “Habilitation à Diriger des Recherches” (1986) both from Paul Sabatier University in Toulouse. He is the co-author of two monographs on fuzzy sets and possibility theory, has contributed a great number of technical papers (including about 200 journal papers) and has co-edited several books. In 2002 he received the Pioneer Award of the IEEE Neural Network Society. He is an IFSA Fellow and an ECCAI Fellow. His current research interests are in uncertainty and preference modeling, non-classical logics, approximate, plausible, and analogical reasoning with applications to artificial intelligence and information systems.



**Professor Xiaofang Zhou** is a Professor of Computer Science at the University of Queensland, and Head of the Data and Knowledge Engineering (DKE) Group. He received his PhD in Computer Science from the University of Queensland. His research focus is to find effective and efficient solutions for managing, integrating and analyzing very large amount of complex data for business, scientific and personal applications. He has been working in the area of spatial and multimedia databases, data quality, high performance query processing, Web information systems and bioinformatics, co-authored over 250 research papers with many published in top journals and conferences such as SIGMOD, VLDB, ICDE, ACM Multimedia, The VLDB Journal, ACM Transactions and IEEE Transactions. Currently he is an Associate Editor of The VLDB Journal, IEEE Transactions on Cloud Computing, World Wide Web Journal, Distributed and Parallel Databases, and IEEE Data Engineering Bulletin.